Published in final edited form as:

*Biol Blood Marrow Transplant*. 2015 August ; 21(8): 1343–1359. doi:10.1016/j.bbmt.2015.05.004.

# National Institutes of Health Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-versus-Host Disease: VI. The 2014 Clinical Trial Design Working Group Report

**Paul J. Martin**[1,*], **Stephanie J. Lee**[1], **Donna Przepiorka**[2], **Mary M. Horowitz**[3], **John Koreth**[4], **Georgia B. Vogelsang**[5], **Irwin Walker**[6], **Paul A. Carpenter**[1], **Linda M. Griffith**[7], **Gorgun Akpek**[8], **Mohamad Mohty**[9], **Daniel Wolff**[10], **Steven Z. Pavletic**[11], and **Corey S. Cutler**[4]

[1]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, Washington
[2]Center for Drug Evaluation Research, Food and Drug Administration, Silver Spring, Maryland
[3]Division of Hematology and Oncology, Medical College of Wisconsin, Milwaukee Wisconsin
[4]Division of Hematologic Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts
[5]Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, Maryland
[6]Department of Medicine, McMaster University, Hamilton, Ontario [7]Division of Allergy, Immunology and Transplantation, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland [8]Stem Cell Transplant Program, Banner MD Anderson Cancer Center, Gilbert, Arizona [9]Clinical Hematology and Cellular Therapy Department, Hôpital Saint-Antoine, University Pierre & Marie Curie, INSERM U938, Paris, France [10]Department of Internal Medicine III, University of Regensburg, Regensburg, Germany [11]Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland

## Abstract

Treatment of chronic GVHD is intended to produce a sustainable benefit by reducing symptom burden, controlling objective manifestations of disease activity, preventing damage and impairment, and improving overall survival without causing disproportionate harms related to the treatment itself. Successful management can control the disease until systemic treatment is no longer needed. The complexity of the disease, the extended duration of follow-up needed to

[*]Correspondence: Paul J. Martin, M.D., Fred Hutchinson Cancer Research Center, D2-100, P.O. Box 19024, Seattle, WA 98109-1024. pmartin@fredhutch.org (P.J. Martin)..



*Financial Disclosure Statement:* The authors have nothing to disclose.

*Disclaimer:* The opinions expressed are those of the authors and do not represent the position of the National Cancer Institute, the National Heart, Lung and Blood Institute, the National Institute of Allergy and Infectious Diseases, the National Institutes of Health, the Food and Drug Administration, or the United States Government.

*Conflict of interest statement:* P.J.M., Scientific Advisory Board Meeting for Pharmacyclics, Inc.; S.J.L., single Scientific Advisory Board Meetings for Kadmon Biotech and Bristol Myers Squibb; J.K., research funding from Prometheus Laborabories, Millenium Pharmaceuticals and Otsuka Pharmaceuticals, Scientific Advisory Boards for Takeda Pharmaceuticals, Amgen and Kadmon Biotech; C.S.C., Pharmacyclics, Inc., Onyx, Inc., Immucor, Inc., Fate Therapeutics.

**SUPPLEMENTARY DATA**

Supplementary data related to this report can be found online at http://www.asbmt.org/?page=PIND124475

observe disease resolution and withdrawal of immunosuppressive treatment, and the lack of fully developed shorter-term endpoints impede progress in the field. Identification and characterization of primary endpoints demonstrating clinical benefit without the need for years of follow-up is an urgent need, with the understanding that clinical benefit encompasses not only the self-evident benefit of the primary endpoint but also any other associated benefits. This report discusses regulatory considerations, eligibility criteria, the value of controlled trial designs, the merits of proposed primary endpoints, and key considerations elaborated from experience and progress during the past decade. The report concludes by mapping an overall approach that could support and lead to maximally informative clinical trials, especially those that seek to demonstrate clinical benefit along a pathway to regulatory review and approval.

### Keywords

Chronic graft-versus-host disease; allogeneic hematopoietic cell transplantation; clinical trials; design; consensus; guidelines

## INTRODUCTION

Effective treatment for chronic GVHD is an urgent unmet clinical need. The number of pharmaceutical companies interested in developing products for chronic GVHD increased markedly since the previous NIH Consensus Conference. In April 2015, ClinicalTrials.gov listed 9 industry-sponsored studies open for recruitment to evaluate systemic products for treatment of chronic GVHD. Progress is hampered by the lack of defined pathways for clinical development and regulatory approval of products intended for treatment of chronic GVHD. Regulatory applications are most likely to be submitted as new indications for products previously approved for other indications, but in certain cases, they could be filed for new products to treat or prevent chronic GVHD. Relevant considerations include a plausible mechanism of action, demonstrated anti-inflammatory, immunosuppressive or anti-fibrotic activity, and the safety profile.

The number of patients available for enrollment in chronic GVHD treatment trials is limited. Among the ~8,000 allogeneic hematopoietic cell transplantations performed each year in the U.S. [1], at least 35% of recipients would be expected to develop chronic GVHD requiring systemic treatment [2], such that the total incidence is approximately 3,000 per year. The total prevalence in the U.S. is estimated at approximately 10,000, after accounting for durable disease resolution and end of treatment together with deaths and recurrent malignancy during treatment [3]. Although the incidence of steroid-refractory chronic GVHD is lower than the overall incidence of chronic GVHD, the prevalence of steroid-resistant or steroid-refractory chronic GVHD is difficult to estimate. This limited pool of potential study candidates suggests that major questions can be addressed most efficiently by multicenter or cooperative group trials.

Controlled designs have been used for only a small minority of chronic GVHD treatment trials. The 6 published randomized trials of first-line treatment for chronic GVHD used a variety of primary endpoints, and none demonstrated superiority of the investigational arm [4-9]. The design of controlled second-line treatment studies is hampered by the lack of a

standard treatment regimen. Only 1 randomized trial of second-line treatment for chronic GVHD has been published [10, 11].

Uncontrolled, single-arm studies of second-line treatment typically show overall response rates of 30 – 70% [12]. In many studies, response criteria are poorly defined, and results are interpreted under the premise that no response would have occurred in the absence of the investigational treatment. This premise might not hold true, especially if the prior trajectory of GVHD is not taken into account, if active topical agents are added, if doses of prior medications are increased or if systemic agents other than the investigational product are added between enrollment and the response assessment. These factors, together with variation in selection criteria, patient characteristics, baseline disease manifestations and assessment time points, make it difficult to interpret the results of single-arm studies of second-line treatment or to establish reliable benchmark response rates for planning future studies [12].

## PURPOSE OF THIS DOCUMENT

The complexity of chronic GVHD and the lack of fully developed research methods make it difficult to design, conduct, and analyze clinical trials involving subjects with this disease, even when promising treatment options are available. The 2006 Design of Clinical Trials Working Group Report [13] offered important recommendations and definitions for investigators in an attempt to overcome these obstacles. The Working Group 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 decade since the previous NIH Consensus Conference yielded improvements in the precision and accuracy of criteria for the diagnosis and staging of chronic GVHD [14], the interpretation of histopathology [15], the discovery and validation of biomarkers for diagnostic and prognostic applications [16], and supportive care for patients with chronic GVHD [17]. These results 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. A separate report describes progress toward the development of the most clinically relevant response criteria [18]. The current report focuses on considerations for the design of clinical trials.

While the original recommendations from the 2006 Design of Clinical Trials Working Group Report [13] were broad-based and grounded in good clinical practice, further improvement is needed. The current report is focused primarily on the development, characterization, validation and selection of primary and secondary endpoints that could be used to demonstrate clinical benefit without requiring years of follow-up, thereby mapping an overall approach that could support regulatory review. Prior recommendations that merit updating or attention based on experience since the previous NIH Consensus Conference are also highlighted and elaborated.

This report does not address considerations for the design of trials for prevention of chronic GVHD. The development of designs and regulatory paths are more advanced for prevention trials than for treatment trials. Sponsors and investigators of chronic GVHD prevention

studies are encouraged to use information from the NIH Consensus Development Project on Criteria for Clinical Trials in Chronic GVHD in defining diagnostic [14] and pathological criteria [15] and in applying biomarkers [16] in clinical trials.

## SUMMARY OF UPDATED RECOMMENDATIONS

1. Primary and secondary endpoints should be selected for their ability to demonstrate clinical benefit, which can be a prolongation of survival or an improvement in the way a patient feels or functions. The overall concept of clinical benefit encompasses not only the self-evident benefit of the primary endpoint per se but also any tangible and measurable benefits in symptom burden, quality of life or other important outcomes demonstrated to be closely associated with the primary endpoint. Patient-reported measures should be incorporated whenever feasible. Standardized and clinically valid measurements should be used. Composite endpoints may be required in some protocols.
2. Many endpoints are clinically meaningful, but the number of patients available to participate in chronic GVHD treatment trials is small. Therefore, studies should be designed to capture as much relevant information as possible from providers and patients, even if the data will not be submitted for regulatory review.
3. Inclusion and exclusion criteria should ensure the ability to interpret results of the study while as much as possible allowing enrollment of patients who might benefit, based on the investigational product's mechanism of action.
4. Baseline evaluations should document eligibility, capture prognostic characteristics, and specifically characterize the condition of subjects at the time of enrollment, so that results of therapy can be interpreted.
5. Within reason, the study protocol should specify or provide guidance regarding the dosing and dose adjustment of all immunosuppressive medications and topical treatments, including the non-investigational products. Reasons for deviations, discontinued administration or use of the study product, and new treatment should be documented in case-report forms.
6. Case-report forms should be calendar driven by the protocol to provide assessment of chronic GVHD and adverse events at regular intervals. Study personnel should conduct real-time cleaning and monitoring of baseline and response assessments to ensure accurate, consistent evaluations.
7. Biostatistical analysis should incorporate considerations of competing events such as recurrent malignancy or non-relapse mortality and any new concomitant systemic or topical therapy started at the time of enrollment or afterwards. The protocol should provide appropriate power calculations and summarize statistical plans for any interim analyses, sensitivity analysis, subset analysis, and missing measurements.
8. Controlled designs are preferred whenever possible, because they allow interpretation of the treatment effect (i.e., efficacy and safety), if prognostic risk factors are balanced between the arms. Stratification can be used to decrease the

risk of imbalanced distribution of risk factors in controlled trials. Treatment effects are very difficult to interpret single-arm studies.

## GOALS OF TREATMENT FOR CHRONIC GVHD

Treatment of chronic GVHD is intended to produce a sustainable benefit by reducing symptom burden, controlling objective manifestations of disease activity, preventing damage and impairment, and improving overall survival without causing disproportionate harms related to the treatment itself. Early experience showed that in the absence of systemic treatment, chronic GVHD progresses inexorably to disability and death [19]. Management of chronic GVHD has relied on corticosteroids as the mainstay of treatment for more than 3 decades, although treatment regimens vary [20-22]. Systemic treatment typically begins with prednisone at 0.5 to 1 mg/kg per day, with or without cyclosporine, tacrolimus or sirolimus. Prolonged treatment with prednisone at high doses causes many adverse effects, making it necessary to taper the dose as soon as GVHD improves. Manifestations of chronic GVHD can wax and wane when efforts are made to reduce or closely calibrate the intensity of immunosuppressive treatment to the minimum needed to control the disease (Figure 1). In a recent prospective study, the average dose of prednisone was tapered to 0.20 – 0.25 mg/kg per day or 0.4 – 0.5 mg/kg every other day within 3 months after starting systemic treatment [8].

