Published in final edited form as:

Urol Oncol. 2015 March ; 33(3): 122–127. doi:10.1016/j.urolonc.2014.11.002.

Measures of survival benefit in cancer drug development and their limitations

Ruth Etzioni¹, Roman Gulati¹, and Daniel W Lin^{1,2}

¹Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington

²University of Washington School of Medicine, Seattle, Washington

Abstract

The investigator comparing outcomes across arms of a drug development trial has available a variety of measures of survival benefit. In this article we systematically review the most common measures of comparative survival used in published studies. We distinguish between relative and absolute survival differences, and measures of instantaneous and cumulative risk. We consider settings in which the endpoint is overall survival as well as those in which disease-specific endpoints are of primary interest. We note that different measures capture different aspects of benefit and some may be more reliable than others or more representative of clinically relevant benefit. Rather than simply using procedures that have become standard, analyses should identify the most clinically relevant measures of impact and apply procedures that reliably estimate these.

Introduction

Suppose that we wish to evaluate a cancer treatment in a clinical trial in which the primary endpoint is the time to an event of interest. In cancer clinical trials, the primary interest is often time to death, but in some settings, we may be interested in cause-specific death or cancer progression.

Time to event as an endpoint raises many issues that have been topics of extensive methodological research. A key issue is that we invariably have incomplete follow-up and/or competing events that preclude the endpoint of interest. Some observations are therefore incomplete or right censored, where the qualifier “right” highlights that the missing information occurs on the right-hand-side of the event timeline.

© 2014 Published by Elsevier Inc.

Correspondence to: Ruth Etzioni, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, M2-B230, P.O. Box 19024, Seattle, WA 98109-1024, USA. Tel: +1.206.667.6561. Fax: +1.206.667.7264. retzioni@fhcrc.org.

Conflict of interest: None

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Because of censoring, we have different ways to summarize a set of observations and make comparisons between treatment groups. We don't use means and variances because unbiased estimates of these quantities are generally impossible to obtain when there is censoring. Instead, we examine two different types of measures.

The first type of measure is the instantaneous risk of the event **at** any point in time. This is referred to as the “hazard.” The second is a cumulative quantity, the chance of the event happening **by** any given point in time. We most frequently use the complement of this cumulative risk, namely the chance that the event has **not** happened **by** a given point in time, also called the “survival.” Most statistical inference procedures used in drug development are based on these measures. Statistical tests include the log-rank test and Cox regression analysis, which compare hazards. Survival is arguably easier to interpret, and most clinically relevant estimates of the impact of treatments, such as the number needed to treat (NNT), are based on survival.

In this article, we examine these measures and explain their limitations. We do not provide theoretical details or formulas for how the different metrics are calculated; the reader is assumed to have basic familiarity with concepts of censoring, survival curves, Kaplan-Meier estimates, and Cox regression. We will focus on concepts, interpretation, and limitations, illustrating these with examples from simulated datasets and published studies. We partition our review into two parts. The first part addresses analysis of all-cause mortality, and the second part addresses analysis of disease-specific mortality, progression, and other non-mortality endpoints. In each part, we will separately discuss comparisons of hazards and of survival.

Comparisons of all-cause mortality

Comparing hazards – relative comparisons of instantaneous risk

Many clinical trials are conducted primarily to evaluate the impact of a treatment on all-cause mortality. Comparison between treatment groups is typically conducted using the log-rank test or Cox regression analysis. These procedures effectively compare hazards between treatment groups, testing the hypothesis that the hazard of death is the same in each group. The analyses are designed to detect differences between groups assuming the treatment group hazard is a constant multiple of the control group hazard. While this multiple is assumed to be constant over time, these analyses make no assumptions about how the control group hazard changes over time.

Results are typically expressed as the ratio of these two hazards, presenting a relative perspective on the comparison of risks. And this ratio is typically a single number, implying that the pattern of risk over time is similar in the two groups, but that the level of risk may be different.

Summarizing the effect of a treatment using a hazard ratio that is constant over time may not be appropriate under long-term follow-up or if there is a delay in the treatment effect. For example, in the European Randomized Study of Screening for Prostate Cancer¹, the relative

risk for prostate-cancer death changed over time, from 0.85 in years 1–9 to 0.62 in years 10–11 due to delayed effects of screening on the risk of this endpoint.

While tests of hazards are most commonly used to evaluate evidence of treatment efficacy, estimates of survival are most commonly used to quantify clinical impact. For this purpose, interest focuses on differences in survival rather than on hazards.

