Published in final edited form as:

Invest Ophthalmol Vis Sci. 2009 February ; 50(2): 560–566. doi:10.1167/iov.08-1881.

Comparison of Optical Coherence Tomography Retinal Thickness Measurements in Diabetic Macular Edema with and without Reading Center Manual Grading from a Clinical Trials Perspective

Adam R. Glassman, M.S.,

Jaeb Center for Health Research, DRCR.net

Roy W. Beck, M.D.,

Jaeb Center for Health Research

David J. Browning, M.D.,

Charlotte Eye Ear Nose and Throat Assoc, PA

Ronald P. Danis, M.D., and

UW-Dept. of Ophthalmology & Visual Sciences

Craig Kollman, Ph.D.

Jaeb Center for Health Research, DRCR.net

for the Diabetic Retinopathy Clinical Research Network Study Group

Abstract

Purpose—To analyze the value of reading center error correction in automated Optical Coherence Tomography (OCT) retinal thickness measurements in eyes with diabetic macular edema (DME).

Methods—6,522 Zeiss Stratus OCT scans obtained in 7 DRCR.net studies were analyzed. The Reading Center evaluated whether the automated center point measurement appeared correct, and when not, it was manually measured using calipers. Center point standard deviation as a percentage of thickness, center point thickness, signal strength, and analysis confidence were evaluated for their association with an automated measurement error (manual measurement needed and exceeded 12% of automated thickness). Curves were constructed for each factor by plotting the error rate against the proportion of scans sent to the reading center. The impact of measurement error on interpretation of clinical trial results and statistical power was also assessed.

Results—Standard deviation was the best predictor of an automated measurement error. The other three variables did not augment the ability to predict an error using standard deviation alone. Based on standard deviation, an error rate of 5% or less could be achieved by sending only 33% of scans to the reading center (those with a standard deviation $\geq 5\%$). Correcting automated errors had no appreciable effect on the interpretation of results from a completed randomized trial and had little impact on a trial's statistical power.

Corresponding Author: Adam R. Glassman, MS, c/o Jaeb Center for Health Research, 15310 Amberly Drive, Suite 350, Tampa, FL 33647; Phone: (813) 975-8690, Fax: (800) 816-7601, DRCR.netX3@jaeb.org.

A current list of the Diabetic Retinopathy Clinical Research Network is available at www.drcr.net

There are no conflicts of interest.

An address for reprints will not be provided.

Conclusions—In DME clinical trials, the error involved with using automated Stratus OCT center point measurements is sufficiently small that results are not likely to be affected if scans are not routinely sent to a reading center, provided adequate quality control measures are in place.

Introduction

Optical coherence tomography (OCT) is a noninvasive method for measuring the thickness of the central retina. It has become a standard tool in the management of patients with diabetic macular edema (DME). The Zeiss Stratus OCT (Carl Zeiss Meditec, Dublin, California) provides automated measurements of mean retinal thickness of the center point, central subfield, and each of 4 inner and 4 outer subfields (in microns), and total macular volume (mm^3). The OCT software identifies the center point as the intersection of 6 radial lines and reports the center point thickness as the average of the 6 measurements of thickness at the center point. A standard deviation of the 6 measurements is computed.

In the DME studies conducted by the Diabetic Retinopathy Clinical Research Network (DRCR.net), OCT scans have been sent to a central reading center for evaluation of quality, morphology, and accuracy of automated thickness measurements. DRCR.net has adopted use of the central subfield for analysis of central retinal thickness data rather than center point because of reduced variability in the measurement of the former compared with the latter.^{1, 2} However, only center point thickness can be easily manually measured when the automated measurement is incorrect. To limit missing data for scans with incorrect automated measurements, the center point manual measurement can be used to compute a value for the central subfield since the two are highly correlated.^{1, 3}

Since most scans in DME have accurate automated measurements, we questioned the cost-effectiveness of sending scans to a reading center to correct errors in the automated measurements compared with just using the automated measurements for analysis of clinical trial data. We used data collected in DRCR.net protocols to address this issue.

