The NEW ENGLAND JOURNAL of MEDICINE

Perspective
JULY 13, 2006

STATISTICS AND MEDICINE

Time-to-Event Analyses for Long-Term Treatments — The APPROVe Trial

Stephen W. Lagakos, Ph.D.

The Adenomatous Polyp Prevention on Vioxx (APPROVe) trial[1] compared rofecoxib with placebo in the prevention of recurrent colorectal polyps, but the researchers also collected data on adverse cardiovascular events, including confirmed serious thrombotic events. Assessment of the cardiovascular data raises important issues about the analysis and interpretation of a time-to-event end point in a randomized, placebo-controlled trial evaluating a long-term treatment. These issues include the appropriate period of follow-up for safety outcomes after the discontinuation of treatment; the purpose and implications of checking the assumption of proportional hazards, which underlies the commonly used log-rank test and Cox model; and what the results of a trial examining long-term use imply about the safety of a drug if it were given for shorter periods.

With regard to the first issue, the distribution of the time to an event is described by the cumulative incidence function, I(t), which for every time t after the start of treatment gives the cumulative probability that the event occurred in a patient. I(t) is usually estimated by the Kaplan–Meier method.

Time-to-event analyses of a safety end point sometimes count only events that occur during the scheduled treatment period, T_s, or during a limited window of time afterward, T_W. For example, in the APPROVe trial, T_s was 36 months and T_W was 14 days, so data on cardiovascular events were scheduled to be collected for a total of 36 months and 14 days after the initiation of treatment. There are several reasons why using such windows might be desirable. First, events occurring during treatment or the subsequent window period might be the most relevant clinically for assessing the safety of the treatment. Second, any increased risk attributable to the treatment might diminish shortly after the discontinuation of treatment, so the power of the log-rank or Cox test might be diluted if events that occurred after the window period were counted. And third, patients might receive other therapy after the discontinuation of the study treatment that could affect their risk of a safety end point.

Two important considerations are the length of T_W and the duration of follow-up for patients who discontinue treatment prematurely. Suppose that all patients continue to receive treatment until the end of the scheduled period or until the outcome event occurs, whichever comes first. Then, the power of the log-rank test to detect an increased risk in the treatment group, as compared with the placebo group, depends on several factors, including the value of T_w and the way in which the relative risk of the outcome changes during treatment and after the discontinuation of treatment.[2] The optimal length of T_w depends in a complex way on these and other factors, but the period will typically end before any elevated risk associated with the treatment disappears entirely. The use of a shorter or longer window period will reduce the statistical power to detect an increased risk.[2]

Suppose, however, that some patients discontinue treatment prematurely and subsequently have a different risk of the outcome event than patients who continue the treatment. Then, if the patients who discontinue treatment prematurely are followed for the outcome event only for a specific T_w after discontinuation, the results of the log-rank or Cox test can be distorted, either masking a real difference or the magnitude of a difference or showing a difference where none exists (a false positive result), especially if the rates of premature discontinuation differ in the treatment and placebo groups. For example, if the treatment (in this case, rofecoxib) causes a side effect that increases the likelihood of both the discontinuation of treatment and the outcome event, then following these patients for a short time (say, a T_w of 14 days) after discontinuation might cause a real difference to be obscured by the differential exclusion of events that occur in the treatment group after the 14-day window. Premature discontinuations can also cause bias in the Kaplan–Meier estimate of cumulative incidence and the estimated relative risk of treatment. Such distortions and biases can be avoided by counting all end points that occur in patients during the scheduled follow-up period, T_s+T_w (in the case of the APPROVe Trial, 36 months plus 14 days), regardless of whether they discontinue treatment prematurely. The power of the resulting log-rank or Cox test would still depend on the value of T_w, as described above.

