Biostatistics Primer

What a Clinician Ought to Know: Hazard Ratios

Helen Barraclough, MSc, Lorinda Simms, MSc, PStat,† and Ramaswamy Govindan, MD‡§*

Abstract: Hazard ratios (HRs) are used commonly to report results from randomized clinical trials in oncology. However, they remain one of the most perplexing concepts for clinicians. A good understanding of HRs is needed to effectively interpret the medical literature to make important treatment decisions. This article provides clear guidelines to clinicians about how to appropriately interpret HRs. While this article focuses on the commonly used methods, the authors acknowledge that other statistical methods exist for analyzing survival data.

Key Words: Biostatistics, Hazard ratio, Proportional hazards, Survival analysis.

(*J Thorac Oncol.* 2011;6: 978–982)

WHAT IS A HAZARD RATIO?

Hazard ratios are frequently used to estimate the treatment effect for time-to-event end points, such as overall survival (OS) and progression-free survival (PFS), in oncology randomized clinical trials (RCTs). A time-to-event analysis (also and from this point on called a survival analysis) analyzes the time from the start of a study (e.g., randomization) to an event (e.g., death for OS). Before discussing hazard ratios (HRs), key concepts related to understanding survival data are reviewed.

A Kaplan-Meier (KM) curve (also called a survival curve or a KM analysis) is used to estimate survival and presents survival data. The curve represents the proportion of patients event-free (e.g., alive for OS) at any time. The curve is not smooth, but actually a series of downward steps occurring each time a patient has an event (e.g., death in an OS curve). At any time point in an OS analysis, patients can have only one of the following events: (i) death, (ii) continue to be monitored, or (iii) stop being monitored (i.e., they are censored). Patients who are censored have not had an event as of the end of the observation period, and it is unknown whether they will have an event in the future. There are several reasons why censoring can occur; typically it is because patients are still alive at the end of the study or are lost to follow-up.[1] Reasons for censoring and censoring patterns should be closely examined as they may impact interpretation of the results. Regardless of how the censoring rules are defined in a given RCT, the intention is to incorporate all data collected on patients while they were participating in the study.

A HR provides an estimate of the ratio of the hazard rates between the experimental group and a control group over the entire study duration. The hazard rate is the rate of patients experiencing the event of interest over a short time interval within each of the treatment arms in the study. This concept can be illustrated by a hypothetical example: an RCT with two treatment arms and a primary end point of OS (Table 1). During the first week, the rate of patients dying is higher in the control arm (0.04) than in the experimental arm (0.03). In the second week, the rate of patients dying is double than that in the first week: 0.08 for patients in the control arm and 0.06 for patients in the experimental arm. The HR (experimental versus control) is calculated for each week by dividing the rate of patients dying in the experimental arm by the rate of patients dying in the control arm. Although the hazard rate changes over time, the HR is approximately constant ($\sim$0.75) for each week (Table 1). Hence, the HR reported for this RCT would be $\sim$0.75 because a constant HR is calculated over the entire duration of the trial.

The HR is usually calculated from a Cox proportional hazards model, which is one of the standard methods for analyzing survival end points in oncology RCTs.[2] A simplistic interpretation is that a HR = 1 means equal efficacy of the experimental and control treatments (Figure 1). Usually, the HR is presented so that if the experimental treatment is (i) better than the control then the HR <1 or (ii) worse than the control then the HR >1.

WHY ARE HAZARD RATIOS USEFUL?

The log-rank and Wilcoxon tests are commonly used to compare the entire survival data over the duration of the trial, between treatment arms; they do not compare the medians or time point estimates. However, these methods only generate a *p* value and not an estimate of the magnitude or direction of the treatment effect (i.e., these methods assess whether survival between the two treatment arms was significantly dif-

* Asia Pacific Statistical Sciences, Eli Lilly and Company, Sydney, Australia; †Eli Lilly Canada Inc., Toronto, Canada; ‡Division of Oncology, Department of Medicine, Washington University School of Medicine, St. Louis, Missouri; and §Alvin J Siteman Cancer Center at Washington University School of Medicine, St. Louis, Missouri.

