A Quantitative System to Evaluate Diabetic Retinopathy From Fundus Photographs

Stephen S. Feman, Thomas C. Leonard-Martin, J. Stevens Andrews, Cecile C. Armbruster, Theresa L. Burdge, Judith D. Debelak, Angela Lanier, and Amy G. Fischer

Purpose. To evaluate a quantitative system to measure the early lesions of diabetic retinopathy seen in stereoscopic fundus photographs.

Methods. Using a quantitative classification system, photographs of 4657 eyes (7 stereo pairs of 35-mm slides per eye) were scored for 16 diabetic lesions. A single severity level (identical to the ETDRS Interim Scale) was calculated for each eye. The reliability of this technique, and its reproducibility by independent examiners, was evaluated for individual lesions and severity levels using percent agreement, kappa, and weighted kappa statistics.

Results. This quantitative technique demonstrated an "almost perfect" agreement (weighted kappa $\geq$ 0.810) on all but one lesion by independent observers. For the severity levels, there was a 95.7% perfect agreement (kappa = 0.9428). The reproducibility of agreement over time was "almost perfect" on all but four lesions; with 88% perfect agreement (kappa = 0.8394) for severity levels.

Conclusions. When used to evaluate the early lesions of diabetic retinopathy, the Vanderbilt Classification System is highly reliable between graders and over time. This system can gather quantitative data and evaluate incremental changes in an accurate, reproducible manner. Invest Ophthalmol Vis Sci. 1995;36:174–181.

Although there have been major advances in the treatment of diabetes[1] and diabetic retinopathy,[2,3] diabetes mellitus remains the leading cause of blindness in adults in the United States.[4] For this reason, diabetic retinopathy continues to be a subject of significant research at many sites. In the past, most clinical research on diabetic retinopathy that required large numbers of patients had graded lesions by comparisons to "standard photographs." One of the first descriptions of this technique, as well as original photographs, was presented in 1968.[5] Although this "Airlie House" classification system has been modified several times,[6,7] it is limited by the difficulty in specifying all the relevant factors in each standard photograph. At the same time, intermediate stages between the standard photographs are blurred by its grading scheme. To overcome this problem, a system was developed that parallels the extension of the Modified Airlie House Classification (xMAHC) used for the Early Treatment Diabetic Retinopathy Study (ETDRS).[7] This Vanderbilt Classification System makes use of counts, or measurements, of lesions instead of comparisons to standard photographs. This report describes the Vanderbilt Classification System (VCS) and its intergrader and intragrader reproducibility.

In the original publication, the Airlie House classification system used five standard photographic fields and classified 14 retinal lesions on a three-step scale: absent, mild to moderate, and severe.[5] Standard photographs were chosen to demarcate the lower limit of each category.[5] When planning the Diabetic Retinopathy Study, it was found that another gradation between absent and severe was needed. Also, the vicissitudes of the actual grading required the addition of a "questionable" category. At the same time, two more standard photographic fields (6 and 7) were defined, and several other lesions were added. The result became known as the Modified Airlie House Classification system.[6]

From the Fundus Photograph Reading Center, Department of Ophthalmology and Visual Sciences, School of Medicine, Vanderbilt University, Nashville, Tennessee.
Supported in part by funds from Wyeth-Ayerst Research, Inc., Philadelphia, Pennsylvania, and by unrestricted funds from Research to Prevent Blindness, Inc., New York, New York.
Submitted for publication May 16, 1994; revised June 28, 1994; accepted July 8, 1994.
Proprietary interest category: C2.
Reprint requests: Stephen S. Feman, 8011 Medical Center East, Vanderbilt University Medical School, Nashville, TN 37232-8808.

