2014 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

Abstract

The complexity of chronic graft-versus-host disease (GVHD) and the lack of fully developed research methods have made it difficult to design, conduct, and analyze clinical trials involving subjects with this disease, even when promising treatment options are available. Recommendations from the 2006 Clinical Trials Working Group Report remain as pertinent and as important as they were when the report was written, but subsequent experience has identified significant gaps and opportunities for further improvement. Major issues addressed in this report include the definition of eligibility criteria, the development, validation and selection of primary and secondary endpoints, the establishment of benchmarks that could be used to set the null hypothesis for the primary endpoint in single-arm studies, and the mapping of development paths that could support regulatory review. Development of a standardized, validated clinical scale for measuring global response is an urgent unmet need in the field. Attention to standards required for regulatory review and approval could improve the design, conduct, documentation, reporting and interpretation of studies that are not intended for regulatory review.

The opinions expressed here are those of the authors and do not represent the official position of the National Institutes of Health, US Food and Drug Administration, or the US Government.

Introduction

The complexity of chronic GVHD and the lack of fully developed research methods have made it difficult to design, conduct, and analyze clinical trials involving subjects with this disease, even when promising treatment options are available. The 2006 Clinical Trials Working Group Report[1] offered several important recommendations for investigators as an approach for overcoming these obstacles. It was agreed that clinical trials in chronic GVHD should adhere to principles of good trial design and practice. Inclusion and exclusion criteria should allow as many subjects to participate as possible without compromising the interpretation of results. Pre-enrollment assessment of chronic GVHD characteristics should be standardized. The protocol should provide clear guidance about administration of study medication and other interventions. Methods of assessing response should be defined and validated in advance. Efficacy endpoints should be selected to reflect clinical benefit. Expert biostatistical support is needed to ensure the validity and reliability of trial results. It was anticipated that the use of consistent standards in clinical trial designs to evaluate agents that have activity in pathogenic pathways could facilitate advances in the treatment of chronic GVHD.

Work during the past decade since the NIH Consensus Conference in 2004 has yielded improvements in the precision and accuracy of criteria for the diagnosis and staging of chronic GVHD, the interpretation of histopathology, the discovery and validation of biomarkers for diagnostic and prognostic applications, and supportive care for patients with chronic GVHD. These results as summarized in previous reports from the 2014 NIH Consensus Conference have set the stage for much needed progress in the definition of response criteria and the design considerations to be applied in clinical trials testing the efficacy and safety of products for treatment of chronic GVHD. Progress toward the development of response criteria has been described in a separate report. The current report is focused on considerations for the design of clinical trials.

While the original recommendations from the 2006 Clinical Trials Working Group Report were broad-based and grounded in good clinical practice, opportunities remain for further improvement. This report addresses the definition of eligibility criteria, the development, validation and selection of primary and secondary endpoints, the establishment of benchmarks that could be used to set the null hypothesis for the primary endpoint in single-arm studies, and the mapping of development paths that could support regulatory review for market approval. Lessons gained from mapping development paths for regulatory review could also be applied in order to improve the design, conduct, documentation, reporting and interpretation of studies that are not intended for regulatory review.

Background

Development of more effective treatments for chronic GVHD is an urgent unmet clinical need. No products have regulatory approval for this indication. Regulatory applications are most likely to come as new indications for approved products, but in certain cases, they could also come as new products for the specific indication of chronic GVHD. The numbers of patients available for enrollment in clinical trials evaluating products for treatment of chronic GVHD is limited. In the U.S., approximately 8,000 allogeneic hematopoietic cells transplants are now done each year.[2] Among these, at least 35% would

be expected to develop chronic GVHD requiring systemic treatment,[3] such that the total incidence is approximately 3,000 per year. Based on reported rates of survival, recurrent malignancy and withdrawal of immunosuppression after resolution of chronic GVHD,[4] the total prevalence in the U.S. is estimated at less than 10,000.

To date, results from eight randomized trials for treatment of chronic GVHD have been published,[5-12] but none of these studies demonstrated superiority of the investigational arm. Much of the published literature has suffered from the absence of any true benchmarks that would enable an informed statistical design in studies of treatment for chronic GVHD.[13] Instead, results focus on response rates assessed by often poorly defined criteria, under the premise, but not the evidence, that no responses would have occurred in the absence of the investigational treatment. This premise might not be true, especially if the prior trajectory of the disease and the effects of other elements in the treatment regimen are taken into account. These include changes in the doses of steroid, calcineurin inhibitors or sirolimus and addition of topically active agents implemented at the same time when the investigational treatment was started or at any time after enrollment but before the assessment of response.

Guidelines and standards emerging from established development paths could improve the quality of studies by academic sponsors even if these studies are not intended to support regulatory review. Most trials are likely to have Phase I or II designs to test previously approved products for the new indication of steroid-refractory GVHD, sometimes focused on a specific manifestation, as exemplified in current studies for treatment of ocular (NCT01616056) and pulmonary involvement (NCT01307462). Advances are likely modest, but results could nonetheless lead to incremental improvements for patient care.