Comparing survival – absolute comparisons of cumulative risk

Figure 1 shows two groups in a hypothetical randomized trial with a 12-month follow-up period. In each panel, the relative risk associated with treatment is 0.80, i.e., at any point in time, patients in the treatment group have a 20% lower risk of death than patients in the control group. In the first panel, 12-month survival in the control group is 40% and in the second panel it is 80%.

In both panels, the survival curves diverge and the difference between them increases over time. Even though the ratio of hazards is constant over time, the difference in survival will continue to increase over time. Thus, survival differences can be highly sensitive to the duration of follow-up. In addition, the magnitude of the survival differences depends on the baseline level of survival in the control group. This is demonstrated in Figure 1, which shows a much greater difference in survival after 12 months in the first panel than in the second panel. This is because there are more deaths in the control group without treatment in the first panel, and consequently there are more lives that can be saved by treatment.

The dependence of survival differences on baseline survival becomes critically important when computing a key measure of treatment benefit, the number needed to treat (NNT) to save one life. Consider the first panel in Figure 1. In the control group, 40 out of 100 cases are alive at the end of follow-up; in the treatment group, 48 out of 100 cases are alive at the end of follow-up. Therefore, there are 8 lives saved per 100 cases treated, and the NNT is 12.5 (100/8). In the second panel in Figure 1, 80 out of 100 cases survive 12 months in the control group compared with 84 out of 100 cases in the treatment group, so the NNT is 25 (100/4). In both panels, the treatment is equally efficacious from a relative hazard perspective, but because the baseline survival is different in the two panels, the two NNTs over the same period are quite different.

Reporting the NNT without reference to baseline survival can make interpretation challenging. In some cases, a high NNT may be incorrectly interpreted as indicating lack of efficacy. Consider, for example, the two randomized trials of the efficacy of radical prostatectomy (RP) compared with watchful waiting. For the endpoint of prostate cancer death, the Scandinavian Prostate Cancer Group Study Number 4 (SPCG-4) trial yielded a hazard ratio associated with RP of 0.62 after a median of 12.8 years² whereas the US Prostate Cancer Intervention versus Observation Trial (PIVOT) yielded a hazard ratio associated with RP of 0.63³ after a median follow-up of 10 years. Although these results suggest similar treatment efficacy in the two settings, the PIVOT trial had a considerably lower number of prostate cancer deaths (8.4% in the control group) and an NNT to save one prostate cancer death of 38. In contrast, the SPCG-4 trial had a higher number of prostate cancer deaths (23.3% in the control group) and an NNT of 16. Thus, even though the two

trials produced similar results in terms of relative hazards, the NNT for the PIVOT trial was considerably higher; the lower baseline risk of death in that trial meant that there were fewer lives to save even under a hazard ratio that was almost the same as in the SPCG-4 trial. A recent study that examined these results and attempted to reconcile them⁴ concluded that differences in the prevalence of PSA screening in the two case populations largely explained the discrepancies between the trials in the observed frequency of prostate cancer deaths and survival.

Given the sensitivity of the survival difference and the NNT to follow-up duration, it is reasonable to ask whether there is a single measure that will characterize differences in the survival experience between two comparison groups.

A commonly used statistic is median survival; the difference in median survival between treatment groups is often cited as a summary measure of efficacy. The median is used because it is a measure of the center of the distribution of survival times; by definition, half of the survival times lie below and half above the median. In survival analysis, the median of a sample corresponds to the point at which the survival curve reaches 50%. Thus, median survival can only be calculated if follow-up is long enough and the risk of death is high enough that the survival curve crosses the 50% mark. This occurs for both groups in the first panel in Figure 1 but not in the second panel. The risk of death is simply not high enough and the follow-up is not adequate to estimate median survival in the second panel. Even if median survival can be estimated in both groups, the difference between the medians may not be a reliable summary of the separation between two survival curves. This can occur if median survival is reached relatively late in the follow-up period, when few patients remain under observation. In the PREVAIL study⁵, a phase 3 trial of enzalutamide versus placebo in men with metastatic, castrate-resistant prostate cancer, the median overall survival was 32.4 months in the enzalutamide group and 30.2 months in the control group. However, the median survival among men on the enzalutamide and placebo arms was reached late in the course of follow-up, when only a handful of patients remained under observation in both groups.

