Published in final edited form as:

Best Pract Res Clin Haematol. 2008 June ; 21(2): 357–372.

Study design and endpoints in graft-versus-host disease

Paul J. Martin, MD

Member, Division of Clinical Research, and Professor of Medicine, University of Washington, Division of Clinical Research, Fred Hutchinson Cancer Research Center, P.O. Box 19024, 1100 Fairview Avenue North, Suite D2-100, Seattle, WA 98109-1024, USA, and Department of Medicine, University of Washington, Seattle, Washington, USA

Abstract

The design of clinical trials for prevention or treatment of acute or chronic graft-versus-host disease poses many challenges. These challenges include the selection of primary and secondary endpoints that demonstrate clinical benefit, and the identification of measures indicating success both for individual patients and groups. Assessment of response in treatment trials should ideally encompass the prior trajectory of change before treatment. The criteria, timing and duration of response should be specified, and the potential effects of concomitant treatment and complications other than GVHD should be taken into account in assessing outcomes. A crucial element in clinical trial design is the pre-specification of the hypothesis to be tested in quantitative terms. Potential barriers to enrollment should be carefully considered in order to ensure timely completion of the trial.

Keywords

graft-versus-host disease; clinical trials; design

A variety of inherent challenges must be considered in designing a clinical trial of prevention or treatment of acute or chronic graft-versus-host disease (GVHD). Clinical trials generally serve to establish the safety and efficacy of a medical product for a specific purpose. In designing any clinical trial, three critically important questions must be considered: (1) who will be enrolled in the trial, (2) how will subjects be treated, and (3) how will results be evaluated? An NIH consensus development project recently developed a set of recommendations for clinical trials in chronic GVHD, as summarized in Table 1.¹ The consensus document on study design was focused mostly on technical details and was intended serve as a useful reference guide for readers who might wish to undertake a clinical trial for patients with chronic GVHD.

This chapter focuses more broadly on the challenges involved in the study design and endpoints in clinical trials for prevention or treatment of acute or chronic GVHD. Rather than repeating previously published technical information, attention is directed toward more ‘philosophical’ personal views that have not been tempered through an iterative process of formal consensus development.

*Tel.: +1-206-667-4798 ; Fax: +1-206-667-5255. E-mail address: pmartin@fhcrc.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

CLINICAL AND REGULATORY MINDSETS

The design of a clinical trial depends to some extent on whether the results will be used ultimately for ‘academic’ or ‘regulatory’ research purposes. Important cultural differences exist between ‘clinical’ and ‘regulatory’ mindsets. A clinical mindset has great interest in comparing available treatment alternatives. Clinicians have a high tolerance for uncertainty, reflecting the frequent need to make at least outwardly confident treatment recommendations, often relying on personal experience when higher standards of formal evidence are not available. In a ‘clinical’ mindset, studies that show no statistically significant difference for the primary endpoint might still be useful, particularly if differences are observed for important secondary endpoints, perhaps akin to a billiards game in which any ball in any pocket counts as success. If robust clinical endpoints are difficult to attain, a clinical mindset might be inclined to accept non-validated surrogate endpoints as evidence of efficacy. With this mindset, results of clinical trials are typically evaluated according to a ‘preponderance of evidence’ standard in the context of the overall literature.

A regulatory mindset is focused on formal evaluation of the efficacy and safety of an investigational product proposed for marketing approval to be used for a specific indication. The basis for approval depends on documentation of overall clinical benefit, as demonstrated by prolonged survival or improved quality of life. A regulatory mindset has low tolerance for uncertainty. Demonstration of statistically significant differences for a pre-specified primary endpoint is of paramount importance in the interpretation of results, and concordant results from two formal studies are typically needed for approval, akin to a billiards game in which only ‘called shots’ count, and successes must be repeated as a demonstration of genuine skill as opposed to luck. Any surrogate endpoints must be validated to show strong correlation with clinical benefit. With this mindset, results of clinical trials are typically evaluated according to a ‘beyond reasonable doubt’ standard in the context of previous approvals.

Trial designs can be greatly enhanced if certain ideals are upheld even when studies are not intended for regulatory review. The Food and Drugs Administration (FDA) has summarized the characteristics of adequate and well-controlled ‘pivotal’ studies needed for regulatory approval in the United States.² These characteristics include: (1) a clear statement of the objectives and methods of analysis, with an implicit emphasis on prospective design; (2) a valid comparison with a control to provide a quantitative assessment of effect; (3) a method of subject selection that provides adequate assurance that they have the disease being studied; (4) a method to avoid bias in the assignment of subjects to treatment and control groups; (5) adequate measures to minimize bias on the part of subjects, observers and analysts of the data; (6) well-defined and reliable methods to assess subjects’ responses; and (7) an analysis that is adequate to assess the effects of the treatment.