Goals of Treatment for Chronic GVHD

Treatment of chronic GVHD is intended to produce a sustained benefit by reducing symptom burden, controlling objective manifestations of disease activity and preventing damage and disability, without causing disproportionate toxicity or harms related to the treatments themselves. The long-term goal of GVHD treatment is the development of immunological tolerance, indicated by successful withdrawal of all immunosuppressive treatment without recurrence or clinically significant exacerbation of disease manifestations. It is not known whether currently available immunosuppressive products can provide a benefit of this type in patients with chronic GVHD, since few, if any, are known to induce tolerance. Even if they do not shorten the time to develop immunological tolerance, however, immunosuppressive products could provide clinical benefit if they reduce symptom burden, control disease activity and prevent damage and disability more effectively, while causing no greater burden of adverse effects than currently available treatments. Alternatively, immunosuppressive products could provide clinical benefit if they are equivalently effective with respect to currently available treatment but cause a lesser burden of adverse effects.

Eligibility Criteria

Well-defined eligibility criteria are needed for all trials. Inclusion criteria depend on the specific medical indication for treatment. For chronic GVHD treatment trials, the possible indications include global

systemic effect, specific systemic effect such as fibrosis, or local effect on specific organs such as pulmonary disease. Exclusion criteria have several purposes, including the protection of patients who could be harmed by participation in the study and elimination of factors that could confound the interpretation of results. At the same time, the eligibility criteria should be designed so that the enrolled patients are representative of patients with the indication.

Patients can be enrolled in clinical trials either with or without the medical necessity of a treatment change. Enrollment in primary and secondary systemic treatment trials is motivated by the immediate need to relieve symptoms, control disease activity, prevent damage and disability, and if possible, promote tolerance induction. New onset of chronic GVHD prompts the need for primary treatment, and unsatisfactory response to previous treatment prompts the needs for secondary treatment. In secondary treatment trials, the minimum dose and duration of prior treatment and the severity or trajectory of worsened disease manifestations must be defined and documented. Standardized definitions would facilitate comparisons between results of different studies. It is also possible that patients with stable manifestations of chronic GVHD could enroll in clinical trials in the absence of any immediate need for a treatment change. The eligibility criteria for such studies would require longitudinal assessment and documentation of symptom burden, disease activity and damage across some minimum time interval in order to demonstrate that manifestations of the disease were truly stable.

The role of biomarkers in defining eligibility for clinical trials has not been established. Validated biomarkers that reliably reflect the severity of chronic GVHD manifestations or predict the likelihood of response to treatment would be very useful in the design and conduct of clinical trials. Objective laboratory-based biomarkers measured with standardized assays would be very useful in comparing the baseline characteristics of patients enrolled in different studies.

Controlled Designs

Adequate and well-controlled studies of investigational products intended for systemic control of chronic GVHD in patients who need an immediate treatment change are feasible. In such a study, one arm would receive the investigational product, and the other arm could receive any other treatment considered within the scope of usual practice, since no standard of care has been established for this indication. Blinded trial designs would be optimal but are not always feasible. Open label trials could be conducted, but highly robust, objective response endpoints would be needed for regulatory review. “Add-on” designs are appropriate for studies of first-line treatment. In such a study, one arm could receive the investigational product plus conventional treatment, and the other arm would receive conventional treatment alone. No precedent has been established for designs in which one arm would receive an investigational product without conventional treatment and the other arm would receive conventional treatment for chronic GVHD, although such an approach has been used in a trial of treatment for mild acute GVHD.[14] Such a design could be used if prior studies indicate that the investigational product has activity in patients with chronic GVHD.

Adequate and well-controlled studies are also feasible when an immediate change of treatment is not needed. In such a study, one arm would receive the investigational product, and the other arm would continue the baseline management. The design of such studies could include "induction" and "maintenance" phases with different doses of the same product or with the sequential use of different products. Adequate and well-controlled studies of investigational products intended for effect at a specific site or on a specific organ or manifestation of chronic GVHD are feasible, particularly if the use of the investigational product can be blinded.

Endpoints

The primary endpoint in a clinical trial represents the major criterion by which success with the use of the investigational product will be determined, but it is far from the only criterion in judging the merits of an intervention. The primary endpoint should reflect clinical benefit, defined as surviving longer or living with fewer symptoms or improved function. Overall success with the primary endpoint is defined in statistical terms, based upon a pre-specified null hypothesis, an alternative, and the corresponding requisite sample size that affords adequate statistical power and a two-sided false-positive rate conventionally set at 5% or less. The null hypothesis is typically set by the standard of care. A list of endpoints most commonly used in clinical trials of chronic GVHD therapy is found in detail in Table 1 of the initial publication of this Working Group.[1]

In successful trials, secondary endpoints provide necessary additional evidence that benefits exceed harms. For example, successful trial results should show that the benefits of a high response rate are not offset by low survival, as has been reported in some studies of treatment for acute GVHD.[15] As discussed below, progress has been made in developing benchmark rates that could be used to set the null hypothesis for some endpoints but not for others.