Investigative Ophthalmology & Visual Science, January 1995, Vol. 36, No. 1

When the ETDRS was designed, the Modified Airlie House Classification system was deemed unsatisfactory for its specific needs. Several features were added[7] to extend the system so it could be used for the planned research. New steps, with accompanying standard photographs, were added to the scale for hard exudates, soft exudates, arteriovenous nicking, retinal elevation, and vitreous hemorrhage. Individual categories were made of some previously pooled items, and several entirely new categories were created.[7]

The Vanderbilt Classification System uses the same seven standard photographic fields and the same lesion categories as does this extension of the Modified Airlie House Classification system. However, some lesions of the xMAHC that have not been associated with progression of diabetic retinopathy[7,8] have been excluded from the Vanderbilt Classification System. Nevertheless, this parallel technique allows the Vanderbilt Classification System to yield a single severity level per eye in a manner similar to the ETDRS Interim Scales, Final Scales, or both.[8] For all individual lesions, the steps above "questionable" are defined in terms of quantities and sizes. Thus, although knowledge of a lesion's appearance or a photograph of a typical lesion may be needed for an understanding of the pathologic disorder under evaluation, no references to standard photographs are required in the application of this system. A detailed analysis of the reproducibility of this technique is included to verify that photographic comparisons are unnecessary.

MATERIALS AND METHODS

To evaluate this system, the procedures of the Diabetic Retinopathy Study and ETDRS were used. This included the use of 30° stereoscopic color photograph pairs of the seven standard fields, mounted in plastic sheets and viewed on light boards using stereo viewers (magnification, ×5). The graders were research personnel, not ophthalmologists, trained to measure diabetic retinopathy using this system. Four graders reported the data. Graders were masked to the scores of other graders and to their own previous scores. Each lesion was graded independently in all or a specified series of fields. Graders were allowed to use overlapping areas of fields to form their evaluations.

Graders were taught to count lesions, such as microaneurysms, in either a scanning or a geographic procedure. In addition, because no exact number had to be recorded once the count reached greater than 20, they did not have to count every lesion. In lesions in which area was measured, the grading scale was divided into ranges. To compare these results to reports using other techniques and to simplify data management, absolute numbers and areas are described in limited fashion; however, the ranges used and summarized in the following section were chosen to match those used with the ETDRS Interim and Final Scales.[7,8] If the level was not obvious, the area covered by the lesion(s) was measured with a transparent, graphic device to determine size.

The photographs used in this report are of patients enrolled in the Tolrestat Trials, sponsored by Wyeth–Ayerst Research (Philadelphia, PA). Patients were recruited for these trials by regional investigators at multiple sites in the United States. Because many medical centers were involved, the clinical investigators at each site had to obtain the approval of the local institutional review board to enroll each patient. Informed consent signed by the patient and approved by the local institutional review board was obtained before photographs could be used. In this manner, the tenets of the Declaration of Helsinki were followed.

To test the reliability of the system between different graders, 2329 sets (two eyes per set minus one eye: $n = 4657$) were scored independently by pairs of graders. To avoid internal bias, a masking technique was used that prevented each grader from having knowledge of other grading activities involving these specific sets of photographs. Reliability over time was measured by having each grader regrade a subset of photographs evaluated previously. Three different groups of photographs (total: 163 sets) were randomly selected 3 to 6 months after their initial grading. Because each pair of eyes was scored initially by two independent graders, the same eyes were graded the second time by the same graders. Thus, 652 comparisons are reported here.

Individual Lesions

Count of microaneurysms per seven fields (ct. MA/global). This was the number of intraretinal red spots with round and sharp margins that measured ≤125 μm at its longest dimension. A red spot measuring >125 μm in its longest dimension was defined as a retinal hemorrhage unless it had a central light reflex and sharp, smooth margins. All other intraretinal red spots were considered retinal hemorrhages. In this global category, the grader identified individual microaneurysms, evaluated portions of the seven fields that overlap, and only counted each microaneurysm in the combined seven-field area once. The grading scale was: 1 = no evidence; 2 = questionable; 3 = number if ≤20; 4 = >20; and 5 = cannot be graded. "Cannot be graded" was used in this and in all other descriptions of lesions if the photographic quality was too poor to evaluate a lesion or if hemorrhage, proliferative tissue, or both obscured the photographic field.

Count of retinal hemorrhages per seven fields (ct. RH/global). This item included all intraretinal red spots

not captured by the above criteria. It therefore represents those lesions that did not meet the definition of a microaneurysm. This included blot, flame, and punctate hemorrhages. As above, the grader only counted a hemorrhage once, even if the lesion appeared in more than one photographic field. The grading scale was: 1 = no evidence; 2 = questionable; 3 = number if ≤20; 4 = >20; and 5 = cannot be graded.