ENDPOINTS

Selection of endpoints

The choice of specific primary and secondary endpoints will be influenced by the nature and scope of claims to be made at the end of the study, the type of intervention, study phase, indication (prevention, primary treatment or secondary treatment), conduct in an academic versus a community setting, and by the presence or absence of blinding. Valid primary and secondary endpoints should focus on the clinical benefits of central importance to patients: survival free of the underlying disease; freedom from GVHD manifestations, treatment and complications; and improvement in symptoms, function and quality of life. Intermediate endpoints that reflect the degree of GVHD activity – such as physician assessment, biochemical tests, and biomarkers³ – are important only to the extent that they may predict or confirm one of the patient-experienced outcomes.^{4,5} In the evaluation of hepatic GVHD, for example, the

degree of hyperbilirubinemia could be used as a surrogate endpoint on the basis of a strong correlation with non-relapse mortality.⁶

Survival as an endpoint for GVHD trials

Improved survival has been a difficult endpoint to demonstrate in GVHD studies. Increased survival rates 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 2 presents hypothetical 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 non-relapse mortality approaches 100%.⁷ The expected frequency of grade-III GVHD was approximately 30%, and for these patients the non-relapse mortality rate is approximately 50%. Grade-II GVHD has little detectable effect on non-relapse mortality after HCT. The expected frequency of GVHD grades 0–II after unrelated transplantation was about 60%, and the combined non-relapse mortality at 1–2 years for these patients is approximately 25%. Under these conditions, the overall expected non-relapse 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 non-relapse mortality from 40% to 29% (Table 2). A randomized prospective clinical trial to detect the difference between 40% and 29% would require enrollment of at least 600 patients.

MEASURES OF SUCCESS

Prevention trials

A variety of endpoints has been used for GVHD prevention trials. Some trials have focused on the occurrence of any GVHD (including grade I), but most often the primary endpoint is related to the occurrence of GVHD grades II–IV, since these grades generally require systemic immunosuppressive treatment and are therefore considered clinically significant. Other trials have focused on GVHD grades III–IV, since these grades are associated with increased risk of non-relapse mortality, whereas grade-II GVHD has little or no detectable effect on risk of non-relapse mortality after HCT for treatment of malignant diseases.

The assessment of GVHD itself as an endpoint for prevention studies has many well-recognized pitfalls. 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 leading 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.⁸ 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.

Strategies for preventing GVHD are intended to avoid clinical manifestations of the disease and the need for systemic immunosuppressive treatment beyond the originally planned prophylaxis regimen. Since actions speak louder than words, and since some degree of uncertainty surrounds the diagnosis of GVHD, a strong case can be made that the use of systemic immunosuppressive treatment for GVHD might serve as the most relevant endpoint for prevention studies (Table 3). Since decisions regarding the need for systemic immunosuppressive treatment for acute GVHD involve some degree of subjective clinical judgment, results of a trial with this primary endpoint would be most persuasive with the use of a double-blind design.

Treatment trials

The distinction between disease activity and damage plays a critical role in the evaluation of response to treatment for GVHD. Disease activity causes manifestations that would be expected to resolve after pathogenic immunological mechanisms are controlled. Certain manifestations generally reflect inflammation and include erythematous rash, oral erythema and lichenoid changes, diarrhea, hyperbilirubinemia, transaminase elevations, and eosinophilia, which may be expected to resolve within weeks after pathogenic immunological mechanisms are controlled (Table 4). Other manifestations that are characteristic of chronic GVHD typically reflect fibrosis and include sclerotic skin changes, fasciitis, and contractures.⁹ Control of disease activity is required in order to halt progression of these manifestations, and resolution occurs very slowly during a period of months to years through a prolonged process of tissue remodeling. Disease damage leads to manifestations that are generally considered to be irreversible.¹⁰ These manifestations reflect tissue destruction caused by chronic GVHD and include ocular and oral sicca, bronchiolitis obliterans and vitiligo. At best, an arrest of disease activity would be expected to halt progression of these manifestations. Measurement of response in clinical trials should account for differences in the expected rates of improvement between inflammatory and fibrotic manifestations of GVHD, and under certain circumstances an arrest of previously progressive damage could be considered as therapeutic benefit, even if the intervention does not provide functional improvement.

Most phase-II studies have relied on GVHD response as the primary endpoint. While 'response' is easily recognized in clinical practice, documentation of response for purposes of clinical trials (i.e., convincing other people who cannot interview or examine the patient) is much more difficult. Definitions of response for chronic GVHD have been proposed by the NIH Consensus Development Project but have not yet been validated,¹¹ and time to response depends on the inflammatory or fibrotic nature of disease manifestations among patients enrolled in the study. Results might be affected by changes in ancillary treatment and supportive care.¹² Response assessments are subject to bias, particularly in open-label, single-arm studies.