Response. Assessment of response compares manifestations of chronic GVHD for each patient at baseline and at one or more defined subsequent time points. Trials using response as an endpoint should be designed to measure and document the durability of response and to determine whether continued treatment is needed in order to maintain response. For a variety of reasons, response at any single time point after enrollment is an incomplete indicator of clinical benefit. Response should be assessed at multiple time points in all studies in order to demonstrate sustained benefit. A "response" cannot be considered as success for an investigational treatment if it occurred only after a subsequent treatment was introduced. This same problem applies for all potential endpoints other than death. Premature treatment changes for reasons other than disease progression or inadequate response could increase the risk of false-negative results in a single-arm study but would not increase the risk of false-positive results. Premature treatment change would not affect results of a controlled, double-blind prospective study.

Most investigational products are likely to be used in conjunction with anti-inflammatory glucocorticoids and other agents. Trials using response as an endpoint should be designed to distinguish the effects of the investigational product from the effects of concomitant treatment. Single arm comparisons between baseline and subsequent symptom burden, disease activity and level of damage and disability

for each patient are generally not sufficient for this purpose. A case could be made if the pre-enrollment trajectory of disease manifestations is clearly documented for each patient to show lack of prior improvement during treatment with the baseline regimen, and if no systemic or topical treatment is concurrently or subsequently increased or added before the assessment of response. Otherwise, designs that include a control group given the concomitant treatment without the investigational product are necessary in order to distinguish the effects of the investigational product from the effects of concomitant treatment.

No standardized, validated global response measure has been developed for studies testing products for treatment of chronic GVHD. Such clinical scales have been developed for regulatory review of other disease indications, including Crohn's disease,[16] rheumatoid arthritis,[17] ankylosing spondylitis,[18] systemic lupus erythematosus,[19] and myositis.[20] Response defined with the use of these clinical scales typically requires improvement in at least 2 defined objective measures, together with improvement in some minimum number of other items from a defined menu of objective, subjective or laboratory measures. Patient reported outcomes represent an important component in each of these clinical scales.

The provisional criteria proposed by 2005 National Institutes of Health (NIH) Consensus Conference for measuring treatment response of chronic GVHD were based on expert opinion,[21] and an Excel spreadsheet tool has been developed to apply these criteria in clinical trials.[22] Responses defined according to the proposed algorithm correlated with improved symptom burden but not with improved quality of life by other measures.[23] Furthermore, agreement between response and physicians' clinical assessment was poor.[24] Response at 6 months correlated with a lower risk of subsequent mortality in a prospective study of 39 patients with steroid-refractory chronic GVHD[22] but not in a prospective, multicenter, observational cohort comprised of 283 chronic GVHD cases.[24] In another study, response by a wide variety of definitions at 6 months did not correlate with subsequent development of tolerance.[25]

Nonetheless, response should be measured, documented and reported in all trials of treatment for chronic GVHD, since response is an important component of clinical benefit. Protocols and study reports should provide criteria used to define the baseline severity of patient-reported symptom burden and physician-assessed disease activity and damage. Protocols and study reports should likewise provide criteria to define the degree of subsequent change in each of these domains required for improvement or worsening. In addition, information regarding the trajectory of changes in pulmonary function tests and other objective measures before enrollment can be used to help interpret changes that occur after enrollment.

Tools for measurement of patient-reported symptom burden have been developed and validated.[26] (See report of the Response Criteria Working Group for summary and other references.) Patient-reported outcomes (PROs) are particularly useful when symptoms and physical disability cannot be reliably graded and documented by objective measures. Global scales assessing chronic GVHD-specific PROs should be used in conjunction with other measures of response in trials of systemic treatment.

Response-based endpoints and PROs are highly appropriate for controlled studies assessing specific manifestations of chronic GVHD. As discussed above, single-arm designs cannot control and account for the myriad other factors that could influence response in a single-arm study with response as the primary endpoint. Moreover, no well-defined, reliable benchmarks have been established for such studies.

The role of biomarkers as a component in the assessment of response has not been defined. Even so, measurement of biomarkers can be very helpful in determining whether a product has an expected biological effect in patients with chronic GVHD. For products evaluated through the accelerated approval pathway, the FDA permits pivotal efficacy trials to use surrogate endpoints that are "reasonably likely" to predict clinical benefit.[27] Hence, changes in biomarkers could be used as a surrogate for response in a clinical trial, if previous studies have demonstrated that the biological effect induced by the product is linked to the pathogenesis of GVHD, and if the effect is sufficient for a sustained clinical response with the use of a product.

