CHRONIC GRAFT-VERSUS-HOST DISEASE

Classification systems for chronic graft-versus-host disease

Stephanie J. Lee
Fred Hutchinson Cancer Research Center, Seattle, WA

Chronic graft versus host disease (GVHD) is a major cause of morbidity and mortality after allogeneic hematopoietic cell transplantation. Clinically, chronic GVHD is a pleiotropic, multiorgan syndrome involving tissue inflammation and fibrosis that often results in permanent organ dysfunction. Chronic GVHD is fundamentally caused by replacement of the host’s immune system with donor cells, although the heterogeneity of clinical manifestations suggests that patient, donor, and transplant factors modulate the phenotype. The diagnosis of chronic GVHD and determination of treatment response largely rely on clinical examination and patient interview. The 2005 and 2014 National Institutes of Health Consensus Development Projects on Criteria for Clinical Trials in Chronic GVHD standardized the terminology around chronic GVHD classification systems to ensure that a common language and procedures are being used in clinical research. This review provides a summary of these recommendations and illustrates how they are being used in clinical research and the potential for their use in clinical care. (Blood. 2017;129(1):30-37)

Introduction

When the hematopoietic system of a nongenetically identical donor is transplanted into a recipient, the resulting inflammation and immune dysregulation can lead to chronic graft-versus-host disease (GVHD). Chronic GVHD is the most common long-term complication after allogeneic hematopoietic cell transplantation (HCT), and it decreases the success of transplantation by increasing the risk of death and disability. Chronic GVHD is the leading cause of nonrelapse mortality in transplant survivors otherwise cured of their diseases, and its adverse effects include physical, functional, and psychosocial deficits, inability to return to work, and poor quality of life (QOL). It is tragic that in the course of trying to cure 1 life-threatening disease, we often cause collateral suffering and death due to a common iatrogenic complication.

Most cases of chronic GVHD are diagnosed within the first year after HCT, but 5% to 10% of affected patients do not develop signs and symptoms until later. Approximately 30% of chronic GVHD is de novo without any preceding acute GVHD. At onset, many patients have an inflammatory skin rash, oral sensitivities or dryness, or dry, irritated eyes. Transaminase elevations and eosinophilia are common. These early manifestations are relatively easy to control with standard corticosteroid-based immunosuppression but often recur with the same or new manifestations when immunosuppression is tapered. Other manifestations that are less common but much more difficult to control include skin sclerosis or fasciitis, bronchiolitis obliterans syndrome, oral ulcers unresponsive to local therapies, severe dry eyes, serositis, and gastrointestinal (GI) involvement. These manifestations respond poorly to standard immunosuppressive therapies, cause significant organ dysfunction, and are often persistent or permanent.

Because there are no accepted diagnostic biomarkers, and pathologic samples may be difficult to obtain, most chronic GVHD evaluations are based on clinical examinations and patient interviews. The previous lack of objective criteria and reliance on clinician reporting meant that clinical research was dependent on the unstructured reports of individual clinicians who varied greatly in their experience and attention to chronic GVHD manifestations. In an effort to standardize reporting, the 2005 and 2014 National Institutes of Health (NIH) Consensus Development Projects on Criteria for Clinical Trials in Chronic GVHD (henceforth called the “NIH Consensus Conferences” for simplicity) made recommendations for diagnostic criteria, severity scoring, and response assessments to be used in clinical trials. These recommendations have largely been adopted for therapeutic clinical trials by cooperative groups and industry. However, they are less commonly reported in observational and retrospective studies because the measurement tools have not been routinely implemented in clinical care.

This review will focus on the classification systems of chronic GVHD, which include the diagnostic criteria, assessments of organ involvement, and methods to document improvement or worsening during treatment. Throughout, I will try to provide historical perspective for these concepts in the context of therapeutic studies and clinical care. Finally, I will summarize my perspective on the persistent gaps in research and practice, and review ongoing efforts to address these issues.

