ASBMT
American Society for Blood and Marrow Transplantation

Poor Agreement between Clinician Response Ratings and Calculated Response Measures in Patients with Chronic Graft-versus-Host Disease

Jeanne M. Palmer,[1] Stephanie J. Lee,[2] Xiaoyu Chai,[2] Barry E. Storer,[2] Mary E. D. Flowers,[2] Kirk R. Schultz,[3] Yoshihiro Inamoto,[2] Corey Cutler,[4] Joseph Pidala,[5] Mukta Arora,[6] David A. Jacobsohn,[7] Paul A. Carpenter,[2] Steven Z. Pavletic,[8] Paul J. Martin[2]

In 2005, a National Institutes of Health consensus conference was held to refine methods for research in patients with chronic graft-versus-host disease, including proposed objective response measures and a provisional algorithm for calculating organ-specific and overall response. In this study, we used weighted kappa statistics to evaluate the level of agreement between clinician response ratings and calculated response categories in patients with chronic graft-versus-host disease. The study included 290 patients who had paired enrollment and follow-up visits. Based on a set of objective measures, 37% of the patients had an overall complete or partial response, whereas clinicians reported an overall complete or partial response rate of 71% (slight to fair agreement, weighted kappa 0.20). Agreement rates between calculated organ-specific responses and clinician-reported changes in skin, mouth, and eyes were fair to moderate (weighted kappa, 0.28-0.54). We conclude that for both overall and organ-specific comparisons, clinician response ratings did not agree well with calculated response categories. Possible reasons for this discrepancy include a high clinical sensitivity for detecting response, a clinical predisposition to recognize selective improvements as overall response, the large change in objective measures proposed to define response, and the high incidence of progressive disease based on new manifestations. Conclusions from prior literature reporting high overall response rates based on clinician judgment would not be supported if the provisional algorithm had been applied to calculate response. Our analysis also highlights the need to define an overall response measure that incorporates both patient-reported and objective measures and accurately reflects the outcome in patients with a mixed response in which one organ or site improves, whereas another shows new involvement.

Biol Blood Marrow Transplant 18: 1649-1655 (2012)

KEY WORDS: Allogeneic hematopoietic cell transplantation, Chronic graft-versus-host disease, Response assessment

INTRODUCTION

Chronic graft-versus-host disease (cGVHD) affects 40% to 70% of patients after allogeneic hematopoietic cell transplantation (HCT) and is associated with significant morbidity and mortality [1]. Historically, cGVHD was defined as any GVHD occurring more than 100 days after HCT [2]. In 2005, the National Institutes of Health (NIH) formed a consensus committee to clarify the diagnosis of cGVHD [3]. Rather than being defined as a function of time after HCT, it was proposed that cGVHD should be defined according to diagnostic signs such as lichen planus-like lesions, sclerosis, or fasciitis, or by distinctive manifestations confirmed by biopsy or laboratory testing [3]. The NIH criteria define mild, moderate, and severe cGVHD according to scores for signs and symptoms involving skin, fascia and joints, eyes, mouth,

From the [1]Medical College of Wisconsin, Milwaukee, Wisconsin; [2]Fred Hutchinson Cancer Research Center, Seattle, Washington; [3]University of British Columbia, Vancouver, British Columbia, Canada; [4]Dana Farber Cancer Institute, Boston, Massachusetts; [5]Moffitt Cancer Center, Tampa, Florida; [6]University of Minnesota, Minneapolis, Minnesota; [7]Children's National Medical Center, Washington, DC; and [8]National Cancer Institute, Bethesda, Maryland.

Financial disclosure: See Acknowledgments on page 1654.

Correspondence and reprint requests: Jeanne M. Palmer, MD, Division of Hematology and Oncology, Medical College of Wisconsin, 9200 W Wisconsin Ave, Milwaukee, WI 53226 (e-mail: sjlee@fhcrc.org).