Failure-free survival. Failure-free survival (FFS) defined by the absence of systemic treatment change, nonrelapse mortality and recurrent malignancy has been proposed as an appropriate primary endpoint for studies of investigational products intended for systemic control of GVHD in patients who are enrolled because an immediate change of treatment is needed.[28-29] The absence of nonrelapse mortality and recurrent malignancy clearly reflect clinical benefit, but the extent to which the absence of systemic treatment change reflects clinical benefit has not yet been determined. FFS is not optimally suited for studies to evaluate a specific systemic effect such as fibrosis, or local effects on specific organs such as pulmonary disease or ocular involvement.

Concerns could be raised about the reliability of using treatment change as an element in the composite FFS endpoint. It should be noted, however, that these concerns also apply in the assessment of response. Introduction of a new systemic treatment in a patient with chronic GVHD can be taken as evidence that current treatment has not provided satisfactory results, because chronic GVHD manifestations are progressing, persisting without improvement, or improving more slowly than desired. This interpretation can be confounded, however, when changes are motivated by toxicity, inconvenience, financial burden, or the availability of a newer alternative that is more attractive to the patient or physician. For example, in a randomized trial with a crossover option, some patients who were evaluated as having a "response" crossed over to the other arm. This pairing between assessment and action might appear to be inconsistent, but the baseline for comparison in clinical practice is typically the most recent clinic visit, not the entry point into the study. A systemic treatment change might be perfectly reasonable for a patient who is better as compared to enrollment but worse as compared to the most recent visit. Reasons for decisions to introduce new systemic treatment after enrollment should be documented in all studies.

FFS could be used as an endpoint in controlled trials. Benchmark rates that could be used as the null hypothesis for the primary endpoint in future single arm trials of first-line and second-line treatment have been reported from retrospective studies at one center,[28-29] and results of first-line treatment were replicated in a small cohort of patients enrolled in a previous controlled trial with a double-blind

design.[29] Further studies have not yet been carried out to determine whether the reported data are representative of results from other centers. Results from the BMT CTN 0801 trial (NCT01106833) could help to address this important question. This prospective trial was designed to evaluate a regimen of steroids and sirolimus with or without a calcineurin inhibitor for initial treatment of chronic GVHD, although this study also allowed enrollment of patients who had an inadequate response to initial treatment.

FFS could be used as the primary endpoint in conjunction with secondary endpoints that measure response in the interim until a global response measure has been fully developed and validated for use as the primary endpoint in clinical trials. Two distinct questions should be addressed, given the uncertain reliability and reproducibility of FFS as an endpoint. 1) What evidence should be summarized in clinical trial reports to persuade well-informed, scientifically trained, but appropriately skeptical physicians that the use of a particular product is worthy of being considered in caring for patients with chronic GVHD? 2) Similarly, what evidence should be summarized in a new drug application to persuade extremely well informed scientists and physicians with regulatory authority that a particular product should have market approval for treatment of chronic GVHD? Large improvements in FFS supported by evidence of reduced symptom burden, decreased disease activity, and an absence of emerging damage or disability should suffice in answer to the first question. Further data are needed in order to determine whether similar results could suffice in answer to the second question.

Steroid-sparing effect. No precedent has been established for designs testing whether an investigational product has a "steroid-sparing effect." Patients with mild manifestations at the initial onset of GVHD would be suitable for such studies. Benchmark serial steroid dose data from a large prospective study of first-line treatment have been reported[10] and could be used as a benchmark for future single arm studies of first-line treatment, but endpoints that could be used to assess clinical benefit from being able to reduce the dose of steroids have not been defined or reported. Controlled trials with this endpoint would have to be designed to distinguish the effects of the investigational product from the effects of concomitant treatment and to determine whether any reduction in steroid-related toxicity is offset by toxicity of the investigational product. Such an approach would require the development of a validated global measure of steroid-related toxicity.

Survival. Mortality rates in large cohorts of patients with chronic GVHD have been reported,[30] and these results could be used as be used as a benchmark for single arm studies. Improved survival from more effective treatment of chronic GVHD is not likely to be measurable until several years after patients are enrolled in a trial, making survival an unattractive primary endpoint for early phase trials. Even so, survival should be measured, documented and reported in all trials of treatment for chronic GVHD in order to determine whether an investigational treatment causes harm.

Permanent withdrawal of all systemic treatment after resolution of GVHD. Shortening of time to develop immunological tolerance as manifested by permanent withdrawal of all immunosuppressive treatment after resolution of reversible chronic GVHD manifestations clearly represents clinical benefit, even if certain irreversible sequelae of chronic GVHD persist. Shortened time to development of tolerance is not likely to be measureable until several years after patients are enrolled in a trial, making

tolerance an unattractive endpoint for early phase trials. Immunological tolerance representing cure of chronic GVHD would represent a very attractive endpoint for late phase trials.