Complete response or strictly defined nearly complete response might serve as the most relevant endpoint for treatment studies (Table 3). For acute GVHD, such a response could be defined as the absence of erythematous rash attributable to GVHD, a total serum bilirubin concentration <2.0 mg/dL in the absence of hepatic complications other than GVHD, or <3.0 mg/dL in the presence of such complications, absence of anorexia, nausea or vomiting attributed to GVHD, and the presence of at least intermittently formed stools. The endpoint should make limited allowances for abnormalities caused by complications other than GVHD, such as hyperbilirubinemia related to hemolysis or calcineurin inhibitor concentrations that are

inadvertently elevated above the therapeutic range, and diarrhea related to infections or oral administration of magnesium.

Qualifiers of success and comparative metrics in GVHD trials

As a general rule, success in treatment trials cannot be claimed for individual participants who receive additional therapy for GVHD before response is evaluated, unless the study is intended to demonstrate that the investigational product improves the response to subsequent treatment. An investigational approach that effectively prevents or eliminates all GVHD manifestations could prove disadvantageous if it increases the risk of infections, non-relapse mortality or recurrent malignancy, or if survival is decreased compared to control results with a standard approach that is less effective for preventing or treating GVHD. For this reason, claims of success in both prevention and treatment trials must be qualified by data demonstrating that the investigational approach is not associated with an increased risk of major adverse events or complications caused by impaired immune reconstitution or loss of immunological effects against malignant cells. Claims of success in a treatment trial must also be qualified by data demonstrating that the elimination of GVHD manifestations was not related to higher doses of steroid treatment given with the investigational approach. In fact, one would expect that the amount of steroid treatment would be decreased if an investigational approach effectively controlled GVHD.

Comparisons between trials would be facilitated if publications included some standard metric of overall success. The proportion of subjects surviving to day 80 after HCT without systemic immunosuppressive treatment for acute GVHD could serve this purpose very well for GVHD prevention trials. Follow-up should be truncated at day 80 after HCT, since longer follow-up results could be confounded by the development of chronic GVHD. With this metric, participants who die without systemic treatment for GVHD before day 80 and those who develop GVHD after withdrawal of immunosuppressive medications in order to treat recurrent malignancy before day 80 would count as failures. The numbers of such events would have to be interpreted in light of pre-transplant risk factors.

The proportion of participants surviving to 6 months after enrollment without resort to a different systemic treatment might serve as a standard metric for GVHD treatment trials. Resort to a different systemic treatment implies that the investigational treatment did not provide the desired benefit. Survival for 6 months with uncontrolled GVHD would be unusual, and most deaths caused by infection related to excessive immunosuppression would be expected to occur within 6 months. Longer follow-up results could be confounded by recurrent malignancy. Deaths within 6 months related to recurrent malignancy would count as failures, and the numbers of such events would have to be interpreted in light of pretransplant risk factors. Since decisions regarding the need for treatment rely on clinical judgment, these measures would be subject to bias in open-label studies, but they could be highly robust indicators of success in blinded trials for prevention or treatment of GVHD.

Challenges with add-on designs

Many trials for prevention or treatment of GVHD employ designs in which a study product is added to a 'standard' regimen. Although this approach can work well for prevention studies, the lack of consensus on starting doses and tapering schedule of steroid administration complicates the design of primary treatment studies, and clinically unavoidable variation in steroid doses at the onset of secondary treatment can confound the interpretation of results from studies of secondary treatment. Given the frequent adverse events associated with high-dose steroid treatment, physicians have a medical and ethical imperative to use the lowest effective steroid dose, and protocols must allow investigators to use case-by-case discretion in the management of steroid treatment. In this situation, physicians will calibrate steroid doses

according to the severity of GVHD manifestations: higher doses will be given to patients with more severe manifestations, while lower doses will be given to those with less severe manifestations. In the end, the difference in steroid administration could obscure any potential differences in the frequency of response between the two arms.

Subjects can be harmed by excessively stringent definitions of response in GVHD treatment studies. For example, response definitions that require complete absence of GVHD-related rash early after starting treatment might discourage appropriate tapering of steroid doses when the rash is improving at a medically acceptable rate. Likewise, appropriate tapering of steroid doses might be discouraged by response definitions that require absence of all GVHD manifestations for a prolonged period of time after complete response is attained. In order to encourage the use of steroids at the lowest effective dose, response definitions should make some allowance for intermittent minor flares of GVHD manifestations.

