STATISTICAL CHALLENGES IN COMPARING CHEMOTHERAPY AND BONE MARROW TRANSPLANTATION AS A TREATMENT FOR LEUKEMIA

JOHN P. KLEIN AND MEI-JIE HANG

The Medical College of Wisconsin

Comparison of survival for patients treated with either post remission chemotherapy or allogeneic bone marrow transplantation (BMT) for leukemias is considered. Two designs for the comparison are considered. The first is a genetic randomized clinical trial. For this type of trial, comparisons can be made either by an intent-to-treat analysis or by a time dependent covariate model. The second design compares data from a multicenter chemotherapy trial with data from a large transplant registry. Here analysis is complicated by the registry only observing patients who are transplanted so adjustments needs to be made for patients who die or relapse while waiting for transplant. Corrections suggested for this source of bias are a matching technique, inclusion of a time dependent covariate and a left truncated Cox model. We examine these techniques through a small Monte Carlo study and compare how much information is lost by using registry data as compared to a genetically randomized trial.

1. Introduction

Both chronic and acute leukemias are treated by one of two treatment modalities: intensive chemotherapy or bone marrow transplantation. Both treatment regimes have shown varying efficacies for different types of leukemia and for different disease states. A obvious question of clinical significance is which of the these two treatments is better. The comparison presents a number of statistical challenges in design and analysis. In this note we shall examine two designs one may use for comparison of a chemotherapy regime (CT) to an allogeneic bone marrow transplant (BMT). These methods are the so-called genetically randomized trial and the comparison of data from multicenter chemotherapy trials to bone marrow transplant data collected by a large registry.

In both types of studies the outcome of interest is the time to some terminal event. Both will typically start with a time origin at a time t_0 where the patient's disease is diagnosed or in remission. Of clinical interest is the time, measured from this point, to recurrence of the leukemia (Relapse), to death without recurrence of the leukemia (Death in Remission) or to the failure of the treatment when a patient either dies or relapses (Leukemia Free Survival, LFS). When comparing relapse rates, patients who die without recurrence of the leukemia are treated as censored observations while when death in remission is the event of interest

patients who relapse are treated as censored. Care must be taken in interpreting analyses based on relapse or death in remission since the censoring times are not independent. Most comparisons will focus on leukemia free survival rates since this best reflects the success rates of the treatments under study. The leukemia free survival rate is usually very close to the overall survival rate since patients tend to die very soon after relapsing.

The two types of studies share common statistical challenges. The first, once the terminal event is chosen, is the choice of an appropriate time scale. For patients with a suitable donor there is a waiting time from t_0 until the transplant is performed. This time may be relatively short if a donor is readily available and the patient is in reasonably good health except for the leukemia. It may be quite long if no donor is immediately available, if the patient needs additional treatment for conditions which preclude a transplant or for chronic leukemias where a patient may stay in a stable phase for a long time allowing transplant to be electively delayed. Some patients who have an available donor and are scheduled for a transplant may die or relapse while waiting for their transplant. Adjustments must be made for this loss in any analysis.

A second challenge is to account for differences in baseline characteristics between patients receiving the two treatments. These characteristics may have the same effect on outcome for both treatments (e.g., disease state, waiting time to remission), have different effects on outcome for the two treatments (e.g., white blood count at t_0) or affect outcome for only one of the treatments (e.g., Donor-recipient sex match for BMT patients). For patients given a bone marrow transplant there may also be a need to make adjustments for intermediate events that occur at random times in the course of a patient's recovery. For example, one may need to adjust for the occurrence of acute or chronic graft-versus-host disease. While these are important concerns, we shall focus on the first challenge of how to handle the different time scales for chemotherapy and transplant patients.

2. Prospective "Randomized" Trials

The "gold" standard for comparison of therapies in medicine is the randomized clinical trial. Here patients are assigned to treatment by some stochastic mechanism. This randomization serves to balance potential risk factors between the two treatments and remove potential physician and patient biases in selecting treatment.

The ideal randomized clinical trial of chemotherapy to allogeneic bone marrow transplantation would be based on a population of patients who had available, at time t_0, an appropriate donor. The patient would then be randomized to a chemotherapy regime or an immediate transplant. This would allow the LFS in the two arms of the trial to be analyzed by conventional statistical methods such as the log rank test or a proportional hazards regression model. It would eliminate the problem of accounting for the waiting time to transplant in the BMT sample. Such a trial would be easily interpretable by clinicians who are used to similar designs in the comparison of chemotherapy trials.