Lessons from Regulatory Review of Treatment for Autoimmune and Musculoskeletal Disorders

Two large-scale reviews of decisions by the United States Food and Drug Administration (FDA) offer insights for the design and conduct of studies intended for regulatory review.[31-32] The first report characterized pivotal efficacy trials that provided the basis for approval of novel therapeutic agents between 2005 and 2012.[31] Among the 448 trials, 36 were intended for indications related to autoimmune and musculoskeletal diseases, the category most closely related to chronic GVHD. All of these trials had randomized control designs, 34 (94%) were double-blinded, 11 (31%) had active comparators and 25 (69%) had placebo comparators, 28 (78%) had clinical scale endpoints, 6 (17%) had surrogate endpoints such as laboratory measures, and 2 (6%) had clinical endpoints such as death, hospitalization, or functional measures. A median of 525 patients were enrolled, and participation extended beyond 6 months in 12 (33%) of the studies. Approvals for the 13 indications in this category were based on studies that enrolled an aggregate median of 1209 patients with an aggregate median of 1955 patients in the safety population. Among the 13 indications, 11 (85%) approvals were based on at least 2 studies, and only 2 (15%) were based on a single trial.

The second report characterized reasons for disapproval of new drug applications between 2000 and 2012.[32] As summarized in an accompanying editorial,[33] the results indicate that in reviewing clinical trials, FDA is looking for evidence of generalizable study populations, adequate sample size, meaningful health outcomes and degree of influence on those outcomes, consistency of multiple endpoints among different trials and sites, improvement over the standard of care, and evidence that benefits exceed harms.

Enrollment a sufficient sample size poses the most difficult challenge in conducting trials for treatment of chronic GVHD. The largest trial to date enrolled 287 patients.[5] The BMT CTN 0801 trial and another recent multicenter trial[10] each took 4 years to enroll 151 patients, even though both adults and children were eligible. Hence, a very large effect size would be needed for successful development of a new treatment for chronic GVHD.

Examples of Possible Development Paths for Investigational Products Intended for Treatment of Chronic GVHD

Development paths leading to regulatory approval for indications related to chronic GVHD have not been established. The small market and lack of an established development path stand as disincentives for industry sponsors. Establishment of development paths could decrease the risks for industry sponsors and increase their interest in chronic GVHD. As examples, we outline development paths for 3 indications related to chronic GVHD: 1) systemic effect in patients with disease manifestations that require immediate intervention, 2) systemic effect in patients with mild manifestations of chronic GVHD at initial diagnosis, and 3) effect in patients with specific manifestations of chronic GVHD. These examples are not intended to be comprehensive, and they do not address all possible contingencies.

Systemic effect in patients requiring immediate intervention. The first trial in this pathway could be designed to test the efficacy of a product for controlling rapidly reversible manifestations of disease activity in patients with chronic GVHD that has not responded adequately to initial treatment, including glucocorticoids (second-line treatment). Manifestations of disease activity include erythematous rash, oral mucosal changes, conjunctival inflammation not caused by dry eye, abnormal liver function tests, and rapidly reversible gastrointestinal manifestations, including nausea, vomiting, diarrhea and weight loss. Baseline treatment with agents other than glucocorticoids may be continued, but no new systemic or topically active agents should be added, other than the investigational product. Improvement in these manifestations of disease activity should generally be evident within 4 weeks, but longer durations of administration will be needed even in the initial studies in order to assess the durability of response, the effects on less rapidly reversible manifestations such as weight loss, sclerosis and fasciitis, the ability to taper concomitant medications, including the dose of glucocorticoids, and the safety of the product in patients with chronic GVHD. Longer durations of administration may also be needed before the response assessment for certain products that are known not to produce prompt improvement (e.g., extra-corporeal photopheresis).

If the first trial shows evidence of efficacy in controlling manifestations of disease activity, a follow-up or extension phase II trial could address the question of whether the product improves FFS in patients with chronic GVHD that has not responded adequately to initial treatment.[28] In all trials using FFS as the primary endpoint, response outcomes should be defined, measured, documented and reported, and serial steroid dose data should be collected. Alternatively, a follow-up study could be conducted by enrolling patients under conditions where an immediate change of treatment is not needed. In such a study, one arm would receive the investigational product, while the other arm would continue the baseline management, and either response or FFS could be used as the primary endpoint in comparing the two arms. A single arm study would not be feasible with this approach, because benchmark data that could be used to set the null hypothesis for the primary endpoint have not been established in this setting. Successful results with either approach could lead to a controlled phase III trial for the same indication.

If the follow-up or extension trial improves FFS, the development path could shift to initial treatment. It is also possible that the development path could bypass the first phase of testing with second-line treatment and begin with initial treatment. For example, it might be appropriate to test the effects of adding the product to initial glucocorticoid treatment, using FFS as the primary endpoint in a single-arm phase II study.[29]

If results of a single-arm phase II study suggest success, a blinded randomized phase II study could be done to determine whether the postulated difference in FFS with and without the product added to initial glucocorticoid treatment could be confirmed. Such a study is likely to be underpowered, unless the difference between arms is very large. Nonetheless, results of this trial would be useful in providing true estimates of the difference between the arms and would provide longer-term safety data with the use of the product in patients with chronic GVHD.

