Arthritis & Rheumatism (Arthritis Care & Research)
Vol. 59, No. 6, June 15, 2008, pp 884–891
DOI 10.1002/art.23707

ORIGINAL ARTICLE

Assessment of the Item Selection and Weighting in the Birmingham Vasculitis Activity Score for Wegener's Granulomatosis

ALFRED D. MAHR,[1] TUHINA NEOGI,[1] MICHAEL P. LAVALLEY,[1] JOHN C. DAVIS,[2] GARY S. HOFFMAN,[3] W. JOSEPH MCCUNE,[4] ULRICH SPECKS,[5] ROBERT F. SPIERA,[6] E. WILLIAM ST.CLAIR,[7] JOHN H. STONE,[8] AND PETER A. MERKEL,[1] FOR THE WEGENER'S GRANULOMATOSIS ETANERCEPT TRIAL RESEARCH GROUP

Objective. **To assess the Birmingham Vasculitis Activity Score for Wegener's Granulomatosis (BVAS/WG) with respect to its selection and weighting of items.**

Methods. **This study used the BVAS/WG data from the Wegener's Granulomatosis Etanercept Trial. The scoring frequencies of the 34 predefined items and any "other" items added by clinicians were calculated. Using linear regression with generalized estimating equations in which the physician global assessment (PGA) of disease activity was the dependent variable, we computed weights for all predefined items. We also created variables for clinical manifestations frequently added as other items, and computed weights for these as well. We searched for the model that included the items and their generated weights yielding an activity score with the highest R^2 to predict the PGA.**

Results. **We analyzed 2,044 BVAS/WG assessments from 180 patients; 734 assessments were scored during active disease. The highest R^2 with the PGA was obtained by scoring WG activity based on the following items: the 25 predefined items rated on ≥5 visits, the 2 newly created fatigue and weight loss variables, the remaining minor other and major other items, and a variable that signified whether new or worse items were present at a specific visit. The weights assigned to the items ranged from 1 to 21. Compared with the original BVAS/WG, this modified score correlated significantly more strongly with the PGA.**

Conclusion. **This study suggests possibilities to enhance the item selection and weighting of the BVAS/WG. These changes may increase this instrument's ability to capture the continuum of disease activity in WG.**

INTRODUCTION

The Birmingham Vasculitis Activity Score for Wegener's Granulomatosis (BVAS/WG) is an instrument designed to measure disease activity (1) and has been used to describe patient populations and to assess treatment efficacy in clinical trials. Much in the same manner as the instrument from which it was derived, the Birmingham Vasculitis Activity Score (BVAS) (2), the BVAS/WG computes disease activity at a given time point as the sum of individual

Supported by Vasculitis Clinical Research Consortium grants (NIH National Center for Research Resources U54-RR-019497 and NIH National Institute of Arthritis and Musculoskeletal and Skin Diseases AR-47785), an NIH National Institute of Arthritis and Musculoskeletal and Skin Diseases Multidisciplinary Clinical Research Center grant (2-P60-AR-047785-06, Boston University), and an NIH National Institute of Arthritis and Musculoskeletal and Skin Diseases grant (5-UO1-AR-51874-03). The Wegener's Granulomatosis Etanercept Trial was supported by an NIH National Institute of Arthritis and Musculoskeletal and Skin Diseases grant (N01-AR-92240), an Office of Orphan Products, FDA grant (FD-R-001652), and General Clinical Research Center grants M01-RRO-00533 (Boston University), M01-RRO-0042 (University of Michigan), M01-RR-30 (Duke University), and M01-RRO-2719 (Johns Hopkins University School of Medicine), from the National Center for Research Resources/NIH. Dr. Merkel is recipient of an NIH National Institute of Arthritis and Musculoskeletal and Skin Diseases Mid-Career Development Award in Clinical Investigation (K24-AR-02224).

[1]Alfred D. Mahr, MD, MPH, Tuhina Neogi, MD, Michael P. LaValley, PhD, Peter A. Merkel, MD, MPH: Boston University, Boston, Massachusetts; [2]John C. Davis, MD, MPH: University of California, San Francisco; [3]Gary S. Hoffman, MD, MS: Cleveland Clinic, Cleveland, Ohio; [4]W. Joseph McCune, MD: University of Michigan, Ann Arbor; [5]Ulrich Specks, MD: Mayo Clinic, Rochester, Minnesota; [6]Robert F. Spiera, MD: Hospital for Special Surgery, New York, New York; [7]E. William St.Clair, MD: Duke University Medical Center, Durham, North Carolina; [8]John H. Stone, MD, MPH: UptoDate, Waltham, Massachusetts.

Dr. St.Clair has received consultant fees (less than $10,000 each) from Genentech, Biogen Idec, Medimmune, Xoma, and Human Genome Sciences.

Address correspondence to Peter A. Merkel, MD, MPH, Vasculitis Center, E5, Boston University School of Medicine, 715 Albany Street, Boston, MA 02118. E-mail: pmerkel@bu.edu.

Submitted for publication September 28, 2007; accepted in revised form December 21, 2007.

organ system manifestations caused by active Wegener's granulomatosis (WG). Such manifestations are collated in a list of 34 predefined items, whereas other unlisted clinical features can be added manually by the clinician. Items carry weights of either 1 or 3 that reflect differences in terms of severity.

The BVAS/WG has been validated for clinical use (1), but still has potential limitations (3). The comprehensiveness of the 34-item list has not been verified in a clinical setting, and it is possible that it does not include all items most relevant to active WG. Moreover, the weights of 1 or 3 were determined empirically, by expert opinion, and they may not accurately reflect the relative disease activity of these manifestations. Finally, the importance, if any, of distinguishing between new/worse as opposed to persistent disease activity remains unclear. Although these potential limitations of the BVAS/WG do not affect the instrument's performance in dichotomizing WG into disease that is either active or inactive, they do call into question the use of the BVAS/WG as a continuous measure.