Historical perspective

Chronic GVHD was first described in 1978 as a wasting syndrome observed in some long-term survivors of allogeneic HCT. Affected patients had severe sclerosis with joint contractures, lung involvement, weight loss, dry eyes, and other organ manifestations reminiscent of autoimmune diseases. Initially, patients did not receive any treatment other than supportive care when they first developed symptoms. In 1981, Sullivan et al reported that treatment with corticosteroid-based therapy controlled symptoms and improved survival. A series of randomized trials over the next 3 decades tried to improve initial treatment of chronic GVHD. In summary, nothing has proven to be
superior to single-agent prednisone for initial treatment of chronic GVHD, and no secondary treatment has proven superior to others once the disease progresses or recurs when steroids are tapered.1,19

Currently, between 10% and 70% of patients develop chronic GVHD depending upon donor and transplant characteristics: multi-center and registry statistics show an aggregate cumulative incidence of 30% to 50%.14,20 Use of bone marrow instead of peripheral blood,21-24 specific acute GVHD prophylaxis regimens (eg, broad or selective T-cell depletion,25-28 posttransplant cyclophosphamide29-31), and donor types (eg, umbilical cord blood20,32-34) appear to decrease the incidence, severity, and treatment refractoriness of chronic GVHD. Although data are limited, chronic GVHD manifestations, severity, and outcomes are not different depending on conditioning regimen intensity. Nonrelapse mortality of patients with chronic GVHD has improved over the decades, likely because of better supportive care.14,35

There are no US Food and Drug Administration (FDA)-approved treatments for chronic GVHD, although ongoing trials are testing several agents for this indication.

**2005 and 2014 NIH Consensus Conferences**

In 2004, the NIH convened experts to identify barriers to progress in chronic GVHD research. The following year, Steven Pavletic and Georgia Vogelsang chaired the 2005 NIH Consensus Conference at which findings and recommendations were presented. Six working groups published their papers in 2005 to 2006, summarizing consensus recommendations for diagnosis and scoring,36 histopathology,37 biomarkers,38 response criteria,39 ancillary and supportive care,40 and clinical trials.41

Between 2005 and 2014, many studies were published using the criteria recommended by the 2005 NIH Consensus Conference. A survey of experts in 2013 confirmed agreement with many of the 2005 recommendations but also identified several areas of controversy based on experience with the criteria in practice. For example, experts thought that active disease and irreversible “fixed” deficits should be distinguished, and organ dysfunction entirely attributed to a nonchronic GVHD etiology should be excluded from severity and response scoring.42 To address these controversies, as well as review progress and update recommendations, a second NIH Chronic GVHD Consensus Conference took place in 2014.43 The 250 participants reaffirmed most of the previous 2005 recommendations, and a new series of 6 papers was published.44,45 The response criteria had the most changes because of empiric data published in the 10 years since the original conference.46

**NIH late acute and overlap chronic GVHD**

Prior to 2005, any alloimmunity that resulted in clinical manifestations before day 100 was called acute GVHD, whereas any clinical alloimmunity after day 100 was considered chronic GVHD.50 The 2005 NIH Consensus Conference abolished the day 100 dividing line and redefined acute and chronic GVHD as distinct clinical syndromes without a time restriction (Figure 1). Classic acute GVHD occurs before day 100 and is staged according to the percentage of body surface area with rash, total bilirubin elevation, and volume of diarrhea. Late acute GVHD occurs after day 100 and is defined as signs and symptoms of acute GVHD without chronic GVHD. Late acute GVHD is further subdivided into “persistent” if it is a continuation of classic acute GVHD, “recurrent” if classic acute GVHD resolves then recurs after day 100, or “de novo” if initial onset is after day 100 without any prior acute GVHD. Some studies show a higher mortality for patients with late acute GVHD compared with patients with classic chronic GVHD50,51 whereas others do not.52,53 Recognition of the late acute GVHD category decreased the incidence of chronic GVHD because many patients with alloimmunity after day 100 do not meet with diagnostic criteria for chronic GVHD based on organ involvement but were previously considered chronic GVHD based solely on time since transplant.54 Although the 2014 NIH Consensus Conference recommended that patients with late acute could be included in chronic GVHD trials with appropriate stratification,45 in practice they are usually excluded.

