Pitfalls in the Design of Clinical Trials for Prevention or Treatment of Acute Graft-versus-Host Disease

Paul J. Martin, Richard A. Nash

Division of Clinical Research, Fred Hutchinson Cancer Research Center and Department of Medicine, University of Washington, Seattle, Washington

Correspondence and reprint requests: Paul J. Martin, MD, Fred Hutchinson Cancer Research Center, PO Box 19024, Seattle, WA 98109-1024 (e-mail: pmartin@fhcrc.org).

Received October 21, 2005; accepted November 7, 2005

ABSTRACT

This review addresses pitfalls of clinical trials to evaluate new approaches for prevention or treatment of graft-versus-host disease. Determination of end points poses a difficult challenge in the design of clinical trials, and examples from previous studies are used to illustrate some of the pitfalls. Also discussed is the need for a new conceptual approach for evaluation of graft-versus-host disease after nonmyeloablative conditioning regimens, because the donor antirecipient alloimmune reaction is the primary mechanism of benefit with this type of treatment. Finally, investigators should be aware of regulatory and socioeconomic pitfalls that apply to all clinical trials.

© 2006 American Society for Blood and Marrow Transplantation

KEY WORDS

Hematopoietic cell transplantation ● Graft versus host disease ● Clinical trials

INTRODUCTION

This review of pitfalls in clinical trials for prevention or treatment of acute graft-versus-host disease (GVHD) reflects the authors’ personal experience and does not represent a global survey of the literature. The merits of different end points for prevention and treatment studies are evaluated, and some of the limitations of phase II studies are outlined. The increased reliance on immune reactions to control malignant disease after nonmyeloablative conditioning regimens leads to suggestions for a new conceptual approach to the evaluation of GVHD. Finally, some of the regulatory and socioeconomic pitfalls that apply to all clinical trial are discussed.

END POINTS FOR PHASE III PREVENTION STUDIES

Survival as an End Point

Survival has been a difficult end point to achieve in GVHD prevention studies. Increased survival related to an intervention to prevent GVHD can be demonstrated only if the risk of GVHD in the study population is very high and the risks of non-GVHD complications are very low and balanced between arms. In addition, the intervention must have a large beneficial effect, with no increased risk of fatal adverse effects that would offset the benefits of preventing GVHD. Throughout the history of hematopoietic cell transplantation (HCT), very few studies have met all of these criteria, and for this reason, very few studies have demonstrated improved survival through interventions designed to prevent GVHD.

Table 1 presents hypothetic data to illustrate the difficulty of demonstrating an improvement in survival related to more effective prevention of acute GVHD. After HCT with unrelated donors in the late 1980s and early 1990s, the expected frequency of grade IV GVHD was approximately 10%, and for these patients, the risk of nonrelapse mortality approaches 100%. The expected frequency of grade III GVHD was approximately 30%, and for these patients, the nonrelapse mortality is approximately 50%. Grade II GVHD has little detectable effect on nonrelapse mortality after HCT. The expected frequency of grades 0 to II GVHD after unrelated transplantation was approximately 60%, and the combined non-

This research was supported by grants CA18029 from the Public Health Service, Department of Health and Human Services. Clinical trials comparing cyclosporine and tacrolimus for prevention of acute graft-versus-host disease were supported in part by grants from Fujisawa Healthcare.
relapse mortality at 1 to 2 years for these patients was approximately 25%. Under these conditions, the overall expected nonrelapse mortality in the control arm of a study to prevent GVHD would be estimated at approximately 40%.

A reasonably effective therapy to prevent GVHD would be expected to prevent grade IV GVHD and possibly decrease the incidence of grade III GVHD from 30% to 15%. Given the relationships between GVHD grade and survival described above, however, this striking reduction in GVHD risk would be expected to decrease nonrelapse mortality from 40% to 29% (Table 1). A randomized prospective clinical trial to detect the difference between 40% and 29% would require enrollment of at least 600 patients.

### Table 1. Hypothetical Effect of Graft-versus-Host Disease on Nonrelapse Mortality

<table>
<thead>
<tr>
<th>GVHD Grade</th>
<th>Standard Prophylaxis</th>
<th></th>
<th>Risk of NRM</th>
<th>Expected NRM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td></td>
<td></td>
<td>Expected</td>
</tr>
<tr>
<td>0–II</td>
<td>0.6</td>
<td>0.25</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>0.30</td>
<td>0.5</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>0.10</td>
<td>1.0</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>0.40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experimental Prophylaxis</th>
</tr>
</thead>
<tbody>
<tr>
<td>GVHD Grade</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>0–II</td>
</tr>
<tr>
<td>III</td>
</tr>
<tr>
<td>IV</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

GVHD indicates graft-versus-host disease, NRM, nonrelapse mortality.