To address these matters, we reevaluated the BVAS/WG item selection and weighting based on the data collected in a large clinical trial that used this instrument as the primary outcome measurement tool.

MATERIALS AND METHODS

Study setting. The Wegener's Granulomatosis Etanercept Trial (WGET) evaluated the investigational medication plus standard care in the induction and maintenance of disease remissions (4). Members of the WGET Research Group are listed in Appendix A. This randomized, double-blind, placebo-controlled trial enrolled 180 patients from 8 centers in the US. Eligibility criteria included newly diagnosed or flaring WG with a baseline disease activity level of $\geq$3 according to the BVAS/WG. We analyzed data on all BVAS/WG forms collected for every patient throughout the trial. This included assessments at baseline, 6 weeks, and 12 weeks, then every 3 months until the common study closeout date, and 3 and 6 months thereafter.

BVAS/WG instrument. The BVAS/WG instrument is a 1-page form comprising 34 predefined items grouped into 9 organ systems. The items included refer to clinical features frequently observed in patients with active WG. For example, the pulmonary system items include pleurisy, nodules or cavities, other infiltrate secondary to WG, endobronchial involvement, alveolar hemorrhage, and respiratory failure. Each item has a specified weight of either 3 or 1, depending on whether it reflects major or minor disease activity. Manifestations detected by the clinician but not listed as original items can be added in a free-text section entitled "other." Other items are also weighted either 3 or 1 based on the rater's judgment.

The total BVAS/WG score is the weighted sum of individual manifestations that are present and believed to be due to active WG. Higher scores reflect more active disease. BVAS/WG scores range from 0 to 64, not including possible other items. Reported items are categorized further as new/worse (i.e., new occurrence or worsening during the previous 28 days) or persistent (i.e., continued presence without worsening since the last assessment), but both categories contribute equally to the final BVAS/WG score. The BVAS/WG also includes a physician's global assessment (PGA) that consists of an undivided 100-mm visual analog scale anchored by remission (score 0) and maximum activity (score 100). The result of the PGA is not included in the BVAS/WG score (1).

Evaluation of BVAS/WG item use. To determine which clinical features are most relevant to the assessment of disease activity in WG, we calculated the frequency of use of the BVAS/WG items and of the various manifestations added as other. Frequencies were calculated per patient (i.e., proportion of patients in which a given item was rated at least once) and per visit (i.e., proportion of trial visits in which a given item was rated). We also evaluated the frequency with which items were categorized as new/worse or persistent.

Generation of data-driven weights for BVAS/WG items and newly derived variables and cross-validation. Data-driven weights for the BVAS/WG items were generated by multiple linear regression that used the PGA values of disease activity as the dependent variable and the individual BVAS/WG items as explanatory variables. Generalized estimating equation (GEE) techniques were used to account for the possibility that measurements were correlated with each other (5). In all primary analyses, the data were clustered by patients, but we also performed secondary analyses accounting for potential clustering of data from any one trial center. The new item weights were assigned based on the beta regression coefficients, rounded to the nearest integer (6). If a given model generated negative regression coefficients for 1 or several explanatory variables, we removed those terms one by one according to a stepwise backward procedure (by eliminating the one with the lowest value) until all of the remaining variables had regression coefficients >0.

To exclude from the model explanatory variables unlikely to produce reliable estimates, we included only those of the 34 predefined BVAS/WG items that had been recorded at $\geq$5 visits. Predefined items rated at 1–4 visits were merged with the major other or minor other items, according to whether the weight initially assigned to them was 3 or 1. Subsequently, among the list of recorded other items, we created new specific items for those clinical features that had been added at $\geq$10 visits. Finally, to assess the impact of new or worse disease (as compared with persistent disease), we defined a new/worse variable that was set to 0 for visits with no item rated as new/worse and to a nonzero value (with various coding formats tested) for visits with $\geq$1 item rated as new/worse. These supplementary items were entered one by one as explanatory variables to the regression model and were retained only if they improved the model's fit to the data.

To evaluate a model's fit to the data, we calculated preliminary activity scores for all trial visits. These scores were based on the precise items included in a given model and the weights produced by the model. The square of

Pearson's correlation coefficient (R^2) was used to evaluate a score's ability to explain the PGA. The model ultimately retained was determined as that with the highest R^2.

Cross-validation was performed to evaluate the replicability of the linear regression model (7,8). The full data set was split into several subsets using 2 distinct approaches. In the first approach, we randomly split the entire data set into 5 subsets of equal size; to ensure that all subgroups were composed of visits displaying a similar range of disease activity levels, each subgroup was assembled by randomly selecting one-fifth of the baseline visits and one-fifth of the nonbaseline visits. In the second approach, the data set was divided according to trial centers; the data of 2 centers with the lowest enrollments were merged to reduce the differences in subset sizes. For both data partitioning approaches, the linear regression analyses (which included the same explanatory variables as those retained in the best-fitting model obtained from the full data set) were repeated by omitting 1 subset of the data at a time. The item weights generated by these derivation models were again equated with the beta coefficients rounded to the nearest integer, but, in order to adopt the most conservative approach, items could also be assigned negative weights. By applying these weighting systems to their corresponding validation subgroups that consisted of visits excluded from the regression model, we recalculated activity scores and computed their R^2 value to predict the PGA. We interpreted significant reductions in the R^2 value (compared with that obtained from the full data set) as an indication of poor replicability.