Methods

The study included 6,522 OCT scans at baseline and follow up from study eyes obtained in seven DRCR.net studies of DME. OCT operators were certified by the DRCR.net Reading Center at the University of Wisconsin, Madison. Some protocols required that central foveal edema was present in the study eyes whereas others did not. In each study, Stratus OCT was obtained following pupil dilation. Scans were 6 mm length and included the 6 radial line pattern (128 scan resolution, fast macular scan option with Stratus OCT) for quantitative measures and the cross hair pattern (6-12 to 9-3 o'clock) at 512 scan resolution for qualitative assessment of retinal morphology¹. OCT operators were instructed that if standard deviation of center point thickness was greater than or equal to 10% of the center point thickness or if signal strength exceeded 5, the scans should be repeated and submitted only if the OCT technician believed that the scans were of adequate quality or better quality measures were not obtainable after repeat scans.

Scans were sent to the DRCR.net Reading Center for evaluation. Scans with center point measurements graded as inaccurate by evaluators at the reading center were manually remeasured from prints of the retina map scans using digital calipers. The center point was manually measured for all scans with a center point standard deviation $\geq 10.5\%$, which was computed as the standard deviation of the 6 automated center point measurements divided by the center point thickness (average of the 6 measurements). This cutpoint was selected based on the reading center's prior experience in evaluating a random sample of images in which all images with center point standard deviation $\geq 10.5\%$ needed manual measurement of the center

point. For all other scans, if decentration (the center of the scan not aligned with the center of the macula) or boundary line artifacts (the instrument software incorrectly identified the internal and external limits of the retina) were identified, the center point thickness was manually measured unless poor scan quality precluded a measurement.

For analysis, each OCT scan was categorized as having a correct or incorrect center point measurement. The measurement was considered to be correct if either the reading center determined that a manual measurement was not needed or a manual measurement was deemed necessary but the automated measurement was within 12% of the manual measurement. Twelve percent was selected because it approximates the 95% confidence interval on an OCT measurement.² An automated measurement error was defined as an automated measurement that differed from a manual measurement by 12% or more. Automated measurements of scans in which poor quality precluded a manual measurement were also considered to be errors.

Four factors were evaluated as possible predictors of an incorrect automated center point measurement error: standard deviation of the center point as a percentage of the center point thickness (SD), signal strength, the confidence of the analysis as reported by the instrument software, and the automated center point thickness (CPT). All four factors are present on or can be calculated from the OCT 3 printout. The correlation of each factor with the others was assessed with Spearman correlation coefficients. The association of each factor with the center point automated measurement error rate was assessed in univariate and multivariable logistic regression models. A cost-benefit type of analysis was used to further evaluate the relative importance of each factor in determining whether automated errors were present as defined above. The cost benefit analyses were performed by calculating the number of scans in which an incorrect center point measurement was detected (benefit) as a function of the proportion of images to be sent to the Reading Center for manual assessment (cost). Parallel analyses using receiver operating curve methods produced similar conclusions (data not shown). Additional cost-benefit curves were constructed for standard deviation stratified by center point thickness, where separate curves were generated for CPT < 250 μ , CPT 250-299 μ , CPT 300-399 μ , CPT 400-499 μ , and CPT $\geq$ 500 μ .

To evaluate the impact of reading center evaluation of the center point automated measurements in a clinical trial, results from a DRRCR.net protocol comparing two laser techniques for DME were used in which the results obtained with reading center evaluation of all scans were compared with the results that would have been obtained using just the automated measurements.⁴ For this analysis, change in CPT from baseline to 12 months was compared between treatment groups using repeated measures least squares regression models adjusted for baseline thickness and accounting for the correlated data from subjects with two study eyes. In parallel fashion, the treatment groups were compared for change in central subfield thickness, with and without replacing invalid automated central subfield values with values imputed from manually measured CPT¹.

The impact on a trial's sample size of using solely automated center point measurements versus having a reading center evaluate the scans was assessed using the fact that sample size is proportional to the overall variance. The overall variance was calculated as the between-subject variation plus measurement error. The percent reduction in variance (and therefore sample size) was calculated assuming that manual measurement would give the correct thickness value (i.e., zero measurement error for scans sent to the reading center). Sample size for a hypothetical trial with 90% power, type 1 error of 5%, assuming an effect size of 50 microns with a standard deviation of 150 microns was then computed varying on the number of images sent to the Reading Center. Statistical analyses were performed using SAS software version 9.1 (Cary, NC).

