Comparability of Proptosis Measurements by Different Techniques

MARIA SEGNI, MD, GEORGE B. BARTLEY, MD, JAMES A. GARRITY, MD, ERIK J. BERGSTRALH, AND COLUM A. GORMAN, MB, BCH, PHD

• PURPOSE: To compare measurements of proptosis obtained by clinicians and computed tomography.
• DESIGN: Cohort study.
• METHODS: In a prospective randomized study of orbital radiotherapy for Graves' ophthalmopathy, measurements of proptosis were made on the same visit by an endocrinologist and an ophthalmologist using the Krahn exophthalmometer and by a technician using orbital computed tomography (CT) scans taken with head fixation to minimize position artifact.
• RESULTS: Both clinical observers recorded proptosis measurements that were greater by 0.6 to 1.6 mm than those observed on the CT scan. This discrepancy resulted in part from the clinical measurements being made to the anterior corneal surface, whereas the CT measurements were made to the posterior corneal surface (a difference of approximately 0.5 mm). The aggregated observations of the clinicians did not vary significantly from each other but wide discrepancies (as much as 5 mm) were noted between single measurements made on the same patient and on the same day by different clinicians.
• CONCLUSIONS: The degree of variance observed in clinical measurements emphasizes the importance of defining reproducibility of the measurement techniques in prospective studies of therapeutic efficacy in patients with Graves' ophthalmopathy. The systematic difference between CT and clinical measurements of proptosis should be noted when results of clinical trials are compared. (Am J Ophthalmol 2002;133:813–818.

Accepted for publication Feb 7, 2002.
InternetAdvance publication at ajo.com March 7, 2002.
From the Department of Ophthalmology, La Sapienza University, Rome, Italy (M.S.); Departments of Ophthalmology (G.B.B., J.A.G.), and Health Sciences Research (E.J.B., C.A.G.), Mayo Clinic and Mayo Foundation, Rochester, Minnesota.
Supported in part by Research to Prevent Blindness, Inc. New York, NY
Correspondence to George B. Bartley, MD, Department of Ophthalmology, Mayo Clinic, 200 First Street, SW, Rochester, MN, 55905; fax: (507) 284-4612; e-mail: gbartley@mayo.edu

REPORTS OF TREATMENT OF GRAVES' OPHTHALMOPAthy describe the severity of the disease process using a variety of measurements. It is important to determine if different clinical techniques that purport to measure the same variable give different results and to define the degree of interobserver variance in each type of measurement.

In this study, we examined proptosis measurements that were made clinically and by computed tomography (CT) in a group of 42 patients with Graves' ophthalmopathy who were examined regularly over a 3-year period as part of a study of orbital radiotherapy.[1] It is hoped that the analysis will aid readers to compare results from groups of patients in various studies in which proptosis has been recorded by different methods.

METHODS

FORTY-TWO PATIENTS WITH MODERATE GRAVES' OPHthalmopathy were examined at 3-month intervals for 1 year and again 2 years later during an orbital radiation therapy program for Graves' ophthalmopathy. On each visit, measurements of orbital volume, proptosis, and the volume of retrobulbar muscle and fat were made on CT scans as primary study endpoints.

Proptosis was independently measured by two experienced clinical observers (an ophthalmologist and an endocrinologist) who used a Krahn exophthalmometer and who recorded their results without knowledge of the other clinician's findings or of the CT results. Two ophthalmologists (G.B.B. and J.A.G.) and three endocrinologists were involved in the study, although one of the endocrinologists (C.A.G.) was responsible for more than 95% of the measurements. We chose the Krahn exophthalmometer (Marco Instrument Co, Jacksonville, Florida, USA) because it contains an internal correction for parallax and it has been the instrument of choice at our institution for many years. The measurements were taken with the patient's head in the primary position and the examiner's eyes at the same level as the patient's eyes. Several measurements to the nearest 1 mm were taken for each eye and the mode value was recorded.