In the APPROVe trial, 32 percent of patients in the rofecoxib group and 25 percent of patients in the placebo group discontinued treatment prematurely, many because of side effects. Since the trial found a significantly higher rate of serious thrombotic events in the rofecoxib group than in the placebo group ($P=0.008$ by the log-rank test), these early discontinuations did not lead to an overall false negative finding. However, the premature discontinuations may have biased the Kaplan–Meier estimates of cumulative incidence (see Figure 2 of the APPROVe study) and the estimated relative risk associated with treatment. When the APPROVe investigators eventually publish information about serious thrombotic events that occurred more than 14 days after the premature discontinuation of the study drug, updated estimates of the cumulative incidence functions and relative risk can be calculated and compared with those in the original report to assess the possible extent of bias and its clinical implications.

The second issue raised by the analysis of the cardiovascular data is that of the assumption of proportional hazards. The log-rank and Cox tests are motivated by this assumption — that is, that the relative risk remains constant over time. Given this assumption, the relative risk provides a simple way of describing the magnitude of the effect of treatment on the end point, and one can infer that the corresponding cumulative incidence curves diverge throughout the entire time range covered. These tests can be well powered to detect some differences between treatment groups that do not satisfy the assumption of proportional hazards, but they can have poor power to detect other differences, including cumulative incidence curves that are initially equal but later diverge and others that initially diverge but later approach one another.[3] When either test yields a nonsignificant difference between the treatment groups, one concern is whether the treatments could differ in a way that is not captured by the test. Thus, the proportional-hazards assumption is tested to determine whether a nonsignificant difference between groups might have been due to a treatment effect that does not satisfy that assumption.

The most common analytic way of testing the proportional-hazards assumption is by fitting a Cox model with one term representing the treatment group and another term representing an interaction between the treatment group and either time or the logarithm of time. These models correspond to a relative risk that changes exponentially (relative $\text{risk}(t)=\gamma e^{\beta t}$) or as a power of time (relative $\text{risk}(t)=\gamma t^{\beta}$). Which of these two interaction tests is more powerful will depend on the nature of the difference between the treatment groups. When ap-

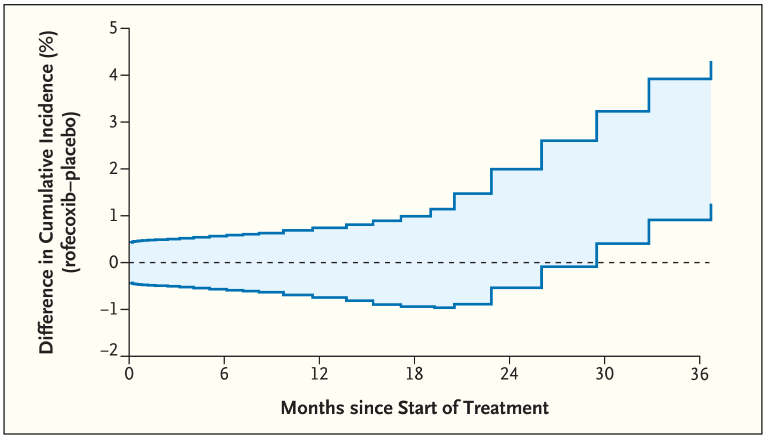

Figure 1. Hypothetical 95 Percent Confidence Band for the Difference, $I_{36}(t) - I_P(t)$, between the Cumulative Incidence Curves for the Rofecoxib (I_{36}) and Placebo (I_P) Groups, Constructed from the Results of the APPROVe Trial.

Differences lying partly or completely outside the shaded region are inconsistent with the data. Differences lying wholly within the shaded region include the following: separation of the cumulative incidence curves in the two groups at times both before and after 18 months, and consistently higher or lower cumulative incidence in the rofecoxib group before 18 months.

plying them, it is important to keep in mind that rejection of the proportional-hazards assumption does not mean that the true relative risk follows the form assumed in an expanded Cox model, nor does the failure to reject the assumption necessarily mean that the assumption holds.