Results

Of the 6,522 OCT scans included in the analysis, 1,243 (19%) required manual measurement of the center point as determined by the Reading Center. For 44 (4%) of these images, the poor quality of the scan precluded a manual measurement and the automated measurement was assumed to be inaccurate. For the 1,199 manually-graded scans where a manual measurement was attainable, the median difference between the automated center point value and the manually graded value was 34 microns (interquartile range = -2 to 65 microns) and the median absolute value of the difference was 48 microns (interquartile range = 23 to 82 microns). The automated measurement was more frequently greater when the retina was thinner and the manual measurement was more frequently greater when the retina was thicker (Figure 1). For 362 (30%) of the 1,199 manually-graded images, the automated measurement was within 12% of the manual measurement. Thus, for 5,641 (86%) of the 6,522 scans, the automated center point measurement was either considered to be correct or was within 12% of the manual measurement and for 881 (14%) the automated measurement was inaccurate by 12% or more or had image quality which precluded manual measurement.

Table 1 shows the frequency of an automated measurement error across the range of values of the four evaluated factors (standard deviation of the center point, signal strength, analysis confidence, and center point thickness). The correlation between these factors was weak, ranging in absolute value from 0.04 to 0.30 (Table 2-journal website). Each of the four factors had a significant association with the automated measurement error rate. The error rate was higher with a larger standard deviation, when the analysis confidence variable was low, when the signal strength was low, and when CPT was low. After adjusting for the effect of the other factors, all were still significant except for the analysis confidence variable (Table 3). However, standard deviation of the center point explained much more of the variance than did the other factors. Adding the other factors to a model with standard deviation did not add to the predictive ability.

Cost-benefit curves were constructed to show the error rate relative to the proportion of scans that would need to be graded by a reading center. It can be seen that for a given percentage of scans sent to the reading center, the error rate is lower for using the standard deviation as the flag to send than for the other three factors (Figure 2) and there is no additive value for these factors. If no scans were sent to a reading center, 14% of scans would be considered to have an automated measurement error and the mean center point value would be 5 microns higher and the mean central subfield value 4 microns higher than the mean using just the automated measurements. Retinal volume mean would be 8.3 mm³ using automated measurements and 8.2 mm³ with inaccurate automated measurements excluded.

Additional cost-benefit curves were constructed for standard deviation stratified by five levels of automated center point thickness; < 250μ, 250-299μ, 300-399μ, 400-499μ and ≥500μ. Figure 3 shows that even after adjusting for standard deviation, the error rates are lower in eyes with thicker maculas. If no scans were sent to the reading center, 15%, 23%, 13%, 7%, and 6% of the images would be considered to have an automated measurement error for the five thickness subgroups respectively.

As seen in figure 4, based on standard deviation, a 5% or less error rate can be achieved by sending only 33% of scans to the reading center (those with a center point standard deviation ≥5%). The center point mean using the automated measurements only (with none sent to the reading center) would be 2 microns higher than the mean when sending the 33% of scans with a center point standard deviation ≥5%. A 5% or less error rate can be achieved by sending 37%, 69%, 34%, 4%, 1% of scans to the reading center for images with automated thicknesses of <250 microns, 250-299 microns, 300-399 microns, 400-499 microns, and ≥ 500 microns

respectively. The distance between standard deviation levels on each curve differs based on the center point thickness stratum.

To illustrate the effect in a randomized trial of using only the automated measurements with no scans evaluated for manual measurement by a reading center, data were reanalyzed from a DRCR.net randomized trial comparing two laser regimens for DME (Table 4).⁴ In the analysis using the automated measurements, the treatment group difference in mean change ($\pm$ standard error) in central subfield thickness from baseline to 1 year was 19 ± 15 microns ($P=0.02$) compared with 29 ± 15 microns ($P=0.01$) using the corrected dataset. Similarly, the treatment group difference in change in center point thickness from baseline to 1 year was 23 ± 17 microns ($P=0.04$) from the automated dataset compared with 34 ± 18 microns ($P=0.01$) using the corrected dataset.

Data from the DRCR.net studies can be used to assess the impact on projected sample size in a hypothetical trial for varying proportion of scans sent to the Reading Center. Figure 5 demonstrates the projected sample size for a hypothetical power calculation on the difference in central subfield thickness having 90% power with an effect size of 50 microns assuming a standard deviation of 150 microns. It can be seen that the proportion of scans sent to the Reading Center has little impact on the projected sample size, estimating 402 subjects if no images are sent to the Reading Center and 378 subjects if all of the images are sent to the Reading Center, an increase of only 6%.