### GVHD as an End Point

Survival is difficult end point for GVHD studies, but the measurement of GVHD itself as an end point for prevention studies also has pitfalls. Qualitative components of the diagnosis are very difficult to document. In most cases, clinicians can confidently distinguish rash caused by GVHD from rash related to other causes, but the case report forms used for clinical trials cannot easily capture the qualitative elements that led to the distinction. The assessment of GVHD severity is often confounded by complications other than GVHD. For example, both GVHD and sinusoidal obstruction syndrome cause hyperbilirubinemia, and the distinction between the two must be based on concomitant findings such as ascites, right upper quadrant tenderness, or rash. Hyperbilirubinemia in the absence of rash is usually not related to GVHD, but hyperbilirubinemia presenting concurrently with rash is usually interpreted as related to GVHD. The principle of parsimony does not always apply, however, and patients can have both GVHD and other complications simultaneously. Finally, the peak severity of GVHD reflects not only the efficacy of the prophylaxis, but also the efficacy of later treatment. In a prevention study, the effects of treatment can blunt differences between the arms that would otherwise be much more apparent.

### Examples from Previous Phase III Prevention Studies

Two important prospective trials [1,2] illustrate some of the problems associated with end-point determination related to GVHD. Both studies were conducted to compare tacrolimus and cyclosporine for prevention of acute GVHD. In one study [1], patients had HLA-matched sibling donors, and in the other [2], patients had HLA-matched unrelated donors. These multicenter studies are among the largest phase III randomized GVHD prevention trials ever done. The primary end point for each study was grade II to IV GVHD, and both had an open-label design.

A variety of methods were used to assess GVHD in these studies. One was clinical assessment by the primary investigators. Because this was an open-label study, the primary investigators knew whether patients were taking cyclosporine or tacrolimus, and this knowledge could have influenced their assessment of GVHD. A computer algorithm was also used to evaluate all-cause morbidity involving the skin, liver, and gastrointestinal tract. This algorithm did not distinguish between morbidity related to GVHD versus other causes. A further, and perhaps more telling, indication of whether a patient had GVHD was steroid administration. In addition to receiving steroids for treatment of GVHD, however, some patients received steroids as substitute prophylaxis when the calcineurin inhibitor could not be given. The study of unrelated HCT in which this end point was analyzed did not distinguish among the different reasons for administration of steroids.

A 3-member end-point evaluation committee independently assessed the acute GVHD end point for all patients participating in the study of unrelated HCT. The committee members were expert clinicians from institutions that were not participants in the clinical trial. Committee members retrospectively reviewed abstracts and flow sheets of the clinical data and pertinent supportive documentation, but they were blinded to the study drug assignment. Biopsy specimens were available for review if subsequent pathologic evaluation was necessary. Flow sheets used by the committee described clinical manifestations at weekly intervals and summarized laboratory results, biopsy reports, and complications other than GVHD that might confound the diagnosis. Compliance with methotrexate administration was recorded. Compliance with administration of the study drug was also reported, together with information indicating blood
who might not share those same biases.

In the study of patients with HLA-matched sibling donors [1], the incidence of GVHD determined by the investigator was 26% in the tacrolimus arm and 41% in the cyclosporine arm (Table 2). The computer algorithm reflecting morbidity from all causes had much higher estimates, 80% and 85%, respectively. In the study with unrelated donors [2], investigator assessment again showed a statistically significant difference between arms (Table 3). The frequency of steroid administration was higher in both arms, in part because of substitute prophylaxis when the calcineurin inhibitor could not be given. The committee underestimated the incidence of GVHD compared with reports by the investigators, possibly because the reviewers did not have information about GVHD treatment. Differences in the estimates of GVHD incidence by the various methods used in these two studies are striking.

In the study with unrelated donors, data from the primary investigator assessments (Table 4) showed an inverse correlation between the number of patients enrolled at each site and the percentage-point difference between the two arms. At the largest center, the difference between the two arms was only 6%. In centers B, C, D, E, and F, the differences between arms were in the vicinity of 21% to 26%. In the smallest centers, enrolling only 10 patients in one center and 7 patients from 3 other centers, the differences between arms were 50% to 60%. Smaller centers with a limited number of physicians could have had strong biases about the efficacy of tacrolimus compared with cyclosporine, whereas in larger centers, the care is spread across many more individuals who might not share those same biases.