The APPROVe investigators planned to use an interaction test with the logarithm of time as the primary basis for testing the proportional-hazards assumption. This test resulted in a P value of 0.07, which did not quite meet the criterion of 0.05 specified for rejecting the assumption. However, the original report of the APPROVe trial[1] mistakenly gave the P value as 0.01, which was actually the result of an interaction test involving untransformed time. (This error is corrected in this issue of the *Journal*.) The investigators noted that the estimated cumulative incidence curves for adjudicated serious thrombotic events in the rofecoxib and placebo groups were similar for approximately the first 18 months of treatment and thereafter diverged. I interpreted this statement as no more than a simple way of describing the visual difference between the Kaplan–Meier curves for the rofecoxib and placebo groups and not as a claim that the cumulative incidence rates were equivalent in the two groups for the first 18 months, since this neither was demonstrated nor follows from the use of either of the interaction models used to test the proportional-hazards assumption.

The estimated relative risk calculated with the use of the Cox model represents a time-averaged hazard ratio and thus may not adequately describe the difference between the treatment and placebo groups when the proportional-hazards assumption does not hold. It may then be of interest to assess how the cumulative incidence curves might plausibly differ over time. Doing so by means of post hoc analyses based on visual inspection of the shapes of the Kaplan–Meier curves for the treatment groups can be misleading and should be avoided. A better approach is to create a confidence band[4] for the difference between the cumulative incidence curves in the treatment and placebo groups — that is, for the excess risk in the treatment group. Confidence bands can be constructed in several ways[4] and for settings in which some observations are informatively censored.[5] The bands are commonly centered around the estimated difference between the treatment groups, so that for a 95 percent band, the 5 percent error is evenly split above and below the band.

To illustrate, Figure 1 shows a hypothetical 95 percent confidence band for the difference between the rofecoxib and placebo groups in the cumulative incidence of confirmed serious thrombotic events in the APPROVe trial. The band is approximately centered around 0 percent for the first 18 months and thereafter increases, reflecting the pattern in Figure 2 of the original APPROVe study.[1] Analogous to the way in which a 95 percent confidence interval for a parameter, such as a relative risk or odds ratio, provides a plausible range of values for that parameter consistent with the data, the shaded region in Figure 1 represents a plausible range of values for the excess risk associated with rofecoxib therapy over time that are consistent with the data. Any difference between groups in the cumulative incidence curves that does not fall wholly within the band is inconsistent with the data. The graph shows that there are many plausible differences, including a separation of the curves at times both before and after 18 months and a consistently higher or lower

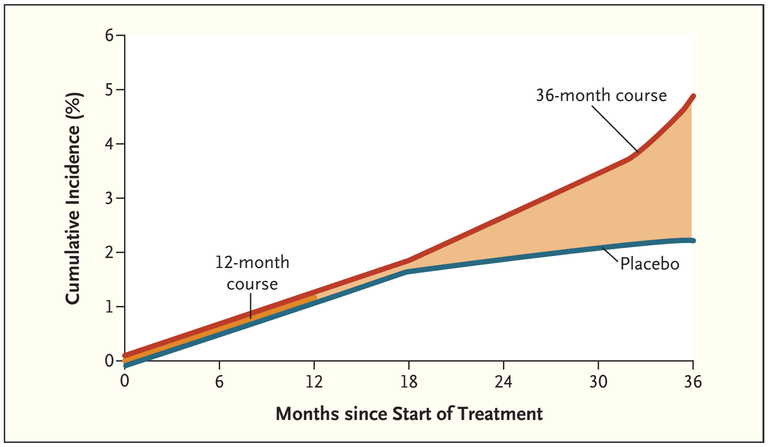

Figure 2. Logical Inferences about the Cumulative Incidence Function, $I_{12}(t)$, for a 12-Month Course of Rofecoxib, Based on Known Values for $I_{36}(t)$ and $I_p(t)$ That Are Identical for 18 Months before Diverging.

If monotonicity is assumed, so that $I_p(t) \leq I_{12}(t) \leq I_{36}(t)$, then $I_{12}(t)$ must equal $I_p(t)$ for first 18 months and be somewhere in the shaded region after 18 months. The lower edge of the shaded region corresponds to the absence of an increased risk with the 12-month course; all other scenarios in the shaded region correspond to an excess risk with the 12-month course that occurs only after the discontinuation of treatment. If monotonicity is not assumed, nothing can be inferred about $I_{12}(t)$ beyond month 12; however, the 12- and 36-month courses are identical for the first 12 months, so that, all other things being equal, $I_{12}(t)$ must equal $I_{36}(t)$, and thus $I_p(t)$, through month 12. Although drawn as separate curves to be visually informative, the inferences are based on the assumption that the cumulative incidence functions overlap for the first 18 months.

cumulative incidence in the rofecoxib group, relative to the placebo group, before 18 months.