What do we do when we cannot reliably estimate median survival in one or both treatment groups? There is no established answer. However, we note that the median is only one of the quantiles of the distribution of survival times, where the k th quantile (or percentile) corresponds to point below which $k\%$ of the survival times lie above it and $(1-k)\%$ lie below it. In the case where median survival is not reached by the end of follow-up, it may be informative to compare a different quantiles across groups; for example, if only 20% of the survival times are estimated to occur by the end of follow-up, comparison of survival times at the 10th percentile may be of value. In the case of the PREVAIL trial⁵, where approximately 50% of patients died during the course of follow-up, comparison of the 20th or 25th percentiles of overall survival in the enzalutamide and control group yields a difference of approximately 5 months, which is considerably greater than the observed difference in median overall survival in the two groups. An alternative would be to compare the fraction surviving beyond a clinically meaning interval determined a priori, e.g., 6-month or 12-month survival estimates in a trial of treatments for advanced or metastatic cancer.

Although standards have not been established for specific quantiles (besides the median) to use in practice, these measures may be informative in many settings. Ultimately, the question of interest in drug development relates to how the drug shifts the distribution of survival times in the treatment group relative to the control group. The nature of this shift can be discerned by comparison of multiple quantiles between the treatment and control groups. Even when medians are observed in both groups, it may be of value to examine more than one quantile (e.g., 10% and 25% in addition to 50%) and determine whether any differences observed are consistent across quantiles. In the third panel of Figure 1, the median survival is similar in the two groups but survival times are longer in the treatment group at the 25th percentile. This suggests that the impact of the treatment is to modestly postpone mortality for the highest-risk patients but not (or less so) for those with more moderate risk. This may also be the case for enzalutamide, given that, in the PREVAIL trial, the difference in overall survival at the 25th percentile considerably exceeded the difference in survival at the 50th percentile.

Comparisons of disease-specific mortality

In many drug development settings, the intervention of interest targets a disease-specific mortality endpoint. Examples include chemotherapies or hormonal treatments directed at specific cancers such as tamoxifen for localized ER-positive breast cancer or enzalutamide for advanced prostate cancer. In the setting of advanced disease, where the majority of patient deaths will be cause-specific, examination of all-cause mortality should provide a clear picture of the activity of the treatment on the disease. However, when other-cause death is common, examination of the cause-specific endpoint can provide insights regarding disease-related activity of the intervention without the signal being drowned by interfering noise from other-cause deaths.

Focusing on disease-specific endpoints leads to a problem known as competing risks, which happens when there are two endpoints and the occurrence of one (e.g., other-cause death) prevents the other (e.g., disease-specific death). Competing risks can also arise when examining the chance of disease progression or recurrence and even when considering risks of treatment complications that are not immediate, like second cancers after primary chemotherapy or radiation. In the presence of competing risks, disease-specific hazards can be calculated at any time point simply by considering the frequency of the endpoint among individuals who are still alive and under observation at that time. However, estimation of disease-specific survival in the presence of competing risks is rather less straightforward⁶, and there are a variety of options for doing so, each with its own caveats and cautions.

Competing risks pose a problem when analyzing the cumulative chance of disease-specific or treatment-related endpoints because other-cause mortality eliminates the opportunity to observe these endpoints. If other-cause death is frequent, for example in an elderly or frail population, then the cumulative chance of disease-specific death may appear to be quite low when in practice it is quite high among those surviving other causes of death. For example, in the PIVOT study⁷, only 8.4% of cases in the control group died of prostate cancer, but the chance of other-cause death was considerably higher among cases recruited into PIVOT

than in the general population. Thus, depending on how the data are analyzed, results may be impacted by the extent of competing risks in the comparison groups.

Comparing hazards

Many studies of disease-specific endpoints use the same approaches for comparing hazards as those used in the absence of competing risks, principally log-rank tests and Cox regression analysis. However, unlike in the case of all-cause mortality, the impact of treatment on the disease-specific hazard as assessed by these methods may not be consistent with the impact of treatment on crude, cause-specific survival. An example of this was evident in the serial publications^{2,8,9} reporting results of the SPCG-4 trial of radical prostatectomy versus watchful waiting for localized prostate cancer. Graphs of the cause-specific cumulative incidence of prostate cancer deaths showed almost zero impact of treatment among older men and a marked treatment benefit among younger men, suggesting differential efficacy by age. But a Cox regression analysis—which included an interaction term to test whether the efficacy of treatment on the hazard of death depended on age group—did not produce a statistically significant result. Specialized hypothesis tests^{10–12} address this potential inconsistency by directly testing the impact of treatment on the cumulative incidence of the disease-specific endpoint.

Comparing survival

There are two main approaches to examining survival when it comes to studying disease-specific death in a competing risks setting. The first relates to the cumulative fraction of individuals who died of disease. Individuals dying of other causes cannot die of disease; therefore, the cumulative fraction is always less than one. This metric for examining disease-specific death is called the “cumulative incidence.” In some applications (e.g., the SEER cancer registry¹³), this approach is reflected in crude survival, which is equal to one minus the cumulative incidence. In the setting of competing risks, the NNT is determined by estimating the lives saved based on the difference between cumulative incidence estimates.

