OXFORD

JNCI J Natl Cancer Inst (2018) 110(3): djx179

doi: 10.1093/jnci/djx179
First published online September 27, 2017
Review

REVIEW

Understanding and Communicating Measures of Treatment Effect on Survival: Can We Do Better?

Everardo D. Saad, John R. Zalcberg, Julien Péron, Elisabeth Coart, Tomasz Burzykowski, Marc Buyse

Affiliations of authors: International Drug Development Institute (IDDI), Louvain-la-Neuve, Belgium (EDS, EC, TB); School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia (JRZ); Department of Medical Oncology, Hospices Civils de Lyon, Pierre-Benite, France (JP); CNRS UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique-Santé, Université de Lyon, Lyon, France (JP); Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Hasselt University, Diepenbeek, Belgium (TB, MB); International Drug Development Institute (IDDI), San Francisco, CA (MB).

Correspondence to: Everardo D. Saad, MD, International Drug Development Institute, Avenue Provinciale 30, 1340 Louvain-la-Neuve, Belgium (e-mail: everardo.saad@iddi.com).

Abstract

Time-to-event end points are the most frequent primary end points in phase III oncology trials, both in the adjuvant and advanced settings. The evaluation of these end points is important to inform clinical practice. However, although different measures can be used to describe the effect of treatment on these end points, we believe that any treatment benefit in a given trial is best reported using various absolute and relative measures. Our goal is to help clinicians understand the strengths and limitations of the traditional and novel measures used to denote the effect of treatment in randomized trials. Although none of these measures can reliably predict the outcome of individual patients, some measures could be added to the commonly used hazard ratio to provide a more patient-oriented assessment of treatment benefit. In particular, the difference of mean survival times quantifies the average survival benefit for a patient receiving a new treatment compared with a patient treated with standard of care, whereas the net benefit quantifies the probability of a patient receiving the new treatment to live longer by at least m months (for any number of months m of interest) than a patient receiving the standard treatment. We encourage statisticians and clinical scientists to include various measures of treatment benefit in the reports of phase III trials, acknowledging that different clinical situations may call for different measures of treatment effect. By using the various available measures, we may better inform ourselves and communicate results to our patients.

End points that assess the time from random assignment to the occurrence of clinically relevant events are widely used in oncology trials and are the most frequent primary end points in phase III trials, both in the adjuvant and the advanced settings (1–3). Clinical trials inform clinical practice, and the clinician often needs to communicate results on these time-to-event end points to patients and their families. However, different measures and approaches can be used to describe the effect of treatment on these end points. It is well known that clinicians are affected by the choice of measures used to report trial results (4), and the way such results are communicated by clinicians may in turn influence the acceptance of different treatments in oncology (5). Thus, a clear understanding of the various possible ways of describing the effect of treatment on time-to-event end points is important for oncologists. Although previous authors have attempted to demystify measures of treatment benefit in oncology (6–8), the recent focus on alternative methods to describe survival curves (9–11) and new statistical developments (12,13) in the field offer new ways of assessing the treatment effect. We will discuss the complementary roles of these measures in fully assessing the results of comparative time-to-event analyses.

Objective and Methods of This Review

Our goal here is to help clinicians understand the strengths and limitations of traditional and novel measures used to denote

Received: April 24, 2017; **Revised:** July 2, 2017; **Accepted:** August 4, 2017

the effect of treatment on a time-to-event end point in randomized trials in oncology. This is done under the assumptions that one single measure is not sufficient in clinical practice and that ideal measures should be well founded from the statistical, clinical, and patient perspectives. We begin with a short introduction on the hazard ratio, a very useful and nearly universally reported measure of the treatment effect. By referring to it, we can more clearly point to advantages and disadvantages of the alternative measure, which we will often contrast with the hazard ratio. We then briefly discuss illustrative examples for which we consider a single time-to-event end point in a trial. In these examples, we introduce some of the absolute and relative measures that will be discussed in further depth below. Novel measures that can also be used in these examples will also be discussed. Table 1 summarizes the advantages and disadvantages of each of the traditional and novel measures discussed in this review. Such measures can be applied to both simple end points consisting of the time to a single event of interest (eg, overall survival [OS], time to progression, duration of response, etc.) and composite end points consisting of the time to the first of a number of events (eg, disease-free survival, progression-free survival [PFS], time to treatment failure, etc.). For the illustrative examples, we assume that hazards are proportional (the meaning of this assumption is provided below). Of note, we do not address the relationship between treatment benefit and the associated toxicity, convenience, or cost of treatment; for the interested reader, we recommend recent reviews on these topics (14–17). Moreover, methods have been proposed to incorporate safety issues into treatment decisions (and more generally several outcomes), but the discussion of these methods is beyond the scope of this review (18,19).

On a more technical note, we should point out that we do not discuss the merits of the actual tests used to assess the statistical significance of the results quantified by the various measures described. Moreover, we assume that the estimation of treatment effect is unbiased (ie, we leave aside considerations about the study design or analysis that might have an impact on the estimated treatment effect by introducing systematic errors), and for simplicity we do not take into account confidence intervals or other measures of precision for the estimated measures of treatment effect. Such measures of precision, which quantify the uncertainty around the estimates obtained from clinical trials, help interpret the expected clinical benefit from treatment, but they add a layer of complexity that is beyond our scope here. Finally, although we point out specific limitations of individual measures that are relevant for clinicians interpreting survival curves, we do not address some of the more technical problems with these measures and the remedial actions that can be taken to correct these problems during the planning or analysis stage of a trial; although important, we believe these issues are beyond the scope of how best to communicate clinical trial results to patients and their families.