There are several problems with implementation of such a trial. First, there are logistical problems. These include, for example, the difficulty of having a pool of patients and/or donors available for an immediate transplant, scheduling problems inherent with the need for BMT patients to spend their initial recovery period in special rooms or beds, and the need, in some case, for attention to other conditions a patient may have at the time of diagnosis or remission. Second, there may be ethical problems associated with such a design. For a physician to put a patient on a randomized study he or she must believe that each

treatment is equally likely to be successful. A final problem is that, even if such studies can be implemented, they will involve small sample sizes that will only allow for detection of gross differences between the two treatments.

An alternative to the ideal randomized trial is a trial based on "genetic" randomization. Here sequential patients who meet the disease criterion are entered on study. Patients with a suitable donor are scheduled for a transplant while those without a donor are assigned to the chemotherapy arm. An assumption is made that the availability or non availability of a donor is sufficiently random that the results of such a trial will mimic a purely randomized trial.

There are two possible ways to analyze such a trial. While any of a variety of statistical methods can be used to compare the survival experience in the two arms (cf. Andersen et al (1993) for a survey) we shall focus on the Cox (1972) proportional hazards model. The most common type of analysis is based on an intent-to-treat analysis. Here patients are assigned to the appropriate arm at time t_0 and treatment is modeled by a fixed time covariate. Patients who die or relapse in the transplant arm without receiving a transplant are counted against transplant. The second type of analysis uses a time dependent covariate, $Z(t)$, with the value 1 after a patient is transplanted and 0 if the patient has yet to be transplanted or is in the chemotherapy arm. This is analogous to the type of analysis done on the Stanford Heart Transplant Study (c.f. Turnbull et al (1974)). Note that here patients with a donor who die or relapse prior to transplant are counted against the chemotherapy arm. For both types of analysis adjustments for possible covariates are made to both arms in the final Cox model.

Which type of analysis to use is open to debate (See Nowak (1994) for a recent discussion of these issues). The intent to treat analysis is simple for clinicians to understand. It uses the same time scale for both arms so that natural estimates of the LFS curves can be constructed. It handles deaths or relapses while waiting for transplant quite simply. The time-dependent covariate approach, on the other hand only puts patients in the transplant group after transplant. Since most transplant patients are treated similarly to chemotherapy patients until the time of transplant this may be appropriate. The approach may be more reasonable when some of the risk factors that need adjustment are clearly time dependent as well. For example the donor-recipient sex match is only relevant for patients actually transplanted not those whom we intend to transplant. In section 4 we shall compare the statistical performance of the two methods based on our Monte Carlo study.

Regardless of the analysis method there are several disadvantages to genetically randomized studies. First, they are typically small, single institutional studies so that only very

of patients from many institutions in a natural way to greatly increase the power of the comparison between the two treatment modalities.

A source of data on bone marrow transplantation worldwide is the International Bone Marrow Transplant Registry (IBMTR). This registry, formed in 1972, collects data on successive transplants at 238 transplant centers in 42 countries. On the basis of surveys conducted by the IBMTR these teams account for about sixty percent of all the teams in

to the risk set as their transplant time occurs and are deleted from the risk set as they experience the event or are censored. This test is the left-truncated version of both the fixed time and time dependent Cox models of the genetically randomized trial.

4. Monte Carlo Comparisons

We report here results of a Monte Carlo study comparing various methods for treatment comparisons. A log logistic model was assumed for the time to death or relapse for patients in the chemotherapy group. That is, the hazard rate for a chemotherapy patient is

$$h_c(t) = \frac{k(t/\theta)^{k-1}}{\theta[1+(t/\theta)^{k-1}]}, \quad \text{for } t,\, q,\, k > 0. \tag{4.1}$$

This model has a hump shaped hazard rate that is typical shape of the hazard rate we see in these types of studies. Note that θ is the median time to death and/or relapse.

For a patient in the transplant group we first generate a random transplant time, X, from the following density function

$$f(x) = \begin{cases} \phi x & \text{if } 0 \le x < 8 \\ \alpha \exp(-\gamma x) & \text{if } x \ge 8 \end{cases} \tag{4.2}$$

Once a transplant time is generated, the LFS time for the transplant patient is generated from the following conditional proportional hazards model:

$$h_T(t|X) = \begin{cases} \exp(\beta_1) h_c(t) & \text{if } t < X \\ \exp(\beta_2) h_c(t) & \text{if } t \ge X \end{cases} \tag{4.3}$$

Here the parameter β_1 models pre-transplant differences between the two samples and β_2 the effect of transplant.