0002-9394/02/$22.00
PII S0002-9394(02)01429-0

FIGURE 1. A plastic mask was used to ensure that head position did not change during computed tomography (CT) scan acquisition.

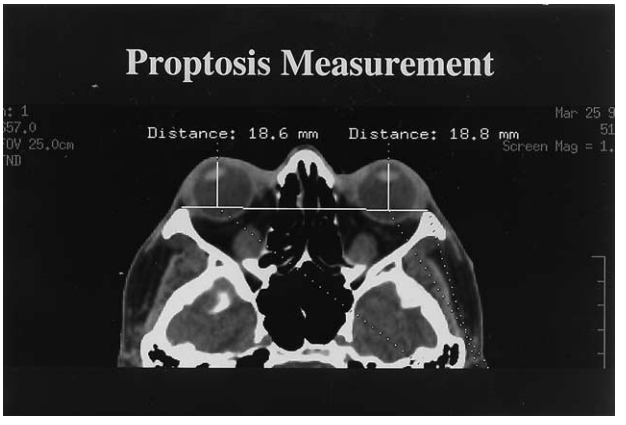

FIGURE 2. Technique for proptosis measurement on computed tomography (CT) scan. On the section that bisects the lens, a line is drawn between the lateral orbital rims. A perpendicular line is erected that passes through the lens and terminates at the posterior surface of the cornea. The length of the perpendicular line is recorded as the proptosis measurement.

Computed tomographic scans were done using 1.5 mm sections through the orbit on a GE9800 machine using Advantage software.[2,3] Each patient had a custom-made plastic mask that was used to secure the position of the head during radiotherapy. To ensure reliability of our CT measurements of proptosis, the same mask was used during diagnostic CT scan acquisition (Figure 1).

At each visit a measurement of proptosis was made on the CT scan by drawing a horizontal line between the lateral orbital rims on the CT section that bisects the lens and then drawing a perpendicular line forward to the posterior (interior) surface of the cornea (Figure 2). The posterior surface was chosen because on the CT images it is difficult at times to define the anterior (outer) surface of the cornea. Closed eyelids may be indistinguishable from the anterior corneal surface.

To assess the reproducibility of proptosis measurements by CT we studied 10 patients who were not included in the main study group. CT scans and proptosis measurements were performed as for the study patients. No mask was used for these patients since we were not attempting to reproduce the method over time. Using the same scan for each patient, we repeated the proptosis measurement by the same observer on three occasions on different days without reference to the previous measurements. Note that this does not account for variability that would be derived from repeated scan acquisition on the same patient. We did not consider it ethical to perform repeated head CT scans without a clinical indication.

The methods of Bland and Altman[4] were used to assess the agreement between proptosis readings done on the same visit by an ophthalmologist, by an endocrinologist, and by a CT technician.[1] Specifically, the three possible differences (ophthalmologist-CT [Ophth-CT], endocrinologist-CT [Endoc-CT], ophthalmologist–endocrinologist [Ophth–Endoc]) were calculated for each of the 440 readings (42 patients × at least 5 visits × 2 eyes). Bias between readers was assessed using the one-sample *t* test (for null hypothesis of true mean difference = 0) for each of the three comparisons. The potential association of the bias with the mean level was assessed by plotting the difference in proptosis readings vs. the average value. Simple linear regression analysis and Pearson correlation (r) were used to test whether the difference was significantly associated with the mean level. Differences in readings done in different months and on different eyes were assumed to be independent.