Count of microaneurysms (ct. MA; fields 1-7). The definition of what constituted a microaneurysm was the same as in the global description above. However, in this case, each individual field was counted separately to avoid duplication caused by overlapping areas of different photographic fields. The scale was: 1 = no evidence; 2 = questionable; 3 = number if ≤20; 4 = >20; and 5 = cannot be graded.

Count of retinal hemorrhages (ct. RH; fields 1-7). This item was parallel to the count of microaneurysms, and the grading scale was the same: 1 = no evidence; 2 = questionable; 3 = number if ≤20; 4 = >20; and 5 = cannot be graded.

Total hemorrhages and microaneurysms (H/MA; fields 1-7). This grouping combined the two categories immediately above. However, the scale is much higher for this item: 1 = no evidence; 2 = questionable; 3 = number if ≤20; 4 = >20 but <100; 5 = >100 but ≤300; 6 = >300; and 7 = cannot be graded. These specific levels were chosen to correspond closely to the discontinuities between the ETDRS standard photographs 1, 2A, and 2B used in the xMAHC.[7]

Soft exudates (SE; fields 1-7). These were superficial pale yellow to white areas with ill-defined, feathery margins indicative of swelling of the nerve cells. The area covered by the lesion(s) was measured with a transparent graphic to determine size. The grading scale was: 1 = no evidence; 2 = questionable; 3 = area smaller than a circle with a diameter measuring 500 μm; 4 = area greater than or equal to a circle with a diameter measuring 500 μm but less than a circle with a diameter measuring 1500 μm; 5 = area greater than or equal to a circle with a diameter measuring 1500 μm; 6 = cannot be graded. These levels represent measurements derived from the anatomic features found in the ETDRS standard photographs 8A and 5.

Intraretinal microvascular abnormalities (IRMA; fields 2-7). These were intraretinal vascular segments of various widths that were convoluted (tortuous) and often appeared as early neovascular growths within the body of the retina. If there were many widely separated fragments, the grader was instructed to combine the areas of IRMA, that is, to measure each area separately and combine the results. Definite neovascularization was not included. No attempt was made to compress the lesion's size or remove the areas of noninvolvement between the segments of individual continuous lesions; therefore, these measurements represent the surface area of the retina involved with these lesions. The grading scale was: 1 = no evidence; 2 = questionable; 3 = area less than a circle measuring 500 μm in diameter; 4 = area greater than or equal to a circle with a diameter measuring 500 μm but less than a circle with a diameter measuring 1500 μm; 5 = area greater than or equal to a circle with a diameter measuring 1500 μm; 6 = cannot be graded. Once again, these levels represented the use of anatomic measurements when applied to the features seen in the ETDRS—in this case, the standard photographs 8A and 5.

Venous beading (VB; fields 3-7). This lesion appeared as localized areas of diffuse increases in venous caliber in which swelling extended evenly on all sides of the vessel. The total number of venous "beads" was counted. The grading scale was: 1 = no evidence; 2 = questionable; 3 = <10; 4 = ≥10 <30; 5 = ≥30; 6 = cannot be graded. These intervals corresponded closely to those seen in ETDRS standard photographs 6A and 6B.

Hard exudates (HE; fields 2-7). These lesions appeared to be yellow to white intraretinal deposits with sharp margins. Often, they appeared to be shiny or waxy and deeper than soft exudates. They could be small, irregular dots or large deposits forming partial or complete rings in areas of retinal edema. The grading scale was: 1 = no evidence; 2 = questionable; 3 = number if <8; 4 = ≥8 but <16; 5 = ≥16 but <80; 6 = ≥80; and 7 = cannot be graded. These levels correspond closely to those of ETDRS standard photographs 3, 4, and 5.

Preretinal hemorrhage (PRH; fields 1-7). Any hemorrhage lying on the retinal surface or in the subhyaloid space was measured in this category. The grading scale was: 1 = no evidence; 2 = questionable; 3 = area less than a circle with a diameter measuring 1500 μm; 4 = area greater than or equal to a circle with a diameter measuring 1500 μm but less than half the photographic field; 5 = area greater than half the photographic field but less than the entire photographic field; 6 = cannot be graded. The levels for this lesion are a combination of the conventional anatomic measure and the ETDRS standard photographs 9 and 13.

Vitreous hemorrhage (VH; fields 1-7). This category included any hemorrhage that has invaded the vitreous cavity. The grading scale was the same as for Preretinal Hemorrhage. In this case, the grading scale was identical to that of the xMAHC.

