Commentary: The Hazards of Survival Comparisons

DONALD BERRY

The University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA

A standard way to present time-to-event analyses—such as comparisons of two survival distributions—is to use a hazard ratio (HR). This is a weighted average (over time) of the corresponding hazard functions. The term “hazard function” stems from reliability theory, where it is also called failure rate. Actuaries use the term “force of mortality” for the same concept.

The annual premium that you pay for life insurance is a bet that you’re going to die in the next year. Appropriate odds on the bet (padded somewhat so the insurance company can make money) are based on the proportion of people like you who are alive at your age and who will die in the next year. That’s your hazard of dying, and it depends on your age. Next year it will be different.

Picture a curve (or function) showing the hazard of dying within the next small interval of time, for each time from birth to age 100. The same kind of curve applies for the occurrence of an event (recurrence, death, or whatever) in a clinical trial for patients in a particular treatment group. An estimate of the population hazard over the next year, say, is the number of events in that year divided by the number of patients at risk for the event at the beginning of the year. Time 0 is the time of randomization, for example.

Comparing two treatment groups means comparing the two curves. One such comparison is the ratio of the two curves, the hazard ratio function (HRF). But that’s still a curve, and it may take on a different value for each time point. The HRF may be bigger than 1.0 sometimes (at times when the denominator treatment is looking better) and at other times it may be smaller than 1.0 (when the numerator treatment is looking better). If so, which treatment is better overall? And by how much?

A way to address such questions is to take the average of the HRF over time and call it the HR. But in the averaging process, one might reasonably give greater weight to values of the HRF where there are more events. A standard way to compute a weighted average is using a Cox proportional hazards regression model. It gives a single number, the estimated HR, and its confidence limits. An approach that gives similar answers is the “event rate ratio” used by the Early Breast Cancer Trialists’ Collaborative Group (EBCTCG).

The HR has a number of deficiencies. The most obvious is that no single number can pretend to carry the information content of an entire curve.

This brings me to the Lewis piece published in this issue of The Oncologist [1]. Dr. Lewis suggests that “The EBCTCG figures and data can be used to help understand where, when, and for how long our adjuvant drug therapies work. For tamoxifen as well as for polychemotherapy, these adjuvant therapies appear to work on soon-to-emerge clinically unrecognized tumors during the period of drug exposure, and the benefit gained lasts at least for 15 years. There is little to support these therapies having an increased beneficial effect after the drugs have been withdrawn.” In effect, he is observing that the HRF is less than 1.0 (adjuvant treatment better) in the early years and quite similar to 1.0 in the later years.

This observation is not new. And it applies in individual trials as well as in the EBCTCG meta-analysis [2–4]. For example, Berry et al. [4] observe that “The risks of recurrence and death vary over time . . . and so do the reductions in risk that are attributable to improved chemotherapy.” And “There was little or no advantage of [improved chemo-
therapy] after three years. However, the benefit was durable in that there was no sign of a rebound in risk . . . in later years, a finding reflected as well by the persistent separation in the corresponding survival curves. . . ” Berry et al. [4] make a similar point about tamoxifen.

Dr. Lewis supports his observation by shifting one cumulative recurrence curve so that the two curves intersect at year 5—see his Figures 1B, 2B, and 3B. This crude method is not appropriate for comparing late risks. The reason is that the sizes of the at-risk populations are different at the joinpoint of the two curves. If the curve that is shifted up happens to coincide identically with the stationary curve beyond the joinpoint, then the former group—the one that was shifted—would have a greater hazard than the latter group in the later years.

Dr. Lewis’s method of shifting of curves to make them agree at a particular time point would have to be modified to correctly reflect the late risk for recurrence. Namely, one would have to change the scales of the curves by dividing by the size of the at-risk populations at the joinpoint. The slopes of the resulting curves would be the hazard functions. The magnitude of Dr. Lewis’s error is the amount he shifts up one of the curves, which seems to be 5%–13% for the three figures in question. Because these shifts are moderate in size, the correctly adjusted curves would also be roughly similar after 5 years. The resulting hazard functions would also be similar from year 5 and beyond.

In any case, the correct way to compare cumulative incidence curves as regards potentially differing relative risks over time is using hazards—the HRF and its standard error, for example.

An important question is whether apparent changes in the HR over time since surgery are real or a statistical artifact. This can be addressed statistically by testing the hypothesis that the HRF is constant over time. However, this is a very knotty problem. One must propose alternative hypotheses. These are likely to be data driven—as in the case at hand—and therefore subject to bias. Exacerbating the problem is the heterogeneity of breast cancer, especially node-positive disease. Aggressive tumors contribute to high hazards in the early postsurgery time frame. But once these aggressive tumors recur, they remove themselves from the at-risk population and the hazard drops. Interestingly, the hazard of node-positive tumors seems to drop to that of node-negative tumors (beyond 5 years, say). And the hazards of estrogen receptor (ER)-negative tumors can drop to (and even below!) those of ER-positive tumors.

My view is that the effect is probably real. The major unanswered scientific question is whether a therapy cures some cancers or delays the growth of a larger number of cancers, or some mixture of the two. In the growth-delay case one would expect to see a rebound in the recurrence curves for the treated patients. As Dr. Lewis indicates, the EBCTCG results give little or no evidence of such rebounds. However, delayed recurrences could be long in coming, making the two cases nearly indistinguishable. A mixture is the safe and likely correct answer. But my examination of many HRFs suggests that cure may be the dominant component of the mixture for chemotherapy while growth delay is the dominant component for tamoxifen. Breast cancer patients shouldn’t care much which it is, especially because there’s no way currently to know for sure whether they were one of the lucky ones who benefitted.

What is the clinical impact of an early benefit to therapy that wanes over time? The HRF is a measure of relative risk. It’s a very handy statistical tool. But it is not appropriate for making clinical decisions. These should be based on comparisons of time-to-event curves. For example, the estimated incremental time of being event-free is the area between the two time-to-event curves.

I have one more small bone to pick with Dr. Lewis. He draws conclusions such as the following without recognizing plausible alternatives: “polychemotherapy impacts soon-to-emerge tumors and has no demonstrable effect on tumors slowly developing that are destined to emerge in later years.” The basis for this statement is his observation that (in my words) polychemotherapy reduces the early hazard but has no effect on the late hazard. Even if this observation is correct, the conclusion may be quite wrong. Indeed, polychemotherapy may have a beneficial effect on every tumor by delaying their emergence. The “soon-to-emerge” tumors become tumors that emerge later and replace (in time) the tumors that would have emerged later but that are also benefited and they emerge even later yet.

**REFERENCES**


www.TheOncologist.com