The 2005 NIH Consensus Conference also recommended a new category called “overlap chronic GVHD” when concurrent acute and chronic GVHD are present, because of the perception that once patients are diagnosed with chronic GVHD, ongoing acute GVHD portends a worse prognosis. Although controversial because of the different interpretations of how nausea, anorexia, diarrhea, erythematous or maculopapular rashes, and elevated liver function tests should be attributed to acute or chronic GVHD, and the clinical observation that most patients have some elements of acute GVHD at some point in their chronic GVHD experience, “overlap chronic GVHD” was retained in the 2014 update.43 Some studies show worse survival with overlap chronic GVHD compared with classic chronic GVHD,55 but others do not. Overlap chronic GVHD is allowed in chronic GVHD clinical trials, although it is recommended that detailed information about concurrent acute GVHD be captured to allow stratification during the analysis.

**NIH diagnostic criteria**

The NIH diagnostic criteria for chronic GVHD changed little between 2005 and 2014. Per the 2014 NIH criteria, the diagnosis of chronic GVHD requires (a) at least 1 diagnostic manifestation or (b) 1 distinctive manifestation confirmed by biopsy or testing of the same or other involved organ.45 “Diagnostic” manifestations sufficient by themselves to establish the diagnosis of chronic GVHD may be found in the skin, mouth, GI tract, lung, fascia, and genitalia (for example, lichen planus or lichen sclerosis, poikiloderma, sclerosis, or esophageal webs) (Figure 2). There are no diagnostic features of the nails, eyes, liver, or other organs. If the lung is the only site of chronic GVHD without a distinctive manifestation elsewhere, then a
lung biopsy showing bronchiolitis obliterans syndrome is required to establish the diagnosis of chronic GVHD for the purpose of clinical trials enrollment. "Distinctive" criteria are clinically suspicious for chronic GVHD but are not sufficient by themselves to establish the diagnosis because other etiologies could account for the signs, so a confirmatory test is required. Examples of distinctive features are papulosquamous lesions, oral ulcers, onycholysis, or dry gritty eyes. Examples of confirmatory tests are tissue biopsies (eg, skin, mouth, lung, liver, GI, genital), organ-specific testing (eg, pulmonary function tests, Schirmer tests), imaging (eg, a barium swallow showing an esophageal ring), or evaluation by a specialist such as an ophthalmologist or gynecologist confirming GVHD. Biopsy evidence showing “likely” GVHD is sufficient; the histopathology does not need to be definitive as long as other potential etiologies are not present. Because the differential diagnosis for liver test abnormalities and GI symptoms in the posttransplant setting is broad, histopathologic confirmation may be especially important as a prior study reported incorrect attribution to GVHD in many cases.

It is important to realize that the NIH diagnostic criteria were devised for clinical trials to ensure that study participants had unequivocal chronic GVHD. Many patients with signs and symptoms encountered in practice will not meet the NIH diagnostic criteria for chronic GVHD but nevertheless have active alloimmunity requiring systemic immunosuppression to improve symptoms and prevent ongoing organ damage. Clinical studies that require participants to meet the NIH diagnostic criteria should collect data to confirm eligibility because a multicenter cooperative group study found that up to 10% of enrolled patients actually did not meet NIH diagnostic criteria for chronic GVHD (Paul Carpenter, Fred Hutchinson Cancer Research Center, oral communication, October 2015).

Until 2016, the Center for International Blood and Marrow Transplant Research (CIBMTR) used the older case report forms that did not reflect NIH recommendations out of concern that trying to collect observational data from medical records according to the new NIH consensus criteria would be impossible given documentation standards. The CIBMTR has now updated their 2016 case report forms to capture chronic GVHD according to NIH recommendations, but it remains to be seen whether data quality will be sufficient for analysis.

### NIH severity scoring

Signs and symptoms of chronic GVHD vary between individuals and in the same individual over time, making determination of GVHD severity challenging. The most frequently involved organs in patients with chronic GVHD are skin, mouth, and liver, with less frequent involvement of eye, lung, GI tract, joint/fascia, and genital tract. Organs are scored on a 0 to 3 scale from no involvement/no symptoms to severe functional compromise. Most organs have a single scale to capture severity; however, the skin and lung have multiple components that contribute to maximum severity. For the skin, the highest score from items about body surface area involvement and severity of sclerosis is used to determine global severity (Figure 3). For the lung, pulmonary function test results are used if available, otherwise the lung score is based on symptoms. Higher organ scores for the skin, lung, GI tract, and liver have been associated with worse survival.