For all models, assumptions of linearity were checked graphically by drawing plots of the observed values versus the predicted values, and of the residuals versus predicted values.

Evaluation of the validity of data-driven item weights and modified item selection. To assess the validity of the changes in item selection and weights, we compared the activity scores obtained by this modified BVAS/WG instrument with the BVAS/WG as originally described. We also compared the modified instrument with a completely unweighted BVAS/WG in which all items were empirically assigned a weight of 1. For all 3 variations of the BVAS/WG instrument, we assessed Pearson's correlation coefficients (r) with the PGA. Hotelling's statistics were used to determine whether the differences between 2 dependent correlations were statistically significant (9). Comparisons were made within the full data set, within the subset of visits in which WG was active, and within the baseline visits alone. In addition, we performed subgroup analyses based on whether the patients had severe or limited WG at trial entry (10).

Statistical analyses. All statistics were computed by the SAS Statistical Software, version 9.1 for Windows (SAS Institute, Cary, NC). For all analyses, a 2-tailed *P* value less than 0.05 was considered significant. Continuous variables were expressed as the mean ± SD and range.

RESULTS

WGET data set and data checking. The 180 WGET participants contributed 2,044 trial visits. The mean ± SD followup (from baseline to last visit) was 30.3 ± 11.7 months (range 0–47.0). The mean number of visits per patient was 11.4 ± 3.8 (range 1–17).

Among the 2,044 BVAS/WG forms analyzed, we reclassified 20 manifestations listed as other items into their appropriate original item designation (for example, orbital pseudotumor and biopsy-proven glomerulonephritis were reclassified as one of the original items: retroorbital mass/proptosis and red cell casts, respectively). Data checking also identified 8 forms in which both hematuria and red cell casts were scored. For those visits, we analyzed only red cell casts, according to the original BVAS/WG provisions (1).

For 1 visit, the BVAS/WG form was incomplete, and 17 forms had missing PGA values. None of these 18 BVAS/WG forms (0.9%) occurred at a baseline visit. Among the 2,026 BVAS/WG forms with no missing data, we defined 734 as representing active disease because the PGA score was >0. Among those 734 forms, 21 had BVAS/WG scores of 0, including 4 with possible outlying PGA values of 11, 21, 34, and 80, respectively (all occurring during followup visits). Among the 1,292 forms with PGA scores of 0, there were only 9 for which the BVAS/WG score was not 0 (range 1–4).

Use of predefined original and other BVAS/WG items. Item use was calculated from all 2,044 BVAS/WG forms. The frequency of use of each individual original item is shown in Table 1. Five predefined items were reported in only 1–4 forms (respiratory failure, pericarditis, mesenteric ischemia, gangrene, and uveitis), and 4 items were never utilized (retinal exudates/hemorrhage, meningitis, cord lesion, and stroke). Table 1 also shows the proportion of study patients for whom each individual item had been rated during at least 1 visit.

A total of 175 other items were added at 149 visits, including 18 rated as major and 157 rated as minor. The most commonly added other item was fatigue, rated in 58 forms; fatigue was rated as the only manifestation in 6 followup visits (corresponding to 3 different subjects). Weight loss was added to 12 visits and was never rated in isolation. The entire list of other manifestations is shown in Table 2. To simplify the presentation, other items were combined, when appropriate, under generic headings.

New/worse and persistent status of items scored. Of the total 1,855 ratings of predefined items, 568 (31%) were classified as persistent and 1,287 (69%) were classified as new/worse. Among the 175 other items, 40 (23%) were classified as persistent, and 135 (77%) as new/worse. In total, 447 (22%) and 305 (15%) BVAS/WG forms had ≥1 and ≥2 items checked as new/worse, respectively. At the baseline visits, 829 (95%) and 47 (5%) of the rated items were new/worse and persistent, respectively, and ≥1 and ≥2 new/worse items were reported for 175 (97%) and 168 (93%) patients, respectively.

Table 1. Use of 34 predefined Birmingham Vasculitis Activity Score for Wegener's Granulomatosis items in 180 participants and 2,044 visits of the Wegener's Granulomatosis Etanercept Trial

Variable	Frequency, no. (%)	
	All 2,044 visits	All 180 participants
Arthralgia/arthritis	351 (17.2)	135 (75.0)
Nasal discharge/crusting	331 (16.2)	122 (67.8)
Sinus involvement	199 (9.7)	98 (54.4)
Nodules or cavities	141 (6.9)	67 (37.2)
Hematuria	115 (5.6)	77 (42.8)
Conductive deafness	95 (4.7)	52 (28.9)
Other infiltrates	83 (4.1)	63 (35.0)
Red blood cell casts*	66 (3.2)	58 (32.2)
Subglottic inflammation	54 (2.6)	22 (12.2)
≥30% creatinine increase*	51 (2.5)	46 (25.6)
Conjunctivitis/episcleritis	47 (2.3)	36 (20.0)
Purpura	43 (2.1)	35 (19.4)
Fever	42 (2.1)	38 (21.1)
Alveolar hemorrhage*	37 (1.8)	35 (19.4)
Mouth ulcers	28 (1.4)	26 (14.4)
Retroorbital mass/proptosis	28 (1.4)	12 (6.7)
Sensorineural deafness*	27 (1.3)	11 (6.1)
Sensory neuropathy*	26 (1.3)	20 (11.1)
Endobronchial involvement	19 (0.9)	11 (6.1)
Pleurisy	18 (0.9)	16 (8.9)
Skin ulcers	17 (0.8)	8 (4.4)
Scleritis*	11 (0.5)	8 (4.4)
Swollen salivary gland	6 (0.3)	6 (3.3)
Motor neuropathy*	6 (0.3)	4 (2.2)
Cranial nerve palsy*	5 (0.2)	5 (2.8)
Respiratory failure*	3 (0.2)	3 (1.7)
Mesenteric ischemia*	2 (0.1)	2 (1.1)
Pericarditis	2 (0.1)	2 (1.1)
Gangrene*	1 (0.1)	1 (0.6)
Uveitis	1 (0.1)	1 (0.6)
Retinal exudates/hemorrhage*	0 (0)	0 (0)
Meningitis*	0 (0)	0 (0)
Cord lesion*	0 (0)	0 (0)
Stroke*	0 (0)	0 (0)

* Major items.

