Getting More Out of Survival Data by Using the Hazard Function

Kenneth R. Hess and Victor A. Levin

Abstract

Information about the survival experience for a group of patients is almost exclusively conveyed using plots of the survival function. However, the hazard function provides information about the survival experience that is not readily evident from inspection of the survival function. The hazard function tracks the instantaneous failure rate over time among the surviving patients and can be readily estimated with available software. We illustrate how these estimates can be used in conjunction with estimates of the survival function to glean clinically relevant information from survival data. Clin Cancer Res; 20(6); 1404–9. ©2014 AACR.

Introduction

The survival experience for a group of patients is most often analyzed using only plots of the survival function (also called the survival curve). The survival function is the probability that an individual will survive to time, \( t \), plotted as a function of \( t \). Kaplan–Meier estimates of the survival function are widely reported in clinical research (1). Plots of the survival function express risk of death in terms of cumulative probabilities. For example, to survive, say, the first 3 years, a patient must survive the first year, then the second year, and finally the third year. We can compute the probability of surviving the first 3 years by multiplying the probabilities of surviving each successive year. In this way, the survival probabilities accumulate over time.

The survival function gives, for each time \( t \), the probability that a patient will survive to time \( t \). Now consider the conditional probability of a patient dying in the short time interval between \( t \) and \( t + x \) given survival to time \( t \). This is just the probability of dying in this short interval among the patients surviving to the start of the interval. Now divide this conditional probability by the size of the interval (\( x \)). This yields a conditional probability per unit of time or, in other words, a conditional rate. If we now take the limit of this rate as the length of the interval gets smaller and smaller (i.e., as \( x \) approaches zero), we get the instantaneous conditional death rate at time \( t \) (ref. 2, p. 11). This is just the instantaneous death rate among the survivors at time \( t \), and we call this the hazard rate at time \( t \). A plot of the hazard rates over time is called the hazard function, and it conveys how the instantaneous death rate among the survivors changes over time.

Thus, while the survival function conveys the probability of surviving to time \( t \) (i.e., the probability of not dying before time \( t \)), the hazard function conveys the instantaneous conditional death rate at time \( t \). The first obvious difference between these measures of risk is that the survival function relates to the probability of not dying, whereas the hazard function relates to the probability of dying such that high survival probabilities suggest low hazard rates and vice versa. A second difference is that the survival function is unconditional, whereas the hazard function is conditional. This just means that as a cohort is being followed over time, the survival function conveys information about risk of death at time \( t \) for the entire cohort, whereas the hazard function conveys information about risk of death at time \( t \) for only those patients remaining at risk at time \( t \). A third difference is that the survival function conveys risk in terms of probability and the hazard function conveys risk in terms of an instantaneous rate. Average death rates over an interval of time measure the risk of death per unit of time for that interval. Essentially, they are conveying the intensity of the death risk in that interval. An instantaneous death rate is just the limiting value of the average death rate as the interval gets smaller and smaller.

Although the values of the survival function have a lower limit of 0 and an upper limit of 1 (as they are probabilities), the values of the hazard function have only a lower limit of 0 (as they are rates). The units of measure of hazard rates are events per unit of time, and thus the values of the hazard function change when the units of time in which the survival times are recorded are changed. If we change the units of time from days to weeks, the values of the hazard rate will increase by a factor of 7 [as the denominator (time) has decreased by a factor of 7].
There is a complex mathematical relationship between the hazard function and the survival function (ref. 2, p. 12). The hazard function is the negative of the first derivative of the logarithm of the survival function. From this relationship, we can see that it is difficult to judge the shape of the hazard function when viewing the survival function and vice versa. Because both functions are informative with respect to the patients’ survival experience, it would seem prudent to estimate and graph both functions to gain insight into the properties of a given dataset.

Estimation

Historically, nonparametric estimates of the hazard function have been produced using the life table approach (3). In this approach, the time scale is divided into discrete intervals and hazard rates are computed for each interval separately. This yields a piecewise constant estimate that is similar in many respects to the familiar histogram used to estimate the density function for uncensored data. A similar approach is to divide the follow-up time into equal intervals and compute the hazard rate in each interval. If we assume that the hazard rate is constant in the interval, we can estimate the hazard rate as the number of events observed in the interval divided by the sum of the follow-up times observed in the interval (ref. 2, p. 160).