Figure 1 illustrates the relationship between intensity of immunosuppressive treatment and survival after treatment for acute GVHD. Insufficient treatment is likely to result in low survival because of multi-organ failure related to GVHD, while excessive treatment is likely to result in low survival because of fatal infections. An effective treatment for GVHD should have a wide therapeutic window that allows GVHD to be controlled while preserving a sufficient immune function to control infections. In the absence of laboratory tests that measure the intensity of immunosuppressive treatment, physicians must be guided by the severity of GVHD manifestations. Immunosuppressive treatments that suppress all manifestations of GVHD for prolonged periods of time are likely to be associated with profound immunodeficiency that first becomes apparent at the onset of a life-threatening opportunistic infection.

DIMENSIONS OF RESPONSE IN TREATMENT TRIALS (TABLE 5)

Trajectory of change before treatment

The interpretation of outcomes after an investigational treatment depends on the 'trajectory' of changes before enrollment in the study. For example, stable FEV₁ values before and after treatment would indicate that the treatment had no effect, whereas treatment-induced stabilization of FEV₁ values after a period of progressive deterioration before treatment could indicate benefit. Eligibility criteria for secondary treatment trials should emphasize not only the condition of the patient at the time of enrollment but also the progression of events leading to the decision that additional treatment was needed. Such decisions could be based on judgments that GVHD manifestations were getting worse, staying the same for too long a duration of time, or not improving at an acceptably rapid rate.

Response criteria

Clinical protocols should define the criteria for response, including the organ systems and sites involved by GVHD, the types of manifestations to be measured, and the units of measurement. Protocols should specify units of measurement as categorical changes, minimum absolute change, or minimum relative changes that qualify as sufficient for response. Categorical changes should be formulated to ensure that differences are clinically meaningful. For example, an improvement from grade-III GVHD to grade-II GVHD might not be meaningful if the change reflected a decrease in the serum total bilirubin concentration from 3.1 mg/dL to 2.9 mg/dL. This problem can be prevented by requiring a minimum categorical change of two grades or stages in order to qualify as a response (e.g. from grade-III GVHD to grade-I GVHD).

Criteria for complete response should include the maximum deviation from normal values that would still qualify as complete or near-complete response, and criteria for partial response should state the minimum amount of change needed to categorize a change as 'response.' If possible, active, reversible disease manifestations should be distinguished from fixed deficits

and irreversible damage. Protocols should define criteria for response not only for each organ or site but also for the disease as a whole.

Timing of assessments

The protocol should define the timing of response assessments that will be used for statistical analysis. These assessments can be done on a single pre-specified occasion, on more than one pre-specified occasion, or as often as weekly. Single response measurements that are made too early (e.g. <2 weeks after starting treatment for acute GVHD, or <2 months after starting treatment for chronic GVHD) might not allow sufficient opportunity for the benefits of an investigational treatment to become apparent, while response measurements that are made too late (e.g. ≥ 6 weeks after starting treatment for acute GVHD, or ≥ 12 months after starting treatment for chronic GVHD) might overlook an accelerated onset of response induced by an investigational treatment.¹³ This problem can be addressed by using time-to-event analyses, as opposed to analysis of proportions at one or more fixed time points.

Duration and continuity of response

The minimum duration of response required to designate a treatment for acute or chronic GVHD as 'successful' will vary according to the intended duration of treatment, trial phase and endpoint. Products intended for use on a single occasion would be expected to produce durable subsequent benefit, while those intended for chronic use would be expected to provide benefit at least for the duration of treatment. For these reasons, protocols should specify requirements for durability of response in the definition of endpoints. Where applicable, protocols should define criteria for failure of primary therapy and failure of secondary treatment.

Concomitant treatment

Concomitant therapy must be controlled when response is the primary endpoint of phase-II studies designed to assess the efficacy of an investigational agent. 'Standard' steroid therapy currently represents a well-established mainstay of initial treatment for both acute and chronic GVHD. Protocols should provide guidelines for tapering of glucocorticoid doses and should specify daily or alternate-day administration. The guidelines should indicate the appropriate starting dose and any adjustments in glucocorticoid doses, depending on toxicity and changes in GVHD severity. A fixed starting dose and taper schedule may be appropriate for clinical trials where administration of glucocorticoids is an integral component of the intervention being tested and for trials to evaluate short-term efficacy of a new regimen. In longer-term trials, however, adherence to a fixed schedule is likely to be impossible because of variation among subjects in persistence of GVHD and tolerability of glucocorticoid-related side effects. Protocols should highlight the possibility of adrenal insufficiency at the end of the glucocorticoid taper and should provide guidelines for reinstitution of glucocorticoid treatment if GVHD recurs after glucocorticoid administration has been discontinued.