Data-driven BVAS/WG item weighting. Linear regression models were based on the 734 complete BVAS/WG forms with active disease. Similar results were obtained when these analyses used all 2,026 complete BVAS/WG forms, because the differences between the samples mostly included noninformative visits with 0 values for both BVAS/WG score and PGA.

A flow diagram of explanatory variables implemented in the linear regression models is shown in Figure 1. The initial linear model included the 25 predefined BVAS/WG items rated in ≥5 forms, the major other variable, and the minor other variable. Using the weights produced by this model, we calculated the activity scores and their ability (R^2) to explain the PGA. As compared with the R^2 value of 0.5693 for the original BVAS/WG (and the R^2 value of 0.5278 for the unweighted BVAS/WG), the R^2 value for this preliminary index was 0.6248. The introduction of the variables fatigue and weight loss into this model improved the R^2 to 0.6275 and 0.6310, respectively. Additional specific variables were also created for skin nodules and otitis/mastoiditis, but insertion of these variables into the model either did not further improve the model's ability to explain the PGA (R^2 = 0.6310) or yielded a negative beta regression coefficient (−1.3181, P = 0.4651). Thus, these latter 2 variables were not retained in the models.

We then added a new/worse variable, based upon whether ≥1 items present at a given visit were classified as either new or worse. We analyzed several formats for this variable, and obtained the following R^2 values: ordinal (exact number of new/worse items at visit): R^2 = 0.6510; dichotomous (0 or ≥1 new/worse items at visit): R^2 =

Table 2. Use of other Birmingham Vasculitis Activity Score for Wegener's Granulomatosis items in 180 participants and 2,044 visits of the Wegener's Granulomatosis Etanercept Trial*

Variable	Frequency, no. (%)	
	All 2,044 visits	All 180 participants
All combined	175 (8.6)	84 (46.7)
Fatigue	58 (2.8)	28 (15.6)
Weight loss	12 (0.6)	12 (6.7)
Skin nodules	12 (0.6)	10 (5.6)
Otitis/mastoiditis	11 (0.5)	7 (3.9)
Chondritis	8 (0.4)	5 (2.8)
Miscellaneous skin items†	7 (0.3)	6 (3.3)
Myositis/myalgia	5 (0.2)	5 (2.8)
Respiratory items†	5 (0.2)	5 (2.8)
Optic neuritis‡	5 (0.2)	2 (1.1)
Eye adnexa disease-related items†	5 (0.2)	4 (2.2)
Laryngotracheal items§	5 (0.2)	4 (2.2)
Rash	5 (0.2)	2 (1.1)
Breast involvement	5 (0.2)	1 (0.6)
Proteinuria	4 (0.2)	4 (2.2)
Gingivitis	3 (0.1)	3 (1.7)
CNS symptoms or disease-related items¶	3 (0.1)	3 (1.7)
Treatment failure‡	3 (0.1)	3 (1.7)
Keratitis	3 (0.1)	2 (1.1)
Lymphadenopathy	2 (0.1)	2 (1.1)
Testicular involvement	2 (0.1)	2 (1.1)
Liver disease	2 (0.1)	1 (0.6)
Orbit and sinus wall osteolysis‡	2 (0.1)	1 (0.6)
Deep venous thrombosis‡	1 (0.05)	1 (0.6)
Hemorrhagic lesions on palate	1 (0.05)	1 (0.6)
Myocarditis	1 (0.05)	1 (0.6)
Neuropathic pain‡	1 (0.05)	1 (0.6)
Night sweats	1 (0.05)	1 (0.6)
Pulmonary artery stenosis‡	1 (0.05)	1 (0.6)
Skull-based mass‡	1 (0.05)	1 (0.6)
Tinnitus	1 (0.05)	1 (0.6)

* CNS = central nervous system.
† Generic variables combining several other items.
‡ Other items rated as major.
§ Generic variable combining several other items; includes 1 rating of cricoarytenoid inflammation as major other.
¶ Generic variable combining several other items; includes ratings of CNS vasculitis and pituitary involvement (each rated once) as major other.