Successful management of chronic GVHD can control the disease until systemic treatment is no longer needed to prevent recurrent or progressive disease activity or exacerbation of any residual damage. After withdrawal of systemic treatment, laboratory testing may detect persistent low-level alloreactivity that is not sufficient to cause progression in any residual clinical manifestations of the disease. Systemic treatment is discontinued in approximately 50% of patients within 7 years after starting systemic treatment. Approximately 10% of patients require continued systemic treatment for an indefinite period beyond 7 years, and the remaining 40% develop recurrent malignancy or die from non-relapse causes while continuing systemic treatment within 7 years after diagnosis [3]. Discontinuation of systemic treatment possible for some patients with far advanced chronic GVHD that has persisted despite the use of multiple immunosuppressive agents for many years. In these circumstances, the goals of treatment are to control symptoms and disease activity, to prevent further damage and impairment, whether from the disease itself or from the medications used for management, and to improve survival.

It is not known whether currently available immunosuppressive products shorten or lengthen the time to withdrawal of treatment. In either case, they provide clinical benefit by controlling disease activity and preventing impairment until systemic treatment can be discontinued. In this context, new products for treatment of chronic GVHD could increase clinical benefit if they are more effective than currently available treatments without causing a disproportionate burden of side effects or if they are as effective as currently available treatment but cause a lesser burden of side effects.

## REGULATORY CONSIDERATIONS FOR CHRONIC GVHD CLINICAL TRIAL DESIGN

Many important clinical trials, including some that changed the standard practice in the field of hematopoietic stem cell transplantation, were not done with the objective of approving a new drug or a new indication for an already approved drug. On the other hand, commercial sponsors are key stakeholders in the development of new therapies for treatment of chronic GVHD, and if progress is to be made in this area, clinical trial designs must address the regulatory requirements that commercial sponsors must meet. Overall survival and survival to durable resolution of chronic GVHD with withdrawal of all systemic treatment are endpoints that clearly indicate clinical benefit in regulatory terms, but the long follow-up time needed to ascertain these endpoints make them challenging for use in chronic GVHD drug development. Alternative shorter-term endpoints considered by the Clinical Trials Design Working Group include clinical response, failure-free survival (FFS), survival without progressive impairment (SWOPI), patient-reported outcomes (PROs), and composite scales that incorporate provider and patient assessments.

In preparing the current report, members of the Clinical Trials Working Group met with representatives of the Food and Drug Administration (FDA) to discuss the regulatory perspectives on proposed endpoints. Briefing materials used for this meeting summarize current knowledge about potential endpoints in chronic GVHD trials (see on-line Supplement). Key general advice provided by FDA included the following.

- Endpoints may differ based on the natural history of the intended population as determined by the eligibility criteria. A survival benefit might need to be demonstrated when the intended patient population has a relatively short expected overall survival, while for patients who live long but with the potential for disability (most patients with chronic GVHD in the modern era), a Clinician-Reported Outcome (CRO) or PRO might be more appropriate.
- A CRO or PRO assessment might be acceptable as a measure of clinical benefit without validation against survival. The tool should be well defined and reliable. A PRO outcome should be supported by an objective clinical measure of drug activity, such as complete plus partial response.
- A composite endpoint would be acceptable if each component could be justified, but sample size considerations or characteristics of the intended patient population may warrant a simpler endpoint or co-primary endpoints instead.
- A composite endpoint that includes efficacy and safety outcomes would not be sufficient to demonstrate efficacy, since differences in safety might obscure differences in efficacy.
- Use of progression-free survival (PFS) types of chronic GVHD clinical endpoints for a regulatory decision must be meaningful for the particular study population. Whether a PFS endpoint is meaningful depends on relevance of the criteria for PFS to direct clinical benefit, the magnitude of the effect, and the risk-benefit of the new treatment.

- Time-to-event endpoints commonly used in randomized trials, such as PFS, are generally not interpretable in single-arm trials, especially with patient populations heterogeneous for factors that may affect the endpoint.

In the sections that follow below, the Clinical Trials Working Group discusses key aspects of eligibility criteria, types of comparators, and the proposed endpoints, and reflects on when and how these regulatory considerations might affect the clinical trial design.

## ELIGIBILITY CRITERIA AND DATA CAPTURE

Well-defined eligibility criteria are needed for all trials. Inclusion criteria depend on the specific medical indication for treatment. For chronic GVHD treatment trials, possible intended uses include global systemic effect, effect on a specific systemic manifestation 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 intended indication.

Standardized assessment forms used at baseline and follow-up should contain sufficient detail to verify the diagnosis of chronic GVHD and establish eligibility. The level of detail should also be sufficient to determine global severity according to updated NIH criteria, since global severity is associated with overall survival and the risk of non-relapse mortality [23]. In addition, refinements proposed by the 2014 Diagnosis and Staging Working Group capture data indicating whether individual organ scores should be attributed to GVHD [14]. Baseline data for CRO and PRO endpoints should be collected before randomization in order to ensure the absence of bias. As a key lesson from recent experience, study personnel should conduct real-time cleaning and monitoring of baseline data and follow-up response assessments. Delayed reconciliation between medical records and case report forms and other data cleaning make it extremely difficult to rectify omissions or inconsistencies across the various multi-organ assessments.

Eligibility criteria for specific trials may vary according to whether or not patients require treatment change. Enrollment in first and second-line systemic treatment trials is motivated by the immediate need to relieve symptoms, control disease activity, prevent damage and impairment, and if possible, shorten the time to withdrawal of systemic treatment. New onset of chronic GVHD prompts the need for first-line treatment, and unsatisfactory response to previous treatment prompts the needs for second-line treatment. In second-line treatment trials, the minimum dose and duration of prior treatment and the minimum severity of unimproved disease manifestations or criteria for worsening must be defined, although medical records may lack optimal documentation of this information. Clinical trials may also be designed for patients with stable chronic GVHD manifestations. For these trials, documentation of stable disease manifestations and treatment across some minimum time interval before enrollment is required.

Standardized definitions of steroid-refractory and steroid-dependent chronic GVHD in determining eligibility for enrollment would facilitate comparisons between results of

different studies, but practices vary considerably, making it difficult to reach consensus. This Working Group offers the following definitions and considerations for second-line treatment trials.

- Steroid-refractory chronic GVHD during first-line treatment may be defined when manifestations progress despite the use of a regimen containing prednisone at ≥1 mg/kg per day for at least 1 week or persist without improvement despite continued treatment with prednisone at ≥0.5 mg/kg per day or 1 mg/kg every other day for at least 4 weeks.
- Steroid-dependent chronic GVHD may be defined when prednisone doses >0.25 mg/kg per day or >0.5 mg/kg every other day are needed to prevent recurrence or progression of manifestations as demonstrated by unsuccessful attempts to taper the dose to lower levels on at least 2 occasions, separated by at least 8 weeks. These suggested dose thresholds match the average doses from 3 months onward in a recent prospective study of first-line treatment [8]. Other definitions may be appropriate depending on the trial context and should be specified in the protocol.

Three important caveats apply for these definitions. First, they are far less relevant to eligibility criteria for trials beyond second-line treatment, since the transition between first-line and second-line treatment has already established that the disease is steroid-refractory or steroid-dependent. In this scenario, appropriate eligibility criteria could be based on clinical judgment that the disease is refractory to the current treatment regimen. Second, they serve as general guidelines in writing eligibility criteria for second-line treatment trials, but they do not completely match clinical practices [24]. For example, some patients advance to second-line treatment after first-line treatment with prednisone that never exceeded 0.5 mg/kg per day or after complete withdrawal of prior systemic treatment. Small numbers of patients begin second-line treatment due to progressive disease after less than 7 days of first-line treatment or due to insufficient improvement after less than 4 weeks of first-line treatment. While no consensus has been reached for these situations, second-line treatment protocols should address the required minimum dose and duration of prior steroid treatment, since these parameters might relate to the lack of an adequate response during first-line treatment. Third, steroid intolerance alone is not a sufficient reason to enroll a patient in a trial intended to evaluate an investigational product for treatment of steroid-refractory or steroid-dependent chronic GVHD.

When a reduction in symptoms based on a PRO is the primary objective, attention should be paid to the minimum burden of symptoms at baseline for eligibility, in order to ensure that a clinically meaningful reduction can be measured. Similarly, for studies that seek to prevent progression of symptoms, the eligibility criteria should ensure that the baseline symptom burden of study participants is not so great that worsening could not be detected.

The role of biomarkers in defining eligibility for clinical trials is not established. Validated biomarkers that reliably reflect the severity of chronic GVHD manifestations, indicate the prognosis for patients with chronic GVHD or predict the likelihood of response to treatment would be very useful in the design and conduct of clinical trials [16]. Objective laboratory-based biomarkers strongly correlated with disease activity and measured with standardized

assays would be very useful in comparing the baseline characteristics of patients enrolled in different studies.

## CONTROLLED DESIGNS

Single-arm studies are not interpretable for regulatory purposes, especially if the population has heterogeneity in prognostic factors. The heterogeneity of the disease process and the patients affected means that differences between single-arm studies may be due solely to population differences rather than treatment effects. Blinded randomized trial designs help to prevent bias in the assessment of such endpoints, but these designs are not always feasible. For open-label trials, a highly robust, well-characterized, objective primary endpoint and related supporting secondary endpoints are generally needed for adequate interpretation.

In controlled studies of investigational products intended for first-line systemic treatment, one arm could receive the investigational product plus conventional treatment, while the other arm receives conventional treatment alone. Designs in which one arm receives an investigational product without conventional treatment while the other arm receives conventional treatment are also feasible [25]. In controlled studies of investigational products intended for second-line systemic treatment in patients who need an immediate treatment change, one arm would receive the investigational product. Since no standard of care has been established for this indication, the other arm could receive any other treatment considered within the scope of usual practice, although some restriction in control treatments might be desirable. In controlled studies of investigational products intended for systemic control of chronic GVHD in patients who do not need an immediate treatment change, one arm could receive the investigational product, while the other arm continues the baseline management.

In any of these approaches, studies could include “induction” and “maintenance” phases with different doses of the same investigational product or with the sequential use of different products. Similar considerations apply in studies of investigational products intended for effect at a specific site or on a specific organ or manifestation of chronic GVHD.

## ENDPOINTS

The primary endpoint in a clinical trial represents the major criterion by which success 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. The overall concept of clinical benefit encompasses not only the self-evident benefit of the primary endpoint *per se* but also any other benefits closely associated with the primary endpoint. For example, a full understanding of the clinical benefit of “response” requires characterization of the extent to which a defined type of response is associated with improvements in the overall burden of symptoms, level of function, overall survival and any other relevant outcomes of importance and value to patients with chronic GVHD. Clinical benefits that are not self-evident in the

primary endpoint could be understood as “collateral” benefits in the sense that they coincide with or serve to support or corroborate the self-evident benefit of the primary endpoint.