Discussion

In this study, we have shown that in a large randomized trial there is little benefit derived from sending all scans to a reading center to evaluate whether the automated measurement of center point thickness is correct in eyes with DME. When sending all scans, approximately 86% will have accurate automated measurements. Manually measuring the center point thickness for the other 14% enhances overall accuracy, but only slightly. In doing this, the mean difference in center point thickness differs by only 5 microns and the mean difference in central subfield thickness differs by only 4 microns compared with having all scans reviewed by a reading center. Similar results are obtained using retinal volume as the outcome measures.

The standard deviation of the center point is more predictive of the probability of an error in the automated measurements than signal strength or the analysis confidence assessment. Both of these measures are provided by the Stratus OCT software as measures of scan quality but neither added much to the center point standard deviation. Although automated errors are more likely for eyes with center involved DME when the fovea is thinner, retinal thickness also added little to the predictive ability of the center point standard deviation alone. If there is a desire to reduce the error rate from the expected 14%, a portion of the scans could be sent to the reading center. Our results indicate that the center point standard deviation is the best measure to use in this regard and an error rate of 5% can be achieved by sending to a reading center scans with a center point standard deviation exceeding 5%. However, this has no appreciable impact on the error present in the mean change in center point thickness.

Evaluation of OCT scan quality by the clinical site OCT technicians and investigators would also increase data accuracy. The standard deviation of the center point may still be less than 5% even if the scan has decentration and/or boundary line errors. By repeating erroneous scans until a satisfactory quality scan is obtained for submission, operator-error issues can be corrected. Specific training regarding OCT quality assessment is required for optimal clinical site performance. Despite this caveat, it is unlikely that the frequency of erroneous scans could be reduced significantly in this study considering in the DRCR.net protocols, OCT technicians

were instructed to repeat the OCT measurement if the standard deviation was $\geq 10\%$ or the signal strength was ≤ 6 (i.e. images of potentially poor quality).

There is a trade off for a clinical trial in reducing or eliminating the number of scans sent to a reading center. Sending scans to the reading center and correcting errors in the center point measurement will reduce the required sample size for a given statistical power. Our data suggest that the increase in sample size is likely to be small, approximately 6%, in a study of center-involved DME powered to assess treatment group differences in the change in central subfield thickness if no scans were sent to a reading center compared with having all scans sent to a reading center. In large trials, even a 6% increase in sample size could lead to a significant increase in study cost. Therefore, in planning a trial, the cost per additional subject recruited versus the cost to have OCT scans sent to a reading center for grading must be considered to determine the most cost-effective strategy for the trial.

When the retina is thicker, the likelihood of a measurement error decreases. Thus, the yield is lower in sending scans to a reading center for evaluation. This is likely due to a detection bias, where a decentered scan is more likely to be identified if the retina has a morphologically distinct center (such as a foveal depression) which is more likely to be observed in a thinner retina compared to a thickened retina with disorganized morphology. In determining whether to send scans to a reading center for a protocol, there may be less reason to send the scans for evaluation in a study in which eligibility requires substantial central retinal thickening than there would be in a protocol in which central thickening was required to be absent and the objective of the intervention was the prevention of the development of DME. In the latter circumstance, the misclassification rate of outcome due to automated errors could necessitate that all scans be evaluated by a reading center.

DRCR.net is utilizing the results of this study to determine for each protocol which scans should be sent to the reading center for evaluation and manual measurement of the center point when indicated. For protocols in which OCT-measured retinal thickness is not the primary outcome measure and morphology grading is not required, in general either no scans are being sent to the reading center or only scans exceeding a specified center point standard deviation value or less than a specified center point thickness are being sent to the reading center.