Neovascularization elsewhere (NVE; fields 1-7). New vessels extending onto or above the surface of the retina were graded as neovascularization elsewhere, unless they were on or within 1500 μm of the margin of the optic disk. If the new vessels were widely separated, the grader was instructed to combine them in

the manner described under IRMA. Compression techniques that would have reduced the size of the involved surface areas were not used. The grading scale was: 1 = no evidence; 2 = questionable; 3 = area less than half a disk area; 4 = greater than or equal to half a disk area but less than four disk areas; 5 = area greater than four disk areas; 6 = cannot be graded. This scale conforms closely to that of the xMAHC using standard photograph 7.

Fibrous proliferation elsewhere (FPE; fields 2-7). Fibrous tissue made of multiple fine lines that are opaque and are on or above the surface of the retina were graded in this category. Fibrous proliferation on or within 1500 μm of the disk margin was excluded. As with IRMA and neovascularization, the grader had to combine widely dispersed fibrous tissue. The grading scale was: 1 = no evidence; 2 = questionable; 3 = area less than half a disk area; 4 = greater than or equal to half a disk area but less than $2^1/_2$ disk areas; 5 = area greater than $2^1/_2$ disk areas; 6 = cannot be graded. This scale was identical to that of the xMAHC because standard photograph 11 in that study was found to have fibrous tissue (excluding that within 1500 μm of the disk) covering approximately $2^1/_2$ times the area of the average disk.

Neovascularization on disk (NVD; field 1). New vessels on or within 1500 μm of the disk margin were recorded here. The grading scale was: 1 = no evidence; 2 = questionable; 3 = area less than one third the disk area; 4 = greater than or equal to one third the disk area but less than four disk areas; 5 = area greater than four disk areas; 6 = cannot be graded. These chosen intervals combine the ETDRS standard photographs 10A and 10C and the anatomic measures.

Fibrous proliferation on disk (FPD; field 1). The definition and grading scale was the same as for Fibrous Proliferation Elsewhere. However, this item captured that fibrous tissue on or within 1500 μm of the margin of the disk. This scale corresponds to Standard Photograph 10B of the xMAHC.

Clinically significant macular edema (CSME; global). For this study, the separate measurements made of retinal edema in field 2 in the xMAHC have been limited to the definitions of Clinically Significant Macular Edema, as defined by the ETDRS. Thus, this item was described as definitely present if one or more of the following conditions was met: thickening of the retina at or within 500 μm of the center of the macula; hard exudates at or within 500 μm of the center of the macula if associated with thickening of adjacent retina; a zone or zones of retinal thickening 1 disk area or larger, any part of which was within 1 disk diameter of the center of the macula. The grading scale for this item was merely: 1 = no evidence; 2 = questionable; 3 = definitely present.

Evaluation of Reproducibility

For each of the above lesions, a summary score could be calculated to yield a single value for the entire eye. For features measured in more than one field, the method described in ETDRS Report No. 10 was used. This procedure assigns a maximum grade and counts the number of nonoverlapping fields 3 to 7 in which that grade occurs.[7] Although results from fields 1 and 2 could be used to establish the maximum grade, they could not be included in the number of fields greater than 1. This procedure results in a five-step scale for each severity grade, which is then reduced to a three-step scale of: maximum grade present in one field (max/1 fld); maximum grade present in two or three fields (max/2–3 flds); and maximum grade present in four or in five fields (max/4–5 flds). This produces summary scales of differing lengths, depending on the number of steps in the lesion grading scale. For example, a count of microaneurysms can yield the following summary scales: 1; 2/1; 2/2–3; 2/4–5; 3/1; 3/2–3; 3/4–5; 4/1; 4/2–3; 4/4–5; 5/1; 5/2–3; 5/4–5. For analysis, a difference between 3/1 and 4/1 is considered a full step difference, whereas a difference between 3/1 and 3/2–3 represents only a one-third step difference. For lesions measured in only one field, however, the score served as the summary grade.

The lesions' summary grades are used to generate single retinopathy severity levels for each eye. For this report, the ETDRS Interim Scale[8] was modified for use with these summary grades. This severity scale is presented in Table 1. Levels are determined by the presence of specified maximum grades/number of fields for particular lesions.