Disclosure: Helen Barraclough and Lorinda Simms are employed by and own stock in Eli Lilly and Company. Ramaswamy Govindan has no conflicts of interest.

Address for correspondence: Ramaswamy Govindan, MD, Division of Medical Oncology, Washington University School of Medicine, 660 S. Euclid, Box 8056, St. Louis, MO 63110. E-mail: rgovinda@im.wustl.edu

ISSN: 1556-0864/11/0606-0978

TABLE 1. Concept of Calculating of a Hazard Ratio

Time Period	Treatment Arm	No. Patients at Risk[a] (B)	No. Died During Time Period (A)	No. Patients Dropped Out During Time Period[b]	Rate of Patients Dying[c] (Calculated as A/B)	Proportion of the Rate of Patients Dying in Experimental Arm Compared with Control Arm[d]
First week	Experimental	100	3	0	0.03	0.75
	Control	100	4	0	0.04	
Second week	Experimental	97	6	3	0.06	0.74
	Control	96	8	1	0.08	
Third week	Experimental	88	9	1	0.10	0.74
	Control	87	12	2	0.14	

To precisely calculate the hazard ratio, the Cox proportional hazards model must be used, which accounts for the censoring times.
[a] Patients who are alive and still in the study at the start of the time period.
[b] Censored patients.
[c] Hazard rate for the time period (results shown are rounded to 2 decimal places).
[d] Hazard ratio (E vs C) for the time period. Please note that results shown are rounded to 2 decimal places, but the calculations used the raw numbers from the previous column ([c]) and therefore give different results than if the rounded numbers were used (eg, 0.06/0.08 = 0.75).

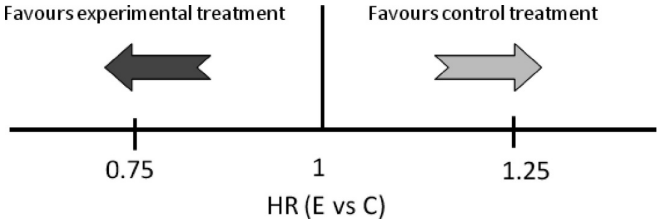

FIGURE 1. Simplistic interpretation of a hazard ratio (HR). HR = 1 means equal efficacy of the experimental (E) and control (C) treatments. If the experimental treatment is better than the control, then the HR (E versus C) <1. If the experimental treatment is worse than the control, then the HR (E versus C) >1.

BOX 1. Hazard Ratio vs Median Survival and Time Point Estimates

Hazard Ratio	Median Survival	Time Point Estimates
Relative efficacy measure	Absolute efficacy measure	Absolute efficacy measure
Summarizes the treatment effect over the entire duration of a RCT, i.e., uses all of the information in the whole KM curve	Compares the survival time at one point on the KM curve only	Compares the survival time at one point on the KM curve only (e.g., 1 yr)
Superiority and noninferiority trial designs are usually based on the HR	Claims of superiority can be misleading. Superiority and noninferiority trial designs are generally based on the HR and not the median	Claims of superiority can be misleading. Superiority and noninferiority trial designs are generally based on the HR and not an estimate at a specific time point
Unadjusted and adjusted HRs can be generated from univariate and multivariate Cox proportional hazards models, respectively	Unadjusted	Unadjusted
Calculated using a Cox proportional hazards model	Calculated by reading the survival time of a KM curve when the survival probability is 50%	Calculated by reading the survival probability on a KM curve at the time point of interest (e.g., 1 yr)

ferent but not how it was different). That is, the log-rank and Wilcoxon tests only determine whether the treatments are different (or not) but do not indicate how much one treatment is better or worse than the other.

The three main ways of estimating the magnitude and direction for survival outcomes from RCTs are (i) the HR, (ii) by reporting the median survival for each treatment arm, and (iii) time point analyses (e.g., proportion of patients alive at 1 and 2 years for each treatment arm in an OS analysis). Medians and time point estimates are often generated from the KM analysis (Box 1). Sometimes differences between treatments in medians, or in the time points of 1 or 2 years, are also presented. From these, the magnitude and direction of the treatment effect can be estimated. However, HRs differ from the other two measures in the following ways.