It is important to recognize that these results apply to cases of DME and not to age-related macular degeneration or other macular conditions in which the automated error rate can be much higher.⁵⁻⁸ In addition, these results apply to the software available with the Zeiss Stratus OCT and not necessarily to software algorithms of other OCT devices. Nevertheless, the principles followed in this analysis may be generalized to other devices and all conditions with regard to determining the most efficient strategy for having a reading center evaluate retinal thickening, whether by OCT or other technology, in a clinical trial. Finally, frequency domain OCT with registration capability to fundus landmarks is expected to reduce decentration errors and better retina sectioning algorithms are expected to reduce boundary line errors in the near future. It appears that for Stratus OCT, manual grading adds little measurement accuracy to studies of DME. Manual grading of thickness measurements of frequency domain OCT images in clinical trials of the future may be even less important. Reading centers provide a mechanism for masked assessment of morphology and will be important for quality control in clinical trials employing only the numeric data from the OCT output. With the common scarcity of resources and the need to reallocate available support to other important aspects of clinical trials, utilization of reading center analyses where necessary, and avoidance where not of major impact will permit cost savings without substantially sacrificing data accuracy.

Acknowledgments

Supported through a cooperative agreement from the National Eye Institute and the National Institute of Diabetes and Digestive and Kidney Diseases EY14231, EY14269, EY14229.

References

1. Browning D, Glassman A, A L, Bressler NM, et al. Optical coherence tomography measurements and analysis methods in optical coherence tomography studies of diabetic macular edema. *Ophthalmol*. 2007 In Press.
2. Diabetic Retinopathy Clinical Research Network. Reproducibility of macular thickness and volume using Zeiss optical coherence tomography in patients with diabetic macular edema. *Ophthalmology* 2007;114:1520–5. [PubMed: 17353052]
3. Diabetic Retinopathy Clinical Research Network. Relationship between optical coherence tomography-measured central retinal thickness and visual acuity in diabetic macular edema. *Ophthalmology* 2007;114:525–36. [PubMed: 17123615]
4. Writing Committee for the Diabetic Retinopathy Clinical Research Network. Comparison of the modified early treatment diabetic retinopathy study and mild macular grid laser photocoagulation strategies for diabetic macular edema. *Arch Ophthalmol* 2007;125:469–80. [PubMed: 17420366]
5. Sadda SR, Wu Z, Walsh AC, et al. Errors in retinal thickness measurements obtained by optical coherence tomography. *Ophthalmology* 2006;113:285–93. [PubMed: 16406542]
6. Ray R, Stinnett SS, Jaffe GJ. Evaluation of image artifact produced by optical coherence tomography of retinal pathology. *Am J Ophthalmol* 2005;139:18–29. [PubMed: 15652824]
7. Sadda SR, Joeres S, Wu Z, et al. Error correction and quantitative subanalysis of optical coherence tomography data using computer-assisted grading. *Invest Ophthalmol Vis Sci* 2007;48:839–48. [PubMed: 17251486]
8. Leung CK, Chan WM, Chong KK, et al. Alignment artifacts in optical coherence tomography analyzed images. *Ophthalmology* 2007;114:263–70. [PubMed: 17123619]

Figure 1.
Relationship of Difference in Automatic and Manually Graded Center Point Thickness vs.
Manually Graded Center Point Thickness (n=1199)

Figure 2. Relationship of Error Rate by Proportion of Scans Sent to Reading Center Based on Predictive Factors of Need for Manual Grading. The error rate is defined as the percentage of scans requiring manual grading and having an automatic measurement differing from the manual measurement by > 12% that were not sent to the Reading Center.

Figure 3.

Relationship of Error Rate by Proportion of Scans Sent to Reading Center Based on Standard Deviation of the Center Point Stratified by Automated Center Point Thickness. Each point on the curve denotes a separate threshold for sending scans to the RC based on standard deviation of the center point. The error rate is defined as the percentage of scans requiring manual grading and having an automatic measurement differing from the manual measurement by $> 12\%$ that were not sent to the Reading Center.

Figure 4.

Relationship of Error Rate by Proportion of Scans Sent to Reading Center Based on Standard Deviation of the Center Point. Each point on the curve denotes a separate threshold for sending scans to the RC based on standard deviation of the center point (SD). For example, using SD $\geq 5\%$ as the criterion results in sending approximately 33% of images to the RC and a corresponding error rate of 5%. The error rate is defined as the percentage of scans requiring manual grading and having an automatic measurement differing from the manual measurement by $> 12\%$ that were not sent to the Reading Center.

Figure 5.
Required Sample Size to Achieve 90% Power to Detect Treatment Effect of 50 Microns with
Standard Deviation 150 Microns, Varied by Number of Images Sent to Reading Center.