The Hazard Ratio

In epidemiology, the most commonly used measure of change in risk—that is, in the probability of an event (sometimes called "cumulative risk")—occurring over a fixed period of observation is the relative risk, or risk ratio. The risk ratio contrasts the risks for the event in exposed and unexposed groups, without regard for the precise time of occurrence of each of those events for individual subjects. In clinical trials, and particularly for events that will happen in most, if not all, patients in both groups, the time of occurrence of the events is of interest. As a consequence, the focus shifts to the hazard rate. The hazard rate at time t is the "instantaneous risk" (probability) of occurrence of an event at time t given that the event has not occurred by t. In general, hazard rate depends on time, and a hazard function describes this dependence. Consequently, in clinical trials, the hazard ratio replaces the relative risk; the hazard ratio is a ratio of hazard rates (or hazard functions). It follows that the hazard ratio captures the reduction in hazard, that is, the reduction in the instantaneous risk of an event.

It is worth noting that the hazard ratio is often, and somewhat confusingly, referred to as a "relative risk." From the above it should be clear that the proper term should rather be the "relative instantaneous risk," and the hazard ratio should be interpreted as capturing the reduction in hazard, not the (cumulative) risk of occurrence of an event over a specified period of time. Figure 1 illustrates these concepts by showing survival probabilities, cumulative risks, and risk reductions for a fictitious example for two populations of patients with constant hazard rates of 0.40 (control) and 0.20 (experimental), and a corresponding constant hazard ratio of 0.50.

As mentioned above, in general, the hazard rates as well as the hazard ratio are functions of time. However, if the hazard rates can be assumed to be proportional between the groups being compared (ie, the ratio of the hazard function for the experimental arm to the hazard function for the control arm is constant over time), the hazard ratio becomes constant in time and can be expressed by a single numerical value. Under this assumption of proportional hazards, the hazard ratio can be estimated using the Cox proportional hazards model. The hazard ratio is a very useful and nearly universally reported relative measure of the treatment effect. On the other hand, the hazard ratio does not allow easy calculation of absolute benefits in terms of survival probabilities at times of interest, such as one or two years. Given that the hazard ratio is a number that cannot be directly related to the expected survival of a given patient, its use as a measure to communicate treatment benefit in clinical practice is limited (8).

It is worth noting that even if the hazard rates are proportional, the estimated hazard ratio can be biased if important covariates are omitted from the proportional hazards model (20). In particular, even in a properly randomized clinical trial in which the distribution of prognostic factors is balanced between the compared treatment arms, the true hazard ratio can be underestimated if the prognostic effect of these factors is not taken into account (20). On the other hand, if one cannot assume that the hazard rates are proportional, the interpretation of the hazard ratio becomes problematic because in this case the hazard ratio is time dependent and should not be represented by a single numerical value. The key assumption of proportional hazards may not be fulfilled in two distinct situations: 1) when the patient population is a mix of patients with differing treatment effects (ie, there are interactions between treatment effects and molecular or other patient characteristics) and 2) when the treatment effect truly varies over time (12,21,22). The Iressa Pan-Asia Study (IPASS) trial provides a notorious example of the former situation, with a crossing of PFS curves due to the subset of patients with epidermal growth factor receptor (EGFR) mutation enjoying a much prolonged PFS compared with patients receiving chemotherapy, and the opposite patterns for patients with wild-type EGFR (23). There are also many examples of the latter situation, with either survival curves converging after an early separation in comparisons between surgical and medical therapy, or with survival curves diverging after

Table 1. Advantages and disadvantages of different measures of treatment effect

Measure	Advantages	Disadvantages
Hazard ratio	Almost always reported Clear interpretation Takes entire survival curve into account	Not practical for patient communication Difficult to interpret for nonproportional hazards
Difference between survival probabilities at different time points (t)	Easy to read off survival curves	Depends on choice(s) of t Loses information
Difference between medians	Easy to read off survival curves Easy to remember	Not directly patient-relevant Not always reached Affected by schedule of assessment for end points other than overall survival Loses information Statistically unstable
Difference between restricted means	Takes entire survival curve (until chosen time t) into account Does not depend on proportional hazards assumption Intuitive interpretation as difference between areas under the survival curves	Almost never reported Difficult interpretation if survival curves are far from 0 at the largest follow-up time t Potential for misunderstanding the key role of truncation time in its computation
Difference between unrestricted means	Easy to remember Takes entire survival curve into account Does not depend on proportional hazards assumption Intuitive interpretation	Almost never reported Estimation requires a parametric distribution assumption if survival curves do not reach 0 Imprecise estimation if data are not mature (survival curves far from 0 at the largest follow-up time t)
Net benefit	Can be readily interpreted as a net probability of benefit Can express benefit in terms of absolute gains in survival time Takes entire survival curve into account Does not depend on proportional hazards assumption Prioritizes the more relevant component of a composite end point	Recently proposed, hence little experience
Ratio of restricted means	Takes entire survival curve (until chosen t) into account Valid even when hazards are nonproportional	Almost never reported Doubtful interpretation if survival curves are far from 0 at time t
Win ratio	Takes entire curve into account Does not depend on proportional hazards assumption Prioritizes the more relevant component of a composite end point	Interpretation not straightforward Recently proposed, hence little experience

being superimposed in comparisons between chemotherapy and modern anticancer immunotherapy. For these reasons, the hazard ratio has come under some criticism in oncology, and other measures have been proposed as alternatives or complements to it (6,11–13,21,24–26). The following examples will be used as a motivation for these other measures.

Illustrative Examples

Practically any randomized clinical trial could be used to illustrate our contention that various measures are usually needed to fully depict the effect of treatment on survival. In order to limit the possibilities, we will focus on two contrasting settings—advanced pancreatic cancer, an incurable condition with short expected follow-up, and early breast cancer, a potentially curable disease with a need for long follow-up—our aim being to provide examples that cover a broad field of application of the concepts discussed (Table 2).