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. In successful trials, secondary endpoints provide necessary additional evidence of benefits, and safety endpoints provide evidence that the overall benefits exceed harms. A successful trial would show that the benefits of a high response rate are not offset by reduced overall survival.

The 2006 NIH Consensus Conference on Chronic GVHD Design of Clinical Trials Working Group Report addressed a variety of technical and quality considerations in the design and conduct of clinical trials testing products for treatment of chronic GVHD [13]. Potential short-term primary and secondary endpoints discussed in the report included GVHD response and PROs. The report noted that scales for measurement of global response were not yet validated and that few sensitive instruments are available for measuring PROs. As summarized in Table 1, GVHD response was considered most appropriate as a primary endpoint in phase 2 studies and possibly in selected phase 3 studies, while PROs were considered appropriate as secondary endpoints. The clinical benefit associated with these endpoints has not been adequately characterized. These shorter-term endpoints are preferable for early phase trials, but longer-term endpoints are needed for late-phase trials in order to demonstrate sustainable benefit. Complete response and successful withdrawal of systemic treatment after resolution of the disease were considered most appropriate as primary endpoints in phase 3 studies, while non-relapse mortality, survival without recurrent malignancy and overall survival were considered appropriate as secondary endpoints. Notably, in certain subsets of patients who have chronic GVHD with a moderately severe global NIH rating, mortality rates are very low (see on-line Supplement) [23, 26], leaving little opportunity to demonstrate improvement in overall survival.

The proposed endpoints of FFS and SWOPI define clinical benefit as the absence of new harm caused by the disease, whereas complete and partial response defines benefit as improvement in manifestations of the disease. PROs capture patient reports of GVHD symptoms and their degree of bother. Composite scales capture clinician assessments, PROs and laboratory or functional measures in a single global scale. All 5 endpoints represent relatively short-term outcomes as compared to the typical 2 to 5 year duration of treatment needed before the disease resolves with currently available treatment. Therefore, an important issue is the extent to which these short-term endpoints predict the durability of benefit. Strengths and weaknesses of each endpoint are summarized in Table 2.

An optimal primary endpoint is based on objective, reliable and verifiable criteria. Endpoints other than overall survival must have face validity indicating that patients live better with fewer symptoms or improved function as evidence of clinical benefit. For patients with chronic GVHD, evidence of clinical benefit can come from data demonstrating that a defined endpoint is associated with decreased burden of symptoms and symptom bother, better function, fewer side effects associated with treatment, shorter time to durable

resolution of the disease and withdrawal of systemic treatment, and improved overall survival. Table 3 summarizes the extent to which each of the 5 endpoints has these characteristics and demonstrable clinical benefit.

**Response—**Assessment of response compares manifestations of chronic GVHD for each patient at baseline and at one or more defined subsequent time points. 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 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. 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.

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 order to determine whether the benefit is sustained. Protocols should specify how response should be categorized when a new local or systemic treatment is added after a patient has been enrolled but before efficacy is assessed. 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. With highly heterogeneous study populations, single-arm designs cannot control and account for the myriad other factors that could influence response in a study with response as the primary endpoint.

Response endpoints should be defined in ways that are consistently associated with demonstrable clinical benefit. In patients with chronic GVHD that is unlikely to be cured, response endpoints should be defined in a way that demonstrates clinically meaningful durable improvement in disease activity and symptoms. For example, resolution of oral lichenoid changes in a patient with persistent diarrhea should not be considered as clinically meaningful improvement, while isolated improvement in the mouth that leads to better nutrition might be considered clinically meaningful even if other less bothersome manifestations persist. Likewise, improvements that are not durable should not be considered as clinically meaningful. For patients with chronic GVHD characterized by a high risk of mortality, an association of a response endpoint with prolonged overall survival could provide evidence of clinical benefit. For example, changes in the 0-3 NIH composite skin score correlated with both clinician and patient perception of improvement or worsening, and worsening skin scores at 6 months were associated with decreased overall survival. For patients with chronic GVHD characterized by a low risk of mortality, an association of a response endpoint with improved PROs could provide evidence of clinical benefit. Whether a response endpoint should be defined as complete response or as complete

or partial response depends on the degree to which the partial response component of the endpoint is associated with the other benefits seen in patients with complete response.

Table 4 summarizes results of previous studies investigating clinician-reported measures as potential indicators of benefit in clinical trials. The provisional criteria proposed by the 2005 NIH Consensus Conference for measuring treatment response of chronic GVHD were based on expert opinion [27], and an Excel spreadsheet tool has been developed to apply these criteria in clinical trials [28]. Responses defined according to the proposed algorithm correlated with improved symptom burden but not with improved quality of life by other measures [29]. Furthermore, agreement between response and physicians' clinical assessment was poor [30]. Response at 6 months correlated with a lower risk of subsequent mortality in a prospective study of 39 patients with steroid-refractory chronic GVHD [28] but not in a prospective, multicenter, observational cohort comprised of 283 chronic GVHD cases [30]. The association of response with subsequent overall survival might depend on patient or disease characteristics or on the context of first-line versus subsequent treatment. In a study of first-line treatment, complete response or complete plus partial response by a wide variety of definitions at 6 months did not correlate with subsequent resolution of the disease and successful withdrawal of systemic treatment [31].

While complete response clearly provides clinical benefit, the extent to which partial response also provides clinical benefit is less clear. Several approaches could be used to increase the likelihood that response is associated with clinical benefit. First, the benefit of partial response could be enhanced if the definition included a requirement to demonstrate improvement in the most severe manifestation of chronic GVHD. Second, the stringency and reliability of criteria for partial response would be enhanced if the definition required an improvement across two categories of severity in a 4-point scale (e.g., from 4 to 2 or from 3 to 1) instead of one (e.g., from 4 to 3, 3 to 2, or 2 to 1). Third, clinical benefit might be more apparent when response is measured at 12 months instead of 6 months. Studies analyzing data prospectively collected in a standardized manner could be used to test these hypotheses.

**Failure-free survival—**For this endpoint, "failure" has been defined as death, recurrent or progressive malignancy, or the initiation of new systemic treatment for chronic GVHD [24, 32]. Increased dosing of existing treatment is not considered as failure. The premise underpinning this endpoint is that chronic GVHD was adequately controlled in cases where no new systemic treatment was given and that GVHD was not adequately controlled in cases where new systemic treatment was given. Results of retrospective studies showed that for both first-line and second-line treatment, the preponderant cause of failure was the initiation of new systemic treatment, while death and recurrent malignancy accounted for only a small proportion of failures.

The absence of death and recurrent malignancy as components of FFS are presumed to reflect clinical benefit. In a landmark analysis, the absence of new treatment within 12 months after first-line treatment or within 6 months after second-line treatment was associated with a higher subsequent probability of cure of chronic GVHD but not with improved subsequent overall survival [24, 32]. A time-dependent Cox proportional hazards analysis showed that administration of second-line treatment is associated with an increased

risk of non-relapse mortality compared to continued first-line treatment [33]. Data from a prospective longitudinal observational study have suggested that patients with FFS at 1 year have measurable overall reductions in symptom burden, disease activity and functional impairment compared to baseline (see on-line Supplement). Other evidence supports the clinical benefit associated with FFS when the prednisone dose at the endpoint assessment time is taken into account. Lower prednisone doses were associated with higher subsequent probabilities of durably controlling and curing the disease [24, 32] and lower death rates (see on-line Supplement).

Several problems remain to be addressed with the use of FFS as the primary endpoint in clinical trials. First, new treatment decisions are not always driven by lack of efficacy. In particular, new treatment introduced as a replacement for an investigational product that has caused toxicity confounds any subsequent assessment of efficacy of the investigational product. Ideally, the use of an investigational product should be developed sufficiently in phase 2 studies in order to minimize the incidence of treatment discontinuations because of toxicity in a pivotal trial. Second, new treatment decisions are subject to bias, making this primary endpoint inadequate for regulatory purposes. As an alternative approach that could address both problems, pre-specified criteria generally accepted as indicating a need for new treatment could be used as an objective endpoint, regardless of whether treatment had been changed or not. Third, additional studies in other cohorts are needed in order confirm the clinical benefit of FFS or any alternative approach using pre-specified criteria as an objective endpoint.

**Survival without progressive impairment—**As discussed above, treatment of chronic GVHD is intended to produce a sustainable benefit by reducing symptom burden, controlling objective manifestations of disease, preventing organ damage and progressive impairment leading to disability, and improving overall survival while avoiding disproportionate toxicity related to treatment. The term “progressive impairment” is intended to capture the emergence of any enduring chronic GVHD-related manifestation that threatens or compromises a patient's physical well-being or function in ways that cannot be easily reversed. Hence, “progressive impairment” indicates inadequately controlled chronic GVHD.

The 2014 Response Criteria Working Group Report defined criteria for progression in the various manifestations of chronic GVHD [14]. Some of these criteria clearly represent progressive impairment, while others do not. The criteria for progression of the NIH skin score, eye score, NIH joint and fascia score, photographic range of motion score, NIH lung symptom score, upper and lower gastrointestinal scores and esophagus score and decrease in percent predicted FEV1 lung function test all represent progressive impairment. Certain other manifestations such as the development of persistent oral ulceration that interferes with oral intake and vaginal involvement that interferes with sexual function could also be considered as progressive impairment. In contrast, progression defined according to skin itching, the chief eye complaint, the oral mucosal scale, oral sensitivity, liver function tests, or global rating scales would not necessarily indicate progressive impairment since they are more easily reversed. In many instances, such progression can be managed by topical treatment or by increasing the dose of prednisone.

The proposed use of survival without progressive impairment (SWOPI) as the primary endpoint in chronic GVHD treatment trials is based on the premise that products cannot prevent progressive impairment unless they also reduce symptom burden and control objective manifestations of chronic GVHD. Conversely, the clinical value of products that reduce symptom burden and control objective manifestations of chronic GVHD in the short term would be considerably diminished if they could not also prevent progressive impairment in the longer term. SWOPI is conceptually similar to “progression-free survival” (PFS) in oncology trials by focusing on the absence of progression as the primary measure of success. This endpoint would be highly relevant for patients with far advanced chronic GVHD that has continued to progress despite the use of multiple systemic treatments for many years. Durable prevention of further impairment without treatment-related toxicity would have considerable value, even if systemic treatment cannot be withdrawn.

Methods for measuring progressive impairment are not fully developed (see on-line Supplement). Rates of provisionally defined SWOPI events in a mixed cohort of currently treated incident and prevalent chronic GVHD cases were high, demonstrating considerable room for improvement (see online Supplement). The advantage of an investigational product could be demonstrated if its use prevents progressive impairment more effectively than the standard of care. The use of SWOPI as an endpoint has the advantage that it is unaffected by temporary improvement or worsening of reversible disease manifestations associated with changes in steroid dose or topical treatment (Figure 1).