Table 2. Incidence of Grades II to IV Graft-versus-Host Disease

<table>
<thead>
<tr>
<th>Method of Evaluation</th>
<th>Tacrolimus, N (%)</th>
<th>Cyclosporine, N (%)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investigator [2]</td>
<td>46 (51)</td>
<td>63 (70)</td>
<td>.01</td>
</tr>
<tr>
<td>Steroid administration</td>
<td>59 (66)</td>
<td>73 (81)</td>
<td>.02</td>
</tr>
<tr>
<td>Committee</td>
<td>18 (20)</td>
<td>29 (32)</td>
<td>.06</td>
</tr>
</tbody>
</table>

GVHD indicates graft-versus-host disease.

*Ratanatharathorn et al [1].
†Nash et al., unpublished data.

Table 4. Inconsistent Results Between Centers

<table>
<thead>
<tr>
<th>Sites(s)</th>
<th>No. of Patients</th>
<th>Difference between Arms, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>75</td>
<td>6</td>
</tr>
<tr>
<td>B</td>
<td>25</td>
<td>21</td>
</tr>
<tr>
<td>C</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>D</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>E</td>
<td>13</td>
<td>26</td>
</tr>
<tr>
<td>F</td>
<td>13</td>
<td>26</td>
</tr>
<tr>
<td>G</td>
<td>10</td>
<td>60</td>
</tr>
<tr>
<td>HIJ</td>
<td>7</td>
<td>50</td>
</tr>
</tbody>
</table>

Based on graft-versus-host disease grades assigned by investigators (Nash et al., unpublished data).

These observations raise the concern that the diagnosis of GVHD was influenced by the awareness of whether the patient was taking cyclosporine or tacrolimus. In these studies, it is unlikely that GVHD was underdiagnosed among patients who received tacrolimus. If GVHD evolved to the point where it needed to be treated, it should have been diagnosed and treated in the interest of good patient care. It is possible, however, that GVHD was overdiagnosed among patients who received cyclosporine, particularly if clinicians considered cyclosporine to be less effective for preventing GVHD. In this case, the increased suspicion of GVHD would lead to earlier treatment than would have been the case for an otherwise identical patient who had received tacrolimus. Patients who received cyclosporine and early treatment were labeled as having GVHD for a condition that could conceivably have resolved spontaneously if treatment had been delayed, as might have been the case for some of the patients in the tacrolimus arm. This concern, however, would not apply for the committee evaluation of GVHD in the study of unrelated HCT, because the committee did not have information about GVHD treatment.

Another way of evaluating outcomes in GVHD prevention trials is to assess overall morbidity across time (Nash et al., unpublished observations). Results from the unrelated transplantation trial [2] showed that the burden of morbidity among patients with grades III to IV GVHD was statistically higher than among those with grades 0 to II GVHD. It was very difficult to demonstrate that the burden of morbidity among patients with II to IV GVHD was higher than among those with grades 0 to I GVHD. The study arms were then compared in proportional odds models to determine whether there was any difference across time in the severity of rash, the degree of hyperbilirubinemia, or the severity of gut symptoms. Two different methods were used to calculate an overall score; one method followed the standard algorithm for grading GVHD and the other assigned less weight to the severity of gut symptoms. The results showed no significant differences between the study arms, but...
the trends suggested a lower overall burden of morbidity in the skin and liver among patients in the tacrolimus arm. Almost all patients had gut symptoms caused by the conditioning regimen, and this finding may explain the absence of differences in gut morbidity between arms.

Recent studies have shown improvements in the consistency of GVHD grading by different methods. Weisdorf et al. [3] showed excellent consistency of GVHD grading when an evaluation committee was informed of the treatment decisions that were made by the clinicians at the bedside. The sophistication of the computer algorithm was also improved, probably by taking into account treatment and not simply evaluating morbidity, so there was close agreement with results of the panel review. The inclusion of treatment decisions in the committee evaluation and the computer algorithm, however, introduces an unavoidable subjective component in the grading that could be affected by the bias of investigators who are aware of the study arm assignments in open-label studies.