The first example is the phase III trial of gemcitabine plus erlotinib vs gemcitabine plus placebo in advanced pancreatic cancer (27). The primary end point in that trial was OS, which was statistically significantly prolonged by the addition of erlotinib to gemcitabine (stratified log-rank test $P = .038$). In relative terms, the benefit of adding erlotinib to gemcitabine was not trivial because there was an 18% reduction in the hazard of death (HR = 0.82). However, the benefit was far less impressive if looked at from an absolute point of view, that is, if expressed in terms of differences on an appropriate scale. For example, the reduction of the risk of death was only 6% after one year in favor of erlotinib. Moreover, many commentators pointed out that the median OS was 6.24 months for erlotinib and 5.91 months for placebo, a difference of only 10 days.

A second example in advanced pancreatic cancer is the phase III trial of fluorouracil, leucovorin, irinotecan and oxaliplatin (FOLFIRINOX) vs gemcitabine (28). Once again, the primary end point was OS, which was statistically significantly prolonged by the use of FOLFIRINOX (stratified log-rank test $P < .001$). The relative benefit from the combination was

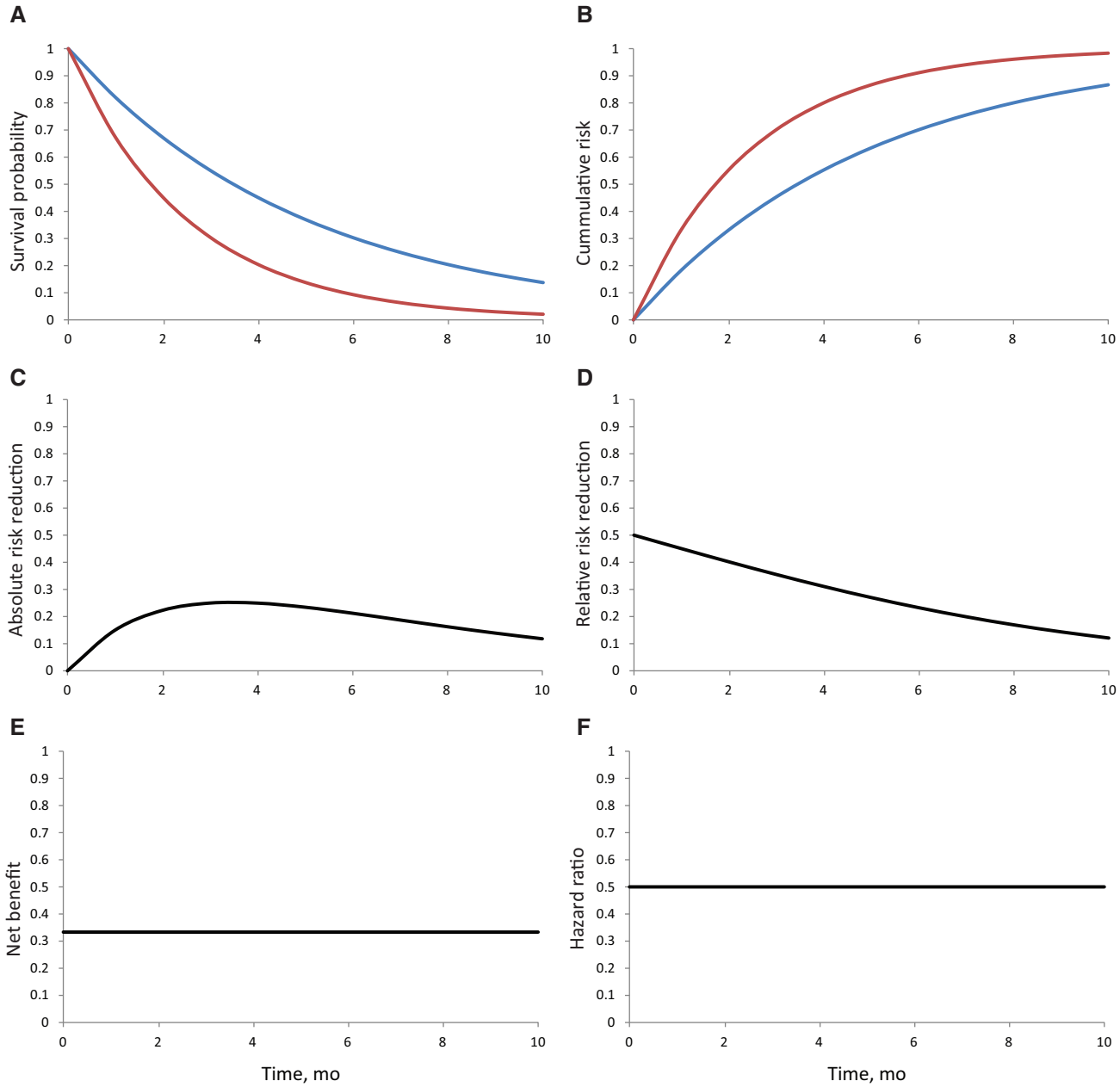

Figure 1. Different views of the treatment effect in a situation of proportional hazards. Survival probability **(A)**, cumulative risk **(B)**, absolute risk reduction **(C)**, relative risk reductions **(D)**, the net benefit **(E)**, and hazard ratio **(F)** for a comparison of two fictitious survival curves of experimental (**blue line**) and control (**red line**) treatments. For simplicity, the distribution of survival is assumed to be exponential (with hazard rates equal to 0.4 and 0.2 for the control and experimental treatments, respectively). The net benefit (Δ) was computed as $\Delta = [1-HR]/[1+HR]$, where HR is the hazard ratio, assuming no censoring. These idealized data illustrate the fact that even with a constant hazard ratio of 0.50, the risk reduction is a function of time and decreases toward 0. Moreover, they illustrate that, with proportional hazards, a reduction by half of the instantaneous risk of death translates into a 33% net benefit, that is, 33% net chance that the survival of a treated patient will be longer than the survival of a control patient.

impressive, with a hazard ratio of 0.57, which indicates a 43% reduction in the hazard of death. In absolute terms, the benefit was also impressive, with a difference in median OS times of 4.3 months and a 21% higher survival probability at one year for FOLFIRINOX patients. Other metrics that can be used to assess the results of these two trials are shown in Table 2 and will be introduced below.

