Prognostic Factors in Clinical Cancer Trials

Elihu Estey
Department of Hematology, University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030

Abstract
Differences in nontreatment-related covariates ("prognostic factors") often account for differences between treatments in clinical cancer trials. This is true even in randomized trials in which the number of patients randomized is less than 200. Hence analysis of prognostic factors is crucial in historically controlled and small randomized trials. However, it is known that factors found prognostic in one series are often not prognostic in another. This is a result of the increased type I error inherent in most methods used to identify the optimal cutpoint of a potential prognostic factor. This report describes new graphical methods, based on Martingale residuals, that can be used to better identify the relationship between outcome and a covariate. For example, use of these methods indicated that the effect of antecedent hematological disorder on survival in acute myelogenous leukemia/myelodysplastic syndrome is continuous (the longer the length of antecedent hematological disorder, the shorter the survival) rather than dichotomous (antecedent hematological disorder present = unfavorable). This report also discusses the use of graphical methods to test the assumption of proportional hazards crucial to the Cox model and to account for any time-varying effects of a prognostic factor. The graphical methods discussed here provide a better fit of a statistical model to the data and provide more reliable estimates of the effect of a particular variable.

Differences between cancer treatments frequently result from differences between the groups given the treatments rather than from the inherent superiority of a particular treatment (1-3). Therefore, the identification of, and subsequent adjustment for, nontreatment-related characteristics that could influence outcome become critically important in the analysis of those trials in which there is no assurance that the groups given different treatments are otherwise identical. Such is the case in historically controlled trials.

There is also no a priori guarantee that the mere act of randomization provides this desideratum (4) and others have pointed out that trials in which fewer than 200 patients are randomized between two treatments can have significant imbalances. Senn (5) has shown that even a nonstatistically significant imbalance in an important prognostic factor can have a very distorting effect on the unadjusted treatment comparison. Thus, if fewer than this number are randomized, adjustment for imbalances in prognostic factors between the groups (so called covariate adjusting) is needed. Review of the 1996 Journal of Clinical Oncology, however, suggests that covariate adjusting is not routinely done. In particular, only 3 of 17 trials randomizing fewer than 200 patients adjusted for covariates. Trials with fewer than 200 patients comprised 39% of randomized studies in the 1996 Journal of Clinical Oncology. Of course, covariate adjusting is impossible if characteristics that determine prognosis are unknown. Simon (1) has pointed out that the usual regression models account for only 20–25% of the variability in outcome. Although this suggests that there is a certain degree of randomness in outcome, it also stresses the need to identify new prognostic factors and further refine analyses of older ones.

In addition to the essential role of prognostic factors in retrospective analyses of historical and small randomized studies, prognostic factors can be used to prospectively assign patients to treatment. For example, at the University of Texas M. D. Anderson Cancer Center, we used the model shown below to assign patients with newly diagnosed AML to receive or not receive granulocyte-macrophage colony-stimulating factor following administration of chemotherapy (6). The model predicts probability (P) of being alive 28 days after initiation of chemotherapy, with the risk of death declining after this time. The model is a typical logistic regression model:

\[ \ln \frac{p}{1-p} = -0.97 + 1.21 \text{ (performance status)} \\
- 0.76 \text{ (bilirubin)} - 0.79 \text{ (age)} + 1.45 \text{ (neutrophils)} \\
+ 1.39 \text{ (fibrinogen)} + 0.61 \text{ (albumin)} \\
+ 1.13 \text{ (hemoglobin)} - 0.57 \text{ (creatinine)} \]

The value 1 is substituted for performance status in the above equation if the patient’s pretreatment performance status is 0–2, and zero is substituted if the performance status is 3–4; 1 is substituted for age if the patient’s age is under 50 years, 2 is substituted if the age is 50–65 years, and 3 is substituted if the age is >65 years. Similarly, 1 is substituted for albumin if the albumin is <3.2, 2 if the albumin is 3.2–3.4, and 3 if the albumin is ≥3.5. Where did these cutoffs come from? Review of the study (6) describing the model indicates that they were essentially arbitrary. This leads us to a discussion of methods for cutpoint identification in prognostic factor studies.

Probably the simplest way to identify a cutpoint is to use


2 To whom requests for reprints should be addressed, at Department of Hematology, Box 61, University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030. Phone: (713) 792-7544; Fax: (713) 794-4297.