Fixed deficits vs active inflammatory or fibrotic disease may be hard to distinguish clinically, and most are not differentiated in severity scoring. For example, dry eyes from destruction of lacrimal glands and joint contractures from sclerosis are often permanent but are still included when scoring their respective organs. In contrast, residual hyperpigmentation, hypopigmentation, and poikiloderma from prior inflammatory skin involvement are no longer counted as part of body surface area involvement.

Patients with chronic GVHD have many concurrent medical issues that may trigger a score on the chronic GVHD assessment form yet be unrelated to chronic GVHD, for example, skin rashes due to drug toxicity or infection, skin erythema from sun exposure, infectious diarrhea, or poor pulmonary function tests that predated transplantation. All dysfunction should be scored on the chronic GVHD assessment sheet with other etiologies noted as appropriate. New with the 2014...
criteria, if an abnormality is entirely due to a nonchronic GVHD cause then the organ is excluded from the calculation of global severity. If chronic GVHD at least partly explains the organ dysfunction, then the score is used in global severity calculation without modification. This compromise was reached because it is impossible to parse out the chronic GVHD component vs other etiologies when there are multiple causes. By noting whether nonchronic GVHD causes contribute to organ dysfunction, investigators can analyze the data depending on the objective of their study. A scientist interested in biomarkers might want to exclude all cases with non-GVHD-contributing causes to obtain a more homogeneous affected population; another scientist might want to include patients with skin rashes from non-GVHD causes as controls to compare with rashes caused by GVHD. One study reported that 78.3% of abnormalities were attributed wholly to chronic GVHD whereas 14.4% were attributable to other causes, especially in the lung, GI tract, and skin; 7.3% were attributed to both chronic GVHD and other causes.62 Exclusion of abnormalities entirely due to other causes decreased global severity by 1 or more categories in 7% of patients.

Global severity scoring is divided into mild, moderate, or severe based on the number and severity of involved organs (Figure 4). Mild disease is 2 or fewer organs with no more than score 1 and no lung involvement. For patients with mild disease, treatment with topical or local therapies may be sufficient, although systemic therapy is often given for patients presenting with high-risk features. Moderate disease is 3 or more organs with score 1, any organ with score 2, or lung with score 1, and usually requires systemic immune-suppressive treatment. Severe disease is any organ with a score of 3 or lung with a score of 2, and means that substantial organ damage already exists. In 1 multicenter prospective study, severity at onset was 19% mild, 53% moderate, and 28% severe.26 Studies show that mild disease is associated with a good prognosis whereas severe disease is associated with higher treatment-related mortality and lower survival.63

Different ethnic groups may have unique chronic GVHD natural histories. For example, studies of ethnic Japanese patients show they differ from Western populations with less severe chronic GVHD and a lower death rate.64 Moderate-severe disease and overlap chronic GVHD also have less prognostic significance for the Japanese population, perhaps because of better overall outcomes.65

### NIH response criteria

#### Individual organs

Defining response criteria that are reliable and sensitive to clinically meaningful changes in chronic GVHD activity has proven challenging. Historically, investigators relied on clinical impressions to determine improvement or worsening. The original 2005 criteria for organ response offered more objective categories but were still based on expert opinion. Patients were scored for 8 organs before introduction of a treatment, and then at calendar-driven time points later. The organs considered in response assessment included skin, mouth, liver, upper GI, lower GI, esophagus, lung, and eye. Genital tract and joint/fascia were not included due to lack of validated measures.

The 2014 Consensus recommendations simplified data collection and scoring based on results of studies performed between the 2 conferences.46 The 2014 criteria eliminated the need for precise body surface area reporting, the Schirmer test,66 and diffusing capacity of the lung for carbon monoxide,67,68 modifications which decreased the burden on patients and clinicians. Two data elements that are easy to assess during the physical examination were added for joint/fascia,69 so now 9 organs contribute to response assessment. The 2014 modification also revised handling of attribution: if organ dysfunction is entirely explained by a nonchronic GVHD cause, that organ is excluded from calculation of the overall response.