Table 1
Frequency of Manual Grading for Factors Potentially Predictive of Center Point Errors *

Factor	N	Error Rate * N(%)
Overall	6522	881 (14%)
Standard Deviation of Center Point		
0%	67	6 (9%)
>0-1%	783	32 (4%)
>1-2%	1130	55 (5%)
>2-3%	1049	84 (8%)
>3-4%	855	69 (8%)
>4-5%	655	67 (10%)
>5-6%	482	45 (9%)
>6-7%	343	55 (16%)
>7-8%	252	44 (17%)
>8-9%	197	33 (17%)
>9-10%	104	22 (21%)
>=11%	605	369 (61%)
Analysis Confidence Low		
Absent	5384	567 (11%)
Present	1138	314 (28%)
Signal Strength		
0	24	16 (67%)
1	56	17 (30%)
2	123	31 (25%)
3	308	53 (17%)
4	714	115 (16%)
5	1268	174 (14%)
6	1577	206 (13%)
7	1260	151 (12%)
8	713	84 (12%)
9	324	28 (9%)
10	155	6 (4%)
Center Point Thickness		
A) <225	2363	312 (13%)
B) 225-249	502	115 (23%)
C) 250-299	871	199 (23%)
D) 300-399	1225	154 (13%)
E) 400-499	742	55 (7%)
F) 500-599	464	19 (4%)
G) 600-699	237	16 (7%)
H) >=700	118	11 (9%)

* A scan was considered to have an error when the automated center point measurement differed from a manual measurement by more than 12% or the automated measurement was determined to need a manual measurement, but the image quality precluded manual measurement

Table 2 (website)

Correlations among the four evaluated factors

	Standard Deviation of Center Point	Signal Strength	Center Point Thickness	Analysis Confidence Low
Standard Deviation of Center Point	--	-0.04	-0.30	0.21
Signal Strength	--	---	-0.29	-0.21
Center Point Thickness	--	--	--	0.21
Analysis Confidence Low	--	--	--	--

Table 3
Potentially Predictive Factors of the Need for Manual Grade in Logistic Regression Models

Factor	Univariate Model of Predictive Factor		Multivariate Model	
	Estimate [Odds Ratio] (95% C.I.)	P-value	Estimate [Odds Ratio] (95% C.I.)	P-value - Final Model
Ratio of Standard Deviation to Center Point Thickness (per 1% increase)	1.20 (1.18, 1.22)	0.001	1.18 (1.17, 1.20)	0.001
OCT Center point Thickness (per 100 micron increase)	0.82 (0.78, 0.87)	0.001	0.80 (0.74, 0.85)	0.001
Signal Strength (per 1 unit increase)	0.85 (0.81, 0.88)	0.001	0.93 (0.89, 0.98)	0.004
Analysis Confidence (Low vs. Not Low)	3.24 (2.77, 3.79)	0.001	1.17 (0.93, 1.47)	0.17

* Cumulative r^2 obtained from forward selection regression, calculating the r^2 after adding factors to the logistic regression model. The numbers in this column represent the r^2 from four models in which the factor in the row has been cumulatively added to the model.

Table 4
Comparison of thickness data from a DRCR.net randomized trial analyzed using the corrected dataset with dataset reanalyzed using only the automated measurements with no scans evaluated for manual measurement by the reading center

	Baseline		1 Year			Change		P-value
	Modified-ETDRS	MMG	Modified-ETDRS	MMG	Modified-ETDRS	MMG		
Automatic Measurements [*]	319±149	322±129	Center Point Measurement		287±135	-63±144	-41±140	0.04
Corrected Dataset [†]	325±156	329±133	Central Subfield		294±153	-71±150	-37±141	0.01
Automatic Measurements [*]	336±131	339±111	Central Subfield		304±115	-60±128	-40±118	0.02
Corrected Dataset [†]	340±134	343±111	Central Subfield		308±129	-66±129	-37±118	0.01

^{*} Analyzed using dataset with thickness from the automated measurement from OCT machine with no scans evaluated for manual measurement by the reading center.

[†] Analyzed using the corrected dataset, i.e. the automated values when the automated values are from a scan not needing manual grading and the manually re-measured values from a scan in which manual grading was necessary.

All measurements are in microns.