At least one pivotal trial would be required if results for a blinded randomized phase II study are encouraging. The pivotal trial could be designed to test a difference of the magnitude observed in the randomized phase II study. A key question here is whether a difference in the FFS endpoint would be considered as sufficient evidence of clinical benefit. If not, then an alternative endpoint would be needed. Resolution of chronic GVHD and withdrawal of all immunosuppressive treatment could serve as an appropriate endpoint, but only for products that are thought to facilitate the induction of immunological tolerance, and only if very long time horizons can be tolerated by the sponsor. The median time from onset of systemic treatment to withdrawal of immunosuppressive therapy after resolution of chronic GVHD exceeds 2 years.[4] A response measure could also be used as the primary endpoint, if it has been previously standardized and validated.

While a major purpose of phase II studies is to identify promising approaches for phase III studies, another important purpose of phase II studies is to identify approaches that do not work, in order to avoid unnecessary investments in large expensive phase III studies that have little chance of success. FFS could be very effectively used multi-stage trial designs. The first stage would enroll patients in a traditional single-arm phase II study with FFS at 6 months or perhaps a year as the primary endpoint compared against historical results for the same indication, with any variety of response measures as secondary endpoints and with other important secondary endpoints such as survival, relapse and non-relapse mortality. A sample size of 40 to 60 patients would suffice for this purpose. Positive results would trigger an immediate second stage. The endpoint here for clinical purposes could still be FFS, but for regulatory purposes, the primary endpoint might have to change to resolution of GVHD and withdrawal of all immunosuppressive treatment (without a prior qualitative change of systemic treatment) if FFS is not accepted as evidence of clinical benefit, even when some measure of response is included. In any case, the phase III stage would have to "start over" with enrollment, excluding results from the first stage in order to avoid bias.

Alternatively, a traditional randomized phase III design could be used with an FFS-based interim stopping rule for futility and either FFS or withdrawal of all immunosuppression (without a prior qualitative change of systemic treatment) as the primary endpoint. As an example, the decision to terminate enrollment in a study of mycophenolate mofetil for initial treatment of chronic GVHD was based on the absence of a difference in the primary endpoint after 4 years.[28] In this study, patients who had prior failure could not have met the primary endpoint, by definition. In fact, the failure rate in the investigational arm was higher than in the control arm from the very beginning of the study. If the trial had been designed to include interim analyses of FFS, this negative study could have been closed for futility much earlier.

<u>Systemic effect in patients with mild manifestations of chronic GVHD at initial diagnosis</u>. Anti-inflammatory glucocorticoids have a long-established, prominent role in the treatment of chronic GVHD. Long-term, high-dose glucocorticoid treatment causes many side effects, some of which are irreversible. These considerations motivate interest in testing treatment regimens that do not contain anti-inflammatory glucocorticoids. Eligibility for such a trial would require the absence of current systemic glucocorticoid treatment. Initial systemic treatment for newly diagnosed chronic GVHD with mild manifestations represents the most likely setting in which the benefits and risks of a glucocorticoid-free

regimen could be tested. Disease manifestations would have to be sufficiently severe to require systemic treatment, but not so severe as to require immediate use of glucocorticoids.

As a first step, it would be appropriate to test the effects of a glucocorticoid-free regimen in a trial with either response or FFS as the primary endpoint. Response should be measured at multiple time points in order to determine whether the benefit is sustained, and improvement should not be counted as a response if systemic or local treatment with glucocorticoids or another product has been added to the regimen before response is assessed. Given the uncertainty of benefit in such a trial, close monitoring at frequent intervals would be needed in order to ensure that symptoms and disease activity are adequately controlled and that no damage is emerging. Although no benchmark response rates have been established in this setting, a high rate of sustained responses would suggest that the product has activity against chronic GVHD, since previous results have indicated that untreated "clinical extensive" chronic GVHD tends to progress inexorably toward disability.[34] FFS could also be used as the primary endpoint in such a trial, as discussed above. In this case, addition of systemic or local treatment with glucocorticoids or another product should be counted as failure.

A pivotal randomized phase III study could be done if results of the first study suggest that the investigational treatment produced response rates or FFS rates better than those expected for the standard of care. Blinding in such a trial would not be feasible, and robust, objective endpoints would be needed in order to demonstrate superior clinical benefit with the use of the investigational product. As discussed above, a multi-stage design could be used to improve the efficiency of the approach. A randomized phase III study could also be done if results of the initial study suggest that the investigational treatment produced response rates or FFS rates equivalent to those expected for the standard of care. In this case, the primary endpoint would be designed to test the hypothesis that the overall burden of adverse effects is lower with the investigational treatment than with the standard of care.

Effect in patients with specific manifestations of chronic GVHD. Certain products might be suitable for treatment of specific systemic manifestations such as fibrosis, or might have local effect on specific organs such as the skin, mouth, eyes, lungs, gastrointestinal tract or genitourinary tract. Eligibility criteria for a trial to test a product for such an indication require careful definition to ensure that the condition is actually caused by chronic GVHD and that other potentially confounding causes are absent.