Further work is needed to establish agreement that each component in a definition of progressive impairment truly indicates reliably measured harm, that chronic GVHD is the most likely cause, and that important components have not been omitted. Patient input should be incorporated into the selection of these components. If a PRO instrument is used, assessment of symptoms by the patient should not include signs or other determinations that would be best made by a clinician, and the clinician's assessment should not include symptoms that would be most reliably reported by the patient. A SWOPI endpoint has the potential to include adverse events that could confound the interpretation of efficacy. Instruments should distinguish impairment caused by the disease per se from those caused by the investigational product or by an interaction of the product with chronic GVHD.

Additional work is also needed to characterize the clinical benefit associated with SWOPI by determining whether progressive impairment is correlated with increased symptom burden and disease activity. Studies should evaluate whether SWOPI predicts improved overall survival or earlier resolution of chronic GVHD and withdrawal of systemic treatment, although this association is not a requirement for determining clinical benefit. Consistency of effect should also be assessed in pre-specified subsets of patients with specific manifestations of chronic GVHD. Data from these studies are needed to identify subsets of patients characterized by higher and lower risks of progressive impairment and to determine the relationship between the duration of follow-up after enrollment and the magnitude of clinical benefit associated with SWOPI. Such data from an early-phase trial would be very useful for sample size considerations in pivotal trials.

As discussed in Guidance for Industry. Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics [34], the frequency of assessment and missing data can complicate the use of a PFS-like endpoint in a time-to-event analysis. Use of PFS as a clinical endpoint for a regulatory decision must be meaningful for the particular study population. Whether this endpoint is meaningful depends on its relevance to the direct clinical benefit, magnitude of the effect, and the risk-benefit of treatment with the investigational product as compared to available therapies.

**Patient reported outcomes—**Incorporation of the patient experience into endpoints for clinical trials addresses the "living better" component of "clinical benefit." For a disease such as chronic GVHD, quality of life and symptoms may reflect disease activity, residual effects of GVHD or the side effects of medications used to treat GVHD. FDA has released draft guidance for qualification of PRO instruments [35]. This guidance outlines steps necessary to consider a PRO instrument adequate to measure clinical benefit for purposes of regulatory approval.

Growing evidence supports the validity of PRO instruments in clinical trials of treatment for chronic GVHD. The Lee Symptom Scale is a 30-item, 7-domain symptom scale that has proven reliable, valid, and sensitive to change. This scale was developed with patient input and was tested in a cohort of 107 patients with active chronic GVHD who completed the questionnaire every 3 or 6 months. Psychometric properties have been published [36]. Subsequent studies have shown that changes in the NIH eye, skin, mouth, GI, and summary scales have correlated with patient- and clinician-reported changes in chronic GVHD activity [29, 37-39]. Although most symptoms are specific to chronic GVHD activity, the interpretation of changes may be confounded by adverse side effects of treatment or side effects of transplantation independent of chronic GVHD. In addition, most trials of chronic GVHD treatment are not blinded, raising concerns about the validity of PROs that can be affected by patient beliefs that an active drug is being administered. Further, the credibility of the analysis may be confounded by missing data and patient dropouts.

The only other chronic GVHD-specific scale is the MD Anderson chronic GVHD symptom scale, published only in abstract form, and modeled after the MD Anderson Symptom Inventory (MDASI) [40]. To date, almost no work in chronic GVHD has used the Patient Reported Outcomes Measurement Information System (PROMIS) instruments [41].

Multi-dimensional health-related quality of life (HR-QOL) instruments such as the MOS SF-36 (Medical Outcomes Study Short Form 36) [42, 43] and the FACT-BMT (Functional assessment of cancer therapy – bone marrow transplantation subscale) [44, 45] have been used in many trials. In general, these instruments are able to detect differences according to the occurrence of chronic GVHD [46], severity of chronic GVHD [47] and change in chronic GVHD activity as reported by patients and clinicians [48], but not when compared with 2005 NIH calculated responses [29]. NIH-calculated response measures capture changes of value or importance to clinicians, but the extent to which they do so for patients has not been defined. Many multi-domain HR-QOL instruments lack sensitivity to changes in specific syndromes associated with disease states. In addition, these instruments are

sensitive to personality traits. Table 5 provides details from studies addressing PROs sensitivity to change in patients with chronic GVHD.

A PRO assessment would be useful in characterizing clinical benefit and might be acceptable as a key secondary or co-primary endpoint to measure the core disease-related symptoms of chronic GVHD. Investigators are encouraged to work closely with regulatory authorities in defining specific PRO measures proposed as key secondary or co-primary endpoints in clinical trials. FDA Guidance for Industry Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims will assist the development, selection, or modification of a well-defined and reliable PRO assessment intended to support labeling claims of treatment benefit [35]. The sample size is driven by the proposed labeling claims. The sample size should therefore account for any key secondary endpoints needed for approval.

**Composite scale—**Validated scales that incorporate clinician assessments (e.g., on a 0-10 or global scale, or organ measures), patient-reported outcomes (e.g., symptoms or quality of life), and laboratory or functional measures (e.g., C-reactive protein) have been used as the primary endpoints in registration trials for other immune-mediated diseases such as lupus [49-55], Crohn's disease [56, 57], ankylosing spondylitis [58] and rheumatoid arthritis [59, 60]. These scales were generally developed by identifying clinical, laboratory and patient-reported parameters associated with reported perceptions of change or changes in management (e.g., adding or decreasing immunosuppressive treatment).

No such composite scale exists for chronic GVHD (see on-line Supplement). The value of including a variety of measures reflecting different aspects of a disease process is codified in the Outcomes Measures in Rheumatology (OMERACT) effort [61, 62]. The OMERACT consensus initiative specifies the process of identifying a core set of measures that should be included in any randomized controlled trial or long-term observational study in a rheumatologic disease, including incorporation of the patient perspective from the start of the process. The framework includes four areas: Death, Life Impact, Resource Utilization, and Pathophysiological Manifestations. Life impact is generally assessed by PROs. Pathophysiologic manifestations are measured by physical exam or laboratory testing. The OMERACT filter requires that one measure in each area be identified as a core measure.

No gold-standard anchor has been defined in assessing a composite scale endpoint for trials of treatment for chronic GVHD. Whether a proposed composite scale endpoint would need to be qualified against a “gold standard” would depend on the individual components, the intended population and the context in which it is to be used. Different composite scales may be needed for different patient populations, and the components in a composite scale might need to be adjusted as new drugs alter the course of the disease. Demonstration of a survival benefit might be expected from effective treatment in a subgroup with relatively short overall survival, while demonstration of clinical benefit through clinical response or a PRO might be more appropriate for patients who live longer with a potential for chronic GVHD-related impairment. Several steps may be needed to reach the ultimate goal of showing that a composite scale correlates with the longer-term goals of preventing disability and controlling the disease until systemic treatment can be withdrawn.

Given the complexity of developing a composite scale, it would be preferable to identify simple endpoints wherever possible and to pre-specify the other measures as additional secondary endpoints to test for internal consistency. A composite endpoint would certainly be acceptable for regulatory purposes if each component could be justified, but sample size considerations or studies of individual patient populations may warrant a simpler endpoint or co-primary endpoints instead. CROs and PROs that are well defined and reliable in the intended population and context may be relevant measures of clinical benefit on their own.

## KEY CONSIDERATIONS FOR THE DESIGN OF CLINICAL TRIALS

Although the definition and characterization of primary endpoints that indicate clinical benefit represent urgent goals for the immediate future, many other key considerations apply in the design of clinical trials for treatment of chronic GVHD. Many of these considerations were addressed in the 2006 Design of Clinical Trials Working Group Report. The following sections address some key considerations that merit further elaboration based on experience and progress during the past decade.

**How are lines of treatment defined?—**First-line treatment is defined as the beginning of systemic treatment for chronic GVHD, typically with NIH global level 2 severity. Treatment generally involves the introduction of prednisone or an increase in the dose to ≥0.5 mg/kg per day, with or without the introduction or continued administration of other agents. Subsequent lines of treatment are most clearly defined by the introduction of any systemic agent not previously used in the regimen for first-line treatment. Dose adjustments of non-steroidal medications used for any given line of treatment are typically not considered as the beginning of the next line of treatment.

The question of whether steroid dose adjustments should be considered as evidence of treatment failure or defined as the beginning of a new line of treatment has not been entirely resolved. In retrospective studies, temporarily increased prednisone doses up to 1 mg/kg per day were not considered as treatment failure or the beginning of a new line of treatment, and pre-specified threshold doses of prednisone at defined time points after starting treatment were used as a component in composite endpoints. Although endpoint results from these studies could be used as benchmarks for early-phase single-arm trials, the interpretation of prospective study results would be confounded by potential bias in the management of steroid dosing.

Compliance with rigid dosing and tapering schedules for administration of steroids is not feasible in GVHD treatment trials. Therefore, clinical protocols should allow some flexibility in the management of steroid administration. For example, trials for first-line treatment have allowed temporary escalation of prednisone doses up to 1 mg/kg per day without necessarily designating such events as treatment failure, even in situations where the disease could be categorized as steroid-refractory or steroid-dependent. Trials for second-line and subsequent treatment have allowed re-escalation up to the dose administered at enrollment in the trial or up to a pre-specified dose that would be considered consistent with standard management principles.

The working group discussed several different approaches for defining treatment failure and the beginning of a new line of treatment based on changes in steroid dosing. In first-line treatment trials, increased prednisone dosing up to 1 mg/kg per day could be allowed without designating these events treatment failures, based on the argument that flares of chronic GVHD are inevitable if steroid doses are tapered too rapidly. In trials for second or subsequent lines of treatment, increased prednisone doses could be allowed as long as they do not exceed the dose at the time of enrollment or do not exceed a threshold specified in the protocol. In trials with a primary endpoint to be assessed at 6 to 12 months after enrollment, a brief pulse of steroid treatment early in the trial could be allowed if needed, but the protocol would have to specify the maximum steroid dose, duration of steroid administration and number of pulses, together with the maximum interval time from enrollment.

The guidelines for determining eligibility for second-line treatment trials based on inadequate response to steroid therapy would logically apply in defining failure of first-line treatment based on steroid dose changes alone. For example, an increase in the prednisone dose because of persistent, manifestations that are not improving despite 4 weeks of treatment at >0.5 mg/kg per day or an increase in the prednisone dose to >0.25 mg/kg per day after two unsuccessful attempts to taper the dose to lower levels could be considered as treatment failures, since these circumstances would make a patient eligible for second-line treatment. Results in trials for second and subsequent lines of treatment would be most informative if no increase in the steroid dose is allowed within a defined period of time before enrollment or at the time of enrollment, and if any subsequent increase in the steroid dose above the baseline is interpreted as treatment failure and the beginning of a new line of treatment.

Taken together, these considerations emphasize the need for clarity in the definitions of eligibility criteria and endpoints with respect to changes in steroid dosing in designing clinical trials. The complex vagaries of decision-making related to steroid dosing emphasize the value and importance of controlled designs with blinding in pivotal trials.

**What specific considerations apply for first-line treatment studies?—**Most first-line trials involve treatment with steroids and an investigational product in single-arm trials and steroids with or without an investigational product in controlled trials. All protocols should specify the following: 1) whether administration of pre-study treatments should be discontinued or continued when patients are enrolled in the study, 2) whether steroid doses may be changed or new topical agents added at the time of enrollment, 3) whether steroid doses may be increased above the baseline dose after enrollment, and 4) whether new topical agents may be added after enrollment. The protocol should define the initial steroid dosing regimen and provide guidelines for tapering the dose of steroids and the sequence in relation to other treatments. The protocol should also provide guidelines for the subsequent withdrawal of other GVHD treatment medications, including the investigational product.