Protocols should provide guidelines for the administration of other treatments that could affect manifestations of GVHD and potentially confound the interpretation of the study. Information about the administration of any such treatments should be recorded for analysis at the end of the study. Study requirements must find a balance between the physician's potential need for flexibility in deviating from a prescribed course versus the study's need for consistent medical management. This balance becomes especially difficult in open-label studies and in trials with longer-term endpoints that involve physician judgment regarding decisions to add or withdraw medications from the immunosuppressive regimen or to adjust the pace of taper.

Accounting for complications other than GVHD

As discussed above, the interpretation of outcomes in patients with GVHD can be confounded by complications other than GVHD. A variety of approaches can be taken to deal with this problem. One approach would be to censor observations with organs affected by complications other than GVHD at the onset of treatment or at the time of evaluation. The pitfall of this approach is that an investigational treatment itself could alter the susceptibility of an organ to complications other than GVHD. For example, it would not be fair to censor the occurrence of diarrhea if an investigational agent increased the risk of diarrhea caused by *Clostridium difficile* infection. A alternative approach would be to include all manifestations that could be caused by GVHD in the assessment, even when other causes of abnormalities are clearly present. With this approach, however, patients with skin GVHD and severe sinusoidal obstruction syndrome of the liver could not be categorized as having a complete response unless both GVHD and veno-occlusive disease resolve. A third approach would be to define clearly articulated rules that make reasonable allowances for complications other than GVHD.¹⁴ Considerations of complications other than GVHD are pertinent primarily in studies that claim efficacy. In this situation it would be useful to demonstrate that the claim of efficacy remains valid regardless of the method used to account for complications other than GVHD.

IMPORTANCE OF A PRE-SPECIFIED HYPOTHESIS

A crucial element in clinical trial design is the pre-specification of the hypothesis to be tested in quantitative terms. This requires a statement of the 'null' hypothesis, indicating the expected proportion of successes if the investigational treatment has no efficacy, and a statement of the 'alternative' hypothesis, indicating the expected proportion of successes if the treatment has the desired level of efficacy. In addition, the statistical design must indicate the acceptable chance of type-1 error or ' α ', indicating the probability of a false-positive result, and the acceptable chance of a type-2 error or ' β ', indicating the probability of a false-negative result. Statistical power is defined as $1 - \beta$. These four specifications determine the number of subjects to be enrolled in the study.

For phase-II studies, robust historical data are needed in to define reasonable expectations for the study group. Care should be taken to ensure that the inclusion and exclusion criteria for selection of historical subjects are identical to those used for the phase-II study when the null and alternative hypotheses are formulated. The interpretation of results should also examine the potential effects of any differences in risk factors between the historical and the study cohorts. Differences could exist, because patients make their own assessments of benefits and risks when deciding to participate in a clinical trial. 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.¹⁵ The patients who enrolled had higher-risk disease and were significantly older than those who declined to participate. Selection biases might go unnoticed unless efforts are made to record the characteristics of eligible patients who decline to participate in clinical trials.

Similarly, the design of phase-III studies should be informed by results of phase-II studies or by robust historical data. With the same null and alternative hypotheses and specifications for α and β , the number of patients needed for a two-arm phase-III study is much larger than the number needed for a one-arm phase-II study. Hence, the commitment in undertaking a phase-III study is much larger than required for a phase-II study. Most large centers can conduct single-institution phase-II studies, but phase-III studies are difficult even for large centers. Since the data from phase-III studies will always be much more persuasive than those from phase-II studies, efforts to organize and conduct phase-III studies should be encouraged as much as possible.

Disease characteristics at the time of enrollment can influence the chance of partial or complete response, and the management of glucocorticoid doses and other concomitant treatment can have marked effects on short-term outcomes. In open-label studies, the evaluation of many endpoints is highly susceptible to bias. In the absence of well-defined response criteria, judgments regarding attainment of complete response are likely to be much more robust than those regarding partial response. However, the short duration of phase-II studies and presence of fixed, irreversible deficits may limit the number of patients who have a complete response. Nonetheless, phase-II studies are useful for screening treatment options and planning future phase-III studies.

Although the pre-specified primary endpoint is of paramount importance in assessing the results of a phase-III clinical trial, other considerations must be taken into account before an overall conclusion can be reached. For example, a secondary analysis of the primary endpoint should demonstrate that a favorable difference in outcome does not reflect imbalance in the distribution of baseline risk factors among arms of the study.

In GVHD trials, differences in the doses of systemic immunosuppressive medications or the administration of topical therapy among the arms of the study could affect the primary endpoint. An effective treatment for chronic GVHD would be expected to reduce the overall level of systemic immunosuppression and the amount of topical therapy. If this is not the case, then a secondary analysis should be carried out in order to correct for possible confounding effects of ancillary treatment and supportive care on the primary endpoint. Finally, favorable results for the primary endpoint should be supported by favorable or neutral results for secondary efficacy endpoints and by the absence of side-effects that outweigh the benefits demonstrated by the efficacy endpoints.