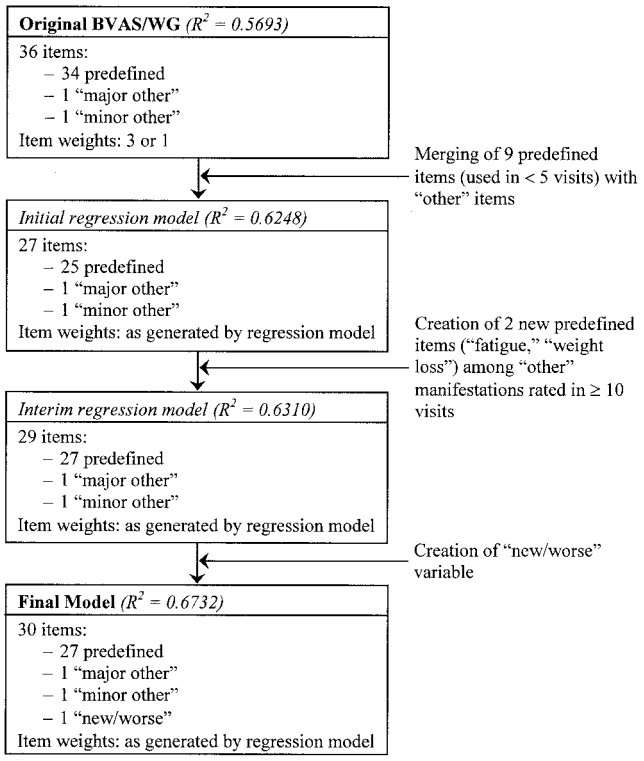

Figure 1. Summary flow diagram of Birmingham Vasculitis Activity Score for Wegener's Granulomatosis (BVAS/WG) items implemented in the linear regression models used to derive a new item weighting system (R^2 values refer to the model's ability to predict the physician's global assessment of disease activity).

0.6608; dichotomous (<2 or ≥2): R^2 = 0.6726; 3-level ordinal (0, 1, or ≥2): R^2 = 0.6732; dichotomous (<3 or ≥3): R^2 = 0.6746; 4-level ordinal (0, 1, 2, or ≥3): R^2 = 0.6796. Ultimately, we selected the 3-level ordinal format (0, 1, or ≥2) because this provided a uniformly high fit when applied to the full data set of 2,026 visits or the 180 visits at baseline.

The results of the final linear regression model are shown in Table 3. No substantial changes in these results were observed in additional analyses, which excluded the 4 BVAS/WG forms with possible outlier PGA values (see above) or included trial center as a covariate, and by GEE modeling that controlled for potential data correlation within trial centers. The variables salivary gland involvement, mouth ulcers, endobronchial disease, and conjunctivitis/episcleritis were dropped from this model because they yielded negative weights or rounded to zero. When assigning to these 4 variables a minimal weight of 1, the R^2 of the model decreased only slightly to 0.6725. The latter index will be referred to as the modified BVAS/WG.

Cross-validation. The cross-validation was also based on the 734 visits with active disease. The weights generated for the 30 items in any of the derivation models (each used either four-fifths of the data set or all data except those from 1 trial center) only marginally differed from the weights that were produced from the full data set. When using these derived weighting systems to calculate the activity scores in the corresponding validation subsample,

Table 3. Results of the linear regression model (based on 734 visits with active disease and using the physician's global assessment as the dependent variable and the Birmingham Vasculitis Activity Score for Wegener's Granulomatosis items as explanatory variables) and derived new item weighting system*

Explanatory variable	Linear regression model			Item weight	
	No.	β coefficient	P	Original	New
Alveolar hemorrhage	36	20.74	< 0.0001	3	21
≥30% creatinine increase	50	17.39	< 0.0001	3	17
Sensory neuropathy	26	16.00	< 0.0001	3	16
Red blood cell casts	63	14.81	< 0.0001	3	15
Weight loss	12	12.89	0.005	NA	13
Fatigue	55	11.48	< 0.0001	NA	11
Nodules or cavities	138	10.20	< 0.0001	1	10
Other infiltrates	81	9.96	< 0.0001	1	10
Hematuria	111	8.97	< 0.0001	1	9
Major other	25†	7.54	0.03	3	8
Sensorineural deafness	27	7.54	0.01	3	8
Cranial nerve palsy	5	7.04	0.24	3	7
Sinus involvement	196	6.89	< 0.0001	1	7
Subglottic inflammation	53	6.28	0.01	1	6
Skin ulcers	17	6.06	0.04	1	6
Pleurisy	18	5.55	0.11	1	6
Purpura	41	5.03	0.03	1	5
Scleritis	11	4.29	0.22	3	4
Retroorbital mass/ proptosis	27	3.95	0.21	1	4
Motor neuropathy	6	3.75	0.52	3	4
Nasal discharge/ crusting	327	3.19	0.01	1	3
Minor other	89‡	3.17	0.06	1	3
Arthralgia/arthritis	339	2.51	0.05	1	3
Fever	41	2.36	0.29	1	2
Conductive deafness	93	1.94	0.24	1	2
Conjunctivitis/ episcleritis	45	0.07	0.98	1	1§
Endobronchial involvement	18	–	–	1	1§
Mouth ulcers	28	–	–	1	1§
Salivary gland involvement	6	–	–	1	1§
New/worse¶					
1 item	140	6.46	< 0.0001	NA	6
≥2 items	303	15.09	< 0.0001	NA	15

* Except for the new/worse item, variables are sorted from highest to lowest new weights. No. = total number of ratings of explanatory variables; NA = not applicable.
† Accounting for 23 visits.
‡ Accounting for 75 visits.
§ Variables with negative regression coefficients in the regression model or values rounded to 0 that were assigned a weight of 1.
¶ Visit with ≥1 item rated new/worse.