**What specific considerations apply for second and subsequent lines of treatment?—**Single-arm phase 1 or 2 trials have been used for the initial evaluation of systemic agents for treatment of chronic GVHD that has not been adequately controlled with steroid treatment. Phase 2 randomized controlled trials may also be considered. Patients in

the control arm should be treated with an accepted standard of care. Eligibility for phase 1 trials depends on the anticipated toxicity and efficacy profiles of the investigational product. For evaluation of potent and potentially toxic immunosuppressive agents, eligibility should be restricted initially to patients with advanced steroid-refractory chronic GVHD.

In addition to the first 4 considerations for first-line treatment studies, all protocols for second and subsequent lines of treatment should specify the following: 1) whether both second and subsequent lines of treatment are allowed, 2) the minimum interval time from the most recent change of systemic treatment to enrollment, and 3) the types and timing of recent treatment changes that are allowed with respect to steroid dosing and the use of topical agents. These considerations are particularly important in studies of patients with sclerotic manifestations, where improvement might not occur until several months after starting treatment.

In single-arm early-phase trials for second-line or subsequent treatment, eligibility criteria may be narrowed in order to improve homogeneity in baseline characteristics of the study cohort, thereby facilitating informal comparisons with results of other single-arm trials. In contrast, eligibility criteria in controlled early-phase trials may be defined more broadly, depending on the anticipated target population for later pivotal trials.

**For response-based endpoints, what reasons for beginning new systemic or topical treatment should be considered as "failure" in the analysis of response-based endpoints, and what reasons should be allowed without being considered as "failure"?—**The protocol should define the extent to which changes in concomitant treatment with systemic and topical agents are allowed at baseline and subsequently. Response-based endpoints are likely to be confounded when such changes in topical or systemic treatment are allowed. Reasons for adding new systemic treatment should always be recorded. The protocol should specify how response would be assessed when such changes are made. Addition of new systemic treatment because of worsening disease manifestations should always be counted as progression in a response endpoint. Pre-emptive addition of new systemic treatment before the response assessment to prevent progression after treatment with an investigational product has been discontinued in a patient with stable disease manifestations should also be counted as progression. Likewise, addition of new systemic treatment before the response assessment because improvement has halted should also be assessed as progression in a response endpoint, although efforts should be made to minimize such changes in therapy, if possible.

**What is the timeframe for expecting responses with various manifestations of chronic GVHD?—**The expected minimal time for response varies and depends on the specific manifestation. Improvement is expected to occur within 4 to 8 weeks for inflammatory manifestations such as erythema, edema, transaminase elevation and diarrhea. Improvement of established sclerosis takes at least 6 months to a year, but may occur within 3 months for early inflammatory fasciitis manifested as edema and tenderness with decreased range of motion without fixed joint contractures.

**What manifestations of chronic GVHD should be considered as "irreversible" for purposes of measuring response?—**Advanced fibrosis, sclerosis, adnexal loss, bronchiolitis obliterans and destruction of lacrimal and salivary glands are often considered irreversible, although complete resolution of advanced cutaneous sclerosis has been reported in some studies [28].

**Would it be acceptable to design a trial that aims only to keep chronic GVHD from progressing or from causing impairment?—**Early experience showed that without treatment, chronic GVHD will progress relentlessly toward disability and death, but prolonged treatment with high-dose glucocorticoids can cause devastating toxicity. Development of a well-tolerated product that could replace prednisone while effectively and reliably preventing newly diagnosed or early stage moderately severe chronic GVHD from progressing or causing irreversible impairment would represent a major step forward in the field, even if this product did not reverse pre-existing manifestations. The high proportion of patients who advance to second-line treatment within the first 2 years of first-line treatment demonstrates that current approaches leave much room for improvement. Similarly, development of a well-tolerated product that could prevent advanced disease from progressing further or causing increased impairment or disability without requiring interminable treatment with high-dose prednisone would represent a major step forward in the field, even if it did not reverse pre-existing manifestations. The high proportion of patients who advance to third-line treatment within 12 months of second-line treatment demonstrates that current approaches are far from satisfactory.

**What are the advantages and disadvantages of controlled trials versus single-arm trials?—**Single-arm trials cannot adequately determine the extent to which trial results were influenced by the disease trajectory before enrollment, the baseline characteristics of the study cohort, or by any concomitant treatment started at enrollment or added between enrollment and the endpoint assessment. Accordingly, the treatment effect (i.e., safety and efficacy) of an investigational product can be difficult to assess in single-arm trials, unless results with a homogeneous population of study patients can be compared to a similarly homogeneous historical group. In most situations, the results of a single-arm phase 2 trial can be used only to determine whether an investigational product has enough activity to warrant further investigation in a phase 3 trial.

Controlled trials make it possible to determine the true treatment effect, if the prior disease trajectory, baseline characteristics and concomitant treatment are similar between the arms. Eligibility criteria can be more flexible in controlled trials, since matching for comparisons with historical experience is not necessary. In controlled trials, stratified randomization decreases the probability of an imbalance in the distribution of risk factors that could affect the primary endpoint. With any given statistical error specification, however, the required sample size is much larger for controlled trials than for single-arm trials. In controlled phase 2 studies intended only to assess the merits of a phase 3 study, this disadvantage could be mitigated by allowing a larger type 1 error specification, and in phase 3 studies, the numbers of patients can be optimized by using group sequential designs. Controlled trials are also more difficult to organize and conduct, because multi-center participation is necessary,

although multi-center participation has the advantage of mitigating possible center-specific effects on trial results. In addition, patients and physicians may be reluctant to participate in controlled trials testing a marketed product if prior experience has suggested that a readily available investigational treatment has advantages over the standard of care or if the known efficacy of standard treatment is limited.

**What are the advantages and disadvantages of crossover designs?—**Crossover designs can be used to compare initial outcomes of treatment with an investigational product versus the standard of care or with two different investigational products. To some extent, crossover designs overcome the limitations of single-arm designs by allowing results with 2 different types of treatment to be compared. Crossover designs also afford all patients an opportunity to be treated with an investigational product. Randomized crossover designs enable a robust interpretation of results up to the crossover point, but the interpretation of outcomes after the crossover point is confounded by the prior treatment. Blinded designs are critically important in order to prevent bias in crossover decisions.

**What are the advantages and disadvantages of delayed start designs?—**Delayed start designs can be used to document the trajectory of disease manifestations before beginning treatment with an investigational product. For example, serial monitoring of pulmonary function test results in a delayed start study could determine whether treatment with an investigational product changes the progression of bronchiolitis obliterans in the absence of a control group. In these studies, the criteria that trigger the onset of treatment must be defined in a way that allows unambiguous demonstration of progression or prolonged stability, without risking harm caused by unduly delayed treatment.

**What are the advantages and disadvantages of composite endpoints?—**Composite endpoints make it possible to encompass several different measures of clinical benefit associated with the primary endpoint of a trial. Each component of a composite endpoint must have demonstrable clinical benefit. The individual components of composite endpoints may have large differences in the extent to which they indicate clinical benefit, thereby making composite endpoints more difficult to interpret as compared to simple endpoints. A composite endpoint affords greater sensitivity to detect treatment failure, especially if the components address different aspects of the disease. Comparisons among different studies could be facilitated by reporting standardized composite endpoints that reflect key aspects of disease activity.

**Do placebos have any role in controlled trials of treatment for chronic GVHD?—**Placebos could be used for 2 purposes. In trials testing the effect of adding a second agent to the standard of care, a “placebo” can be used to blind of the arm assignments. In this situation, the blinded study product is not a “placebo” in the true sense of the word, since patients in the control arm are treated with an active standard of care. Placebos could also be used in trials testing the effects of treatment in patients with stable disease manifestations that do not need immediate intervention.

**What specific considerations apply when clinical trial results will be submitted for regulatory review?—**Trials submitted for regulatory review require a

meticulous statistical analysis plan that will support the proposed labeling claims, together with extensive detail in documenting adverse events and the use of concomitant medications. In other respects, the design and conduct of clinical trials should be based on good clinical science and not be influenced by plans for regulatory review.

## LESSONS FROM REGULATORY REVIEW OF TREATMENT FOR OTHER DISEASES

Two large-scale reviews of decisions by the United States FDA offer insights for the design and conduct of studies intended for regulatory review [63, 64]. The first report characterized pivotal efficacy trials that provided the basis for approval of novel therapeutic agents between 2005 and 2012 [63]. Among the 448 trials, 36 were intended for 13 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 [64]. As summarized in an accompanying editorial [65], 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 of sufficient sample size poses the most difficult challenge in conducting trials for treatment of chronic GVHD. The largest trial to date enrolled 287 patients [9]. Two recent multicenter trials took 4 years to enroll 151 patients in each (Paul Carpenter, personal communication; April, 2014) [8], even though both adults and children were eligible. Both were stopped early for futility. Hence, a very large effect size would be needed for rapid progress in developing a new treatment for chronic GVHD.

## CLINICAL DEVELOPMENT PATHS

In the absence of approval of any drug for treatment of chronic GVHD, no precedent for development paths leading to regulatory approval for this indication has been established. Even so, some general principles have emerged from the considerations summarized above. In this context, phase 1 studies are intended primarily to identify a safe dose of an investigational product specifically in patients with chronic GVHD. As might be expected, the side effects of marketed products in patients with chronic GVHD are generally similar to those observed in patients with the approved indication, but careful consideration must be given to the implications of differences in the concomitant medications that are typically used in patients with chronic GVHD as compared to those used in patients with the

approved indication. In studies testing marketed products for chronic GVHD as a new indication, the initial doses and schedules of administration can be based on those for the approved indication, but preliminary dose finding studies with assessment of pharmacokinetics, pharmacodynamics, potential drug interactions and adverse events may be needed. Studies testing products in humans for the first time would have to follow the usual approach for determining the initial dose, frequency of administration, and dose escalation in phase 1 studies.

Initial studies of treatment for chronic GVHD always include some measure of clinical activity in controlling the disease. For this purpose, shorter-term endpoints are preferable to longer-term endpoints. For example, it would be reasonable to expect that an active product could improve cutaneous erythema and readily reversible oral, gastrointestinal and hepatic manifestations of chronic GVHD within 4 to 8 weeks after starting treatment. Much longer follow-up is needed to determine whether a product could prevent or reverse sclerotic manifestations of chronic GVHD. Systemic treatment would not be expected to reverse bronchiolitis obliterans or destruction of lacrimal and salivary glands caused by GVHD, although certain products could relieve symptoms caused by such damage.