ENROLLMENT BARRIERS (TABLE 6)

Over-restrictive eligibility criteria

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.

Differences in practices among participating institutions

The numbers of patients with GVHD in any single institution are generally not sufficient for studies intended to establish safety and efficacy for investigational products in a reasonable time frame. Studies that enroll patients at several institutions are more likely to meet accrual targets and provide the necessary data to analyze outcomes in a robust manner. A major challenge is to coordinate and align the practices for the purpose of an efficient trial design and unbiased analysis at institutions that have evolved divergent standards based on their own resources and experience. For multi-institutional studies, investigators should ensure that (1) inclusion criteria are defined as broadly as possible; (2) exclusion criteria are defined as narrowly as possible; (3) protocol requirements are limited to those necessary for successful conduct of the study; and (4) non-essential supportive care provisions are formulated as flexible guidance that allows variation according to the existing practices of participating institutions.

Multi-institutional studies provide faster enrollment and allow wider generalization of results. Protocols for multi-institutional studies should address standardization of routine management

strategies, monitoring, and endpoint assessment, since differences in management can confound the assessment of patient-reported outcomes and quality of life. Audits and centralized training of individuals who are to assess patients, follow protocol-mandated procedures, examine histopathology, perform chronic GVHD grading, and assess clinical trial endpoints may improve compliance and reproducibility. Likewise, centralized review of histopathology, clinical grading and response assessments may improve reproducibility and uniformity among institutions. These quality enhancements add greatly to the costs of conducting clinical trials.

Concerns about academic credit and reimbursement

When it comes to promotions, the number of publications is considered, together with an interpretation regarding the role played by 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. Enrollment in clinical trials could potentially be improved by designs that leave the smallest possible 'footprint' on clinical practice and by ensuring that reimbursement for costs is commensurate with the amount of work required for the study.

Absence of clinical equipoise

Enrollment rates in clinical trials can be extremely rapid in studies where a novel and potentially effective investigational agent is available only through participation in a clinical trial. Enrollment rates are predictably much slower in clinical trials with products that are readily available as a result of approval for another indication. Physicians who believe that a product is inadequately effective for GVHD on the basis of personal experience, case reports or limited phase-II studies have little clinical motivation to enroll patients, while those who believe that a product is effective can prescribe it at their own discretion, thereby avoiding the burdens and restrictions imposed by participating in the trial. Enrollment will be slow, unless physicians can set aside such personal beliefs in the interest of improving the quality of evidence regarding the efficacy of products for treatment of GVHD.

Concern about loss of graft-versus-malignancy (GVM) effects

Many retrospective studies have demonstrated inverse correlations between the occurrence or severity of acute or chronic GVHD and the risk of recurrent malignancy after HCT, and numerous reports have described the disappearance of malignant cells after donor lymphocyte infusion for treatment of recurrent malignancy after HCT.¹⁶ The recent advent of non-myeloablative conditioning regimens has further increased awareness of the important role played by donor T and NK cells in eliminating malignant cells in the recipient after HCT.¹⁷ Many physicians are willing to accept some level of GVHD as necessary for optimal control of malignancy. Concerns about possible loss of GVM effects can cause reluctance to participate in studies designed to prevent or treat GVHD. To some extent, the historical experience with studies of T-cell depletion and results from a study with high versus standard doses of cyclosporine justify such concerns.^{18,19} Studies directed toward prevention or treatment of acute or chronic GVHD should include mechanisms to ensure that any increased risk of recurrent malignancy can be detected as early as possible.

Uninteresting hypothesis

Most clinical trials for GVHD have tested products previously approved for another indication. Improvements in clinical trial design can facilitate advances in the treatment of GVHD only if these products or new therapeutic agents currently under development have activity against physiologic pathways leading to development or progression of the disease. A sophisticated

and detailed understanding of the pathophysiology leading to acute GVHD has emerged from experimental studies with mice.^{20,21} Current understanding of the pathophysiology leading to chronic GVHD is limited, but deeper insight could come from preclinical models and from studies of other diseases that result in similar clinical manifestations, such as scleroderma and Sjogren's syndrome. A more sophisticated understanding of the pathogenesis of GVHD, combined with the availability of agents directed at specific targets in pathways leading to development of the disease, could enhance the participation of investigators and patients in clinical trials.