Reproducibility of three repeated proptosis readings of the same CT film in 10 patients was assessed using the intraclass correlation or reliability coefficient (R).[5] The reliability coefficient ranges from 0 to 1 and is a measure of the proportion of total variability due to differences between patients and eyes within patients. This implies that 1-R is the proportion of total variability due to repeat readings of the same scan. Hence, large values of R indicate high reproducibility in repeat readings of the same scan. The reliability coefficient was estimated by modeling the total variability (using nested random effects analysis of variance [ANOVA]) as the sum of components due to patients, eyes within patients, and readings within eyes within patients.[6]

RESULTS

OVERALL THE MEAN (STANDARD DEVIATION [SD]) PROPTOsis by CT, ophthalmologist, and endocrinologist was 21.3(2.8) mm, 22.6(2.9) mm, and 22.3(3.0) mm, respectively (Table 1). Differences between readers were highest for Ophth–CT at 1.2(1.2) mm and Endoc–CT at 1.0(1.6) ($P < .001$ for both). The average difference between

TABLE 1. Proptosis Readings by CT Technician, Ophthalmologist, and Endocrinologist

	No. Readings*	Mean (SD)	*P* Value
Proptosis by reader:			
CT technician	440	21.3 (2.8)	—
Ophthalmologist	440	22.6 (2.9)	—
Endocrinologist	430	22.3 (3.0)	—
Reader differences:			
Ophthal-CT	440	1.2 (1.2)	<.001
Endoc-CT	430	1.0 (1.6)	<.001
Ophthal-Endoc	430	0.2 (1.5)	.006

CT = computed tomography.
*Readings from both eyes of 42 patients with five or six visits per patient.

TABLE 2. Distribution of Differences in Proptosis Readings

	Comparison, % of Readings		
Differences in Proptosis (mm)*	Ophthal-CT (n=440)	Endoc-CT (n=430)	Ophthal-Endoc (n=430)
≥5.0	—	0.5	0.9
4.0 to 4.9	1.1	1.4	1.4
3.0 to 3.9	7.3	7.2	5.6
2.0 to 2.9	19.8	17.4	9.5
1.0 to 1.9	29.3	25.8	20.3
0.1 to 0.9	24.3	21.4	0.2
0	4.8	4.4	32.1
−0.1 to −0.9	10.2	9.6	1.2
−1.0 to −1.9	2.3	7.9	20.2
−2.0 to −2.9	0.4	3.0	6.0
−3.0 to −3.9	0.5	0.9	1.9
−4.0 to −4.9	—	0.5	0.5
≥−5.0	—	—	0.2
Within ±1.9	70.9	69.1	74

CT = computed tomobraphy.
*Proptosis by CT measured to the closest 0.1 mm. Proptosis by clinicians measured to the closest whole mm.

clinicians (Ophth–Endoc) was less at 0.2(1.5) mm and was significant only when pooling all readings ($P = .006$). These differences were consistent over both visits and eyes (data not shown). Positive (and negative) differences of 3 mm or more were observed in 8.4% (and 0.5%) of Ophth–CT, 9.1% (and 1.4%) of Endoc–CT, and 7.9% (and 2.6%) of Ophth–Endoc comparisons (Table 2). Seventy four percent of the readings by clinicians agreed within (less than) 2 mm and 89.5% were within (less than) 3 mm. Differences in readings tended to show little or no association with average levels of proptosis values (Figures 3A, B, and C). Slightly stronger correlation ($r = .91$) was observed between the findings of the ophthalmologist and the CT scan than between endocrinologist and CT scan ($r = .86$) or than was noted between the results from the ophthalmologist and the endocrinologist ($r = .87$).

CT proptosis readings by patient, eye, and replication are graphed in Figure 4. It is obvious that variability between patients is large compared with variability within patients. The intraclass correlation for replicate readings of the same scan within an eye relative to the total variability was 0.988, indicating excellent reproducibility. The estimated variability (standard deviation) due to differences between patients was 2.12 mm, differences between eyes within patients was 1.07 mm, and differences between replicate readings within eyes and patients was 0.26 mm. The average range of readings within an eye was 0.42 mm, and in no case did the range of repeat readings within an eye exceed 0.7 mm.