First, the HR summarizes all the information in the entire KM survival curves, and hence summarizes the treatment effect over the entire duration of a RCT (Figure 2). In contrast, median survival focuses on only one point on the survival curve for each treatment arm (Figure 2). In reality, the median survival is the expected (not observed) time when half of the patients will have had an event (i.e., died for OS) and half will not (i.e., still be alive for OS). The median OS estimates from prospective RCTs are unfortunately sometimes interpreted by individual patients as their precise estimated life expectancy with no corresponding interval for the expected variability in survival around that.[3] It is important to stress that median survival represents at best a "group average" and is overly simplistic as a measure of an individual patient's duration of disease control or OS. Hence, two lines on a KM curve can have the same median survival but look different before and after the median (Figure 3*C*).

Second, the HR provides an estimate of the relative efficacy between the treatment arms (e.g., HR = 0.75 for an OS end point means on average approximately a 25% lower risk of death on the experimental treatment than the control, Box 2). In contrast, comparing medians for each treatment

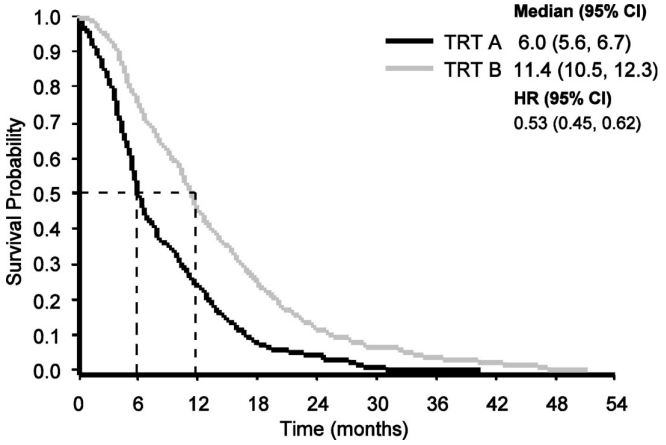

FIGURE 2. Median survival versus the hazard ratio (HR). Kaplan-Meier (KM) curves are used to graphically describe time-to-event (survival) outcomes in randomized clinical trials (RCTs). Above is a KM curve from a hypothetical oncology RCT of two treatment (Trt) arms (Trt A and Trt B) which evaluated overall survival (OS) as the primary end point. The KM curves plot the "events" that occur in each treatment arm. For OS, death is the event so the curve drops down a "step" each time a patient dies to reflect the new probability of still being alive at that point in time in each treatment arm. The HR summarizes all the time-to-event information described by the KM survival curve. In contrast, the median survival focuses on only one point on the KM curve, that is the survival time when survival probability is 50% (shown by the dotted lines).

arm provides an absolute measure of any improvement in efficacy (e.g., 2 months difference in the median survival) relative to a particular absolute median for the control group. Other absolute measures of improved efficacy between the treatment arms include the proportion surviving at defined time points (e.g., such as 1-year OS probability). Third, due to the two properties outlined above, claims of superiority and noninferiority are recommended to be made based on the HR and not the median. Time point analyses also encounter these three issues as described for the median survival.

Finally, adjusted and unadjusted HRs can be calculated. The aim of randomization is for both known and unknown prognostic factors to be balanced between the treatment arms. However, even when well-designed randomization techniques are conducted, imbalances may occur by chance alone. In such situations, analyses adjusted for known prognostic factors are recommended to account for the imbalances.[4] If there are no major imbalances in known prognostic factors, the adjusted and unadjusted analyses should produce similar results. Of note, although KM survival curves can be presented for subsets of patients with particular baseline prognostic factors, they are generally not adjusted for baseline prognostic factors (statistical methods are available for generating adjusted survival curves). Therefore, the median survival and survival probabilities at defined time points (e.g., 1-year OS probability) derived from KM survival curves are also unadjusted. The unadjusted HR summarizes all the information displayed graphically in the KM survival curves