The first example in early breast cancer is the Herceptin Adjuvant (HERA) trial, which had disease-free survival (DFS) as its primary end point (29). This trial showed that one year of adjuvant trastuzumab therapy improves DFS (log-rank test $P < .0001$), in comparison with observation; in relative terms, this benefit is captured by a hazard ratio of 0.54 at the first interim analysis (notwithstanding a decrease in such benefit in updated results for this trial) (30,31). A second example is the phase III trial of adjuvant neratinib vs placebo in patients with early-stage, human epidermal growth factor receptor 2 (HER2)–positive breast cancer previously treated with a trastuzumab-based

Table 2. Results of different measures of treatment effect within trials*

Measure	Advanced pancreatic cancer (27)	Advanced pancreatic cancer (28)	Early breast cancer (29)	Early breast cancer (32)
Treatment comparisons	Gemcitabine plus erlotinib vs gemcitabine plus placebo	FOLFIRINOX vs gemcitabine	Trastuzumab vs observation	Neratinib vs placebo
Summary result for primary end point	Gemcitabine plus erlotinib superior for overall survival	FOLFIRINOX superior for overall survival	Trastuzumab superior for disease-free survival	Neratinib superior for invasive disease-free survival
Hazard ratio	0.82	0.57	0.54	0.67
Difference between survival probabilities	6% at 12 mo	20.7% at 12 mo	8.5% at 24 mo	2.3% at 24 mo
Difference between medians	10 d	4.3 mo	Not estimable nonparametrically (medians not reached)	Not estimable nonparametrically (medians not reached)
Difference between restricted means	0.5 mo with restriction at 18 mo	3.3 mo with restriction at 18 mo	1.2 mo with restriction at 24 mo	0.5 mo with restriction at 24 mo
Difference between unrestricted means (Weibull model–based)	1.1 mo	4.0 mo	100 mo (with extrapolation)	30 mo (with extrapolation)
Net chance of a longer survival	Around 5% for differences in survival of up to 6 mo (18)	Around 25% for differences in survival of up to 6 mo (47)	11% for differences in disease-free survival of any magnitude	8% for differences in invasive disease-free survival of any magnitude

*The restricted mean survival times were not initially reported for any of the trials. For the neratinib trial, the difference in restricted mean survival was later reported by Chan et al (43). For the other three trials (27–29), we determined restricted mean survival times by digitizing the published survival curves and estimating their area under the curves using the trapezoidal method. For the erlotinib trial (27), the computed mean overall survival times restricted at 18 months were 7.5 months for gemcitabine plus erlotinib and 7.0 months for gemcitabine plus placebo. For the FOLFIRINOX trial (28), the computed mean overall survival times restricted at 18 months were 10.6 months for gemcitabine plus erlotinib and 7.3 months for gemcitabine plus placebo. For the trastuzumab trial (29), the computed mean disease-free survival times restricted at two years were 22.5 months for trastuzumab and 21.3 months for observation. The net benefit was likewise unavailable for the two trials on breast cancer (29,32). We therefore computed the net chance of a longer disease-free survival using the formula $\Delta = (1-c)^2 \cdot [1-HR]/[1+HR]$, where c is the proportion of censored observations (46). Note that this formula assumes a situation of proportional hazards, an assumption that would not have been required had individual patient data been available. FOLFIRINOX = fluorouracil, leucovorin, irinotecan, and oxaliplatin.

therapy (32). The primary end point in this case was invasive DFS at two years. Treatment with neratinib statistically significantly increased the invasive DFS (stratified log-rank test $P = .0091$). The relative benefit of extending adjuvant therapy by using neratinib was evident because there was a 33% reduction of the hazard of invasive breast cancer events or death (HR = 0.67). However, the absolute benefit in survival probability was even smaller than the one seen for erlotinib in advanced pancreatic cancer: the probability of staying alive and free from the invasive breast cancer events of interest after two years was larger by 2.3% for the neratinib arm. The survival curves had not reached below 50%, and hence median survival times could not be estimated nonparametrically (ie, using the Kaplan-Meier method) in these two trials that studied early-stage disease.

The contrast between the measures of treatment effect in these two different settings is made with no judgment as to the comparative benefit accrued by patients, regardless of the primary end points. However, the comparison makes the point that a better view of trial results can be obtained by considering various measures. As shown in Table 2, not all measures are readily available, either because they have not been reported by the original authors (eg, the [restricted] mean survival time) or because they cannot be estimated nonparametrically in some cases (eg, the median survival). This contrast also exemplifies the differences between the advanced-disease and adjuvant settings with regard to the relationship between measures. Novel measures can be used to assess these and other trials, as will be shown below.

Measures to Quantify Treatment Benefit, Other Than the Hazard Ratio

Absolute Measures

Absolute measures contrast the survival of two groups of patients by computing a difference. These measures attempt to indicate the actual benefit a new patient can expect from the use of an intervention that proves superior to another in a randomized trial. It is generally accepted that most absolute measures can inform individual decisions in clinical practice more accurately than relative measures by conveying results in a manner that improves patient understanding of trial results (33). However, even absolute measures describe the "average" experience of a patient in the trial, and as such they may not apply to any given individual.