CONCLUSIONS

Unlike solid-organ transplantation, which requires life-long immunosuppressive treatment, allogeneic HCT ultimately results in state of immunological tolerance in which donor cells persist without causing immune-mediated injury in the recipient even in the absence of immunosuppressive medications, while immune defenses against pathogens are preserved. Although this review has emphasized 'response' as a proximal measure of success after treatment for acute GVHD, development of immunological tolerance remains the ultimate goal and distal benchmark of success in preventing or treating GVHD. Hence, reports of clinical trials for prevention or treatment of acute or chronic GVHD would be enhanced by including data regarding time to withdrawal of all immunosuppressive treatment after resolution of GVHD, although extended follow-up is required in order to ascertain this endpoint.

REFERENCES

1. Martin PJ, Weisdorf DW, Przepiorka D, et al. National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: VI. Design of Clinical Trials Working Group Report. *Biol Blood and Marrow Transplant* 2006;12:491–505. [PubMed: 16635784]
2. 21 CFR 314.126(b)
3. Schultz KR, Miklos DB, Fowler D, et al. Toward biomarkers for chronic graft-versus-host disease: National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: III. Biomarker Working Group report. *Biol Blood and Marrow Transplant* 2006;12:126–137. [PubMed: 16443511]
4. Wittes RE. Antineoplastic agents and FDA regulations: square pegs for round holes? *Cancer Treatment Reports* 1987;71:795–806. [PubMed: 3621212]
5. Anonymouns. Outcomes of cancer treatment for technology assessment and cancer treatment guidelines. American Society of Clinical Oncology. *J Clin Oncology* 1996;14:671–679.
6. Gooley TA, Rajvanshi P, Schoch HG, McDonald GB. Serum bilirubin levels and mortality after myeloablative allogeneic hematopoietic cell transplantation. *Hepatology* 2005;41:345–352. [PubMed: 15660388]
7. Martin PJ, Schoch G, Gooley T, et al. Methods for assessment of graft-versus-host disease (letter). *Blood* 1998;92:3479–3481. [PubMed: 9787194]
8. Strasser, SI.; McDonald, GB. Gastrointestinal complications. In: Blume, KG.; Forman, SJ.; Appelbaum, FR., editors. *Thomas' Hematopoietic Cell Transplantation*. Third Edition. Malden MA: Blackwell; 2004. p. 769-810.
9. Filipovich AH, Weisdorf D, Pavletic, et al. National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: I. Diagnosis and Staging Working Group report. *Biol Blood and Marrow Transplant* 2005;11:945–955. [PubMed: 16338616]
10. Shulman HM, Kleiner D, Lee SJ, et al. Histopathologic diagnosis of chronic graft-versus-host disease: National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: II. Pathology Working Group report. *Biol Blood and Marrow Transplant* 2006;12:31–47. [PubMed: 16399567]
11. Pavletic SZ, Martin P, Lee SJ, et al. Measuring therapeutic response in chronic graft-versus-host disease: National Institutes of Health consensus development project on criteria for clinical trials in

- chronic graft-versus-host disease: VI. Response Criteria Working Group report. *Biol Blood and Marrow Transplant* 2006;12:252–266. [PubMed: 16503494]
12. Couriel D, Carpenter P, Cutler C, et al. Ancillary Therapy and Supportive Care of chronic graft-versus-host disease: National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: V. Ancillary Therapy and Supportive Care Working Group report. *Biol Blood and Marrow Transplant* 2006;12:375–396. [PubMed: 16545722]
 13. Martin PJ, Nelson BJ, Appelbaum FR, et al. Evaluation of a CD5-specific immunotoxin for treatment of acute graft-versus-host disease after allogeneic marrow transplantation. *Blood* 1996;88:824–830. [PubMed: 8704237]
 14. Przepiorka D, Weisdorf D, Martin P, et al. 1994 consensus conference of acute GVHD grading. *Bone Marrow Transplant* 1995;15:S98–S103.
 15. Martin PJ, Pei J, Gooley T, et al. Evaluation of a CD25-specific immunotoxin for prevention of graft-versus-host disease after unrelated marrow transplantation. *Biol Blood Marrow Transplant* 2004;10:552–560. [PubMed: 15282533]
 16. Fefer, A. Graft-versus-tumor responses. In: Blume, KG.; Forman, SJ.; Appelbaum, FR., editors. *Thomas' Hematopoietic Cell Transplantation*. Third Edition. Malden MA: Blackwell; 2004. p. 369-379.
 17. Sandmaier, BM.; Storb, R. Nonmyeloablative therapy and hematopoietic cell transplantation for hematologic disorders. In: Blume, KG.; Forman, SJ.; Appelbaum, FR., editors. *Thomas' Hematopoietic Cell Transplantation*. Third Edition. Malden MA: Blackwell; 2004. p. 1164-1176.
 18. Marmont AM, Horowitz MM, Gale RP, et al. T-cell depletion of HLA-identical transplantation in leukemia. *Blood* 1991;78:2120–21301. [PubMed: 1912589]
 19. Bacigalupo A, Van Lint MT, Occhini D, et al. Increased risk of leukemia relapse with high-dose cyclosporine A after allogeneic marrow transplantation for acute leukemia. *Blood* 1991;77:1423–1428. [PubMed: 2009366]
 20. Ferrara JL, Reddy P. Pathophysiology of graft-versus-host disease. *Seminars in Hematol* 2006;43:3–10.
 21. Shlomchik WD. Graft-versus-host disease. *Nature Rev Immunol* 2007;7:340–352. [PubMed: 17438575]