The first trial in the pathway could be designed to test the efficacy of a product for controlling rapidly reversible manifestations of disease activity, using an objective or subjective response measure as the primary endpoint. Claims that improvements after enrollment are related to the investigational treatment would be credible if the baseline systemic treatment is not changed at enrollment, and if no new systemic treatments are added before the assessment of response. In such a trial, the addition of a local or topical therapy before the assessment of response would count as failure. The first trial in the pathway could also be designed to test the efficacy of a product for preventing progression of less reversible manifestations of chronic GVHD. In this case, however, the trajectory of progression would have to be thoroughly documented before enrollment in order to determine whether a change occurred after enrollment in the study.

A pivotal randomized phase III study could be done if results of the first study suggest that the investigational treatment produced unambiguous sustained responses or prevented progression. In such a trial, one arm would be treated with the investigational product while continuing prior treatment without change, and the other arm would continue prior treatment without change. Such a trial should have a blinded design, if possible, in order to minimize bias. With blinding, a crossover design could be used to improve enrollment and motivate adherence to the protocol.

Conclusions

Robust endpoints and efficient phase II study designs are needed in order to identify promising drugs that could be tested in phase III studies. Validated benchmarks that could be used to set the null hypothesis for response rates have not been reported in any setting, and it will likely take another 4 to 5 years before a validated clinical scale has been established to measure global response in patients with chronic GVHD. Until then, FFS could serve as a useful basis for designing studies and interpreting clinical trial results, if several caveats are kept in mind. Additional studies are needed in order to determine whether the FFS data reported to date are representative of results in other centers and to establish benchmarks for contexts other than first or second-line systemic treatment. The proposed use of FFS has been criticized because medical providers sometimes show inconsistent behavior in deciding to change treatment, but this concern applies for all endpoints other than death. In making such decisions, investigators must demonstrate consistency, fair play, integrity, self-discipline and willingness to set aside personal biases, all in the interest determining the truth about the merits of an investigational treatment. Since the FFS endpoint does not give any direct information about improvements in GVHD-related symptoms, activity, damage or disability, studies using FFS as the primary endpoint should also measure changes in these domains as secondary endpoints. In the future, it should be possible to replace FFS with standardized and validated measures of sustained response as the primary endpoint in order to measure clinical benefit in a more direct way.

References

1. Martin PJ, Weisdorf D, 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 Marrow Transplant.* 2006;12(5):491-505.
2. CIBMTR: http://www.cibmtr.org/ReferenceCenter/SlidesReports/SummarySlides/pages/index.aspx
3. Flowers MED, Inamoto Y, Carpenter PA, et al. Comparative analysis of risk factors for acute and for chronic graft-versus-host disease according to National Institute of Health consensus criteria. *Blood.* 2011;117(11):3214-3219.
4. Vigorito AC, Campregher PV, Storer BE, et al. Evaluation of NIH consensus criteria for classification of late acute and chronic GVHD. *Blood*. 2009;114(3):702-708.
5. Sullivan KM, Witherspoon RP, Storb R, et al. Prednisone and azathioprine compared with prednisone and placebo for treatment of chronic graft-v-host disease: prognostic influence of prolonged thrombocytopenia after allogeneic marrow transplantation. *Blood*. 1988;72(2):546-554.
6. Koc S, Leisenring W, Flowers ME, et al. Thalidomide for treatment of patients with chronic graft-versus-host disease. *Blood*. 2000;96(12):3995-3996.
7. Arora M, Wagner JE, Davies SM, et al. Randomized clinical trial of thalidomide, cyclosporine, and prednisone versus cyclosporine and prednisone as initial therapy for chronic graft-versus-host disease. *Biol Blood Marrow Transplant*. 2001;7(5):265-273.
8. Koc S, Leisenring W, Flowers ME, et al. Therapy for chronic graft-versus-host disease: a randomized trial comparing cyclosporine plus prednisone versus prednisone alone. *Blood*. 2002;100(1):48-51.
9. Flowers ME, Apperley JF, van Besien K, et al. A multicenter prospective phase 2 randomized study of extracorporeal photopheresis for treatment of chronic graft-versus-host disease. *Blood*. 2008;112(7):2667-2674.
10. Martin PJ, Storer BE, Rowley SD, et al. Evaluation of mycophenolate mofetil for initial treatment of chronic graft-versus-host disease. *Blood*. 2009;113(21):5074-5082.
11. Greinix HT, van Besien K, Elmaagacli AH, et al. Progressive Improvement in Cutaneous and Extracutaneous Chronic Graft-versus-Host Disease after a 24-Week Course of Extracorporeal Photopheresis-Results of a Crossover Randomized Study. *Biol Blood Marrow Transplant*. 2011;17(12):1775-1782.
12. Gilman AL, Schultz KR, Goldman FD, et al. Randomized trial of hydroxychloroquine for newly diagnosed chronic graft-versus-host disease in children: a Children's Oncology Group study. *Biol Blood Marrow Transplant*. 2012;18(1):84-91.
13. Martin PJ, Inamoto Y, Carpenter PA, Lee SJ, Flowers MED. Treatment of chronic graft-versus-host disease: past, present, and future. *Korean J Hematol*. 2011;46(3):153-163.
14. Pidala J, Tomblyn M, Nishihori T, et al. Sirolimus demonstrates activity in the primary therapy of acute graft-versus-host disease without systemic glucocorticoids. *Haematologica*. 2011;96(9):1351-1356.
15. McCaul KG, Nevill TJ, Barnett MJ. Treatment of steroid-resistant acute graft-versus-host disease with rabbit antithymocyte globulin. *J Hematotherapy & Stem Cell Research.* 2000;9(3):367-374.