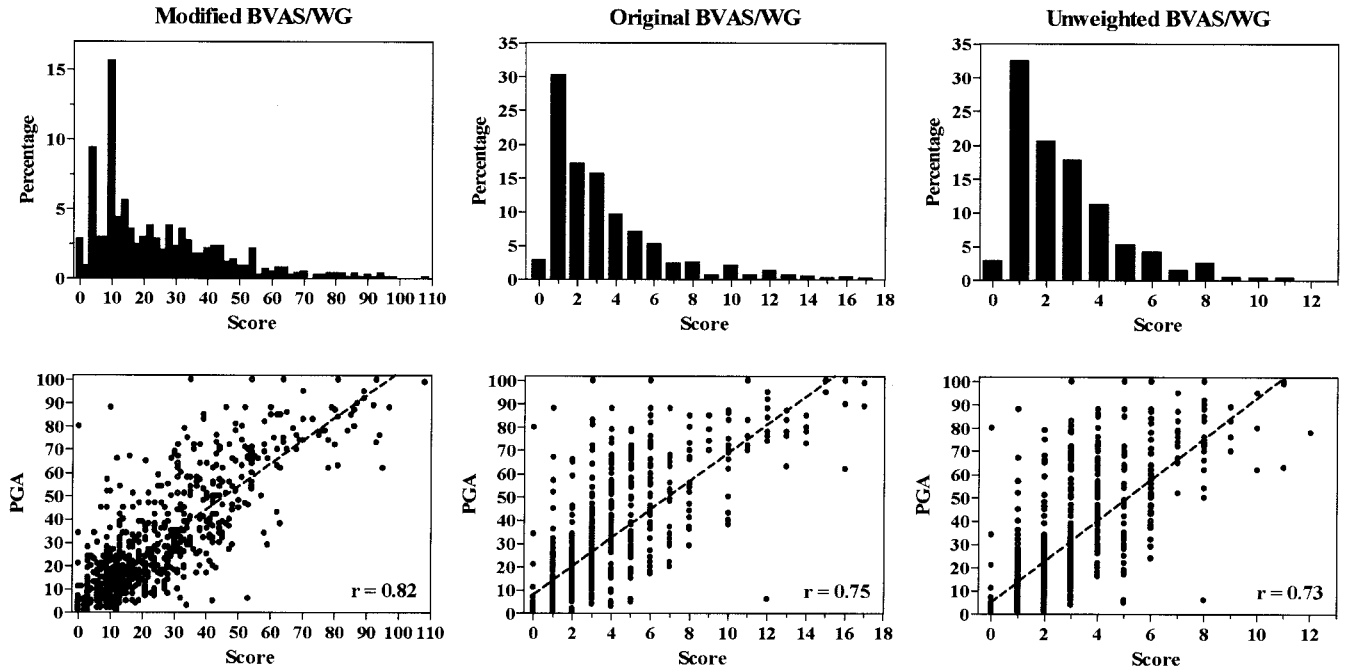

Figure 2. Distributions and scatter plots (plotted against the physician's global assessment [PGA]) of the disease activity scores for the modified Birmingham Vasculitis Activity Score for Wegener's Granulomatosis (BVAS/WG), original BVAS/WG, and unweighted BVAS/WG based on 734 trial visits assessing active Wegener's granulomatosis (i.e., PGA >0).

the R^2 to predict the PGA within the reassembled full data set was only slightly lower than the R^2 obtained by the full model: 0.6444 (for the first approach that split the data set into 5 random subsets) and 0.6194 (for the second approach of data partitioning according to trial centers), respectively. Therefore, the cross-validation findings suggested replicability of both the generated item weights and the increased ability of the modified BVAS/WG to explain the PGA.

Performance characteristics of the modified BVAS/WG. The modified BVAS/WG had item weights ranging from 1 to 21 (Table 3). The possible range of scores was 0 to 204, plus other items.

Figure 2 shows the observed distributions of scores of the modified BVAS/WG as compared with the original BVAS/WG and a completely unweighted BVAS/WG for the 734 visits with active disease. Mean ± SD scores were 24.1 ± 20.4 (range 0–108) for the modified BVAS/WG, 3.4 ± 3.0 (range 0–17) for the original BVAS/WG, and 2.7 ± 2.0 (range 0–12) for the unweighted BVAS/WG. For the 180 trial participants at baseline (study entry), the mean modified BVAS/WG, original BVAS/WG, and unweighted BVAS/WG scores were 49.6 ± 19.1 (range 10–108), 6.9 ± 3.4 (range 2–17), and 5.0 ± 2.1 (range 1–12), respectively.

Table 4 shows the correlation coefficients of the PGA with these 3 scores for all 734 visits with active disease,

Table 4. Pearson's correlation coefficients of the modified Birmingham Vasculitis Activity Score for Wegener's Granulomatosis (BVAS/WG), original BVAS/WG, and unweighted BVAS/WG with physician's global assessment of disease activity

Visits	No.	BVAS/WG version			*P**
		Modified	Original	Unweighted	
All active visits	734	0.82†	0.75†	0.73†	< 0.0001
Limited WG‡	277	0.79†	0.65†	0.67†	< 0.0001
Severe WG‡	457	0.82†	0.78†	0.75†	< 0.0001
Baseline visits	180	0.71†	0.62†	0.51†	< 0.0001
Limited WG	52	0.66†	0.45§	0.46§	< 0.005
Severe WG	128	0.68†	0.58†	0.49†	0.0005
All visits	2,026	0.90†	0.87†	0.86†	< 0.0001

* For comparisons of correlation coefficients of modified BVAS/WG and BVAS/WG (as originally reported) calculated by Hotelling's statistics (9).
† $P < 0.0001$.
‡ According to WG disease classification as determined at study entry.
§ $P = 0.001$.

the 180 baseline visits, and all 2,026 visits with complete BVAS/WG information. Predictably, the modified BVAS/WG correlated more strongly with the PGA than the original BVAS/WG or the unweighted BVAS/WG. Analyses stratified by disease pattern indicated more substantial differences among patients originally classified as having limited as opposed to severe WG. In all subsets, the differences between the correlation coefficients of the modified BVAS/WG and the original BVAS/WG were statistically significant ($P < 0.005$).

DISCUSSION

We reevaluated the BVAS/WG using a data-driven approach. Our results suggest possible ways to enhance the item selection and weighting of the instrument that could improve performance characteristics substantially. When expressed as the ability (R^2) to predict the PGA for disease activity, these modifications increased the R^2 from 57% for the original BVAS/WG to 67% for the modified BVAS/WG, a sizeable difference when considering that a completely unweighted BVAS/WG yielded a value of 53%. Cross-validation demonstrated that our findings were internally replicable and generalizable across the various centers involved in the study.