More recently, however, we have been using kernel smoothing methods to produce smooth estimates of the hazard function from right-censored data (ref. 4; ref. 5, p. 154; ref. 2, p. 32). The kernel estimator of the hazard function at a given time \( t \) is essentially a weighted average of the data in the neighborhood around \( t \). The size of the neighborhood around \( t \) is called the bandwidth. A critical factor in the performance of the kernel estimator, the bandwidth determines the degree of smoothness of the estimate. The larger the bandwidth used, the greater the smoothness. More smoothness leads to lower variability but also generally leads to increased bias (4).
On the basis of our experience, we start with a bandwidth equal to about 20% of the range of the survival times and then select the final bandwidth by trial and error, picking the value that gives the best compromise between over-smoothing and under-smoothing. Sometimes it is helpful to plot estimates computed with multiple bandwidths. We are careful not to plot the hazard function estimates when fewer than 15 to 20 patients remain at risk. Kernel estimates of the hazard function are available in many statistical programs, including SAS, Stata, and R/S+ (using the muhaz library). Confidence intervals (CI) for these estimates can be constructed using bootstrap resampling (6).

Examples

Figure 1A shows the Kaplan–Meier estimate of the survival function of 686 patients enrolled in the German Breast Cancer Study (7). A total of 171 deaths occurred after a median of 4.5 years of follow-up. The survival curve is relatively flat for the first year and then drops with a fairly constant slope between years 1 and 6. After year 6, the curve flattens out again. Figure 1B shows a piecewise constant estimate of the corresponding hazard function. During the first year, there are 12 deaths experienced with a total of 665 person-years of follow-up for an average annual hazard rate of 0.018 deaths per person year of follow-up. At the beginning of the second year, 646 patients remain at risk. During the second year, 57 deaths are experienced with a total of 604 person-years of follow-up for an average annual hazard rate of 0.094 deaths per year, indicating a sharp increase in the hazard function between years 1 and 2. The hazard rate is relatively stable in years 2, 3, 4, and 5. At the beginning of the sixth year, 170 patients remain at risk with 14 deaths occurring during a total of 104 person-years of follow-up for an average annual hazard rate of 0.135 deaths per year, which is about 30% higher than the preceding few years. At the beginning of year 7, 53 patients remain at risk with 1 death occurring in 25 person-years of follow-up for an
average annual hazard rate of 0.040 deaths per year of follow-up.

Figure 1C shows a kernel-based estimate of the hazard function computed using a bandwidth of 1 year. It bears a striking resemblance to a smoothed version of Fig. 1B. In fact, kernel estimation is essentially equivalent to computing a series of piecewise constant estimates with the same bin width but slightly different starting points and then averaging these estimates at each time point (ref. 8, p. 121). Figure 1D shows kernel-based estimates of the hazard function using bandwidths of 0.5, 1.0, and 2.0 years. Clearly, as the bandwidth increases, the hazard estimate becomes smoother. The estimate corresponding to a bandwidth of 0.5 years fluctuates considerably, but the larger bandwidths used for the other estimates act to smooth out the highs and lows of these fluctuations.

Figure 2A shows the Kaplan–Meier estimate of the survival function (with 95% CIs) of 686 patients enrolled in the German Breast Cancer Study for the endpoint of recurrence-free survival (299 events total). This curve is similar in many respects to the survival curve for overall survival shown in Fig. 1A, but it begins to drop sooner and drops faster. Figure 2B is the kernel-based estimate of the corresponding hazard function computed using a bandwidth of 1 year. This estimate starts low, increases quickly to an initial peak toward the end of the second year, and then decreases—first quickly and then less quickly reaching a nadir around the middle of year 5. Figure 2C shows kernel-based estimates of the recurrence-free survival hazard function using bandwidths of 0.5, 1.0, and 2.0 years. As the bandwidth increases, the initial peak becomes lower and shifts to the right while the second rising phase becomes more subtle. Figure 2D shows the kernel-based estimate for a 2-year bandwidth along with estimated 95% CIs computed using bootstrap resampling (6). The CIs widen noticeably as fewer patients remain at risk. At 3 years, at which 331 patients remain at risk, the estimated hazard rate is 0.146 with 95% CI, 0.125–0.169, while at 6 years,
at which 36 patients remain at risk, the estimated hazard rate is 0.147 with 95% CI, 0.073–0.269. The CI at 6 years is more than four times as wide as that at 3 years.