In this type of comparison of categoric scales between expert graders, the percentage of perfect agreement can be misleading if a single score (for example, "no evidence") predominates. Thus, although agreement and selected amounts of disagreements are useful, kappa and weighted kappa[9–11] statistics show the general level of agreement across the entire range of scores. Weighted kappa was developed to recognize that disagreements of single steps may not be as serious as disagreements of two or more steps. For this reason, different weights are assigned to different amounts of disagreement. Similarly, disagreements between "no evidence" and "questionable" are given less weight than all other disagreements. In this study, the method of the ETDRS Report No. 10 was used; disagreements between "no evidence" and "present" received a weight of 0 (no credit), whereas disagreements between "no evidence" and "questionable" received a weight of .75 (partial credit). In all other cases, the weights were as follows: 1.0 for perfect agreement (full credit); 0.75

TABLE 1. Modified ETDRS Interim Retinopathy Severity Scale

Level	*Definition*
10	Ma and all other lesions absent
14	HE, SE, or IRMA definite, Ma absent
15	RH definite, Ma absent
20	Ma definite, no other lesions present
30	Ma plus SE, IRMA, or VB questionable; Ret. Hem. present, H/Ma < 5/1*; or HE definite
	Levels 41 and above require Ma ≥ 3/1
41	IRMA ≥ 3/1-3; or SE ≥ 3/1-3
45	IRMA ≥ 3/4-5; or SE ≥ 3/4-5; or VB definite; or H/Ma ≥ 5/1-3
51	H/Ma ≥ 5/4-5; or VB ≥ definite/2-3; or Combination of SE ≥ 4/4-5, IRMA ≥ 3/2-3, and H/Ma ≥ 5/1
55	IRMA ≥ 4/2-3; or VB ≥ def/2-3, plus 2 other P2 lesions†; or 4 P2 lesions; or H/Ma ≥ 5/4-5 plus 2 other P2 lesions
61	FPE, or FPD definite, NVE or NVD absent; or NVE definite NVE definite
65	Either NVE ≥ 4/1; or NVD ≥ 3/1; or VH or PRH ≥ 3/1, plus NVE ≥ 3/1
71	VH or PRH ≥ 4/1; or NVE ≥ 4/1, and VH or PRH ≥ 3/1; or NVD ≥ 3/1, and VH or PRH ≥ 3/1; or NVD ≥ 4/1
75	NVD ≥ 4/1, and VH or PRH ≥ 3/1

Ma = Microaneurysms; HE = hard exudates; SE = soft exudates; IRMA = intraretinal microvascular abnormalities; RH = retinal hemorrhage; VB = venous beading; FPE = fibrous proliferation elsewhere; FPD = fibrous proliferation on disc; NVE = neovascularization elsewhere; NVD = neovascularization on disc; VH = vitreous hemorrhage; PRH = preretinal hemorrhage.
* Numbers indicate summary scores of maximum grade/number of fields with maximum.
† $P2_2$ lesions = SE ≥ 3/2-3; IRMA ≥ 3/2-3; VB ≥ 3/2-3; H/Ma ≥ 5/1[8].

for one-step disagreement (partial credit), and 0 for all other disagreements (no credit).[7]

RESULTS

Retinopathy Level Per Eye

Tables 2 and 3 show the agreements for retinopathy levels per eye between graders and with the same grader over time, respectively. In these tables, the first column is the percent of perfect agreement, the second column reveals the percent agreement if one-level disagreements are included with perfect agreements. The next two columns show the unweighted and weighted kappa statistics. The pairs of graders agreed exactly with each other 95.77% of the time. When disagreements of only one level were added, the agreement was 98.65%. Unweighted kappa was .9428 and weighted kappa was .9671. Table 3 presents the agreement on retinopathy level between the two gradings over time by the same grader. The graders agreed exactly with their earlier scoring 88.04% of the time, and agreement went up to 94.79% when disagreements of one level were added.

TABLE 2. Retinopathy Level, Per Eye: First Graders Versus Second Graders (n = 4657)

*Agreement**		*Kappa*	
Perfect	*Within One Level*	*Unweighted (SE)*	*Weighted*
95.77	98.65	.9428 (0.004)	.9671

* Percent of eyes with specified levels of agreement.