16. Sandborn WJ, Feagan BG, Hanauer SB, et al. A review of activity indices and efficacy endpoints for clinical trials of medical therapy in adults with Crohn's disease. *Gastroenterology*. 2002;122(2):512-530.
17. Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C *et al.* American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. *Arthritis and Rheumatism* 1995;38(6)**:**727-735.
18. Anderson JJ, Baron G, van der Heijde D, Felson DT, Dougados M. Ankylosing spondylitis assessment group preliminary definition of short-term improvement in ankylosing spondylitis. *Arthritis & Rheumatism.* 2001;44(8):1876-1886.
19. Luijten KMAC, Tekstra J, Bijlsma JWJ, Bijl M. The systemic lupus erythematosus responder index (SRI); a new SLE disease activity assessment. *Autoimmunity Reviews*. 2012;11(5):326-329.
20. Rider LG, Giannini EH, Harris-Love M, et al. Defining clinical improvement in adult and juvenile myositis. *J Rheumatol*. 2003;30(3):603-617.
21. 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: IV. Response Criteria Working Group report. *Biol Blood Marrow Transplant*. 2006;12(3):252-266.
22. Olivieri A, Cimminiello M, Corradini P, et al. Long-term outcome and prospective validation of NIH response criteria in 39 patients receiving imatinib for steroid-refractory chronic GVHD. *Blood*. 2013;122(25):4111-4118.
23. Inamoto Y, Martin PJ, Chai X, et al. Clinical Benefit of Response in Chronic Graft-versus-Host Disease. *Biol Blood Marrow Transplant*. 2012;18(10):1517-1524.
24. Palmer J, Lee SJ, Chai X, et al. Poor Agreement between Clinician Response Ratings and Calculated Response Measures in Patients with Chronic Graft-versus-host Disease. *Biol Blood Marrow Transplant*. 2012;18(11):1649-1655.
25. Martin PJ, Storer BE, Carpenter PA, et al. Comparison of short-term response and long-term outcomes after initial systemic treatment of chronic graft-versus-host disease. *Biol Blood Marrow Transplant*. 2011;17(1):124-132.
26. Lee SJ, Cook EF, Soiffer RJ, Antin JH. Development and validation of a scale to measure symptoms of chronic graft-versus-host disease. *Biol Blood Marrow Transplant*. 2002;8():444-452.
27. US Food and Drug Administration (FDA). Applications for Food and Drug Administration Approval to Market a New Drug: Accelerated Approval of New Drugs for Serious or Life-Threatening Illnesses, 21 CFR §314h (2012). FDA website. http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=314. Accessed <date>.
28. Inamoto Y, Storer BE, Lee SJ, et al. Failure-free survival after second-line systemic treatment of chronic graft-versus-host disease. *Blood*. 2013;121(12):2340-2346.
29. Inamoto Y, Flowers MED, Sandmaier BM, Aki SZ, Carpenter PA, Lee SJ, Storer BE, Martin PJ. Failure-free survival after initial systemic treatment of chronic graft-versus-host disease. *Blood.* 2014; in press.
30. Arora M, Klein JP, Weisdorf DJ, et al. Chronic GVHD risk score: a Center for International Blood and Marrow Transplant Research analysis. *Blood*. 2011;117(24):6714-6720.

31. Downing NS, Jenerius AA, Shah ND, Krumholz HM, Ross JS. Clinical trial evidence supporting FDA approval of novel therapeutic agents, 2005-2012. *JAMA.*2014;311(4)368-377.
32. Sacks LV, Shamsuddin HH, Yasinkaya YI, Bouri K, Lanthier ML, Sherman RE. Scientific and regulatory reasons for delay and denial of FDA approval of initial applications for new drugs, 2000-2012. *JAMA*.2014:311(4):378-384.
33. Goodman SN, Redberg RF. Opening the FDA Black Box. *JAMA.*2014;311(4):361-363.
34. Sullivan KM, Shulman HM, Storb R, et al. Chronic graft-versus-host disease in 52 patients: adverse natural course and successful treatment with combination immunosuppression. *Blood* 1981;57(2):267-276.