Figure 3A shows the Kaplan–Meier estimates of the survival functions (with 95% CIs) for relapse-free survival according to estrogen receptor (ER) status (where positive status is defined as number of ERs > 50). There were 392 patients with ER-negative tumors experiencing a total of 180 events and 294 patients with ER-positive tumors experiencing a total of 119 events. The survival curves are together initially and then diverge with separation until year 5 (and with nonoverlapping CIs for the first 2 years). During the sixth year, the survival functions cross but with little separation thereafter. Figure 3B shows piecewise constant estimates of the corresponding hazard functions demonstrating that patients with ER-negative tumors have higher hazard rates for the first 2 years. The hazard rates are overlapping during years 3 and 4, but thereafter it seems that patients with ER-positive tumors have higher hazard rates. Figure 3C shows kernel-based estimates using a bandwidth of 1.5 year showing the same features as Fig. 3B. The hazard rates for both groups start low and increase to an initial peak around the end of the second year, but the hazard rates for patients with ER-negative tumors are roughly twice as high as those for patients with ER-positive tumors. The hazard functions converge at the end of the third year but while the hazard rates for patients with ER-negative tumors continue to decrease, the hazard rates for patients with ER-positive tumors begin to increase such that after year 4, patients with ER-positive tumors seem to have higher hazard rates than those with ER-negative tumors. Figure 3D shows the kernel-based hazard estimates from Fig. 3C but with 95% CIs superimposed. The CIs are nonoverlapping for those with ER-negative tumors. Figure 3D indicates considerable statistical uncertainty

...the hazard functions cross after the third year. This is due to the fact that the hazard functions represent instantaneous rates, whereas the survival functions represent cumulative probabilities. Arguably, the point of the crossing hazard functions is more clinically relevant as this is the point at which the failure rates among the surviving patients equalize. The crossing point can change with the bandwidth used in the estimation process and is subject to sampling variability, and thus care should be taken when assessing the crossing point of the hazard functions.

Discussion

Although the survival function plays a fundamental role in conveying how risk changes over time in terms of cumulative probabilities, the hazard function provides additional useful insight over and above what can be easily gotten from the survival function. It can be used to obtain more detailed information about the failure process for a single group of patients or when comparing groups. The hazard function conveys how risk changes over time in terms of the instantaneous death rate among the survivors. This instantaneous nature means that hazard functions respond more quickly to changes in risk associated with the failure process under investigation. However, their instantaneous nature makes hazard functions more difficult to estimate reliably.

Although hazard function estimation requires investigators to specify the degree of smoothness, it can be insightful to generate estimates of varying degrees of smoothness. Estimates that are less smooth emphasize local features of the hazard function, whereas smoother estimates emphasize more global features. Hazard function estimates can be unstable (i.e., highly variable) when too few patients remain at risk, and thus it is important to present the estimates along with CIs or to avoid plotting estimates when fewer than 15 to 20 patients remain at risk. Nonetheless, the hazard function can provide important clinical insight and strengthen conclusions conveyed by typical survival analyses.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors’ Contributions

Conception and design: K.R. Hess
Development of methodology: K.R. Hess
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): K.R. Hess, V.A. Levin
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): K.R. Hess
Writing, review, and/or revision of the manuscript: K.R. Hess, V.A. Levin
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): K.R. Hess
Study supervision: K.R. Hess

Grant Support

This study was supported by NIH grant CA16672 (to K.R. Hess).

Received August 1, 2013; revised January 10, 2014; accepted January 17, 2014; published OnlineFirst February 5, 2014.

References


Clinical Cancer Research

Getting More Out of Survival Data by Using the Hazard Function

Kenneth R. Hess and Victor A. Levin


Updated version

Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-13-2125

E-mail alerts

Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions

To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions

To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.