Individual Lesions

Table 4 shows the results of comparing the scores of different graders for each individual lesion. These results include comparisons of all possible combinations of pairs of the four graders. The first two columns list the percent of the eyes with perfect agreement between two graders, with the first column reporting those eyes for which a lesion was graded absent by both graders (A/A) and the second column reporting perfect agreement on all other scores.

For lesions measured in multiple fields, the next two columns display the percent of eyes with agreements within a one-third step on the summary score scale and within one full step, respectively. A full step in the summary scales for lesions measured in multiple fields is defined as a change in the maximum grade with the same number of fields, for example, grade 3 in one field to grade 4 in one field, or grade 3 in two to three fields to grade 4 in two to three fields. Within each step, there are three increments, representing an increasing number of fields in which the maximum grade is reported. Thus, a one-third step is defined as the difference between any two of these increments.

For global measurements and those measured

TABLE 3. Retinopathy Level, Per Eye: Time 1 Versus Time 2, Same Grader (n = 652)

*Agreement**		*Kappa*	
Perfect	*Within 1 Level*	*Unweighted (SE)*	*Weighted*
88.04	94.79	.8394 (0.017)	.8904

* Percent of eyes with specified levels of agreement.

TABLE 4. Intergrader Reliability for Individual Lesions

	Maximum Grade/Number of Fields With Maximum					
	Complete Agreement		*Agree Within*	*Agree Within*	*Kappa*	
Lesion	*A/A**	*Other*	*1/3 Step*	*1 Full Step*	*Unweighted (standard error)*	*Weighted*
			Lesions Measured in Multiple Fields			
Ma	33.20	49.52	97.40	99.33	0.7784 (0.007)	0.9201
RH	40.83	42.55	97.94	99.40	0.7689 (0.007)	0.9177
H/Ma	31.22	51.19	97.51	99.68	0.7810 (0.007)	0.9263
SE	72.66	17.26	97.04	99.51	0.7606 (0.010)	0.8853
IRMA	78.70	11.90	97.12	99.40	0.7187 (0.011)	0.8560
VB	98.17	0.56	99.53	99.85	0.5088 (0.051)	0.6799
HE	83.96	10.44	95.90	97.85	0.7899 (0.011)	0.8287
PRH	99.91	0.06	100.00	100.00	0.8571 (0.141)	0.9600
VH	99.94	0.04	100.00	100.00	0.7999 (0.182)	0.9412
NVE	99.08	0.73	99.91	100.00	0.8869 (0.039)	0.9327
FPE	99.98	0.02	100.00	100.00	1.0 (0.000)	1.0
			Lesions Measured in One Field			
NVD	99.94	0.06		100.00	1.0 (0.000)	1.0
FPD	100.00	0		Not present in sample		
CSME	95.85	2.62		99.81	0.7757 (0.025)	0.8626
			Global Lesions			
Ct. Ma	33.31	55.58		99.33	0.8408 (0.006)	0.8410
Ct. RH	41.97	47.34		99.51	0.8322 (0.007)	0.9396

N = 4657.
* A/A = both graders scored lesion as absent.
Ma = microaneurysms; HE= hard exudates; SE = soft exudates; IRMA = Intraretinal microvascular abnormalities; RH = retinal hemorrhages; VB = venous beading; FPE = fibrous proliferation elsewhere; FPD = fibrous proliferation on disk; NVE = neovascularization elsewhere; NVD = neovascularization on disk; VH = vitreous hemorrhage; PRH = preretinal hemorrhage.

only in one field, the summary scale is the same as the grading scale and, thus, is not split into one-third step increments.

In Table 4, the combined total of perfect agreement was more than 90% for all lesions, with totals for all lesions more than 95% and 97% when differences of one-third step and one full step were added, respectively. The unweighted kappa statistic was >.610 for all except venous beading lesions, and it was >.810 for nine of those lesions. The weighted kappa statistic was >.810 for all except venous beading lesions. Although several of the more severe lesions did not occur often, as shown by the large percent of eyes both graders scored as absent, unweighted and weighted kappa statistics indicated that agreement was high when the lesions did occur.

Table 5 compares the results of each grader's first score with the results of each grader's later scores concerning the same set of photographs. The first two columns show the percent of eyes with perfect agreement between two evaluations by the same grader; the first column reported those eyes for which that lesion was graded as absent both times, and the second column included agreement on all other scores. The third through sixth columns are the same as in Table 4.