Finally, the third question revolves around the implications of analyses of long-term use for the safety of shorter-term use. The results of the APPROVe trial have been misinterpreted by some to mean that treatment with rofecoxib for less than 18 months poses no excess cardiovascular risk. Consider what can be inferred about the cumulative incidence associated with a shorter course of rofecoxib — say, 12 months — from the data for a 36-month course and the data for a placebo group. These cumulative incidence functions are denoted by $I_{12}(t)$, $I_{36}(t)$, and $I_p(t)$, respectively. We first discuss what can be logically inferred about $I_{12}(t)$ when $I_{36}(t)$ and $I_p(t)$ are known and the longer course increases risk, and then discuss what can be statistically inferred about $I_{12}(t)$ when $I_{36}(t)$ and $I_p(t)$ must be estimated. To answer these questions, I assume the monotonic condition that at any given time t, the cumulative probability of the outcome event for the 12-month course lies somewhere between that for the placebo group and that for the 36-month course — that is, $I_p(t) \leq I_{12}(t) \leq I_{36}(t)$ when the 36-month course increases risk.

Suppose that $I_{36}(t)$ and $I_p(t)$ are known and are identical for all times less than 18 months and thereafter diverge (see Figure 2). Then, given the assumption of monotonicity, it must follow that $I_{12}(t)$ equals $I_p(t)$ for all times less than 18 months and must lie somewhere between $I_p(t)$ and $I_{36}(t)$ for times of 18 months or longer (shown as the shaded area in Figure 2). The most optimistic scenario among the many possibilities in the shaded region is given by the lower boundary, which corresponds to the absence, during the entire follow-up period, of an increased risk associated with the 12-month course. For this extreme case to hold, the 12-month course cannot have any effects on patients during the treatment period that are associated with the risk of the outcome event after 18 months. For all other possibilities represented in the shaded region, the 12-month course increases risk, but only after the treatment has ended. Such delayed effects would occur when the latency period between an intervention's initial insult and subsequent clinical outcomes exceeds the duration of treatment. Easily understood examples of such behavior include the increased risk of certain cancers after exposure to ionizing radiation or chemotherapy for another cancer.

In practice, $I_{36}(t)$ and $I_p(t)$ are never known but rather are estimated from the randomized trial in which they were evaluated. Without the assumption of proportional hazards, plausible values for the excess risk associated with the 36-month course are provided by a confidence band (shown in Figure 1) for the difference between $I_{36}(t)$ and $I_p(t)$. Suppose the upper edge of the shaded region in Figure 3 represents an upper 95 percent confidence bound for the excess risk — that is, $I_{36}(t) - I_p(t)$ — of the 36-month course of rofecoxib, constructed from a trial like APPROVe. Upper confidence bounds are simi-

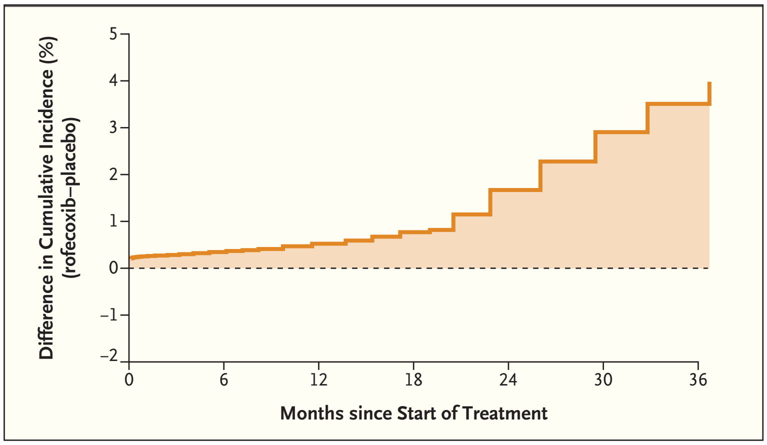

Figure 3. Statistical Inferences about the Excess Risk, $I_{12}(t) - I_p(t)$, Associated with a 12-Month Course of Rofecoxib, Based on the Hypothetical Results of a Trial Comparing a 36-Month Course of Rofecoxib with Placebo.