3 The abbreviations used are: AML, acute myelogenous leukemia. AHD, antecedent hematological disorder.
the median value for the variable of interest in the population under study and then to compare the outcome in the group of patients with values for the variable above this cutpoint and the group with values below the cutpoint. However, this method obviously leads to potential loss of information. The same is true with other arbitrary grouping methods. A more objective method to identify a cutpoint is to evaluate the effect of taking various cutpoints and choosing the cutpoint that corresponds to the most significant relationship with outcome. Essentially, the cutpoint associated with the lowest \( P \) obtained when comparing the high-risk group (e.g., with values above the cutpoint) with the low-risk group (e.g., with values below the cutpoint) is chosen. This method has been called the optimal \( P \) approach. Altman et al. (7) described the use of this method in studying the effect of the percentage of cells in the S phase of the cell cycle on survival in breast cancer. They noted that different authors each described different cutpoints as corresponding to the optimal \( P \). Clearly, not all of these could in reality be optimal. Hilsenbeck et al. (8) discuss reasons that factors (or cutpoints) found prognostic in one study are often not prognostic when other data sets are used. They simulated data sets of 250 or 500 cases and divided these into a training set used to obtain the best prognostic factor cutpoint and a validation set used to confirm the cutpoint. As expected, testing multiple cutpoints markedly increased the probability of type I error. Regardless, however, of the number of cutpoints tested on the training sets, the type I error observed in the validation set was stable. Furthermore, the power of the validation set to detect true differences was unrelated to the number of cutpoints. Thus, one solution to the problem of increased type I error consequent to examining multiple cutpoints to select the optimal one is to use an independent validation set. Failing this, other adjustment factors should be used; all have the effect of raising \( P \) from its nominal value (9).

A graphical method to assess the relationship between a presumed prognostic factor and outcome has been described by Therneau and Grambsch (10) and Grambsch et al. (11). The method entails use of smoothed Martingale residual plots. This method will be applied to the prognostic factor AHD. AHD has been defined as a documented abnormality in blood count for at least one month prior to presentation at M. D. Anderson with AML or myelodysplasia. Patients with an AHD are known to have worse outcomes than patients without an AHD (12). We inquired whether the 1-month cutoff for AHD was rational. We applied the optimal \( P \) method to various AHD lengths observed in 530 patients with newly diagnosed AML or myelodysplasia who received AML-type chemotherapy and found that lowest the \( P \) was associated with an AHD cutoff of 9 months rather than 1 month, i.e., patients with an AHD of \( \geq 9 \) months had worse outcomes relative to patients with an AHD of \(<9 \) months than did patients with an AHD of \( \geq 1 \) month relative to patients with an AHD of less than 1 month. We next compared this result with that obtained using the method of Martingale residuals. On the X-axis were plotted lengths of AHDs (in months). On the Y-axis were plotted the residuals corresponding to the various AHD lengths. A residual of zero corresponds to the average outcome (survival) for all 530 patients determined by fitting a Cox model without any covariates. This is the baseline hazard rate for the population. Residuals \( >0 \) denote an excess number of deaths compared to the no-covariates model, and residuals \( <0 \) denote fewer deaths, again compared to the no-covariates model. Higher values for the residuals correspond to greater deviations from the model (a greater excess or deficit of deaths at a particular AHD length). Examination of the smoothed Martingale residual plot indicated that the relationship between AHD length and outcome was best described not by a single cutpoint, but rather as AHD length increased above zero, outcome became worse until a plateau was reached at an AHD length of 10–20 months. This type of relationship would have been difficult to detect without use of the smoothed residual plot. In turn, the relationship between AHD and outcome calls into question the often-used practice of excluding patients from chemotherapy studies if they have an AHD beyond a particular length, usually 1 or 3 months.

The most commonly used model for assessing the effect of covariates on time-to-event outcomes such as remission duration, disease-free survival, or survival is the Cox model (13). An assumption of this model as generally applied is that of proportional hazards. This term means that the relative effect of a covariate is identical over time. For example, if increasing age were linearly and inversely related to survival, the risk associated with age 60 years relative to age 50 years would be the same between years 4 and 5 after beginning treatment as it had been between years 0 and 1. This assumption is frequently not verified, despite the likelihood that in some situations it might appear tenuous. For example in AML, clinicians might suspect that increasing age is primarily a risk for death occurring during the first 1–4 weeks of therapy rather than later on. Grambsch and Therneau (14) have described a graphical method to assess the possible time-varying effect of a prognostic factor such as age. When we used this method in our analysis of factors influencing treatment outcome in the 530-patient AML/myelodysplastic syndrome data set referred to above, we found that the effect of the great majority of prognostic factors, including age, performance status, treatment in a laminar air flow room, and cytogenetics, varied as a function of time from start of treatment. Clearly, the proportional hazards assumption underlying the Cox model is not tenable in many cases.

The model illustrated above that predicts risk of early death in AML does not consider interactions between the covariates. Thus, the effect of, for example, age \( >65 \) years is assumed to be the same in patients with performance status 0–2 and patients with performance status 3–4. It is unlikely that this “no-interaction” assumption is universally true. Modern statistical techniques are capable of detecting interactions and should be applied routinely (15).

In summary, newer methods for analyzing the effect of potential prognostic factors have become available in the last 5–10 years. These include methods for analyzing the precise relationship between a prognostic factor and outcome, for analyzing time-dependent effects of prognostic factors, and for testing for interactions between prognostic factors. These methods provide a better fit of a statistical model to the data and provide more reliable estimates of the effect of a particular variable.
References