Although the cumulative incidence is interpreted as a disease-specific measure, it depends on the risk of other-cause death as well as on the risk of disease-specific death. Thus, caution in interpreting the cumulative incidence is in order when the intervention affects both disease-specific and other-cause death. For example, suppose there is a toxic intervention that is quite effective in suppressing the growth of aggressive disease but it increases the risk of cardiovascular mortality. Then, in the intervention group, there will be fewer disease-specific deaths but more cardiovascular deaths. This will cause the cumulative incidence curve for disease-specific death in this group to appear even more favorable because cardiovascular deaths will reduce the number of patients observed to die of disease by more than would be expected from the impact of the intervention on the disease itself. Figure 2 demonstrates this phenomenon.

How, then, can one obtain an estimate of cumulative mortality risk (or survival) that pertains only to the disease-specific endpoint? Can one simply consider patients to be censored at the time of their other-cause death and use the corresponding Kaplan-Meier survival curve for the endpoint of disease-specific death? This question has been closely examined in the

statistical literature^{6,14}, and the resulting survival estimates are termed “net” (as opposed to “crude”) survival¹⁵. The consensus conclusion is that use of net survival can lead to biased results in certain settings, specifically, when patients who die of other causes happen to be those who are at higher or lower risk of disease-specific mortality than those who survive. The phenomenon is known as informative censoring. Figure 3 illustrates how informative censoring can lead to biased estimates of disease-specific survival when the net survival estimates (obtained by censoring the competing risks) are used.

Informative censoring can be problematic, particularly when considering endpoints such as complications of treatment, disease relapse, or receipt of secondary therapies, all of which are subject to competing risk of death. In the case of treatment complications, patients who are able to tolerate higher treatment doses will be more susceptible to complications, but they may also have a lower risk of death due to better control of their disease. In the case of disease relapse detected via surveillance testing (e.g., PSA failure in prostate cancer), patients observed to relapse will be more likely to have attended scheduled surveillance tests and may represent cases whose general health is better than those unable to attend scheduled tests. Similarly, patients who are able to undergo secondary therapies will likely be in better health than those who do not initiate these treatments. Censoring deaths when considering time-to-event analyses for these endpoints could yield severely biased results⁶.

In the case of disease progression or relapse, a third approach, progression-free survival, is commonly used to characterize the impact of a new treatment. Patients are considered to have failed if they either progress or die. This approach overcomes the competing risks issues that are present when considering disease progression as an endpoint. This may be the reason why it is so commonly used in practice. However, the problem with this approach is that it conflates the disease-specific endpoint with other-cause death and can only give an estimate of their combined risk. Consequently, despite its widespread uses, the clinical meaningfulness of progression-free survival remains questionable¹⁶.

In summary, a variety of metrics and methods are available to quantify the impact of a new treatment in the drug development setting. There are different survival measures and types of endpoints. In settings where there is a high risk of disease-specific death or the treatment has an adverse impact on patient health, all-cause mortality is likely to be the endpoint of choice. However, when disease-specific mortality is less frequent, the endpoint may need to be re-assessed. In the case of cancer screening, for example, where only a small minority of the population dies of disease even in the absence of screening and screening has little impact on non-cancer mortality, analysis of all-cause mortality is severely handicapped as other-cause deaths will almost surely overwhelm and obscure any impact of the screening test on disease-specific mortality.

Regardless of the endpoint, different procedures provide different perspectives on the impact of treatment. Cox regression models give a sense of relative differences (in terms of hazards) while estimates of the NNT relate to absolute differences in survival. Estimates from Cox models that pertain to hazards may not always be consistent with estimates that pertain to survival or cumulative risk of the endpoint of interest, particularly in the case of competing risks. Every procedure is subject to assumptions and caveats, and these should be

investigated to assure validity of the results. Rather than simply using procedures that have become standard, the analysis should determine the most clinically relevant measures of impact and apply procedures that validly estimate these.

Acknowledgments

Financial support: This work was supported by the National Cancer Institute at the National Institutes of Health (P50 CA097186).