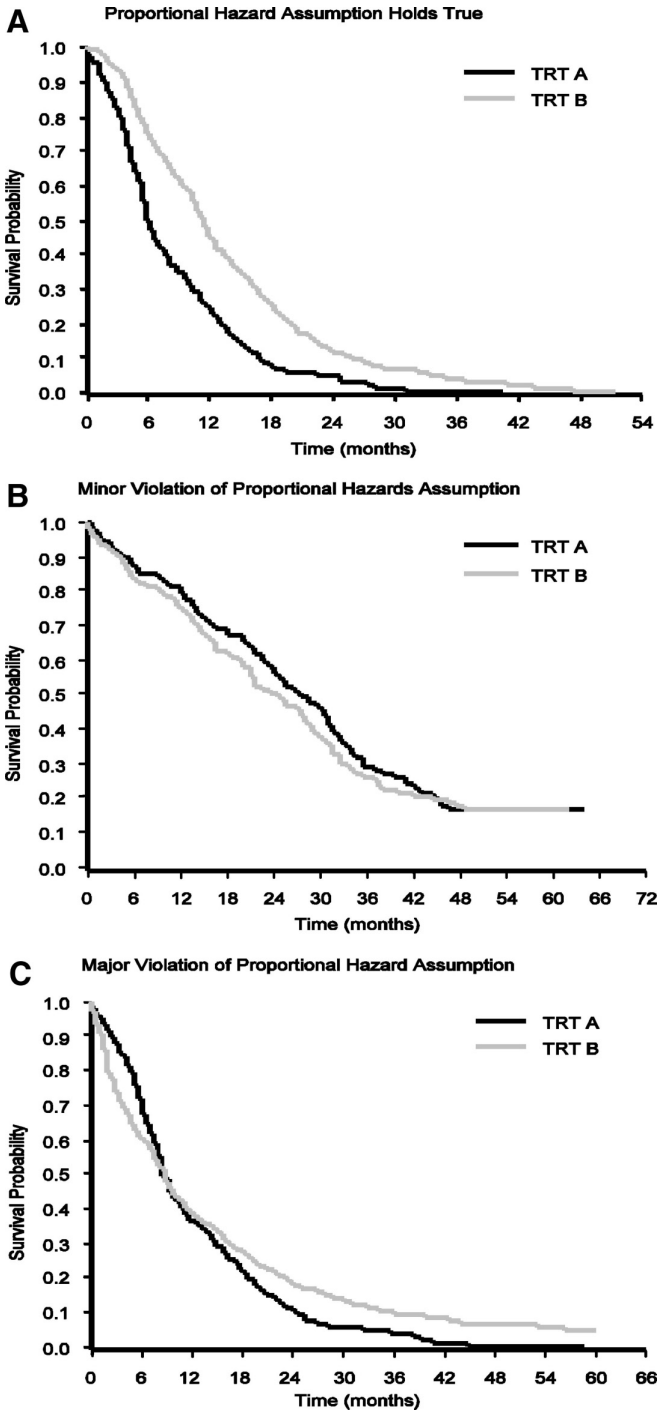

FIGURE 3. The Proportional Hazards (PH) assumption. Proportional hazards assumption holds true (*A*), minor violation of the PH assumption (*B*), and major violation of the PH assumption (*C*). TRT, treatment.

(Figure 2). An unadjusted HR is calculated from a univariate Cox proportional hazards model that only contains treatment (e.g., E versus C) as a covariate, whereas adjusted HRs are typically performed using the multivariate Cox model, i.e., it also contains the covariates that will be adjusted for, such as

BOX 2. Interpretation of a Hazard Ratio

HR (E vs C) = 0.75 for an overall survival end point

This means on average, under an exponential distribution, approximately

- a 25% lower risk of death (25% as 1 − 0.75 = 0.25)
- a 33% increase in survival time (33% as 1/0.75 = 1.33)

On the experimental treatment compared with the control at any point during the trial

age, gender, disease stage, and performance status.[2] If baseline prognostic factors are not balanced, then adjusted survival curves can be constructed to display the expected survival in each treatment arm. There are several available methods to generate the adjusted survival curves, such as the "average covariate" method.[5] These are distinct to the KM method and beyond the scope of this article.

WHAT ARE THE LIMITATIONS?