DISCUSSION

OUR RESULTS SHOW THAT TWO EXPERIENCED CLINICAL observers who are endeavoring to make accurate clinical measurements of proptosis as part of a prospective study may record values that vary widely from each other and from CT-derived measurements. This variance persists throughout the full range of exophthalmometry measurements. To illustrate, exophthalmometry readings of 22 mm by the endocrinologist were recorded as 17 to 25 mm by the ophthalmologist whereas readings of 22 mm by the ophthalmologist were recorded as 19 to 25 mm by the endocrinologist. Only 74% of clinicians' readings agreed within (less than) 2 mm. These results suggest the limitations of clinical proptosis readings as outcome measures for clinical trials.

Other authors have compared the proptosis readings on CT scan with clinical exophthalmometer results using the Hertel or the Krahn instrument. Given-Wilson[7] found a correlation coefficient between CT and Hertel readings of 0.91. The Hallin[8] correlation coefficient for the same parameter was 0.73. In our study, the correlation coefficient was 0.91.

Gibson has defined the CT parameters for proptosis measurement.[9] In that study, the exterior surface of the cornea was used. We have found it difficult on CT images to distinguish between the cornea and the closed eyelids. Alternatively, the posterior surface of the cornea is always clearly visible so we chose that point to define the anterior margin of the proptosis readings on CT. Gibson found that repetitive measurements on the same slice evoked a total measurement error of 0.08 mm.[9] Our measurements are of comparable accuracy (Figure 4).

We found that proptosis readings from CT scans tended to be lower than the clinical exophthalmometer readings. Corneal thickness is approximately 0.5 mm.[10] The lower values on the computer scan are, in part, attributable to this observational difference. In addition, measurements of