One area of ongoing confusion is why the response criteria and the severity scoring criteria for organs are not identical. The rationale for using different tools is that the goals are different. Measures to document severity are broader and designed to be used in clinic by nonchronic GVHD specialists who lack specific training. Categories are simplified to ensure complete capture of reliable information. The global severity score is multidimensional and designed to document cross-sectional assessments. There are ceiling effects where patients scoring in the highest category do not have room to worsen, and step effects where slight changes can result in placement in different categories. In contrast, the response assessment tools are more detailed and capture granular chronic GVHD disease activity. The ability to detect changes along unidimensional or linear scales is emphasized. Nevertheless, the severity scoring and response measures are more

---

**Figure 3. 2014 NIH skin score.** BSA, body surface area.
complexity adds additional barriers to chronic GVHD treatment trials. Transplant version (FACT-BMT)\(^8\) plus the Human Activity Pro

Functional Assessment of Chronic Illness Therapy Bone Marrow

guidance about the qualifi

Lee Symptom Scale,\(^82\,84\) and either the Short Form-36 (SF-36)\(^8\) or the

exacerbations has been qualifi

To date, only 1 instrument for chronic obstructive pulmonary disease

pursuing labeling claims. However, the requirements are very stringent.

other measures, suggesting that even though interpretation of PRO data

requires consideration of response bias and measurement error, PROs

offer unique and important clinical information.\(^9\)

NIH overall response

Based on individual organ measures, responses are classified as complete response (CR; no manifestations of chronic GVHD, including “fixed” defects), partial response (PR; clinically meaningful improvement in 1 or more organs without clinically meaningful progression in any other organ), and disease progression (clinically meaningful worsening in 1 or more organs regardless of improvement in other organs). Cases not meeting the definition of CR, PR, or disease progression are considered stable disease. Note that once the pretreatment and posttreatment organ measures are known, the overall response is calculated and does not rely on clinician-reported interpretations of response. Different clinicians may perform the pretreatment and posttreatment assessments, although confidence is enhanced if the same clinician performs both evaluations to eliminate interrater variation.

Both CR and PR are considered meaningful short-term responses in clinical trials. The FDA has indicated that objective GVHD response is the most appropriate primary end point in phase 2 and possibly phase 3 trials, paired with PROs as secondary end points.\(^4\) The very long-term significance of CR/PR is unclear because data are mixed about their association with survival\(^3\,39\) or eventual successful discontinuation of immunosuppression.\(^31\) However, if chronic GVHD is viewed as a truly chronic alloimmune syndrome akin to autoimmune diseases, then it might be unrealistic to expect any current treatments will result in CR and ability to stop immunosuppression without some ongoing symptoms; rather, the goals of treatment should perhaps be preservation of function and QOL with the least toxicity, anticipating the potential need for prolonged or lifelong treatment.

Alternate end points besides chronic GVHD response and PROs have been suggested. For example, because addition of a new systemic chronic GVHD treatment is considered treatment failure in clinical trials,\(^4\,6\) some investigators have argued that the best measure of treatment success is not having to change to another treatment, so called “failure-free survival” (FFS).\(^9,93\) Although FFS is intuitively attractive and data are easy to capture, concerns about the lack of standardized practice approaches to managing chronic GVHD, including the thresholds for changing treatments, and external influences such as

| Mild | • 1 or 2 organs or sites (except lung) with score 1  
• Mild oral symptoms, no decrease in oral intake  
• Mild dry eyes, lubricant eyedrops ≤ 3x/day |
|---|---|
| Moderate | • 3 or more organs with score 1  
• At least 1 organ or site with score 2  
• 19-50% body surface area involved or superficial sclerosis  
• Moderate dry eyes, eyelid incisions ≥ 3x/day or punctal plugs  
• Lung score 1 (FEV1 60-79% or dyspnea with stairs) |
| Severe | • At least 1 organ or site with score 3  
• > 50% body surface area involved  
• Deep sclerosis, impaired mobility or ulceration  
• Severe oral symptoms with major limitation in oral intake  
• Severe dry eyes affecting ADL  
• Lung score 2 (FEV1 40-59% or dyspnea walking on flat ground) |
the availability of alternative treatments has prevented FFS from gaining traction as a meaningful end point for FDA consideration.