Figure 1. Relationship between intensity of immunosuppression and probability of survival. Low intensity or inadequate duration of treatment is associated with low survival because GVHD is not adequately controlled, while excessive intensity or duration of treatment causes fatal infections.

Table 1
Summary of recommendations.¹

Who will be enrolled?

- Inclusion and exclusion criteria should encompass as many people with chronic GVHD as possible without compromising the ability to interpret results of the study
- Baseline evaluations should document the need for therapy, identify prognostic characteristics, and specifically characterize the condition of subjects at the time of enrollment, so that results of therapy can be interpreted

How will participants be treated?

- Within reason, the study protocol should specify or provide guidance regarding the dosing and dose adjustment of all immunosuppressive medications, including the study drug; reasons for deviations should be noted in case-report forms

How will results be evaluated?

- Case-report forms should be calendar-driven by the protocol to provide assessment of chronic GVHD and adverse events at regular intervals. Patient-reported measures should be incorporated whenever feasible. Standardized and clinically validated measures should be used
 - Primary and secondary endpoints should be selected for their ability to demonstrate clinical benefit, which can be an improvement in the way a patient feels or functions or a prolongation of survival. Endpoints should account for competing outcomes such as death or recurrent malignancy. Composite endpoints may be required in some specific protocols
 - Biostatistical analysis should incorporate considerations of competing events and concomitant therapy, appropriate power calculations, interim analyses, sensitivity analysis and missing measurements
-

Table 2
Hypothetical effect of GVHD on non-relapse mortality (NRM)

GVHD grade	Expected frequency	Risk of NRM	Expected NRM
Standard prophylaxis:			
0–II	0.6	0.25	0.15
III	0.30	0.5	0.15
IV	0.10	1.0	0.10
Total			0.40
Experimental prophylaxis:			
0–II	0.85	0.25	0.21
III	0.15	0.5	0.08
IV	0	1.0	0
Total			0.29

Table 3

Measures of success in preventing or treating graft-versus-host disease (GVHD).

Trial type	Primary measure of success for individual participants	Qualifiers of success for the study group as a whole	Comparative metric for the study group as a whole
Prevention	<i>No systemic immunosuppressive treatment for GVHD</i>	No increase infections No increase non-relapse mortality No increase recurrent malignancy No decrease survival	Proportion of participants surviving to day 80 <i>without systemic immunosuppressive treatment for GVHD</i>
Treatment	Complete or nearly complete resolution of GVHD manifestations (preferred) ^a Complete or partial resolution (alternative) ^a	No increase infections No increase non-relapse mortality No increase recurrent malignancy No decrease survival Early or rapid steroid withdrawal	Proportion of participants surviving for 6 months <i>without resort to a different treatment</i>

Italic identifies criteria that are susceptible to bias.

^aWithout additional therapy for GVHD before response is evaluated.

Table 4

Time to onset of response, according to pathophysiology of disease manifestations.

Pathophysiology	Onset of response	Manifestations
Inflammation	Weeks	Cutaneous erythema Oral manifestations Diarrhea Abnormal liver function
Fibrosis	Months	Cutaneous sclerosis Fasciitis Panniculitis
Destruction	Not seen	Oral and ocular sicca Bronchiolitis obliterans Vitiligo

Table 5
Dimensions of response.

-
- Trajectory of change before enrollment: deterioration versus stability versus slow improvement
 - Criteria: categorical versus minimal absolute change versus minimal relative change versus stabilization
 - Timing of assessments: single occasion versus multiple occasions versus continuous measures
 - Duration and continuity of response
 - Concomitant treatment
 - Accounting for complications other than GVHD: before enrollment and after enrollment
-

Table 6
Barriers to enrollment.

-
- Over-restrictive eligibility criteria
 - Differences in practices among participating institutions
 - Concerns about academic credit
 - Reimbursement not commensurate with costs
 - Absence of clinical equipoise
 - Concerns about loss of graft-versus-malignancy effects
 - Uninteresting hypothesis
-