Phase 2 studies should be designed to determine whether the short-term safety and activity of the product can be confirmed in a larger and potentially more diverse cohort of patients and to assess the safety and activity of the product with respect to the longer-term goals of providing a sustainable benefit. The optimal primary efficacy endpoint for these studies has not yet been defined and characterized. Response definitions associated with sustainable improvements in the most bothersome symptoms and overall symptom burden, reduced disease activity, absence of progressive impairment related to chronic GVHD, and improved survival would offer evidence of clinical benefit. An important goal of phase 2 studies is to estimate the size of effects on the primary endpoint in order to support the design of phase 3 studies. The secondary efficacy endpoints in phase 2 studies should be designed to explore and help characterize the clinical benefit that may be associated with the primary endpoint both in the overall cohort and in subsets of patients with specific manifestations of chronic GVHD. The use of standardized instruments and time points for assessment of efficacy is essential in order to enable comparison of results across multiple studies. Safety endpoints should be designed to assess the long-term tolerability of the investigational product and to identify any potential drug interactions and dose adjustments to be considered and incorporated in a pivotal trial.

The most appropriate primary endpoints for a pivotal trial remain to be defined. The low mortality risk in many patients with chronic GVHD would make it difficult to demonstrate survival benefits in pivotal trials, given the number of patients available for such studies, and a minimum follow-up of at least 2 – 3 years would be needed to demonstrate improved cure rates. These considerations highlight the importance of current efforts to characterize the clinical benefit associated with shorter-term endpoints that could be used in future pivotal trials.

## CONCLUSIONS

Challenges in the design of chronic GVHD treatment trials are much more clearly defined than they were in 2005. As emphasized throughout this report, the identification and characterization of primary endpoints that indicate clinical benefit represent the most urgent goals to be accomplished within the next several years. Prospectively collected data from well-designed observational studies and clinical trials should be used to characterize the clinical benefit associated with a variety of proposed endpoints assessed at specific time points. The most informative results are likely to come from replicated analyses of cases representative of an intended treatment population, anchored to a treatment change and having a well-documented baseline for assessment of response. Incident cases may have less heterogeneity and fewer irreversible disease manifestations compared to prevalent cases, but prevalent cases are more frequent than incident cases.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGMENTS

This project was supported by the National Institutes of Health (NIH): National Cancer Institute (NCI), Center for Cancer Research, Intramural Research Program and Division of Cancer Treatment and Diagnosis, Cancer Therapy Evaluation Program; Office of Rare Disease Research (ORD), National Center for Advancing Translational Sciences; Division of Allergy, Immunology and Transplantation, National Institute of Allergy and Infectious Diseases; National Heart Lung and Blood Institute, Division of Blood Diseases and Resources. The authors acknowledge the following organizations that, by their participation, made this project possible: American Society for Blood and Marrow Transplantation, Center for International Bone and Marrow Transplant Research, US Chronic GVHD Consortium (supported by ORD and NCI), German-Austrian-Swiss GVHD Consortium, National Marrow Donor Program, the Health Resources and Services Administration, Division of Transplantation, US Department of Human Health and Services, Canadian Blood and Marrow Transplant Group, European Group for Blood and Marrow Transplantation, Pediatric Blood and Marrow Transplant Consortium and Deutsche José Carreras Leukämie-Stiftung. The organizers are indebted to patients and patient and research advocacy groups, who made this process much more meaningful by their engagement. Acknowledgement goes to the Meredith A. Cowden GVHD foundation for facilitating the initial planning meeting in Cleveland in November, 2013 in conjunction with the National GVHD Symposium. The project group also recognizes the contributions of numerous colleagues in the field of blood and marrow transplantation in the US and internationally, medical specialists and consultants, the pharmaceutical industry, and the NIH and US Food and Drug Administration professional staff for their intellectual input, dedication, and enthusiasm on the road to completion of these documents. For expert contributions to this 2014 NIH Consensus Clinical Trials Working Group report, special acknowledgement goes to Dr. William D. Merritt. Project participants also recognize Dr. Joseph H. Antin and Dr. Gérard Socié for their independent expert reviews and comments on documents during the June 2014 meeting.

Work by P.J.M. and S.J.L. was supported by grants CA118953 and CA18029 from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services.

## APPENDIX: NATIONAL INSTITUTES OF HEALTH CONSENSUS-DEVELOPMENT PROJECT ON CRITERIA FOR CLINICAL TRIALS IN CHRONIC GVHD STEERING COMMITTEE

Members of this committee included: Steven Z. Pavletic, Georgia B. Vogelsang and Stephanie J. Lee (project chairs), Mary E.D. Flowers and Madan Jagasia (Diagnosis and Staging), David E. Kleiner and Howard M. Shulman (Histopathology), Kirk R. Schultz and Sophie Paczesny (Biomarkers), Stephanie J. Lee and Steven Z. Pavletic (Response Criteria),

Daniel R. Couriel and Paul A. Carpenter (Ancillary and Supportive Care), Paul J. Martin and Corey S. Cutler (Design of Clinical Trials), Kenneth R. Cooke and David B. Miklos (Chronic GVHD Biology), Roy Wu, William D. Merritt, Linda M. Griffith, Nancy L. DiFronzo, Myra Jacobs, Susan K. Stewart, Meredith A. Cowden (members).

## REFERENCES


1. CIBMTR. http/www.cibmtr.org/ReferenceCenter/SlidesReports/SummarySlides/pages/index.aspx
2. Flowers ME, Inamoto Y, Carpenter PA, et al. Comparative analysis of risk factors for acute graft-versus-host disease and for chronic graft-versus-host disease according to National Institutes of Health consensus criteria. Blood. 2011; 117:3214–3219. [PubMed: 21263156]
3. Vigorito AC, Campregher PV, Storer BE, et al. Evaluation of NIH consensus criteria for classification of late acute and chronic GVHD. Blood. 2009; 114:702–708. [PubMed: 19470693]
4. 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; 7:265–273. [PubMed: 11400948]
5. 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; 18:84–91. [PubMed: 21689773]
6. Koc S, Leisenring W, Flowers ME, et al. Thalidomide for treatment of patients with chronic graft-versus-host disease. Blood. 2000; 96:3995–3996. [PubMed: 11090092]
7. Koc S, Leisenring W, Flowers MED, et al. Therapy for chronic graft-versus-host disease: a randomized trial comparing cyclosporine plus prednisone versus prednisone alone. Blood. 2002; 100:48–51. [PubMed: 12070007]
8. Martin PJ, Storer BE, Rowley SD, et al. Evaluation of mycophenolate mofetil for initial treatment of chronic graft-versus-host disease. Blood. 2009; 113:5074–5082. [PubMed: 19270260]
9. 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; 72:546–554. [PubMed: 3042041]
10. 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; 112:2667–2674. [PubMed: 18621929]
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; 17:1775–1782. [PubMed: 21621629]
12. Martin PJ, Inamoto Y, Carpenter PA, Lee SJ, Flowers ME. Treatment of chronic graft-versus-host disease: Past, present and future. Korean J Hematol. 2011; 46:153–163. [PubMed: 22065969]
13. 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; 12:491–505. [PubMed: 16635784]
14. Jagasia MH, Greinix HT, Arora M, et al. National Institutes of Health Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-versus-Host Disease: I. The 2014 Diagnosis and Staging Working Group Report. Biology of Blood and Marrow Transplantation. 2015; 21:389–401. [PubMed: 25529383]
15. Shulman HM, Cardona DM, Greenson JK, et al. NIH Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-versus-Host Disease: II. The 2014 Pathology Working Group Report. Biology of Blood and Marrow Transplantation. 2015; 21:589–603. [PubMed: 25639770]
16. Paczesny S, Hakim FT, Pidala J, et al. National Institutes of Health Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-versus-Host Disease: III. The 2014

Biomarker Working Group Report. Biol Blood Marrow Transplant. 2015 doi.org/10.1016/j.bbmt.2015.01.003.

17. Carpenter PA, Kitko CL, Elad S, et al. National Institutes of Health Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-versus-Host Disease: V. The 2014 Ancillary Therapy and Supportive Care Working Group Report. Biology of Blood and Marrow Transplantation. 2015 doi.org/10.1016/j.bbmt.2015.03.024.
18. Lee SJ, Wolff D, Kitko C, 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. The 2014 Response Criteria Working Group Report. Biol Blood Marrow Transplant. 2015 doi.org/10.1016/j.bbmt.2015.01.003.
19. 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; 57:267–276. [PubMed: 7004534]
20. Wolff D, Gerbitz A, Ayuk F, et al. Consensus conference on clinical practice in chronic graft-versus-host disease (GVHD): first-line and topical treatment of chronic GVHD. Biol Blood Marrow Transplant. 2010; 16:1611–1628. [PubMed: 20601036]
21. Wolff D, Schleuning M, von Harsdorf S, et al. Consensus Conference on Clinical Practice in Chronic GVHD: Second-Line Treatment of Chronic Graft-versus-Host Disease. Biol Blood Marrow Transplant. 2011; 17:1–17. [PubMed: 20685255]
22. Flowers MED, Martin PJ. How we treat chronic graft-versus-host disease. Blood. 2015; 125:606–615. [PubMed: 25398933]
23. Arai S, Jagasia M, Storer B, et al. Global and organ-specific chronic graft-versus-host disease severity according to the 2005 NIH Consensus Criteria. Blood. 2011; 118:4242–4249. [PubMed: 21791424]
24. Inamoto Y, Storer BE, Lee SJ, et al. Failure-free survival after second-line systemic treatment of chronic graft-versus-host disease. Blood. 2013; 121:2340–2346. [PubMed: 23321253]
25. 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; 96:1351–1356. [PubMed: 21565902]
26. Inamoto Y, Martin PJ, Storer BE, et al. Association of severity of organ involvement with mortality and recurrent malignancy in patients with chronic graft-versus-host disease. Haematologica. 2014; 99:1618–1623. [PubMed: 24997150]
27. 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; 12:252–266. [PubMed: 16503494]
28. 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; 122:4111–4118. [PubMed: 24152907]
29. Inamoto Y, Martin PJ, Chai X, et al. Clinical benefit of response in chronic graft-versus-host disease. Biol Blood Marrow Transplant. 2012; 18:1517–1524. [PubMed: 22683612]
30. Palmer JM, 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; 18:1649–1655. [PubMed: 22691695]
31. 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. 2010; 17:124–132. [PubMed: 20601033]
32. Inamoto Y, Flowers ME, Sandmaier BM, et al. Failure-free survival after initial systemic treatment of chronic graft-versus-host disease. Blood. 2014; 124:1363–1371. [PubMed: 24876566]
33. Flowers ME, Storer B, Carpenter P, et al. Treatment change as a predictor of outcome among patients with classic chronic graft-versus-host disease. Biol Blood Marrow Transplant. 2008; 14:1380–1384. [PubMed: 19041060]
34. Guidance for Industry: Clinical Trial Endpoints the Approval of Cancer Drugs and Biologics. http://www.fda.gov/downloads/Drugs/Guidances/ucm071590.pdf

35. Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf
36. Lee S, Cook EF, Soiffer R, 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. [PubMed: 12234170]
37. Inamoto Y, Chai X, Kurland BF, et al. Validation of measurement scales in ocular graft-versus-host disease. Ophthalmology. 2012; 119:487–493. [PubMed: 22153706]
38. Jacobsohn DA, Rademaker A, Kaup M, Vogelsang GB. Skin response using NIH consensus criteria vs Hopkins scale in a phase II study for steroid-refractory chronic GVHD. Bone Marrow Transplant. 2009; 44:813–819. [PubMed: 19430498]
39. Treister N, Chai X, Kurland B, et al. Measurement of oral chronic GVHD: results from the Chronic GVHD Consortium. Bone Marrow Transplant. 2013
40. Williams LA, Couriel DR, Mendoza TR, et al. A New Measure Of Symptom Burden In Chronic Graft-Versus-Host Disease. Biology of Blood and Marrow Transplantation. 2010; 16:S177. abstract.
41. PROMIS. [February 23, 2014] Dynamic Tools to Measure Health Outcomes from the Patient Perspective. www.nihpromis.org
42. Ware, JE.; Kosinski, M.; Keller, SD. SF-36 physical and mental health summary scales: a user's manual. The Health Institute, New England Medical Center; Boston: 1994.
43. Ware, JE.; Snow, KK.; Kosinski, M.; Gandek, B. SF-36 Health Survey: a manual and interpretation guide. The Health Institute, New England Medical Center; Boston: 1993.
44. McQuellon RP, Russell GB, Cella DF, et al. Quality of life measurement in bone marrow transplantation: development of the Functional Assessment of Cancer Therapy-Bone Marrow Transplant (FACT-BMT) scale. Bone Marrow Transplant. 1997; 19:357–368. [PubMed: 9051246]
45. McQuellon RP, Russell GB, Rambo TD, et al. Quality of life and psychological distress of bone marrow transplant recipients: the 'time trajectory' to recovery over the first year. Bone Marrow Transplant. 1998; 21:477–486. [PubMed: 9535040]
46. Lee SJ, Kim HT, Ho VT, et al. Quality of life associated with acute and chronic graft-versus-host disease. Bone Marrow Transplant. 2006; 38:305–310. [PubMed: 16819438]
47. Pidala J, Kurland B, Chai X, et al. Patient-reported quality of life is associated with severity of chronic graft-versus-host disease as measured by NIH criteria: report on baseline data from the Chronic GVHD Consortium. Blood. 2011; 117:4651–4657. [PubMed: 21355084]
48. Pidala J, Kurland BF, Chai X, et al. Sensitivity of changes in chronic graft-versus-host disease activity to changes in patient-reported quality of life: results from the Chronic Graft-versus-Host Disease Consortium. Haematologica. 2011; 96:1528–1535. [PubMed: 21685473]
49. Bombardier C, Gladman DD, Urowitz MB, Caron D, Chang CH. Derivation of the SLEDAI. A disease activity index for lupus patients. The Committee on Prognosis Studies in SLE. Arthritis Rheum. 1992; 35:630–640. [PubMed: 1599520]
50. Furie R, Petri M, Zamani O, et al. A phase III, randomized, placebo-controlled study of belimumab, a monoclonal antibody that inhibits B lymphocyte stimulator, in patients with systemic lupus erythematosus. Arthritis Rheum. 2011; 63:3918–3930. [PubMed: 22127708]
51. Gladman DD, Ibanez D, Urowitz MB. Systemic lupus erythematosus disease activity index 2000. J Rheumatol. 2002; 29:288–291. [PubMed: 11838846]
52. Hay EM, Bacon PA, Gordon C, et al. The BILAG index: a reliable and valid instrument for measuring clinical disease activity in systemic lupus erythematosus. Q J Med. 1993; 86:447–458. [PubMed: 8210301]
53. Liang MH, Socher SA, Larson MG, Schur PH. Reliability and validity of six systems for the clinical assessment of disease activity in systemic lupus erythematosus. Arthritis Rheum. 1989; 32:1107–1118. [PubMed: 2775320]
54. Luijten KM, Tekstra J, Bijlsma JW, Bijl M. The Systemic Lupus Erythematosus Responder Index (SRI); a new SLE disease activity assessment. Autoimmunity reviews. 2012; 11:326–329. [PubMed: 21958603]

55. Vitali C, Bencivelli W, Isenberg DA, et al. Disease activity in systemic lupus erythematosus: report of the Consensus Study Group of the European Workshop for Rheumatology Research. II. Identification of the variables indicative of disease activity and their use in the development of an activity score. The European Consensus Study Group for Disease Activity in SLE. Clin Exp Rheumatol. 1992; 10:541–547. [PubMed: 1458710]
56. Best WR, Becktel JM, Singleton JW, Kern F Jr. Development of a Crohn's disease activity index. National Cooperative Crohn's Disease Study. Gastroenterology. 1976; 70:439–444. [PubMed: 1248701]
57. 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; 122:512–530. [PubMed: 11832465]
58. 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 Rheum. 2001; 44:1876–1886. [PubMed: 11508441]
59. Felson DT, Anderson JJ, Boers M, et al. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum. 1995; 38:727–735. [PubMed: 7779114]
60. Felson DT, Smolen JS, Wells G, et al. American College of Rheumatology/European League against Rheumatism provisional definition of remission in rheumatoid arthritis for clinical trials. Annals of the rheumatic diseases. 2011; 70:404–413. [PubMed: 21292833]
61. Boers M, Idzerda L, Kirwan JR, et al. Toward a generalized framework of core measurement areas in clinical trials: a position paper for OMERACT 11. J Rheumatol. 2014; 41:978–985. [PubMed: 24584922]
62. Kirwan JR, Boers M, Tugwell P. Updating the OMERACT filter at OMERACT 11. J Rheumatol. 2014; 41:975–977. [PubMed: 24788466]
63. Downing NS, Aminawung JA, Shah ND, Krumholz HM, Ross JS. Clinical trial evidence supporting FDA approval of novel therapeutic agents, 2005-2012. JAMA : the journal of the American Medical Association. 2014; 311:368–377. [PubMed: 24449315]
64. Sacks LV, Shamsuddin HH, Yasinskaya 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 : the journal of the American Medical Association. 2014; 311:378–384. [PubMed: 24449316]
65. Goodman SN, Redberg RF. Opening the FDA black box. JAMA : the journal of the American Medical Association. 2014; 311:361–363. [PubMed: 24449313]
66. Mitchell SA, Jacobsohn D, Thormann Powers KE, et al. A Multicenter Pilot Evaluation of the National Institutes of Health Chronic Graft-versus-Host Disease (cGVHD) Therapeutic Response Measures: Feasibility, Interrater Reliability, and Minimum Detectable Change. Biol Blood Marrow Transplant. 2011
67. Jacobsohn DA, Kurland BF, Pidala J, et al. Correlation between NIH composite skin score, patient-reported skin score, and outcome: results from the Chronic GVHD Consortium. Blood. 2012; 120:2545–2552. quiz 2774. [PubMed: 22773386]
68. Palmer J, Williams K, Inamoto Y, et al. Pulmonary symptoms measured by the national institutes of health lung score predict overall survival, nonrelapse mortality, and patient-reported outcomes in chronic graft-versus-host disease. Biol Blood Marrow Transplant. 2014; 20:337–344. [PubMed: 24315845]
69. Inamoto Y, Pidala J, Chai X, et al. Assessment of joint and fascia manifestations in chronic graft-versus-host disease. Arthritis and rheumatism. 2014; 66(4):1044–1052.
70. Bassim CW, Fassil H, Mays JW, et al. Validation of the National Institutes of Health chronic GVHD Oral Mucosal Score using component-specific measures. Bone Marrow Transplant. 2014; 49:116–121. [PubMed: 23995099]
71. Curtis LM, Grkovic L, Mitchell SA, et al. NIH response criteria measures are associated with important parameters of disease severity in patients with chronic GVHD. Bone Marrow Transplant. 2014; 49:1513–1520. [PubMed: 25153693]

72. Yanik GA, Mineishi S, Levine JE, et al. Soluble tumor necrosis factor receptor: enbrel (etanercept) for subacute pulmonary dysfunction following allogeneic stem cell transplantation. Biol Blood Marrow Transplant. 2012; 18:1044–1054. [PubMed: 22155140]

73. 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

74. Walker I, Schultz KR, Toze CL, et al. Thymoglobulin decreases the need for immunosuppression at 12 monhs after myeloablative and nonmyeloablative unrelated donor transplantation: CBMTG 0801, a randomized, controlled trial. Blood. 2014; 124(21):38. abstract.

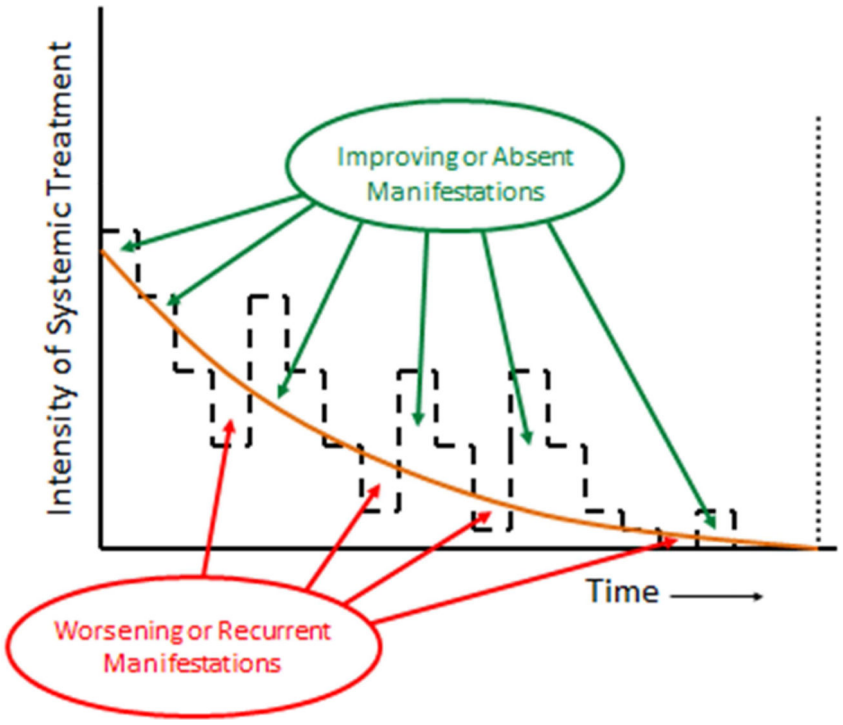


**Figure 1.**

Appropriate management of chronic GVHD requires continuous recalibration of immunosuppressive treatment in order to avoid over- or under-treatment. The intensity of treatment required to control the disease decreases across time. Manifestations of chronic GVHD improve or are absent when the intensity of treatment (- - -) is above the threshold shown as the orange curve, and they worsen or recur when the intensity of treatment is below the threshold. The slope of the threshold varies among patients and can be determined only by serial attempts to decrease the intensity of treatment. Successful management of chronic GVHD can control the disease until systemic treatment is no longer needed to prevent recurrent or progressive disease activity or exacerbation of any residual damage (■ ■ ■ ■ ■ ■ ■).