Some investigators have advocated combining clinician-reported, patient-reported and laboratory testing into a composite disease activity scale, similar to scales developed for autoimmune diseases such as the Crohn Disease Activity Index (CDAI). However, the FDA has expressed skepticism about this type of end point for chronic GVHD, citing the complexity of the different measures and concerns about the attribution of separate contributions.

### Gaps in research and practice

The 2005 and 2014 NIH Consensus Conferences have standardized criteria for clinical trials and removed a major barrier to industry interest in testing new agents for chronic GVHD. The payoff is clear as clinical trials activity in the chronic GVHD space has increased dramatically, including exciting novel agents targeting specific biologic pathways. I believe the next major frontier is to identify patients who are destined to develop hard-to-treat phenotypes, for example, sclerosis/fasciitis, bronchiolitis obliterans syndrome, unresponsive oral ulcers, severe dry eyes, serositis, and GI involvement, and to intervene early and effectively.

Most practitioners view using the NIH chronic GVHD recommendations in their entirety as too burdensome for use in routine clinical practice. Although some aspects such as the response criteria are primarily designed for research use, other classification systems such as severity scoring offer a straightforward systematic approach to assessing organ dysfunction. The PRO instruments measure patient symptoms and QOL using a concise battery. Use of these tools in the clinic might help ensure consistent evaluation of all potentially involved organs, contributing to optimal clinical care, even for patients not participating in clinical trials. Careful screening over time might detect evolving chronic GVHD earlier so that treatment can be started before permanent organ dysfunction develops. These are hypotheses to be tested.

It can be very challenging to conduct clinical trials in chronic GVHD. The population is small, scattered, and heterogeneous. Their clinical manifestations are a mixture of reversible and fixed deficits. Patients receive a variety of potent immunosuppressive agents over the years, and often have a background of chronic illness that leads to frequent infections and disability. However, the unmet need for effective therapies is very great, and trying to enroll patients into clinical trials if possible will help ensure promising approaches can be evaluated expeditiously.

### Summary

Chronic GVHD remains a formidable barrier to successful allogeneic transplantation. Efforts to prevent or ameliorate its clinical significance have been more successful in the last 5 years, primarily though alterations in graft sources and acute GVHD prophylaxis. Lingering concerns about whether the graft-versus-malignancy effect is compromised if GVHD is prevented await more randomized trials. Targeted immunotherapy that does not risk chronic GVHD is another potential solution to the chronic GVHD problem for some patients. Chimeric antigen receptor T cells or other narrow spectrum cellular populations and other targeted immunologic approaches will have lower or negligible risks of alloimmune side effects. Approaches that separate GVHD from the immunologic benefits of allogeneic cells are likely where the ultimate solution to chronic GVHD will come from: preventing it in the first place. There is tangible progress in this direction.

The last decade has seen a dramatic increase in the interest and attention given to chronic GVHD. Among HCT practitioners, there seems to be a greater reluctance to accept chronic GVHD as an inevitable long-term complication of allogeneic transplantation. Many novel agents are being tested, with clinical trials built on the NIH Consensus recommendations. Active collaborations between clinical and laboratory scientists are identifying promising biomarkers. I am optimistic that between better prevention and better treatment, future generations of transplant survivors will suffer less from the devastation of chronic GVHD.

### Acknowledgments

The author gives special thanks to Paul Martin, Mary Flowers, Yoshihiro Inamoto, Joseph Pidala, and 3 anonymous reviewers for helpful comments.

This work was supported by grants CA163438 and CA118953 from the National Institutes of Health National Cancer Institute. The Chronic GVHD Consortium (U54 CA163438) is part of the National Center for Advancing Translational Sciences (NCATS) Rare Diseases Clinical Research Network (RDCRN). RDCRN is an initiative of the Office of Rare Disease Research (ORDR), NCATS, funded through collaboration between NCATS and the National Cancer Institute.

### Authorship

Contribution: S.J.L. wrote the paper.

Conflict-of-interest disclosure: The author declares no competing financial interests.

ORCID iDs: S.J.L., 0000-0003-2600-6390.

Correspondence: Stephanie J. Lee, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, D5-290, Seattle, WA 98109; e-mail: sjlee@fredhutch.org.

### References


6. Goldman JM, Majhail NS, Klein JP, et al. Relapse and late mortality in 5-year survivors of