Type I censoring was used in the study. Patients were entered into the study at a date E generated from a uniform [0,8] distribution. Patients were censored if $T+E$ was greater than 92 units where T is their LFS time. This insures that all patients have at least six units of follow-up.

Two sets of parameters are reported in this note. In model I we have $\phi = 0.009375, \gamma = 0.1071, \alpha = 0.17\ 7, k = 2$ and $\theta = 10$. In model II we have $\phi = 0.015\ 25, \gamma = 0.25, \alpha = 0.923\ , k = 3$ and $\theta = 20$. Model I corresponds to a long waiting time to transplant with 30% of the transplants taking place prior to 8 time units, while model II corresponds to more early transplants with a median time to transplant of 8. When there is no difference in efficacy between the transplant and chemotherapy cohorts Model I has 11% censoring in both samples and 55% of the event times in the BMT group occurring prior to transplan

Table 2 compares the null performance of five possible test statistics. The first two are based on the complete sample and the remaining three on the truncated sample where patients with events prior to transplanww o

fi5000(truncated)TJflT 000(where)]8JflTfl(patiecomplignor)-1400lesamp7remafacted

Table 3 (Continued): Percent Of 1000 Samples Which Reject H_0 Based On A 5% Level Test

							COMPL T SAMPL		TRUNCAT D SAMPL	
n_b	n_c	Percent Censored Chemo	Percent Censored BMT	Percent Truncated	β_1	β_2	Intent To Treat	Time Dependent Covariate	Left Truncated Cox Model	Match Pairs
100	50	11	19	5	.000	-. 93	27.2	79.9	73.2	45.1
100	50	21	35	13	.000	-. 93	84.0	92.2	91.3	52.
100	50	11	7	5	.000	. 93	31.2	91.8	84.5	33.2
100	50	21	13	13	.000	. 93	8 .3	9 .0	93.	53.
100	50	11	1	39	-. 93	.000	48.3	34.1	4.8	4.
100	50	20	21	9	-. 93	.000	. .93			

the survival experience of the transplant group pretransplant is comparable to that of the chemotherapy group. The most powerful analysis of studies of this type is that based on a method that accounts for delayed entry of BMT patients in the risk set at the time of transplant and not on matching. In fact, matching, if done inappropriately, may lead to erroneous conclusions with a rather high probability.

For complete samples we see that the time dependent covariate approach has the best power if the two groups mortality experience is similar prior to transplant. In discussing our Monte Carlo model with investigators in this area we were told that, after adjustments for initial covariates, the pretransplant hazard rates should be similar in the two groups. Which analysis to use depends on the assumptions to be made by the investigator. Note that in complete samples these are testable assumption.

In our Monte Carlo study we ignored other possible covariates that need to be adjusted for. We believe that after these adjustments similar conclusions should hold.

A picture or a survival curve is often worth as much to clinical investigators as a formal test. A product-limit estimator of survival curve can be computed using the left truncated data from a registry. This curve is an estimator of the conditional survival of a patient who was transplanted (see Andersen et al (1993) for details). The product-limit estimator based on the chemotherapy data is an estimator of an unconditional survival curve. An other summary survival curve is due to Begg et al (1984) which provides an estimator of the conditional probability of survival for a chemo patient given this time is larger that a randomly selected transplant time. This method while it has merits ignores the right truncated nature of the time to transplant in the BMT group. Further investigation into the merits of these estimates or into alternative methods of summarizing this data is warranted.

Acknowledgments

This research was supported by Grants 1 R01 CA5470 -03 from the National Cancer Institute and P01-CA-40053 from the National Cancer Institute, the National Institutes of Allergy and Infectious Diseases and The National Heart, Lung and Blood Institute. Thanks to Mary M. Horowitz, MD and Philip A. Rowlings, MD of the IBMTR for their medical insight into this problem and helpful comments.

References

Andersen, P. K., Borgan Ø, Gill, R. D. and Keiding, N. (1993), *Statistical Methods For Counting Processes*, Springer-Verlag, New York.

Begg, C. B., McGlave, P. B., Bennett, J. M., Cassileth, P. A. and Oken, M. M. (1984), "A Critical Comparison Of Allogeneic Bone Marrow Transplantation And Conventional Chemotherapy As Treatment For Acute Nonlymphocytic Leukemia," *J. Clin. Oncol.*, 2, 3 9-78.

Bortin, M. M., Horowitz, M. M. and Rimm, A. A. (1992), "Progress Report From The International Bone Marrow Transplant Registry," *Bone Marrow Trans.*, 10,113-122.