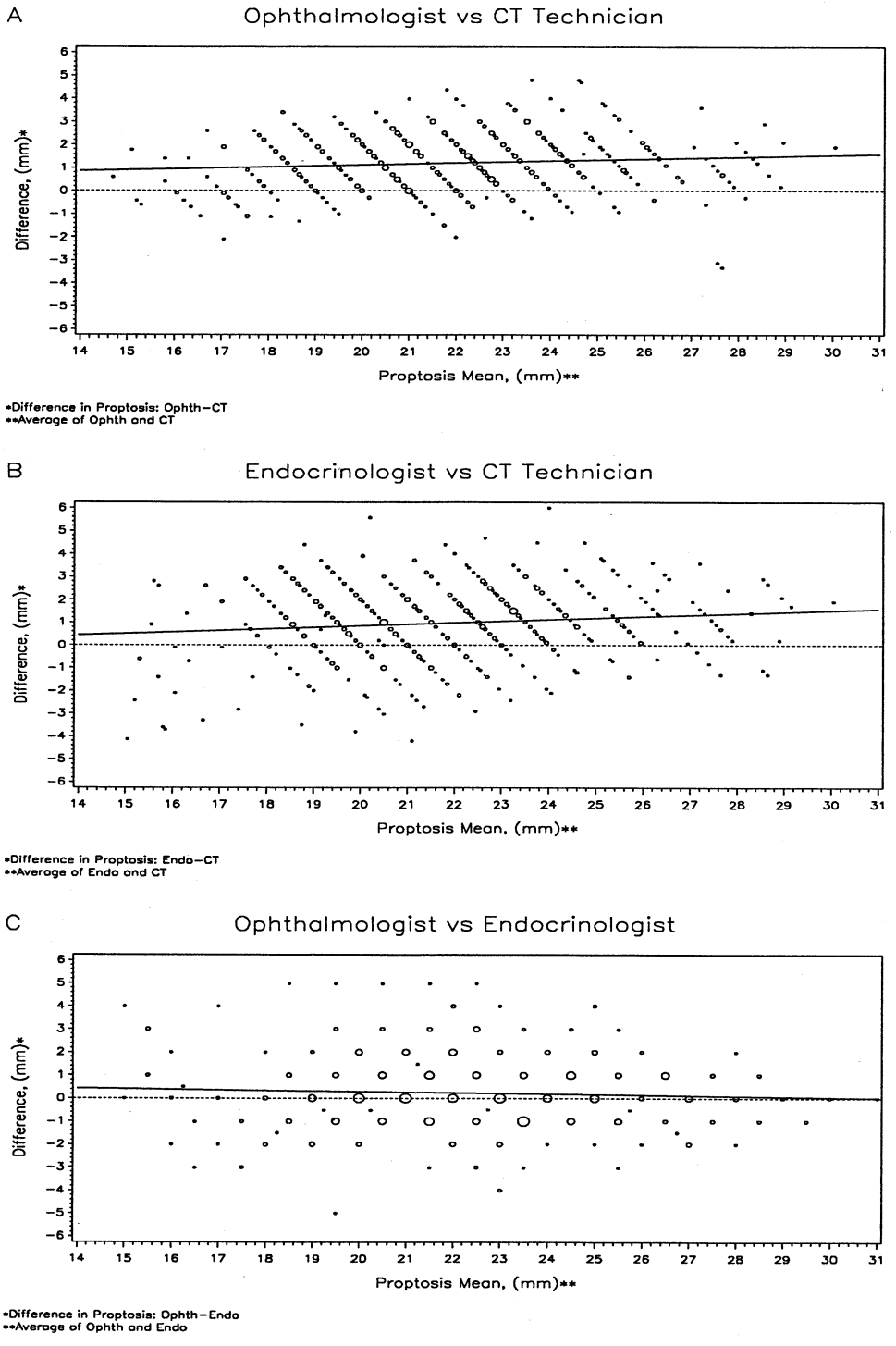

FIGURE 3. Plot of difference in proptosis between types of readers vs. mean value for readers: ophthalmologist vs. computed tomography (CT) technician (panel A), endocrinologist vs. computed tomography (CT) technician (panel B), and ophthalmologist vs. endocrinologist (panel C). In general, the difference between readers does not depend on the mean level, as evidenced by regression lines (solid) with near-zero slopes. Note that CT measurements are on average lower than clinician readings through the entire range of proptosis measurements. Further, agreement between clinicians is good and consistent through the entire range of proptosis measurements. The size of the plotting symbol is proportional to the number of readings plotted at that position.