In Table 5, the combined total of perfect agreement was >76% for all lesions, with totals for all lesions >94% and 96% when differences of one-third step and one full step, respectively, were added. The unweighted kappa statistic was >.610 for all except venous beading and fibrous proliferation elsewhere lesions, with two lesions scoring >.810. The weighted kappa statistic was >.810 for 9 of 13 lesions graded as present, with all except venous beading and fibrous proliferation elsewhere lesions >.610. Two of the lesions present in the larger sample (one preretinal hemorrhage and one vitreous hemorrhage) were graded as absent both times and, thus, were not represented in the subsample.

DISCUSSION

Various measures of reproducibility have been presented because each has its strengths and weaknesses.

TABLE 5. Intragrader Reliability for Individual Lesions

	Maximum Grade/Number of Fields With Maximum					
	Complete Agreement		*Agree Within*	*Agree Within*	*Kappa*	
Lesion	*A/A**	*Other*	*1/3 Step*	*1 Full Step*	*Unweighted (standard error)*	*Weighted*
			Lesions Measured in Multiple Fields			
Ma	33.79	44.95	94.19	96.94	0.7243 (0.020)	0.8634
RH	45.41	33.33	95.72	97.25	0.6907 (0.022)	0.8633
H/Ma	28.75	47.86	95.26	97.55	0.7091 (0.020)	0.8753
SE	73.09	17.13	96.33	98.62	0.7673 (0.025)	0.8574
IRMA	81.65	8.87	96.18	98.78	0.6618 (0.035)	0.7962
VB	97.86	0.31	99.24	99.85	0.3948 (0.104)	0.5503
HE	84.1	9.79	95.87	97.09	0.7673 (0.032)	0.8219
PRH	100.00	0		Not present in subsample		
VH	100.00	0		Not present in subsample		
NVE	98.78	0.61	99.69	100.00	0.7312 (0.089)	0.8691
FPE	99.69	0	99.85	100.00	0.0 (0.000)	0.0
			Lesions Measured in One Field			
NVD	99.39	0.31		100.00	0.7486 (0.125)	0.7488
FPD	100.00	0		Not present in subsample		
CSME	96.48	2.29		99.85	0.7954 (0.071)	0.9754
			Global Lesions			
Ct. Ma	32.74	54.43		97.40	0.8135 (0.018)	0.8077
Ct. RH	45.11	43.88		97.25	0.8226 (0.019)	0.8238

N = 652.
* A/A = both graders scored lesion as absent.
Ma = microaneurysms; HE = hard exudates; SE = soft exudates; IRMA = intraretinal. microvascular abnormalities: RH = retinal hemorrhages; VB = venous beading; FPE = fibrous proliferation elsewhere; FPD = fibrous proliferation on disk; NVE = neovascularization elsewhere; NVD = neovascularization on disk; VH = vitreous hemorrhage; PRH = preretinal hemorrhage.

When a feature does not occur often, a high percentage of perfect agreement may be misleading if there is disagreement when the feature does occur. The kappa statistic tends to control for this bias. Perfect agreement and unweighted kappa, however, treat large and small disagreements similarly. In any sort of skilled human activity, such as grading diabetic retinopathy, slight disagreements about borderline cases occur more often than large disagreements if the classification system is reliable. Thus, by reporting agreements with one-third step, one full step, and one level, and by use of the weighted kappa statistic, these minor disagreements can be distinguished from severe disagreements. Percent agreement, whether perfect or within some range, is simple to interpret. Unweighted and weighted kappas are less straightforward. Landis and Koch[11] recommend the following scale for unweighted kappa statistics: 0 to 0.20 = slight agreement; 0.21 to .40 = fair agreement; 0.41 to 0.60 = moderate agreement; 0.61 to 0.80 = substantial agreement; and >0.81 = almost perfect agreement. ETDRS Report No. 10 recommended that this scale be extended to the weighted kappa statistic.[7]

Some of the lesions associated with more severe diabetic retinopathy were uncommon in the population studied. However, the current focus of clinical research in diabetic retinopathy appears to be aimed at a more precise evaluation of nonproliferative lesions. For this reason, no special effort was made to add photographs with more severe retinopathy. To make clear to the reader how common a given lesion was in this sample, the proportion of photographs where the grades were reported as "no evidence" for each individual lesion is provided in the first columns of Tables 4 and 5.