END POINTS FOR PHASE III TREATMENT STUDIES

Compared with the large number of studies assessing GVHD prophylaxis, relatively few studies of GVHD treatment have been reported, in part because of the difficulty and complexity of evaluating the results of GVHD treatment. Minimal criteria for improvement or progression of GVHD in each organ must be defined, and few published studies have given explicit consideration to the effects of other posttransplantation complications involving the gut or liver. End-point definition is also a problem, and the boundaries between GVHD grades are arbitrary to some extent. The timing of assessments can be crucial, because GVHD can show a waxing and waning course. Furthermore, an individual organ such as the skin or gut can demonstrate regional differences in response to treatment, and improvement of GVHD in one organ can be accompanied by worsening in another.

Competing risks and the complexity of add-on designs can cause other pitfalls. At any given time point, some patients may have already received additional therapy because of lack of satisfactory response to initial treatment, and others may have died. If the primary end point is GVHD severity across time, the analysis must account for patients who have died or returned home. In studies with an add-on design, the underlying standard treatment given to both arms can blunt any differences between arms.

Another difficulty is assessment of improvement. For example, the outcome for a patient could be classified as improvement because the GVHD has changed from grade III to grade II severity, but this change in grade may have been diagnosed on the basis of a decrease in the total serum bilirubin concentration from 3.1 to 2.9 mg/dL. To circumvent problems of this type, we have used a more sophisticated algorithm for assessing outcomes after treatment of acute GVHD [4]. The definition of skin improvement, for example, required a more than 25 percentage-point reduction in the total body surface affected by rash. Progression was the opposite, a greater than 25 percentage-point increase, and no change was the absence of either improvement or progression. We validated the definitions by showing that patients who had complete response, as we defined it at the time, had a lower incidence of nonrelapse mortality than those who had a partial response or mixed response (improvement in one organ but worsening in another). The patients with the worst outcome had progression or no change.

We assessed response with this algorithm in a double-blind prospective clinical trial of a CD5-specific immunotoxin plus steroids compared with steroids alone for treatment of GVHD [5]. We collected data weekly and categorized patients as having a partial or complete response. The results showed that GVHD improved within the first few weeks in most patients. The main end point of interest was the proportion of patients who had complete responses. There was a 2- to 4-week advantage in time to resolution of GVHD in the patients who received the immunotoxin, but the end results at 6 weeks were similar in the two arms. This study illustrates the problem of blunting, in that steroid treatment alone produced a high complete response rate, making it difficult to demonstrate any improvement resulting from additional treatment.

AVOIDING PITFALL IN PHASE III STUDIES

Several methods can be used to circumvent the pitfalls of GVHD trials. Blinding can be an effective way to lessen the impact of investigator subjectivity, but blinding can be difficult to carry out. Blinding was suggested for the trials comparing tacrolimus and cyclosporine, but such a design would have required the collaboration of two competing pharmaceutical companies. A blinded design would likely have increased the number of capsules that patients would have to take, and it would have been necessary to normalize the reporting of drug levels to preserve the blind while allowing physicians to use this information in the clinical treatment of patients.

Another crucial element is prespecification of the end points and the timing of assessments. The primary end point has to be selected to demonstrate clinical benefit, either by prolonging life or by reducing GVHD-related symptoms such as rash or nausea.
and vomiting. Clinical benefit is more difficult to define for hepatic GVHD, because hyperbilirubinemia usually does not cause symptoms, although the degree of hyperbilirubinemia could be used as a surrogate end point that is strongly correlated with nonrelapse mortality [6].

Adequate statistical power for the primary end point is essential. To that end, a phase II study or robust historic data are needed to define expectations for the study arm and the control arm in a phase III clinical trial, so that the power analysis is informed by data and not by mere guesses. Balancing risk factors between arms is also important, as demonstrated by the trial comparing cyclosporine and tacrolimus among patients with sibling donors [1,7].

**LIMITATIONS OF PHASE II STUDIES**

Awareness of problems in GVHD assessment leads to a critical examination of phase II studies in which there is lack of blinding and a poorly defined primary end point. In many phase II studies the timing of assessments and the statistical analysis are not prespecified. The eligibility criteria for a clinical trial reflect a balance between several conflicting interests that can drastically influence the results. The investigator’s desire to demonstrate success leads to tight eligibility criteria that exclude patients who are at high risk of complications, whereas the desire to complete the study may lead to a relaxed set of eligibility criteria, potentially making it more difficult to demonstrate success. In the provider role, the investigator is motivated by a desire to help patients and may encourage enrollment out of hope that the study intervention will be effective. Finding the appropriate balance among these interests can be extremely difficult.