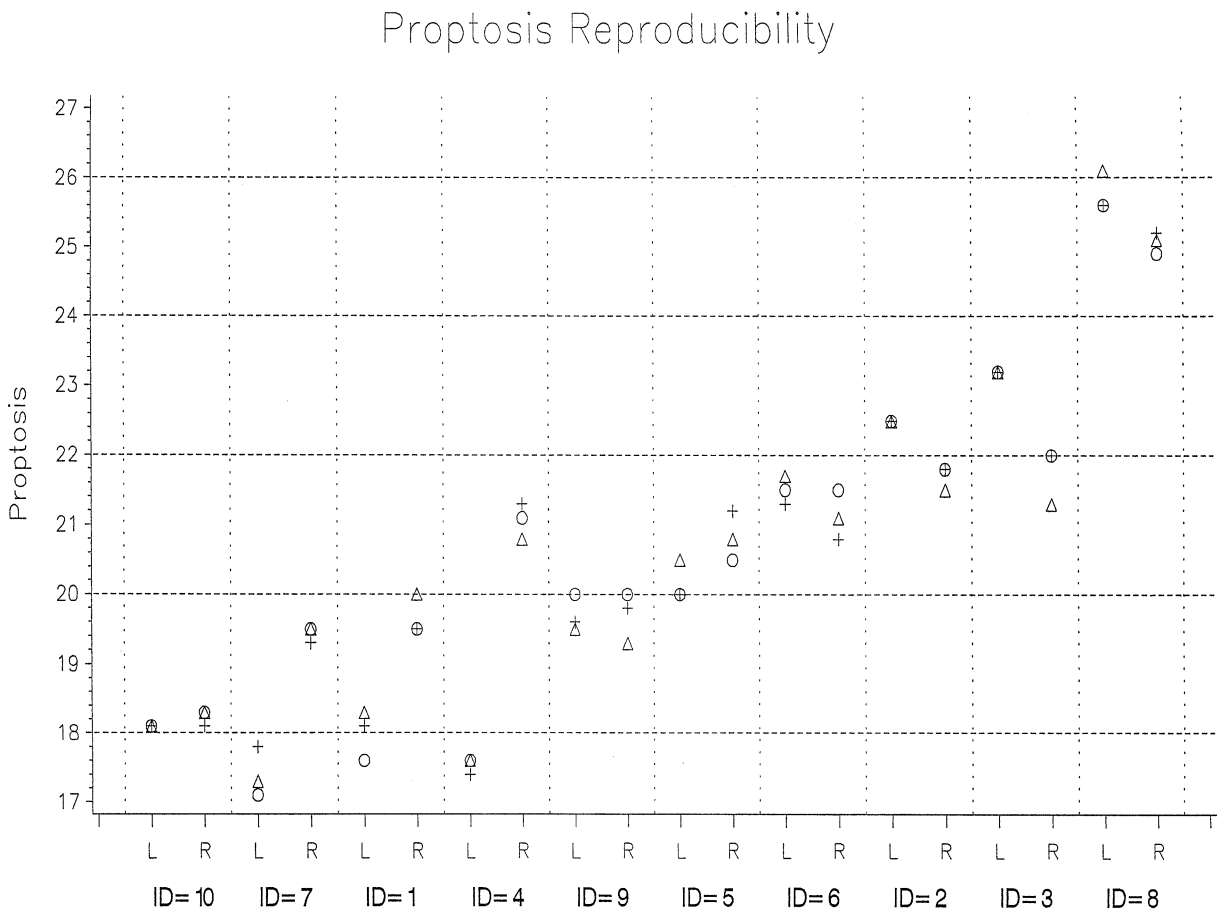

FIGURE 4. Proptosis readings reproducibility on the same computed tomography (CT) scan in 10 patients. The plotting symbols are plus for day 1, circle for day 2, and triangle for day 3. Note the high reliability coefficient for repeat readings within an eye (0.988).

proptosis on CT scan are made with the patient supine. Exophthalmometer readings are made in the seated or standing position. Postural proptosis (ranging from 0.7–1.7 mm) when patients are in the standing position has been described by Duke-Elder and MacFaul.[11]

The technique used to estimate proptosis from the same CT scan was found to be highly reproducible (R = 0.988). While this finding is encouraging, it does not assess variability between technicians or variability due to repeated scan acquisitions from the same patient. We had two additional advantages in achieving reproducible CT results: the patient's head position was fixed in a plastic face mask for each CT scan, and the same technologist measured volumes and proptosis on all scans.

Measurements of proptosis using various instruments have been made since at least 1865.[12–14] A detailed study of the principles of exophthalmometry was conducted by Davanger,[15] who devised an instrument that corrects for parallax and achieved standard error of his measurements of ± 0.3 mm. The Krahn exophthalmometer that we used also corrects for parallax, but the Hertel instrument embodies no parallax correction. Musch and associates,[16] using the Hertel exophthalmometer, demonstrated systematic variances between readings by different observers in an ophthalmic clinic and found that more than 25% of the observations were greater than 1 mm different from the readings of the senior observer, which were taken as the gold standard.

The definition of an abnormal range for proptosis measurements is also problematic. Ethnic differences have been described by Barretto and Mathog,[17] who confirmed the observations of earlier workers[18] that Black patients have greater ocular protrusion than do White patients. Exophthalmometry readings in Asians are lower than in Whites.[19] Values for normal children have been reported by Gerber[20] and by Nucci.[21]

The wide variability in clinical readings of proptosis made by experienced clinicians gives cause for concern about the sensitivity of clinical proptosis measurements as indicators of effective therapeutic interventions. In light of the variance shown by different observers whose readings were recorded within several hours of each other, it is probable that readings separated by several months will show even greater variance.

