ELSEVIER

CrossMark

Urologic Oncology: Seminars and Original Investigations 33 (2015) 122–127

UROLOGIC ONCOLOGY

Seminars article

# Measures of survival benefit in cancer drug development and their limitations

Ruth Etzioni, Ph.D.[a,*], Roman Gulati, M.S.[a], Daniel W. Lin, M.D.[a,b]

[a] *Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA*
[b] *University of Washington School of Medicine, Seattle, WA*

Received 11 October 2014; accepted 5 November 2014

**Abstract**

**Background:** A variety of measures of survival benefit are available to an investigator comparing outcomes across the various arms of a drug development trial. In this article, we systematically review the most common measures of comparative survival used in published studies.

**Materials:** We distinguish between relative and absolute survival differences, and measures of instantaneous and cumulative risk. We consider settings in which the end point is overall survival as well as those in which disease-specific end points are of primary interest.

**Results:** 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.

**Conclusions:** Rather than simply using procedures that have become standard, analyses should identify the most clinically relevant measures of effect and apply procedures that reliably estimate these. 

*Keywords:* Survival analysis; Competing risks; Disease-specific mortality

## Introduction

In certain cases, we may wish to evaluate a cancer treatment in a clinical trial in which the primary end point 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 an event as an end point raises many issues that have been topics of extensive methodological research. A key issue is that we invariably have incomplete follow-up or competing events that preclude the end point 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.

Because of censoring, we have different ways to summarize a set of observations and make comparisons between treatment groups. We do not use means and variances because unbiased estimates of these quantities are generally impossible to obtain when there is censoring. Instead, we examine 2 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 the Cox regression analysis, which compare hazards. Survival is arguably easier to interpret, and most clinically relevant estimates of the effect 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 the concepts of censoring, survival curves, Kaplan-Meier estimates, and

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

* Corresponding author. Tel.: +1-206-667-6561; fax: +1-206-667-7264.
*E-mail address:* retzioni@fhcrc.org (R. Etzioni).

http://dx.doi.org/10.1016/j.urolonc.2014.11.002
1078-1439/

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 2 parts. The first part addresses analysis of all-cause mortality, and the second part addresses analysis of disease-specific mortality, progression, and other nonmortality end points. In each part, we 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 effect 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. Although this multiple is assumed to be constant over time, these analyses make no assumptions about how the control group hazard changes over time.

The results are typically expressed as the ratio of these 2 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 both the groups but 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 for 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 Cancer [1], the relative risk for prostate cancer–related death changed over time, from 0.85 (1–9 y) to 0.62 (10 and 11 y), because of the delayed effects of screening on the risk of this end point.

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

### *Comparing survival—Absolute comparisons of cumulative risk*

Fig. 1 shows 2 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. The 12-month survival in the control group is 40% in the first panel and 80% in the second panel.

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 continues 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 Fig. 1, which shows a much greater difference in survival after 12 months in the first panel than seen 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 NNT to save 1 life. Consider the first panel in Fig. 1. In the control group, 40 of 100 patients are alive at the end of follow-up; in the treatment group, 48 of 100 patients are alive at the end of follow-up. Therefore, there are 8 lives saved per 100 cases

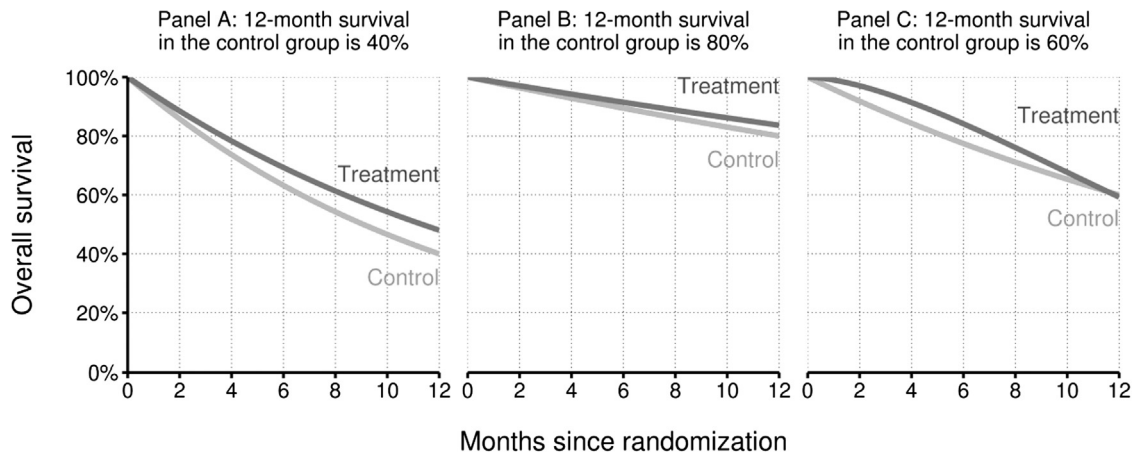


Fig. 1. Overall survival in a hypothetical randomized trial. (A) Treatment lowers the hazard of death by 40% and 12-month survival in the control group is 40%. (B) Treatment lowers the hazard of death by 40% and 12-month survival in the control group is 80%. (C) Median survival in the 2 treatment groups is similar but survival times in the treatment group are longer at the 25th percentile.