The Difference in Survival Probabilities at Time t

A survival probability is sometimes called an event rate, but we will not use this term here to avoid confusion with its connotation in epidemiology (with patient-years used in the denominator) or other fields. Given a chosen point in time (t, for simplicity) of the Kaplan-Meier survival curve (hereafter, just "survival curve"), the survival probability in each group that corresponds to t can be read on the y-axis. This technique, which is one of the versions of the "snapshot method" (10), is straightforward and describes the probability of survival at time t for patients in the trial. Hence, the choice of t influences the

difference in survival probabilities when two treatments are compared. Also, by considering only a single or a few points in time, there is loss of information, and one may not be able to cover the entire range of differences in the survival probabilities compatible with the hazard ratio (see Figure 1). Moreover, depending on the study follow-up period and the hazard ratio, one may or may not be able to observe the maximum survival-probability difference. The problem becomes even more acute when hazard rates are nonproportional, an apparently increasing phenomenon in an age of molecularly defined subsets of patients and immunomodulating agents (22). The loss of information may be eliminated by looking at the whole survival experience. In this regard, "risk-difference curves" may provide a longitudinal assessment of the expected absolute benefit from an intervention over time (21). Risk-difference curves plot the absolute differences between the survival probabilities over time (as in Figure 1C), and 95% confidence bands can be computed for such curves. We will not discuss here the "number needed to treat" (to avoid one event by time *t*), which is simply the inverse of the absolute difference between the survival probabilities at time *t*.

On a more technical note, it is important to realize that the relationship between the difference in survival probabilities and the hazard ratio is a complex one. In particular, even if hazards are proportional, the absolute value (in the mathematical sense) of the difference initially increases from 0 up to a maximum value, and then decreases again toward 0 (see Figure 1).

The Difference in Median Survival Times

Another version of the snapshot method is the computation of differences in median survival times, with the median represented by time t_{med} on the survival curve that corresponds to the 50% survival probability on the *y*-axis for each treatment. The oncology community clearly values this measure of treatment effect (15–17,34–36). Although easy to compute, intuitive for clinicians, and probably the easiest measure to remember, the difference in medians suffers from the same disadvantages highlighted for absolute differences in survival probabilities, and its use leads to loss of information (see Table 1). Moreover, even if the hazards are proportional, the ratio of median times is equal to the hazard ratio only when the survival distribution is exponential. Thus, the relationship between the difference in medians and the hazard ratio is, in general, complex. Consequently, even in the presence of proportional hazards, the magnitude of the difference in medians may not adequately correspond to the magnitude of the overall treatment effect, that is, the hazard reduction implied by the hazard ratio, unless the survival distribution is exponential (37). Moreover, the difference in medians suffers from additional limitations. Unless the survival curve drops below 50%, the median cannot be estimated from the curve. Hence, the estimate is not always available, especially for more indolent tumors, such as chronic lymphocytic leukemia (38), or in the adjuvant setting (32). Also, for end points other than OS, the difference in medians may be affected by the schedule of assessment. This is common in the assessment of PFS in second- or third-line therapy, when an active agent (as judged by statistically significant hazard ratios) displays a median PFS that is nearly identical, for example, to that of placebo/best supportive care (39,40). Finally, the median survival estimate is overly affected by heavy censoring of patients with short follow-up and is generally statistically unstable because its standard error is quite large for commonly used sample sizes of a few hundred patients, a fact that is almost always ignored when the median is reported.

Given the fact that the median is a single point on the curve, Kiely et al. have used multiples of the median survival to derive best case, typical, and worst case scenarios that could be used as a tool to discuss prognosis with breast cancer patients (36). This method is an improvement over snapshot methods, but it relies on the median and is therefore also subject to the problems just discussed.

The Difference Between Restricted Mean Survival Times

Because survival time generally has a skewed distribution and because the median can easily be read off survival curves, the mean survival time has long been neglected as a measure of central tendency in survival analysis. If the survival curve reaches 0 (ie, 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 restricted mean survival time by restricting (or truncating) the follow-up to a given time *t* and analyzing the data only up to that point (9,10,26). This method, first proposed in 1949, in fact measures the average time survived by patients over the period of interest. The restricted time *t* can be chosen arbitrarily, but it is usually taken equal to the minimum of the largest observed event time on each of the two groups in order to make full use of all the information available. The restricted mean survival for each group is the area under its survival curve through time *t* (10,41). Once the restricted means in both groups are computed, they may be contrasted by subtraction; the difference between restricted means is the area between the two survival curves through time *t*. As a result, the difference of restricted mean survival times measures the mean gain in life expectancy through time *t* associated with the superior treatment. For this reason, it is a relevant measure for patients because it tells how much longer a patient receiving the superior treatment is expected to live, on average, through time *t*, than a patient treated with the inferior intervention. On the other hand, the difference of restricted mean survival times may lead to cumbersome discussions with patients, who may not easily grasp the fact that the expected (or restricted mean) survival only applies to a fixed time horizon.

One may also compare the ratio between the areas under the survival curves for the two interventions (see below). Importantly, the use and interpretation of the difference of restricted mean survival times does not depend on whether hazards are proportional or not (42). On the other hand, restricted means are very seldom reported, and their interpretation is difficult when survival probabilities are far from 0 at time *t*. Several authors have found that differences of restricted means can add an absolute dimension to the treatment effect estimated by the hazard ratio, thus being advantageous for clinical decisions (6,9,26). Moreover, the ratio of restricted mean survival times should be equal to the hazard ratio when the survival distribution is exponential, but it has been shown empirically in a review of 54 trials that the hazard ratio tends to overestimate the ratio of restricted mean survival times, thus suggesting an artificially larger treatment benefit when hazards are not proportional (9).

The difference of restricted means between the neratinib and placebo arms in the trial discussed above was equal to only 0.5 months at two years (24) (Table 2). Although this difference appears small, it should be viewed in the context of the

maximum achievable benefit that would be accrued in theory had all patients in the trial been alive and disease-free at two years (43). This maximum achievable benefit was only 1.06 months; hence, the 0.5-month difference in restricted means represents nearly 50% of the maximum possible gain at this follow-up time of two years.