Patients make their own assessments of benefits and risks when deciding to participate in a clinical trial. For example, in a recent study to test the efficacy of a CD25 immunotoxin to prevent acute GVHD, the number of patients who declined to participate in a study was larger than the number who enrolled in the study [8]. The patients who enrolled had higher-risk disease and were significantly older than those who declined to participate. Such selection biases may go unnoticed unless efforts are made to record the characteristics of patients who decline to participate in clinical trials.

**EVALUATION OF GVHD AFTER NONMYELOABLATIVE CONDITIONING REGIMENS**

Regimen-related toxicity can be decreased by the use of nonmyeloablative conditioning regimens, and there is an increasing reliance on immunologic effects of donor cells for efficacy in HCT. A complete reversal of priorities in regard to GVHD may occur for a patient who has had recurrent or progressive malignancy, and this change will affect GVHD analysis. As a result, there is an increased willingness to accept some GVHD as a necessary cost of treatment.

Consistency remains, however, in the desire to avoid severe adverse consequences related either to acute or chronic GVHD—“too much” GVHD [9]. A typical patient with acute myeloid leukemia may undergo extended hospitalization of 60 to 90 days in a given year in the course of treatment. If this length of hospitalization would be considered tolerable for management of acute leukemia, it may also be considered tolerable for a patient with GVHD. On the other hand, persistent or significant disability or incapacity related to bronchiolitis obliterans, contractures, or corneal disease might not be considered acceptable as the price of overcoming leukemia. By the same token, recurrent life-threatening infections requiring hospitalization or causing death related to GVHD or immunosuppressive treatment are obviously situations we want to avoid. Most would agree that these outcomes exemplify too much GVHD. Perhaps this judgment could be used as an end point for clinical trials, testing the level of success in allowing enough immunologic activity to control the malignancy, while at the same time avoiding the consequences of too much GVHD.

**REGULATORY AND SOCIOECONOMIC PITFALLS**

Awareness of regulations related to the conduct of clinical trials is vitally important for investigators. Formal training programs are now becoming available to familiarize academic clinicians with the Code of Federal Regulations, the International Conference on Harmonization guidelines, and Food and Drug Administration Good Clinical Practice guidelines.

Concerns about academic credit can impede the development of multicenter clinical trials. When it comes to promotions, the number of publications is considered together with an interpretation regarding the role of the candidate, often based on the position within the list of authors. In smaller centers that cannot contribute large numbers of patients to a clinical trial, the contribution will most likely result in listing as a middle author, which will have less impact than listing as first, second, or last author.

As an intrinsic part of the scientific method, investigators do not have completely neutral views when a clinical trial is designed. Belief in the potential superiority of a new treatment underpins the design of any clinical trial, as expressed in the statistical considerations that specify the quantitative benchmarks of success to be used at the end of the trial. Regardless of any expectations or hopes of success at the beginning.
of a clinical trial, the investigator must step back to make disinterested and unbiased judgments at the end of the trial, even though it would be more rewarding to publish positive results than negative results.

Financial entanglements can pose difficult problems for academic investigators. Some believe that academic investigators should have a completely neutral judicial role, in which any funding or any benefit coming directly from someone who has an advocacy role in the advancement of a product must be avoided. Medical journals now deal with this problem by providing disclosure statements, but these statements generally do not provide enough information to determine the extent to which objectivity might have been compromised by any financial support or other benefits.

Differences between corporate culture and academic culture can cause unanticipated problems. Corporate success is measured by profit, while academic success is measured by publication, and finding common ground can be very difficult. For example, a decision by a pharmaceutical company to close a clinical trial prematurely for financial reasons would represent a complete loss for an academic investigator who could not publish the results of the study. Likewise, the two cultures might conflict when it comes to publication of negative results that might be harmful to the company but beneficial to the academic investigator and the public.

CONCLUSIONS

The complexity of GVHD poses many challenges in the design of clinical trials to evaluate new approaches for prevention or treatment of GVHD. Awareness of potential pitfalls leads to recognition of the necessity for clearly defined and realistic end points that reflect genuine clinical benefit. Clinical investigators must balance the demands of regulatory agencies, corporate interests, and academic advancement with their mission to provide patient care. The biases inherent in a scientific hypothesis-generating, hypothesis-testing mind-set at the beginning of a clinical trial must be replaced by scrupulously protected objectivity when the results are judged at the end of a clinical trial.

REFERENCES