The upper edge of the shaded region represents an upper 95 percent bound for $I_{36}(t) - I_p(t)$, constructed from the trial results. If monotonicity is assumed, this edge also represents an (at least) 95 percent upper bound for $I_{12}(t) - I_p(t)$. The assumption of monotonicity also implies that $I_{12}(t) - I_p(t) \geq 0$, so that the shaded region represents an (at least) 95 percent confidence band for $I_{12}(t) - I_p(t)$. If monotonicity is not assumed, nothing can be inferred about $I_{12}(t) - I_p(t)$ beyond month 12; however, since the 12- and 36-month courses are identical for the first 12 months, the first 12 months of the confidence band in Figure 1 also represents, all other things being equal, a confidence band for $I_{12}(t) - I_p(t)$ over this period.

lar to confidence bands, except that they are one-sided. Thus, the true excess risk associated with the 36-month course is, with 95 percent confidence, on or below this upper edge. If monotonicity is assumed, the excess risk associated with the 12-month course must be no greater than that associated with the 36-month course, so this edge also provides an upper (at least) 95 percent bound for the excess risk associated with the 12-month course. Since the assumption of monotonicity also implies that $I_{12}(t)$ can be no less than $I_p(t)$, the shaded area between the horizontal axis and the upper bound in Figure 3 provides an (at least) 95 percent confidence band for the excess risk associated with the 12-month course. Excess risks not wholly contained in this shaded area are not consistent with the data.

The lower edge of the shaded area corresponds to the absence of increased risk with the 12-month course. However, unlike the range of possible risks depicted in Figure 2, the shaded region of Figure 3 will typically include scenarios in which there is excess risk associated with the 12-month course both during treatment and after its discontinuation. Examples of such treatment effects are abundant — for example, the increased risk of stroke associated with hormone-replacement therapy and the increased risk of cardiac events associated with trastuzumab therapy for breast cancer. Without the assumption of monotonicity, nothing can be inferred about the excess risk associated with the 12-month course after month 12; however, the first 12 months of the confidence band shown in Figure 1 also apply to the 12-month course because all other things being equal, the 12- and 36-month courses are identical during this period.

When applied to the data from the APPROVe trial, a confidence band analogous to that in Figure 3 would provide a plausible range of excess risks associated with a shorter (less than 18 months) course of rofecoxib. If all the differences represented in this band were clinically unimportant, one could conclude that the data were inconsistent with a clinically important increase in risk for the shorter course of rofecoxib. However, since the band would necessarily include the estimated excess risk associated with the 36-month course reflected in Figure 2 of the original APPROVe trial, one could not conclude from the data that a shorter course of rofecoxib is safe.

This article was published at www.nejm.org on June 26, 2006.

Dr. Lagakos is a professor of biostatistics at the Harvard School of Public Health, Boston, and a statistical consultant to the *Journal*.

1. Bresalier RS, Sandler RS, Quan H, et al. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. N Engl J Med 2005;352:1092-102.
2. Griffin BA, Lagakos SW. Analysis of failure time data arising from studies with alternating treatment schedules. J Am Stat Assoc 2006;101:510-20.
3. Lagakos SW, Schoenfeld DA. Properties of proportional-hazards score tests under misspecified regression models. Biometrics 1984;40:1037-48.
4. Parzen MI, Wei LJ, Ying Z. Simultaneous confidence interval for the difference of 2 survival functions. Scand J Stat 1997;24:309-14.
5. Park Y, Tian L, Wei LJ. One- and two-sample nonparametric inference procedures in the presence of a mixture of independent and dependent censoring. Biostatistics 2006;7:252-67.