The Difference Between (Unrestricted) Mean Survival Times

As has been already mentioned, nonparametric estimation of the mean survival time requires that the survival curve reaches 0. Because this rarely happens, nonparametric estimates of the mean survival time are usually not reported. However, the mean can be estimated if one is willing to make a parametric assumption regarding the distribution of the survival time (eg, that the time follows a Weibull distribution). Parametric analyses are a standard approach for uncensored continuous data. For censored data, they are also frequently used, for example, in engineering (where the survival analysis is termed reliability analysis). In contrast, in medicine, nonparametric analysis has become the standard, due largely to the availability of methods like the Kaplan-Meier survival curve, the log-rank test, and the Cox proportional hazards model. However, the use of parametric models offers several advantages over nonparametric analyses, such as an increase in power (44) or direct availability of estimates for various characteristics (mean, median, etc.) of the survival-time distribution. Moreover, parametric survival time models include models that do not require the proportional hazards assumption, which is an attractive feature given the previously mentioned issues with that assumption (22). On the other hand, the tail of the survival distribution is often not observed because of censoring, so the model fit can only be assessed to the data thus far, and the (unrestricted) mean depends on extrapolation.

The important issue in this type of analysis is the choice of the parametric distribution. However, there is a range of diagnostic tools that can be used to select a suitable parametric model and check its fit to the data (44,45). If a suitable model is selected, the difference in the mean survival times can be easily estimated. The difference has an intuitive interpretation as the mean gain in life expectancy associated with the superior treatment. It is thus a relevant measure for individual patients because it tells how much longer a patient receiving the superior treatment is expected to live, on average, as compared with a patient treated with the inferior intervention.

For the four trials reported in Table 2, the Weibull model was found to fit well to the reported survival curves. Table 2 presents the model-estimated mean differences. For the two pancreatic cancer trials, the mean difference is close to the difference in the means restricted to 18 months. This is due to the fact that both trials had sufficiently long follow-up. Hence, the observed survival curves are close to 0 at 18 months. The (unrestricted) mean differences provide unambiguous and easily interpretable information: for example, for the gemcitabine trial (27), the expected gain in survival time is about one month for patients treated with the combination of erlotinib and gemcitabine as compared with patients treated with gemcitabine alone. On the other hand, for the breast cancer trials, the unrestricted mean differences are remarkably dissimilar from the differences in the means restricted to 24 months. This is due to the fact that in the two trials, the follow-up period was short with regard to the natural history of these conditions. As a result, there were very few observed events and the survival curves are still very far from 0 by 24 months. Consequently, the restricted-mean differences account only for a small portion of the difference over the entire life span of the patients. It is also worth noting that, in this case, despite their considerable magnitude, the unrestricted mean differences are rather imprecisely estimated and, in fact, statistically not significant. This, again, is due to the short follow-up and low number of events. Given this limited amount of information, it appears that alternative parametric models (eg, log-logistic or log-normal) provide a very similar fit to the survival curves as compared with the Weibull model (data not shown). These alternative models would yield quite different (unrestricted) mean differences. Nevertheless, in all cases, the differences are not statistically significant (data not shown). In this respect, the two breast cancer studies illustrate difficulties in inference regarding the long-term survival time characteristics based on nonmature data rather than issues with the suitability of treatment-effect measures such as differences in mean or median survival times.

The Net Benefit

The net benefit, also called "the net chance of a longer survival" when survival is the only outcome of interest, is a recently proposed measure of treatment effect. It is denoted by Δ and defined as the probability that a random patient in one of the treatment groups survives longer (or longer by at least an amount of time considered to be clinically relevant) than a random patient in the other group minus the probability of the opposite situation (12). Δ is equal to 0 if there is no treatment effect, it would be equal to $+1$ (100% in favor of treatment) if all patients in the treatment group fared better than all patients in the control group, and it would be equal to -1 (100% in favor of control) if all patients in the control group fared better than all patients in the treatment group. Hence positive and negative signs indicate the direction of the effect, with a positive Δ indicating that the new treatment is better than control. For example, if Δ is estimated to be 0.10 in a comparison between a new treatment and control, a random patient in the treatment group would have a 10% higher probability of longer survival than a random patient in the control group. Interestingly, Δ can be computed directly from the hazard ratio in an idealized situation with no censoring and proportional hazards ($\Delta = [1-HR]/[1+HR]$) (12). This simple relationship can easily be adjusted in the presence of censoring, but not when hazards are nonproportional (in which case the hazard ratio is a function of time). It has been suggested that this simple transformation of the hazard ratio (equivalent to the net benefit in situations of proportional hazards) can be used to enhance communication with clinicians (37,46). Figure 1E displays the net benefit in a fictitious example with a constant hazard ratio of 0.5, which illustrates that this probabilistic measure of treatment effect is constant over time. Importantly, the net benefit remains interpretable when hazards are nonproportional (12).

The net benefit may be more relevant for an individual patient than other absolute measures because it directly answers the question "What is my net chance of surviving longer with treatment A than with treatment B (by a chosen amount of time)?" For example, in the erlotinib trial in advanced pancreatic cancer, the net benefit was around 5% for differences in survival of up to six months (Table 2) (18). A patient who is told that their chance of surviving six months longer from the addition of erlotinib to gemcitabine is only around 5% may not perceive this treatment as worthwhile, especially in view of its toxicity. Conversely, the net benefit hovered around 25% in favor of FOLFIRINOX (vs gemcitabine) for differences in OS of up

to six months (47). But here too, a complete assessment of the clinical value of adding erlotinib to gemcitabine requires consideration of other end points, especially the added treatment toxicity (27). This can be done using a more sophisticated version of the net benefit that allows for assessing several outcomes of interest in a single analysis; although this discussion is beyond our scope here, analyses considering the net benefit (longer survival or, failing an improvement in survival, lower toxicity) favored placebo over erlotinib (18), but continued to greatly favor FOLFIRINOX over gemcitabine (47).

Relative Measures

Relative measures use ratios to describe differences in the survival experience of two groups of patients, taking into account the totality of the data and not individual arbitrarily chosen time points or survival probabilities. The hazard ratio, already discussed, is the most commonly used relative measure in clinical trials, and others are discussed below (but not calculated in Table 2).