We propose that studies claiming to prove efficacy of a specific therapy in reducing proptosis should be accompanied by data showing the variance of proptosis readings among and within the study observers. Despite the expense, CT readings of proptosis, muscle volume, and fat volume may be needed if subtle improvement or regression in Graves' ophthalmopathy is to be demonstrated.[22] Authors using CT measurements should specify if their proptosis readings are to the inner or outer corneal surface. Comparisons of results between institutions should be made with specific knowledge of the CT techniques involved.

REFERENCES

1. Gorman CA, Garrity JA, Fatourechi V, et al. A prospective, randomized, double-blind, placebo-controlled study of orbital radiotherapy for Graves' ophthalmopathy. Ophthalmology 2001;108:1523–1534.
2. Forbes G, Gorman CA, Brennan MD, Gehring DG, Ilstrup DM, Earnst F. Ophthalmopathy of Graves' disease: computerized volume measurements of orbital fat and muscle. Am J Neuroradiology 1986;7:651–656.
3. Forbes G, Gehring DG, Gorman CA, Brennan MD, Jackson IT. Volume measurement of normal orbital structures by computed tomographic analysis. Am J Neuroradiology 1985; 6:419–424.
4. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–310.
5. Fleiss JL. The design and analysis of clinical experiments. New York, NY: Wiley, 1986:11.
6. Snedecor GW, Cochran WG. Statistical methods. 6th edition. Ames, IA: Iowa State University Press, 1967:285.
7. Given-Wilson R, Pope RM, Michell MJ, Cannon R, Mc Gregor AM. The use of real-time orbital ultrasound in Graves' ophthalmopathy: a comparison with computed tomography. Br J Radiol 1989;62:705–709.
8. Hallin ES, Feldon SE. Graves' ophthalmopathy. II. Correlation of clinical signs with measures derived from computed tomography. Br J Ophthalmol 1988;72:674–677.
9. Gibson RD. Measurement of proptosis (exophthalmos) by computerized tomography. Ausralas Radiol 1984;28:9–11.
10. Liesegang TJ. Disorders of the cornea, conjunctiva and lens. In: Bartley GB, Liesegang TJ, editors. Essentials of ophthalmology. Philadelphia, PA: JB Lippincott, 1992:50.
11. Duke-Elder S, MacFaul PA. Lacrimal, orbital and para orbital diseases In: Duke-Elder S, editor. System of ophthalmology. St. Louis, MO: CV Mosby, 1974:781.
12. Cohn H. Jahrensberd Sehles Ges F Kultur 1865;43:156.
13. Bertelsen TI. On 720 measurements with Lueddes exophthalmometer. Acta Ophthalmol 1953;32:589–595.
14. Knudtzon K. On exophthalmometry: The result of 724 measurements with Hertel's exophthalmometer on normal adult individuals. Acta Psychiatr Neurol 1949;24:523–537.
15. Davanger M. Principles and sources of error in exophthalmometry. A new exophthalmometer. Acta Ophthalmol 1970;48:625–633.
16. Musch DC, Freuh BR, Landis JR. The reliability of Hertel exopthalmometry. Observer variation between physician and lay readers. Ophthalmology 1985;92:1177–1180.
17. Barretto R, Mathog RH. Orbital measurement in black and white populations. Laryngoscope 1999;109:1051–1054.
18. Migliori ME, Gladstone GJ. Determination of the normal range of exophthalmometric values for black and white adults. Am J Ophthalmol 1984;98:438–442.
19. De Juan E Jr, Hurley DP, Sapira JD. Racial differences in normal values of proptosis. Arch Internal Med 1980;140: 1229–1231.
20. Gerber FR, Taylor FH, deLevie M, Drash AL, Kenny FM. Normal standard for exophthalmometry in children 10–14 years of age. Relation to age, height, weight, and sexual maturation. J Pediatr 1972;81:327–329.
21. Nucci P, Brancato R, Bandello F, Alfarano R, Bianchi S. Normal exophthalmometric values in children. Am J Ophthalmol 1989;108:582–584.
22. Gorman CA. The measurement of change in Graves' ophthalmopathy. Thyroid 1998;8:539–543.