**Table 1**

Endpoint Recommendations in the 2005 Working Group Report

| Time horizon | Primary Endpoint | Secondary Endpoints |
|---|---|---|
| Short | GVHD response | Patient-reported outcomes |
| Long | Complete response | Non-relapse mortality |
| | End of systemic treatment * | Survival without recurrent malignancy |
| | | Overall survival |

* resolution of chronic GVHD and durable withdrawal of systemic treatment without subsequent recurrence or progression of disease activity or exacerbation of any residual damage

## Table 2

Strengths and Weaknesses of Five Proposed Endpoints in Chronic GVHD Therapy Trials

| Proposed Endpoint | Definition | Statistical Considerations | Strengths | Weaknesses |
|---|---|---|---|---|
| GVHD Response | Complete plus partial response based on clinician-reported measures | • Comparison of proportions with treatment response at a specific time point | • Direct measure of success<br>• Lengthy follow-up not needed<br>• Easily applied | • Scales not fully qualified |
| Failure free survival | Survival for a defined period without new systemic treatment, death or recurrent malignancy | • Time-to-event, or<br>• Comparison of proportions with failure-free survival at a specific time point | • Benchmarks available for 1st and 2nd-line treatment<br>• Correlates with overall improvement reported by providers and patients<br>• Correlates with ability to discontinue systemic treatment | • Indirect measure of failure<br>• Improvement is not measured (i.e., GVHD manifestations may persist)<br>• New treatment decisions are subject to bias and inconsistency<br>• Not accepted for regulatory approvals |
| Survival without progressive impairment | Survival without an enduring chronic GVHD-related effect that threatens or compromises physical well-being or function in ways that cannot be easily reversed | • Time-to-event, or<br>• Comparison of proportions surviving without progressive impairment at a specific time point | • Failure directly measured<br>• Correlates with overall improvement reported by providers and patients | • Improvement is not measured (i.e., GVHD manifestations may persist)<br>• Impairment is not yet fully defined<br>• Some impairment measures might not be entirely specific for chronic GVHD<br>• Impairment can be caused by adverse events |
| Patient-reported outcomes | Self-reported patient information on symptoms and multi-dimensional quality of life | • Comparison of proportions with clinically meaningful improvement at a specific time point<br>• Comparison of distributions between study arms | • Captures the patient perspective<br>• Lengthy follow-up not needed<br>• Easily applied | • Subject to respondent biases<br>• Missing data difficult to control<br>• Claims limited to PROs |
| Composite scale | Selected measures from provider and patient | • Comparison of proportions with clinically meaningful improvement at a specific time point<br>• Comparison of distributions between study arms | • Aggregates data from multiple perspectives | • Scale not developed or qualified |

**Table 3**

Characteristics and Benefits Demonstrated to be Associated with Proposed Endpoints

| Proposed endpoint | Characteristics | | Associated Benefits | | | |
|---|---|---|---|---|---|---|
| | Objective, Reliable, Verifiable | Simplicity | Decreased Symptom Burden | Better Function | Shorter Time to Durable GVHD Resolution and Withdrawal of Systemic Treatment | Improved Overall Survival |
| Response | Yes | No | Yes | Yes | No | Inconsistent |
| Failure-free survival | No† | Yes | Mixed | NT | Yes | Yes |
| Survival without progressive impairment | Yes | No | Yes | Yes | NT | NT |
| Patient reported outcomes | Yes | Yes | Yes | Yes | NT | Some |
| Composite scale | Yes | No | Yes | Yes | NT | NT |

* NT: not tested

† Decisions to change treatment are not adequately objective or reliable.

## Table 4

Clinician-reported Measures as Potential Indicators of Benefit in Clinical Trials

| Reference | Clinician - Reported Measures | Gold Standard | Study Design Comments | Results |
|---|---|---|---|---|
| Mitchell [66] | Full 2005 NIH spectrum of measures – by transplant clinicians | Subspecialty experts | N=25 children and adults with chronic GVHD (4 consecutive pilot trials) | Supports feasibility of the NIH measures.<br>Inter-rater agreement for skin and oral was satisfactory except for moveable sclerosis and moderate to substantial for functional capacity, GI and global rating measures. |
| Jacobsohn [67] | NIH skin score | Clinician and patient perception of skin improvement or worsening, Overall survival | N=458 prospective multicenter longitudinal observational cohort study | The 0-3 NIH composite skin score correlated with both clinician and patient perception of improvement or worsening. Worsening skin score at 6 months was associated with worse survival. |
| Inamoto [37] | NIH eye score | Clinician and patient perception of eye symptom change | N=387 prospective multicenter longitudinal observational cohort study | Among all scales, changes in the NIH eye scores showed the greatest sensitivity to symptom change reported by clinicians or patients. Schirmer's test did not correlate. |
| Treister [39] | NIH oral score and modified OMRS (0-15) | Patient and clinician-reported change in oral chronic GVHD | N=458 prospective multicenter longitudinal observational cohort study | The clinician-reported measurement changes most predictive of perceived change by clinicians and patients were erythema, extent of lichenoid changes, and NIH severity score. |
| Palmer [68] | NIH lung score symptom scale | Non-relapse mortality (NRM), Overall survival (OS), Patient-reported lung symptoms | N=496 prospective multicenter longitudinal observational cohort study | The NIH symptom-based lung score was associated with NRM, OS, patient-reported symptoms, and functional status. Worsening of NIH symptom-based lung score over time was associated with higher NRM and lower survival. |
| Inamoto [69] | NIH joint-fascia score, Hopkins scale, Photographic (P-ROM) | Clinician and patient perception of change | N=567 prospective multicenter longitudinal observational cohort study | Changes in the NIH scale correlated with both clinician- and patient-perceived improvement. Changes in all 3 scales correlated with clinician- and patient-perceived worsening, but the P-ROM scale was the most sensitive. |
| Bassim [70] | NIH modified OMRS (0-15) | Established measures of oral pain, oral function, oral related QOL, nutrition and laboratory parameters. | N=198 prospective cross-sectional observational cohort study (moderate-to-severe chronic GVHD) | This study supports the use of the OMRS and its components (erythema, lichenoid and ulcerations) to measure clinician-reported severity of oral chronic GVHD.<br>No associations were found between mucoceles and any indicator evaluated. |
| Curtis [71] | 18 clinician-reported (‘Form A’) measures | Concurrent parameters: NIH global score, chronic GVHD activity, Lee symptom score and SF36 PCS | N=193 prospective cross-sectional observational cohort study (moderate-to-severe chronic GVHD) | 4-point and 11-point clinician reported global symptom severity scores are associated with the majority of concurrent outcomes. Skin erythema is a potentially reversible sign of chronic GVHD that is associated with survival. |

| Reference | Clinician - Reported Measures | Gold Standard | Study Design Comments | Results |
|---|---|---|---|---|
| Yanik [72] | Response was defined as 10% FEV1 or FVC improvement | 5-year survival | N=34 patients with subacute pulmonary dysfunction (25 obstructive) received etanercept therapy | 5-year survival 90% (95% CI, 73%-100%) for 10 patients who responded to therapy, compared with 55% (95% CI, 37%-83%) for the 21 patients who did not meet response criteria (P = 0.07) |
| Olivieri [73] | NIH criteria, NIH organ score, Couriel criteria | Overall survival | N=40, phase 2 prospective study of imatinib for steroid-refractory chronic GVHD | The 3-year OS was 94% for patients responding at 6 months and 58% for non-responders according to NIH response criteria (P = 0.007) |
| BMT CTN 0801 (unpublished) | NIH criteria | Clinician assessed overall CR+PR | N=151, randomized phase 2 multicenter trial | AUC for organs (lichenoid mouth, joint score) plus clinician assessed 0-10 global rating scale = 0.79 |

## Table 5

Patient-reported Outcome Sensitivity to Change

| Reference | Patient-Reported Measure | Gold Standard | Study Design Comments | Results |
|---|---|---|---|---|
| Global | | | | |
| Pidala [48] | SF-36, FACT-BMT | Change in global severity, clinician-reported, patient-reported change | N=336, correlation of change scores with response measures in an observational study | Patient-reported severity change was associated with all QOL measures. Change in NIH and clinician-reported chronic GVHD severity did not correlate well with patient-reported QOL changes. |
| Inamoto [29] | SF-36, FACT-BMT, Lee symptom scale | NIH-calculated overall response | N=258, correlation of change scores with NIH-calculated overall response in an observational study | NIH calculated overall responses were associated with patient-reported symptoms in patients enrolled within 3 months of chronic GVHD onset but not in patients enrolled more than 3 months after onset. SF-36 and FACT-BMT changes were not associated with NIH-calculated responses regardless of time since onset. |
| Walker [74] | SF-36, FACT-BMT, Lee symptom scale | N/A | N=203, randomized, unblinded study of thymoglobulin vs. no thymoglobulin, comparing PROs between randomized groups | The study met its primary endpoint: freedom from immunosuppressive treatment at 12 months (37.4% vs. 16.5%, p=0.001). GVHD symptoms were lower in patients randomized to Thymoglobulin (14.95 vs. 20.93, p=0.017). The difference was also clinically meaningful, defined via the distribution method as 0.5 SD. |
| Organ-specific | | | | |
| Inamoto [37] | 0-10 eye symptom, Lee eye symptom score, ocular surface disease index (OSDI) | Patient and clinician-reported change in eye chronic GVHD (8-point scale) | N=387, correlation of PRO change scores with reported response in an observational study | Change in the Lee eye symptom score, 0-10 eye symptom, and OSDI correlated with patient- and clinician-reported change |
| Jacobsohn [67] | Lee skin symptom score | Non-relapse mortality, overall survival, patient-and clinician-reported change (8 point scale) | N=458, correlation with outcomes and reported change in an observational study | Change in the Lee skin symptom score correlated with patient and clinician-perceived changes. Improvement in the Lee skin symptoms score at 6 months |

| Reference | Patient-Reported Measure | Gold Standard | Study Design Comments | Results |
|---|---|---|---|---|
| | | | | was associated with lower NRM and better OS |
| Treister [39] | Lee mouth and nutrition symptom scores, patient mouth sensitivity, pain, dryness 0-10 | Patient- and clinician-reported change in oral chronic GVHD (8 point scale) | N=458, correlation with reported change in an observational study | In multivariate modeling, change in patient-reported Lee mouth symptom score was associated with patient- and clinician-reported change |
| Inamoto [69] | Lee muscle/joint symptom score, global GVHD severity 0-10, SF-36, FACT-BMT | Patient- and clinician-reported change in joint chronic GVHD (8 point scale) | N=567, correlation with reported change in an observational study | Change in the Lee muscle/joint symptom score, overall symptom score and 0-10 global score correlated with patient-reported improvement and worsening of joint GVHD and clinician-reported worsening of joint GVHD. SF-36 PCS correlated with patient- and clinician-reported improvement in joint GVHD; FACT-G correlated with patient- and clinician-reported worsening in joint GVHD; |
| Inamoto [29] | Lee symptom scale, mouth, eye, skin 0-10 symptoms | NIH-calculated organ-specific change | N=258, correlation with NIH-calculated organ changes in an observational study | NIH calculated organ responses were associated with patient-reported symptom change in skin, eye, mouth and GI (nutrition). |

SF-36, Medical Outcomes Study Short Form-36; FACT-BMT, Functional Assessment of Cancer Therapy, Bone Marrow Transplantation subscale; NRM, non-relapse mortality; OS, overall survival