The Ratio of Restricted Mean Survival Times

As mentioned, it is possible to compare the survival experience of two groups by computing the ratio of their restricted means (ie, the ratio of the areas under their survival curves) (9). Essentially, most of the advantages and disadvantages identified for differences in restricted means apply to their ratio (see Table 1). Arguably, however, the ratio of restricted means is not directly relevant to individual patients.

The Win Ratio

The win ratio has been proposed as an alternative measure of treatment effect based on a similar approach as the one used for the net benefit. The net benefit expresses the treatment effect on an absolute scale (difference between the probability of a better outcome on treatment minus the probability of a better outcome on control), while the win ratio expresses the treatment effect as a ratio (ratio of those two probabilities) (13). The win ratio was primarily proposed to compare treatments of cardiovascular disease as an alternative measure of treatment effect when several outcome measures are combined to form composite end points. As discussed above, the net benefit can be similarly extended when several outcome measures are simultaneously of interest (12).

Relevance for the Individual Patient

When comparing different measures, one of the salient issues is the extent to which they are relevant for individual patients. None of the existing measures is reliable in predicting the outcomes of individual patients because doing that would entail at least the use of covariates that are prognostic for the events of interest, thus allowing a prediction of the outcome of a given patient by comparing this with "matched" individuals. Alternatively, prognostic nomograms provide this type of individualized information, but they are often derived from observational data, without incorporating information on the treatment effects of specific interventions. If the issue of covariates is left aside, however, absolute measures can be considered relevant to inform clinical practice; although they have the caveat of describing the expected ("average") survival of a patient, they incorporate information about the time and probability of survival. This kind of information is arguably easier to interpret than relative measures that are not directly related to the probabilities of survival at given time points. Among the absolute measures, the recently proposed net benefit is the closest to addressing a patient-centered question because it incorporates both a time dimension (the difference in survival thought to be of interest, which may differ from patient to patient) and the net probability of enjoying a survival longer by at least this difference if treated (12). The net benefit can be computed using individual patient data from randomized clinical trials (48).

Conclusion

The treatment benefit in a given trial, even when only the primary end point is considered, can be difficult to understand by analyzing the results using a single measure (6,26,42,49). There are many examples, including the two we discuss above (18,32), in which analyses using different methods have provided different perspectives when compared with the primary results of a randomized trial. It is hoped that decisions can be made more rationally if such different perspectives are taken into account. Moreover, the caveats of using relative measures, especially the hazard ratio, have to be acknowledged (4,26,33). As pointed out by Royston and Parmar, the hazard ratio is incomplete as an outcome measure because it lacks an absolute component, and thus needs to be complemented by other measures, such as median survival, survival probabilities at specific time points, or the restricted mean (26). Unfortunately, means (whether restricted or not) and the more recently proposed measures are not usually available in publications and cannot be readily derived by busy clinicians. Although none of the existing measures are able to predict outcomes for individual patients, absolute measures can be more relevant than relative measures to inform clinical practice, and the net benefit is the closest to addressing a patient-centered question in treatment decisions. As have others in the field (13,26,42), we would encourage statisticians and clinical scientists to include novel measures in the analysis and reports of phase III trials, including the mean survival times and the net benefit.

Funding

There was no funding source for this work.

Notes

The authors have no conflicts of interest to disclose.

References

1. Punt CJ, Buyse M, Kohne CH, et al. Endpoints in adjuvant treatment trials: a systematic review of the literature in colon cancer and proposed definitions for future trials. *J Natl Cancer Inst* 2007;99(13):998–1003.
2. Saad ED, Katz A, Buyse M. Overall survival and post-progression survival in advanced breast cancer: a review of recent randomized clinical trials. *J Clin Oncol* 2010;28(11):1958–62.
3. Subramanian J, Madadi AR, Dandona M, et al. Review of ongoing clinical trials in non-small cell lung cancer: a status report for 2009 from the ClinicalTrials.gov website. *J Thorac Oncol* 2010;5(8):1116–9.
4. Bobbio M, Demichelis B, Giustetto G. Completeness of reporting trial results: effect on physicians' willingness to prescribe. *Lancet* 1994;343(8907):1209–11.
5. Chao C, Studts JL, Abell T, et al. Adjuvant chemotherapy for breast cancer: how presentation of recurrence risk influences decision-making. *J Clin Oncol* 2003;21(23):4299–305.
6. Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol* 2014;32(22):2380–5.