Correct interpretation of a HR is based on the assumption that the ratio of the hazard rates at each time interval is approximately constant during the study. This is also known as the "Proportional Hazards" (PH) assumption. In the hypothetical example described above (Table 1), the PH assumption held true as the ratio of the hazard rates (~0.75) was approximately constant for each time interval over the duration of the trial. Hence, the HR can be correctly interpreted as patients on the experimental arm having on average approximately a 25% lower risk of death than those in the control arm at any point in time during the trial (Box 2). Determining whether or not the PH assumption holds true can be determined by formal statistical tests and plots, e.g., Martingale residuals, Schoenfeld residuals versus survival time,[6] and log-negative-log plots. However, it can usually be reliably established from reviewing the shape of the KM survival curves, and thus the results of formal tests are rarely reported in the literature. If the separation between the curves is maintained over time, then the PH assumption is likely to hold true (Figure 3*A*). Mild decrease or increase in separation over time is likely to be a minor violation of the PH assumption. One possibility when the KM curve is separated throughout most of the trial but comes together at the end is that by this time the survival data are very mature because there are few patients still at risk as most of the patients have already died or been censored. Given the poor survival in most cancers, if a trial continues long enough, the curves will often come together as advanced cancers are usually not curable, regardless of what therapy is received. In contrast, in other cancers in which a substantial proportion of patients do survive for longer than the follow-up period, the survival curves may plateau because after this time patients are unlikely to die of the primary cancer (e.g., treatment for high-grade lymphoma). Fortunately, most oncology clinical trials produce KM survival curves that are fairly consistent with the PH assumption. Goodness-of-fit assessment for the Cox model should also be performed.

HOW TO INTERPRET A HAZARD RATIO

As discussed earlier, a simplistic interpretation is that if the HR (E versus C) is <1, then the experimental treatment is better than the control and vice versa if HR (E versus C) >1. The following examples illustrate more detailed explanations and common pitfalls.

Appropriate Interpretation

Suppose the HR (E versus C) = 0.75 from a trial evaluating OS where the PH assumption held true (Figure 3*A*). This can be interpreted as either (i) on average approximately a 25% lower risk of death (25% as 1 − 0.75 = 0.25) or (b) on average approximately a 33% improvement in survival time (33% as 1/0.75 = 1.33) in the experimental arm compared with the control arm at any point during the trial (Box 2). Note that this is on average (assuming an exponential distribution of the survival data), so any such improvement or reduction in survival should be interpreted in the context of the KM curve as a whole.

The reduction in the risk of death is lower than the percent increase in survival time (i.e., a ratio). If the log HR is taken, then the same magnitude of benefit is observed for both interpretations. For example, log 0.75 = −0.125 and log 1.33 = 0.125, so the log magnitude of benefit (i.e., 0.125) is the same for both interpretations.

Inappropriate Interpretation and Common Mistakes

Crossing Survival Curves

If the HR (i.e., the treatment effect) varies over time and changes in magnitude but not in direction (Figure 3*B*), then this is likely a minor violation of the PH assumption. The Cox proportional hazards model is widely accepted as being robust to minor violations of the PH assumption, and the overall HR can be interpreted as an average HR over time.

If there is a major violation of the PH assumption (Figure 3*C*), then it is inappropriate to interpret the overall HR, even as an average over time, because the HR varies so significantly over time (e.g., the treatment effect changes direction). In such cases, there is interest in understanding for which patients the experimental treatment is better than the control and for which patients the experimental treatment is worse than the control, and/or whether there is a time-dependent treatment effect. For example in Figure 3*C*, the treatment A is better than treatment B for the first 9 months of the trial when the curves cross. Hence, the HR (A versus B) is < 1 up to this point. However, after the curves cross, the treatment A is worse than treatment B and so the HR (A versus B) is > 1 after 9 months. In this situation, it should also be investigated by performing subgroup analyses if there is a qualitative interaction driving the curves to cross for the population as a whole (e.g., is the HR in males in the opposite direction and statistically significantly different to the HR in females?)[7] If a significant qualitative interaction is identified, then the KM curves for each level of the subgroup (e.g., males and females) should be reviewed to identify whether the PH assumption holds true in these subsets. If this is the case, then interpretation should be focused on these subsets of

patients in light of existing evidence for this subgroup effect, rather than on the all randomized patient population. In addition, efficacy claims for the all randomized patients population should also be avoided. In this case, as with any subgroup analysis, claims of superiority within a subset of patients cannot be made unless the subgroup analysis was prespecified, a statistically significant interaction was observed, and there is sufficient confirmatory evidence to validate the subgroup effect.