The content validity of the BVAS/WG relies in large part on the comprehensiveness of the list of predefined disease descriptors to ensure that manifestations most pertinent to disease activity assessment are rated consistently. Our findings indicate that fatigue and weight loss should be included as specific items, because both were added frequently as other items and because they improve the instrument's ability to predict the PGA. The weights of 11 and 13 generated for these 2 variables emphasize that physicians highly value constitutional symptoms as indicators of disease activity in patients with WG. The great variety of the remaining other manifestations (Table 2) reflects the protean nature of WG, and the weights generated for the pooled minor other and major other categories underscore the utility of accommodating and counting unlisted manifestations. In contrast, 9 predefined items that refer to less common features of WG were used in $\leq$5 visits, suggesting that the BVAS/WG item list could be streamlined without resulting in untoward effects.

The PGA of disease activity included in the BVAS/WG form provided an unprecedented opportunity to compute an item weighting system. The calculated range of weights of 1–21 cover a much broader spectrum than the original, expert opinion-based weights used in either the BVAS/WG (weights 1 or 3) (1) or its precursor, the BVAS (weights 1–9) (2). In addition, the new weights add substantially to the ability of the index to explain the PGA and appear to have face validity. Indeed, alveolar hemorrhage, items related to renal disease, and sensory neuropathy were the highest ranked weights, and major other items were ranked higher than minor other items.

Although the generated weights offer new insight into the extent to which individual features of WG contribute to physicians' perception of overall disease activity, they must be interpreted with several caveats. First, the estimates of some infrequently rated items may lack precision, as suggested by nonsignificant *P* values of the beta coefficients. Also, as is common in multiple regression models, the accuracy of the estimates may have been hampered by collinearity between variables. However, bivariate analyses among all explanatory variables did not detect correlations exceeding 0.40 (results not shown), suggesting that collinearity was not a major issue. A slightly different situation occurs for manifestations that seldom (or never) present in isolation, and whose assigned weights are therefore connected to weights of related items. Motor neuropathy, for example, was assigned a seemingly low weight of 4, but this manifestation was scored concomitantly with sensory neuropathy in all cases except one, thereby contributing to a combined weight of 20 for the variable sensory-motor neuropathy.

An additional significant finding pertains to the new/worse item, a newly created variable that informed whether 1 or several manifestations had been rated as new/worse at a given visit. It is well acknowledged that once treated, patients with WG may undergo a substantial and durable improvement without achieving remission. In terms of disease activity scoring, this is a tricky problem in that persistent manifestations likely reflect a lower level of disease activity. Pertinently, our data demonstrate that physicians' ratings of disease activity were independently higher when new/worse manifestations were present and suggest that such a weighted new/worse variable may constitute a valuable means to take account of the different disease activity levels of new/worsening and persistent WG. Our findings further indicate that different levels of gradation exist for this variable: visits associated with 1 or $\geq$2 new/worse items were assigned weights of 6 and 15, respectively. Whether more widespread gradations within variables (accounting for different intensities of clinical manifestations) could further improve the performance of the BVAS/WG is beyond the scope of this study.

The present findings have a number of important implications. Owing to the enhanced item selection, the implementation of a data-derived weighting system, and the addition of a new/worse variable, the modified BVAS/WG may have greater ability to capture the continuum of disease activity in patients with WG. The modified BVAS/WG could permit more accurate comparisons of the absolute scores across patients and more precise measurements of the relative amount in disease activity changes at different time points. Such a modified scale could also facilitate the translation of BVAS/WG scores into specific states of disease activity. In particular, there is a recognized need to delineate the state of "grumbling," low-level disease activity that may be a possible therapeutic target in clinical trials.

Our study has several strengths. This analysis took advantage of a large, prospectively collected data set. Because the data set was derived from participants in a multicenter clinical trial with broad eligibility criteria, our results are extrapolated more easily to clinical settings compared with alternative study designs that might have been used, such as investigations based on written summaries of hypothetical patients or a convenience sample of real patients from only 1 center. Moreover, the trial in-

volved 8 centers with expertise in WG and investigators previously trained to use the BVAS/WG, thereby ensuring a high quality in the BVAS/WG assessments and PGA ratings. The consistency in PGA ratings was also indicated by our observation that adjustment of the regression for the trial centers did not affect the values of the derived item weights substantially.

There are 2 factors, both related to the PGA, that potentially limit our findings. First, although the BVAS/WG was originally validated against the PGA (1), one may question the use of the PGA as a reference value to derive the new item weights and to evaluate the modified BVAS/WG. Nonetheless, because there is no other current gold standard for the assessment of disease activity in WG, the PGA represents the most suitable universal means for quantifying WG activity. Of note, the PGA has also been selected as a core set measure for disease activity assessment in a number of other rheumatic diseases (11–14). Second, it could be claimed that because they were not performed independently, the PGA ratings simply reproduced the scores reached by the BVAS/WG. However, this possibility seems countered by the substantial differences between the original and modified item weights and the apparently reasonable weights generated for the newly implemented new/worse variable.

In conclusion, this study clarifies the selection and weighting of the clinical features that are most relevant to measure disease activity in patients with WG. The proposed modifications should be considered for further evaluation and use in clinical research and viewed as a stage in the iterative process of refining the BVAS/WG and perhaps other related instruments.

AUTHOR CONTRIBUTIONS

Dr. Merkel had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study design. Mahr, Neogi, Davis, McCune, St.Clair, Stone, Merkel.