7. Case LD, Kimmick G, Paskett ED, et al. Interpreting measures of treatment effect in cancer clinical trials. *Oncologist* 2002;7(3):181–7.
8. Blagoev KB, Wilkerson J, Fojo T. Hazard ratios in cancer clinical trials–a primer. *Nat Rev Clin Oncol* 2012;9(3):178–83.
9. Trinquart L, Jacot J, Conner SC, et al. 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* 2016;34(15):1813–9.
10. Seruga B, Pond GR, Hertz PC, et al. Comparison of absolute benefits of anticancer therapies determined by snapshot and area methods. *Ann Oncol* 2012; 23(11):2977–82.
11. Royston P, Parmar MK, Altman DG. Visualizing length of survival in time-to-event studies: a complement to Kaplan-Meier plots. *J Natl Cancer Inst* 2008; 100(2):92–7.
12. Peron J, Roy P, Ozenne B, et al. The Net Chance of a Longer Survival as a Patient-Oriented Measure of Treatment Benefit in Randomized Clinical Trials. *JAMA Oncol* 2016;2(7):901–5.
13. Pocock SJ, Ariti CA, Collier TJ, et al. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur Heart J* 2012;33(2):176–82.
14. Ocana A, Tannock IF. When are "positive" clinical trials in oncology truly positive? *J Natl Cancer Inst* 2011;103(1):16–20.
15. Sobrero AF, Pastorino A, Sargent DJ, et al. Raising the bar for antineoplastic agents: how to choose threshold values for superiority trials in advanced solid tumors. *Clin Cancer Res* 2015;21(5):1036–43.
16. Schnipper LE, Davidson NE, Wollins DS, et al. American Society of Clinical Oncology Statement: A Conceptual Framework to Assess the Value of Cancer Treatment Options. *J Clin Oncol* 2015;33(23):2563–77.
17. Cherny NI, Sullivan R, Dafni U, et al. A standardised, generic, validated approach to stratify the magnitude of clinical benefit that can be anticipated from anti-cancer therapies: the European Society for Medical Oncology Magnitude of Clinical Benefit Scale (ESMO-MCBS). *Ann Oncol* 2015;26(8):1547–73.
18. Peron J, Roy P, Ding K, et al. Assessing the benefit-risk of new treatments using generalised pairwise comparisons: the case of erlotinib in pancreatic cancer. *Br J Cancer* 2015;112(6):971–6.
19. Pocock SJ, Stone GW, Mehran R, et al. Individualizing treatment choices using quantitative methods. *Am Heart J* 2014;168(5):607–10.
20. Lin NX, Logan S, Henley WE. Bias and sensitivity analysis when estimating treatment effects from the cox model with omitted covariates. *Biometrics* 2013;69(4):850–60.
21. Coory M, Lamb KE, Sorich M. Risk-difference curves can be used to communicate time-dependent effects of adjuvant therapies for early stage cancer. *J Clin Epidemiol* 2014;67(9):966–72.
22. Royston P, Parmar MK. An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect. *Trials* 2014;15:314.
23. Mok TS, Wu YL, Thongprasert S, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med* 2009;361(10):947–57.
24. Hasegawa T, Uno H, Wei LJ. Neratinib after trastuzumab in patients with HER2-positive breast cancer. *Lancet Oncol* 2016;17(5):e176.
25. Hernan MA. The hazards of hazard ratios. *Epidemiology* 2010;21(1):13–5.
26. Royston P, Parmar MK. 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* 2013;13:152.
27. Moore MJ, Goldstein D, Hamm J, et al. Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: a phase III trial of the National Cancer Institute of Canada Clinical Trials Group. *J Clin Oncol* 2007;25(15):1960–6.
28. Conroy T, Desseigne F, Ychou M, et al. FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. *N Engl J Med* 2011;364(19):1817–25.
29. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, et al. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *N Engl J Med* 2005;353(16): 1659–72.
30. Gianni L, Dafni U, Gelber RD, et al. Treatment with trastuzumab for 1 year after adjuvant chemotherapy in patients with HER2-positive early breast cancer: a 4-year follow-up of a randomised controlled trial. *Lancet Oncol* 2011; 12(3):236–44.
31. Smith I, Procter M, Gelber RD, et al. 2-year follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer: a randomised controlled trial. *Lancet* 2007;369(9555):29–36.
32. Chan A, Delaloge S, Holmes FA, et al. Neratinib after trastuzumab-based adjuvant therapy in patients with HER2-positive breast cancer (ExteNET): a multicentre, randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet Oncol* 2016;17(3):367–77.
33. Zipkin DA, Umscheid CA, Keating NL, et al. Evidence-based risk communication: a systematic review. *Ann Intern Med* 2014;161(4):270–80.
34. Saltz LB. Progress in cancer care: the hope, the hype, and the gap between reality and perception. *J Clin Oncol* 2008;26(31):5020–1.
35. Ajani JA. The area between the curves gets no respect: is it because of the median madness? *J Clin Oncol* 2007;25(34):5531.
36. Kiely BE, Soon YY, Tattersall MH, et al. How long have I got? Estimating typical, best-case, and worst-case scenarios for patients starting first-line chemotherapy for metastatic breast cancer: a systematic review of recent randomized trials. *J Clin Oncol* 2011;29(4):456–63.
37. Moser BK, McCann MH. Reformulating the hazard ratio to enhance communication with clinical investigators. *Clin Trials* 2008;5(3):248–52.
38. Burger JA, Tedeschi A, Barr PM, et al. Ibrutinib as Initial Therapy for Patients with Chronic Lymphocytic Leukemia. *N Engl J Med* 2015;373(25):2425–37.
39. Van Cutsem E, Peeters M, Siena S, et al. Open-label phase III trial of panitumumab plus best supportive care compared with best supportive care alone in patients with chemotherapy-refractory metastatic colorectal cancer. *J Clin Oncol* 2007;25(13):1658–64.
40. Grothey A, Van Cutsem E, Sobrero A, et al. Regorafenib monotherapy for previously treated metastatic colorectal cancer (CORRECT): an international, multicentre, randomised, placebo-controlled, phase 3 trial. *Lancet* 2013; 381(9863):303–12.
41. Guyot P, Ades AE, Ouwens MJ, et al. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol* 2012;12:9.
42. A'Hern RP. Restricted Mean Survival Time: An Obligatory End Point for Time-to-Event Analysis in Cancer Trials? *J Clin Oncol* 2016;34(28):3474–6.
43. Chan A, Buyse M, Yao B. Neratinib after trastuzumab in patients with HER2-positive breast cancer - Author's reply. *Lancet Oncol* 2016;17(5):e176–7.
44. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data (Second Edition)*. Hoboken: Wiley, 2002.
45. Collett D: *Modelling Survival Data in Medical Research (Second Edition)*. Boca Raton: Chapman & Hall/CRC, 2003.
46. Buyse M. Reformulating the hazard ratio to enhance communication with clinical investigators. *Clin Trials* 2008;5(6):641–2.
47. Peron J, Roy P, Conroy T, et al. An assessment of the benefit-risk balance of FOLFIRINOX in metastatic pancreatic adenocarcinoma. *Oncotarget* 2016; 13; 7(50):82953–82960.
48. Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Stat Med* 2010;29(30):3245–57.
49. Chappell R, Zhu X. Describing Differences in Survival Curves. *JAMA Oncol* 2016;2(7):906–7.