Magnitude Matters

The HR is a relative measure. Hence, a statistically significant p value ($p < 0.05$) associated with the HR (E versus C) = 0.75 may be obtained, meaning that (i) the experimental treatment is superior or inferior to the control arm and (ii) there is at most a 5% chance of observing an effect of this magnitude or more extreme (e.g., <0.75) by chance if there is no difference between the treatments. This would seem a positive result for patients. However, whether this is clinically significant also needs to be evaluated. To do this, clinicians will likely need to assess the absolute improvement in survival time by examining absolute measures such as the survival rates at fixed time points (e.g., 1-, 2-, and 3-year survival probabilities) and the median survival. These absolute measures only focus on one point of the survival curve and so can be misleading if they are considered individually. However, collectively they describe the KM curve. Hence, clinicians should look for consistent clinically meaningful improvements across these absolute measures. For example, if a HR = 0.75 corresponds to an increase in the 1- and 2-year OS rate of 10% and 20%, respectively, between the treatment arms in an advanced non-small cell lung cancer trial, these might be deemed clinically meaningful improvements. If the median survival difference between arms is next considered, then an improvement of 50 days may also be regarded as clinically significant, whereas an improvement of approximately 10 days may not. Only if the descriptive absolute measures reveal a fairly consistent pattern of clinically meaningful improvements should a statistically significant HR be hailed as a clinical advancement.

Extrapolation of The HR Beyond The Study Duration

Using the HR reported for a RCT to predict what happens after the duration of the study should be done with great caution and is generally not recommended. This is because HRs summarize the treatment effect over the observed duration of a RCT. The duration of a clinical trial is from when the first patient is randomized to the last patient visit. In the absence of subsequent information, it is not possible to conclusively determine whether the proportional hazards assumption would continue to hold true. In addition, the effect of subsequent interventions on the treatment effect beyond the trial duration is also not accounted for. This is pertinent in the metastatic setting where subsequent lines of therapy or palliative care will heavily influence a patient's survival probability. Statistical techniques exist for extrapolating the HR beyond the duration of a RCT, which can be used in cost-effectiveness analyses, but these techniques rely on strong assumptions and are beyond the scope of this article.

SUMMARY

The Cox proportional hazards model is used to analyze survival data. It provides a HR to assess the relative efficacy of the experimental treatment compared with the control treatment over the duration of the RCT. As with everything else in clinical medicine, results of testing statistical hypotheses and estimating treatment effect should always be interpreted keeping the clinical significance in mind. These statistical tests after all are only means to the noble goal of uncovering the truth and improving the lives of our patients by providing useful treatment and avoiding needless exposure to ineffective or questionably effective therapies.

ACKNOWLEDGMENTS

The authors thank Nancy Iturria for generating simulated survival data to create the survival curve figures and Jonathon Denne and Mauro Orlando for critically reviewing the article and providing helpful comments.

REFERENCES

1. Altman DG, Bland JM. Time to event (survival) data. *BMJ* 1998;317:468–469.
2. Cox DR. Regression models and life-tables. *J R Stat Soc Series B Stat Methodol* 1972;34:187–220.
3. Gould SJ. The median isn't the message. *Discover* 1985;6:40–42.
4. European Agency for the Evaluation of Medicinal Products (EMEA), Committee for Proprietary Medicinal Products (CPMP). Points to consider on adjustment for baseline covariates. CPMP/EWP/283/99, 2003.
5. Nieto FJ, Coresh J. Adjusting survival curves for confounders: a review and a new method. *Am J Epidemiol* 1996;143:1059–1068.
6. Schoenfeld D. Partial residuals for the proportional hazards regression model. *Biometrika* 1982;69:239–241.
7. Barraclough H, Govindan R. Biostatistics primer: what a clinician ought to know: subgroup analyses. *J Thorac Oncol* 2010;5:741–746.