As shown in Tables 2 through 5, the Vanderbilt Classification System has a consistently high reproducibility between graders and over time on all measures reported. Only two types of lesions, venous beading and fibrous proliferation elsewhere, when compared over time, dropped below the "substantial agreement" level on unweighted kappa. Fourteen of the 15 lesions present in the original sample had weighted kappas in the "almost perfect agreement" category, whereas 11 of 15 lesions represented in the subset graded twice by the same graders had weighted kappas

in the "almost perfect" range. For all the individual lesions measured in both the xMAHC and the Vanderbilt systems, kappa and weighted kappa values were higher for the Vanderbilt system when compared to ETDRS Report No. 10.[7] Thus, the Vanderbilt Classification System has been more reliable and reproducible in this type of evaluation of early diabetic retinopathy.

An examination of the data (Tables 2 and 3) implies that the Vanderbilt Classification System is more reliable and more reproducible than the system used in the ETDRS. That may be true, but a test to evaluate that feature was not included in the original research plan. Because this study's population sample had a more restricted pathologic focus, that may not be an appropriate comparison. Nevertheless, the most important feature is that the "weighted" and "unweighted" kappa values are well within the desired ranges for research such as this. The agreement levels between graders is better than the agreement levels found with the same grader over time. Because all four graders are on the same Quality Assessment cycle and participate in a continuous educational program, all measured improvements should parallel one another. As stated earlier, the most important feature is that the "weighted" and "unweighted" kappa values of each analysis were within the desired range.

All systems for classifying diabetic retinopathy rely on skilled human behavior and judgment. The Modified Airlie House Classification system requires trained graders to compare new photographs of patients to a set of standard photographs. The Vanderbilt Classification System depends on trained graders to count and measure individual lesions and to resolve disagreements by simple measurements. The increased benefit to research of this quantitative system is that it permits data to be gathered on specific incremental changes in lesions. It is easy to train graders to use this system in a reliable and reproducible manner. It should be possible for a wide range of investigators to use this technique without the need to compare each set of photographs to a series of standardized photographs. In addition, this quantitative system can be used to program computerized expert systems to screen digitized images for diabetic retinopathy.

Key Words

diabetic retinopathy, quantitative analysis, fundus photographs, Vanderbilt classification system, progression of diabetes

References

1. The Diabetic Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin dependent diabetes mellitus. *N Engl J Med.* 1993;329:977–986.
2. Diabetic Retinopathy Study Research Group. Photocoagulation treatment of proliferative diabetic retinopathy. *Ophthalmology.* 1978;85:82–106.
3. Early Treatment Diabetic Retinopathy Study Group. Photocoagulation of diabetic macular edema: Early treatment: ETDRS Report Number 1. *Arch Ophthalmol.* 1985;103:1796–1806.
4. Klein R, Klein BEK. Vision disorders in diabetes. In Hammon R, Harris MWH, eds. *Diabetes in America.* Bethesda, MD: National Institutes of Arthritis, Diabetes, Digestive and Kidney Diseases; 1985:1–36. Washington, DC: US Dept Health and Human Services; NIH Publication 85-1468.
5. Davis MD, Norton EWD, Meyers FL. The Airlie classification of diabetic retinopathy. In: Goldberg MF, Fine SL, eds. *Symposium on the Treatment of Diabetic Retinopathy.* Washington, DC: US Government Printing Office; 1969:7–22. USPHS Publication #1890.
6. The Diabetic Retinopathy Study Research Group. A modification of the Airlie House classification of diabetic retinopathy: DRS Report 7. *Invest Ophthalmol Vis Sci.* 1981;21:210–226.
7. Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs: An extension of the modified Airlie house classification: ETDRS Report Number 10. *Ophthalmology.* 1991;98:786–806.
8. Early Treatment Diabetic Retinopathy Study Research Group. Fundus photographic risk factors for progression of diabetic retinopathy: ETDRS Report Number 12. *Ophthalmology.* 1991;98:823–833.
9. Cohen J. A coefficient of agreement for nominal scales. *Educational Psychology Measurement.* 1960;20:37–46.
10. Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull.* 1968;33:213–220.
11. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159–174.