Received March 23, 2012; accepted May 9, 2012

1083-8791/$36.00
http://dx.doi.org/10.1016/j.bbmt.2012.05.005

gastrointestinal tract, liver, genital tract, and lungs [3]. These definitions are predictive of overall survival and quality of life [4,5]. No widely accepted gold standard is currently available for determining activity of cGVHD or the response to treatment. Many clinical trials have evaluated treatment of cGVHD, but most relied on the overall judgment of the clinician in reporting response, and response criteria were not precisely defined [6-12]. Only a single randomized study relied on an entirely objective primary endpoint for measuring treatment success [13]. Generally, complete resolution of all signs or symptoms has been classified as complete response (CR), and any significant reduction in signs or symptoms without resolution of all manifestations has been classified as partial response (PR). Progressive disease (PD) has been defined as significant worsening of symptoms or development of new organ involvement, while stable disease has been defined as the absence of significant improvement or worsening.

The 2005 NIH consensus conference addressed the complex considerations involved in assessing response to treatment in patients with cGVHD. In its report [14], the Response Criteria Working Group recommended that the measures used to assess response should be practical for use both by transplantation and nontransplantation medical providers, adaptable for use in adults and in children, and focused on the most important cGVHD manifestations. The measures should also give preference to quantitative, rather than semiquantitative measures, capture information regarding signs, symptoms, and function separately from each other, and use validated scales whenever possible to demonstrate improved patient outcomes to meet requirements for regulatory approval of novel agents. The Working Group proposed a set of objective measures to be considered for use in clinical trials, and forms for data collection were developed. Provisional algorithms for calculating CR, PR, and PD were proposed for each organ and for overall response, based on the objective response measures [14]. The provisional algorithms to calculate response categories were based on expert consensus opinion and were intended to improve consistency in the conduct and reporting of cGVHD trials. These definitions were not intended to be implemented without validation, and infact are currently being assessed in a large multicenter clinical study (BMT CTN 0801: NCT01106833) [15].

In the current analysis, we used data from an observational study to compare response assessed by clinicians versus the calculated response categories. Understanding this relationship will be helpful for evaluating current therapies when results are analyzed according to the provisional response categories in comparison to previously studied therapies in which response was determined by the clinician's overall judgment.

METHODS

cGVHD Consortium: Description of Study Cohort and Cohort for This Analysis

A cohort of HCT recipients affected by cGVHD was prospectively assembled in a multicenter observational study [15]. The protocol was approved by the institutional review board at each site, and all patients provided written informed consent. Patients enrolled in the cohort were allogeneic HCT recipients at least 2 years of age with cGVHD requiring systemic immunosuppressive therapy, including both those with classic cGVHD and those with overlap syndrome. Cases were classified as incident (enrollment less than 3 months after cGVHD diagnosis) or prevalent (enrollment 3 or more months but less than 3 years after transplantation). Primary disease relapse, inability to comply with study procedures, and anticipated survival of less than 6 months were exclusion criteria. At enrollment and every 6 months thereafter, clinicians and patients reported standardized information summarizing cGVHD organ involvement and symptoms after a clinic visit. Patients were allowed to take the self-administered survey home and return it by mail. At each time point, the same clinician reported all the clinician-reported items. Incident cases had an additional assessment time point at 3 months after enrollment. Objective medical data including ancillary testing and laboratory results, medical complications, and medication profiles were abstracted through standardized medical chart review after each visit. Clinicians were not given the calculated NIH overall or organ-specific response at any time during this study, nor were they provided with previously completed study forms.

Measures

Several response measures were used in the comparisons. (1) Objective response measures and calculated overall and organ-specific responses. The objective measures used in this study were derived from recommendations suggested in part IV of the NIH Consensus Development Project on Criteria for Clinical Trials in cGVHD [14]. For example, the skin assessment was based on the extent of involved body surface, the eye assessment was based on the Schirmer test, the oral assessment was based on the 15-point Schubert score, and the liver assessment was based on laboratory values relative to the upper limit of normal. Evaluation of the gastrointestinal (GI) tract used the NIH response scales separately for the esophagus, upper GI tract, and lower GI tract. Provisional overall response categories were assigned based on change from enrollment to follow-up, with complete resolution of organ dysfunction considered a CR. The definition of PR was based on the principle