Acquisition of data. Mahr, Davis, Hoffman, McCune, Specks, Spiera, St.Clair, Stone, Merkel.

Analysis and interpretation of data. Mahr, Neogi, LaValley, Davis, McCune, St.Clair, Stone, Merkel.

Manuscript preparation. Mahr, Neogi, LaValley, Davis, St.Clair, Stone, Merkel.

Statistical analysis. Mahr, Neogi, LaValley, Merkel.

Attained research funding. Merkel.

REFERENCES

1. Stone JH, Hoffman GS, Merkel PA, Min YI, Uhlfelder ML, Hellmann DB, et al, for the International Network for the Study of the Systemic Vasculitides (INSSYS). A disease-specific activity index for Wegener's granulomatosis: modification of the Birmingham Vasculitis Activity Score. Arthritis Rheum 2001;44:912–20.
2. Luqmani RA, Bacon PA, Moots RJ, Janssen BA, Pall A, Emery P, et al. Birmingham Vasculitis Activity Score (BVAS) in systemic necrotizing vasculitis. QJM 1994;87:671–8.
3. Merkel PA, Seo P, Aries P, Neogi T, Villa-Forte A, Boers M, et al, for the Vasculitis Clinical Research Consortium. Current status of outcome measures in vasculitis: focus on Wegener's granulomatosis and microscopic polyangiitis. Report from OMERACT 7. J Rheumatol 2005;32:2488–95.
4. The Wegener's Granulomatosis Etanercept Trial (WGET) Research Group. Etanercept plus standard therapy for Wegener's granulomatosis. N Engl J Med 2005;352:351–61.
5. Hanley JA, Negassa A, Edwardes MD, Forrester JE. Statistical analysis of correlated data using generalized estimating equations: an orientation. Am J Epidemiol 2003;157:364–75.
6. Moons KG, Harrell FE, Steyerberg EW. Should scoring rules be based on odds ratios or regression coefficients? [letter]. J Clin Epidemiol 2002;55:1054–5.
7. Efron B, Tibshirani RJ. Cross-validation. In: Efron B, Tibshirani RJ, editors. An introduction to the bootstrap. New York: Chapman & Hall/CRC; 1993. p. 239–41.
8. Stone M. Cross-validatory choice and assessment of statistical predictions. J Royal Stat Soc 1974;36:111–47.
9. Glass GV, Stanley JC. Statistical methods in education and psychology. Englewood Cliffs (NJ): Prentice Hall; 1970. p. 310–1.
10. The Wegener's Granulomatosis Etanercept Trial Research Group. Limited versus severe Wegener's granulomatosis: baseline data on patients in the Wegener's granulomatosis etanercept trial. Arthritis Rheum 2003;48:2299–309.
11. Smolen JS, Breedveld FC, Schiff MH, Kalden JR, Emery P, Eberl G, et al. A simplified disease activity index for rheumatoid arthritis for use in clinical practice. Rheumatology (Oxford) 2003;42:244–57.
12. Aletaha D, Nell VP, Stamm T, Uffmann M, Pflugbeil S, Machold K, et al. Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score. Arthritis Res Ther 2005;7:R796–806.
13. Leeb BF, Bird HA. A disease activity score for polymyalgia rheumatica. Ann Rheum Dis 2004;63:1279–83.
14. Giannini EH, Ruperto N, Ravelli A, Lovell DJ, Felson DT, Martini A. Preliminary definition of improvement in juvenile arthritis. Arthritis Rheum 1997;40:1202–9.

APPENDIX A: THE WEGENER'S GRANULOMATOSIS ETANERCEPT TRIAL (WGET) RESEARCH GROUP

Members of the WGET Research Group are as follows: John H. Stone (Chairman, The Johns Hopkins Vasculitis Center); Gary S. Hoffman (Co-Chairman, The Cleveland Clinic Foundation Center for Vasculitis Research and Care); Janet T. Holbrook, Curtis L. Meinert, John Dodge, Jessica Donithan, Nancy Min, Laurel Murrow, Jacki Smith, Andrea K. Tibbs, Mark Van Natta (The Johns Hopkins University Center for Clinical Trials); Robert Spiera, Rosanne Berman, Sandy Enuha (The Beth Israel Medical Center, New York); Peter A. Merkel, Rondi Gelbard, Melynn Nuite, Aileen Schiller (Boston University); Gary S. Hoffman, David Blumenthal, Debora Bork, Tiffany Clark, Sonya L. Crook, Leonard H. Calabrese, Sharon Farkas, Sudhakar Sridharan, Kimberly Strom, William Wilke (The Cleveland Clinic Foundation); E. William St. Clair, Nancy B. Allen, Karen Rodin, Edna Scarlett (Duke University); John H. Stone, David B. Hellmann, Amanda M. Moore, Lourdes Pinachos, Michael J. Regan, Misty L. Uhlfelder (Johns Hopkins University); Ulrich Specks, Kristin Bradt, Kimberly Carlson, Susan Fisher, Boleyn Hammel, Kathy Mieras, Steven Ytterberg (The Mayo Clinic); John C. Davis, Maureen Fitzpatrick, Ken Fye, Steve Lund (University of California, San Francisco); Joseph McCune, Billie Jo Coomer, Barbara Gilson, Hilary Haftel, Ana Morrel-Samuels, Sandra Neckel (University of Michigan); Noel R. Rose, C. Lynne Burek, Jobert Barin, Monica Talor (The Johns Hopkins University Immune Diseases Laboratory); Paul L. Canner (Maryland Medical Research Institute); Doyt L. Conn (Emory University); Jack H. Klippel (Arthritis Foundation); J. Richard Landis (University of Pennsylvania).