Bibliography

- Schroder FH. Prostate-cancer mortality at 11 years of follow-up. *N Engl J Med.* 2012; 366:981–90. [PubMed: 22417251]
- Schroder FH. Prostate-cancer mortality at 11 years of follow-up. *N Engl J Med.* 2012; 366:981–90. [PubMed: 22417251]
- Wilt T, Brawer MK, Jones K, Barry M. Radical Prostatectomy versus Observation for Localized Prostate Cancer. *N Engl J Med.* 2012; 367:203–13. [PubMed: 22808955]
- Xia J, Gulati R, Au M, Gore JL, Lin DW, Etzioni R. Effects of screening on radical prostatectomy efficacy: the prostate cancer intervention versus observation trial. *J Natl Cancer Inst.* 2013; 105:546–50. [PubMed: 23411592]
- Beer TM, Armstrong AJ, Rathkopf DE, Loriot Y, Sternberg CN, Higano CS, Iversen P, Bhattacharya S, Carles J, Chowdhury S, Davis ID, de Bono JS, Evans CP, Fizazi K, Joshua AM, Kim CS, Kimura G, Mainwaring P, Mansbach H, Miller K, Noonberg SB, Perabo F, Phung D, Saad F, Scher HI, Taplin ME, Venner PM, Tombal B, Investigators P. Enzalutamide in metastatic prostate cancer before chemotherapy. *N Engl J Med.* 2014; 371:424–33. [PubMed: 24881730]
- Gooley TA, Leisenring W, Crowley J, Storer BE. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Stat Med.* 1999; 18:695–706. [PubMed: 10204198]
- Wilt, T. The VA/NCI/AHRQ CSP #407: Prostate Cancer Intervention Versus Observation Trial (PIVOT): Main Results From a Randomized Trial Comparing Radical Prostatectomy to Watchful Waiting in Men With Clinically Localized Prostate Cancer. *AUA Annual Meeting*; May 14–19; 2011;
- Bill-Axelson A, Holmberg L, Filen F, Ruutu M, Garmo H, Busch C, Nordling S, Haggman M, Andersson SO, Bratell S, Spangberg A, Palmgren J, Adami HO, Johansson JE. Radical prostatectomy versus watchful waiting in localized prostate cancer: the Scandinavian prostate cancer group-4 randomized trial. *J Natl Cancer Inst.* 2008; 100:1144–54. [PubMed: 18695132]
- Bill-Axelson A, Holmberg L, Garmo H, Rider JR, Taari K, Busch C, Nordling S, Haggman M, Andersson SO, Spangberg A, Andren O, Palmgren J, Steineck G, Adami HO, Johansson JE. Radical prostatectomy or watchful waiting in early prostate cancer. *N Engl J Med.* 2014; 370:932–42. [PubMed: 24597866]
- Dignam JJ, Kocherginsky MN. Choice and interpretation of statistical tests used when competing risks are present. *J Clin Oncol.* 2008; 26:4027–34. [PubMed: 18711194]
- Dignam JJ, Zhang Q, Kocherginsky M. The use and interpretation of competing risks regression models. *Clin Cancer Res.* 2012; 18:2301–8. [PubMed: 22282466]
- Kim HT. Cumulative incidence in competing risks data and competing risks regression analysis. *Clin Cancer Res.* 2007; 13:559–65. [PubMed: 17255278]
- Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER 9 Regs Limited-Use, Nov 2012 Sub (1973–2010), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2013, based on the November 2011 submission. (Accessed at
- Pepe MS, Mori M. Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Stat Med.* 1993; 12:737–51. [PubMed: 8516591]

15. National Cancer Institute. [Accessed August 29, 2014] Surveillance Research page on measures of survival. at <http://surveillance.cancer.gov/survival/measures.html>.)
16. Booth CM, Eisenhauer EA. Progression-free survival: meaningful or simply measurable? J Clin Oncol. 2012; 30:1030–3. [PubMed: 22370321]

Figure 1.

Overall survival in a hypothetical randomized trial. (A) Treatment lowers the hazard of death by 40% and 12-month survival in the control group is 40%. (B) Treatment lowers the hazard of death by 40% and 12-month survival in the control group is 80%. (C) Median survival in the two treatment groups is similar but survival times in the treatment group are longer at the 25th percentile.

Figure 2. Cumulative incidence of cancer death in a hypothetical randomized trial. (A) Competing risk of death is the same in both groups. (B) Competing risk of death is higher in the treatment group making the number of deaths observed in that group lower.

Figure 3.

Disease-specific survival in a hypothetical randomized trial when risks of cancer-specific and other-cause death are positively correlated. (A) True underlying disease-specific survival. (B) Kaplan-Meier estimates of disease-specific death when competing deaths are censored. Because of the positive correlation, the competing deaths are precisely those at higher risk of disease-specific death. The individuals remaining in the sample are at a lower risk of disease-specific death, thus the estimated survival curves are more favorable than the true ones.