Supplemental material

Reliability and reproducibility of measurement of carotid plaque burden

Methods:

We conducted a systematic review and meta-analysis according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and GRRAS guidelines (Guidelines for Reporting Reliability and Agreement Studies) (Registration # CRD42020210453).

Literature Search: We performed a comprehensive literature search as of May 19, 2021 using the following motor engines: MEDLINE/PubMed, Scopus, Embase, and Web of Science. The following combinations of keywords and MeSH terms were used to identify studies that have investigated the accuracy and reproducibility of carotid plaque measurement with 2D 3D ultrasound were searched using the following combination of keywords: [("Carotid Stenosis" OR "Carotid plaque" OR "Plaque Burden" OR "Total Plaque Area" OR "Total Plaque Volume" OR "TPV" OR "TPA") AND ("Inter-rater reliability" OR "Inter-rater reliability" OR "inter-rater agreement" OR "inter-rater concordance" OR "inter-observer reliability" OR "Intra-rater reliability" OR "Kappa value" OR "Weighted kappa coefficients" OR "Intraclass correlation coefficient" OR "Correlation coefficient" OR "Kappa coefficients" OR "Cohen's kappa" OR "Pearson's correlation coefficient")]. We used EndNote X9 software to manage the literature findings and delete the duplicate studies. References listed in the relevant studies and reviews were checked to identify additional or missing studies.

Inclusion and Exclusion Criteria: We searched all relevant papers reporting the accuracy and reproducibility (e.g. intraclass correlation, correlation coefficient, and Cohen's kappa) of carotid plaque measurement in 2D and 3D ultrasound. Reproducibility of carotid ultrasound measurements was expressed as an inter-rater and intra-rater reliability. According to the GRRAS [1], important inclusion criteria for reproducibility of carotid plaque include studies reporting sample size, number of raters and subjects, sampling method, rating process, statistical analysis and reliability coefficients. For inter-rater reliability and intra-rater reliability, we gathered all available statistical information including Kappa coefficient (weighted and un-weighted), intraclass correlation coefficient, Pearson correlation coefficient and Spearman rank correlation coefficient. If one study reported both correlation coefficient and the kappa, we only used correlation coefficient values for the final pooling and meta-analysis.

We excluded studies which did not confirm ethics approval, animal studies, short reports or letters without adequate information, case studies, retracted/withdrawn articles; and previous reviews, and studies with insufficient information. Two authors (MRA and RT) independently assessed the search findings based on our inclusion criteria. Any discrepancies were resolved though consensus.

Data extraction: To assess reproducibility of carotid plaque, we extracted the following further information using a predefined spreadsheet in Excel: sample size, number of raters and subjects, sampling method, rating process, statistical analysis and reliability coefficients. Selected items qualified for extraction if described in sufficient detail. We considered each

sample as a unit of analysis; included it once in Excel if the same sample was reported in more than two studies; and if several samples were reported in one selected study, we included each of them separately as a unit of analysis.

Quality assessment: Each item selected for the reproducibility of carotid plaque was considered as a quality score. We defined qualified studies as those with a score >6 out of 8.[2]

Statistical analysis: We pooled data for all different types of reliability measurements. In studies reported a reliability coefficient as the kappa statistic, we represent these data as an r-type coefficient ranging from −1.00 to +1.00. Standard agreement was categorized as poor (κ = 0.00–0.20), fair (κ = 0.21–0.40), moderate (κ = 0.41–0.60), substantial (κ = 0.61–0.80) or almost perfect (κ = 0.81–1.00).[3] Of note, in meta-analyses, the kappa statistic can be treated as a correlation coefficient.[4] Therefore, we transformed all reliability coefficients into Fisher's Z for variance stabilization and normalization of data distributions in pooling data for reproducibility of carotid ultrasound measurements expressing as an inter-rater and intra-rater reliability coefficient with their 95% confidence interval (CI). Using a random-effects model, we performed a meta-analysis for each of reliability coefficients. In order to facilitate the interpretation, we finally transformed pooled results from Fisher's Z to reliability coefficients. All statistical analyses were performed using a 'meta' package in R software.

Results:

We retrieved a total of 1421 papers to identify reproducibility of carotid plaque assessments. Figure 1 shows the flowchart process of our search strategies for the identification and selection articles. After removing duplicates, reviewing abstract/title and full texts we finally selected 7 articles for current review on the reproducibility of carotid plaque assessments. M.R.A. and R.T. had a near perfect agreement (κ = 0.89) regarding the selection of papers for final meta-analysis. Of these, 6 articles were included for meta-analysis on inter-rater reliability and 5 articles 6 studies on intra-rater reliability. The detailed main characteristics of included studies are summarized in Supplemental Table 1. Quality of the selected studies are summarized in Supplemental Table 2. All selected studies reported correlation coefficient and we did not enter any Kappa values.

Inter- and intra-observer reliability for measurement of total plaque area: Data regarding accuracy of TPA measurements were available in Table 1. Reliability of TPA measurement was excellent, with an inter-observer reliability of 0.94 (95% CI: 0.83-0.99) and intra-observer reliability of 0.96 (95%CI: 0.94- 0.97) (Supplemental Figure 2 & 3).

Inter- and intra-observer reliability for measurement of total plaque volume:

Inter-observer reliability of TPV was 0.91 (95% CI: 0.62-0.98) and intra-observer reliability was 0.91 (95%CI: 0.83- 0.95) (Supplemental Figure 2 & 3).

Discussion:

In this systematic review and meta-analysis, we found an almost perfect inter and intra-observer reliability for measurement of carotid plaque burden using TPA or volume. The accuracy of plaque measurement assessed by phantoms and by histology was reported by Lopez-Melgar et al.[5] Plaque measurement with ultrasound is thus an accurate and

reproducible method, and should be used widely in vascular prevention clinics for targeted management of vascular diseases.

Discussion:

In this systematic review and meta-analysis of 7 studies we found an almost perfect inter and intra-observer reliability for measurement of carotid plaque burden using TPA or volume. The accuracy of plaque measurement assessed by phantoms and by histology was reported by Lopez-Melgar et al.[5] Plaque measurement with ultrasound is thus an accurate and reproducible method, and should be used widely in vascular prevention clinics for targeted management of vascular diseases.

One of the mistaken criticisms of measuring plaque burden has been the assertion [6, 7] that there are problems with reliability of measurement. In the current study, we found almost perfect inter and intra-observer reliability for both 2D and 3D ultrasound; inter-observer reliability for TPA was 0.95 (95%CI: 0.83- 0.99). Measurement of TPA is reliable and easy to teach and learn. Most experienced and skillful vascular ultrasound technologists can learn to measure TPA reliably in a day. Unlike IMT, the quantities measured (TPA, plaque volume and vessel wall volume) are so much larger than the spatial resolution of ultrasound that errors in measurement represent a very small fraction of the quantity, i.e. the coefficient of variation is very small. The interobserver coefficient of variation of TPA (SD/mean) is thus very low; approximately 1.5%.

In the NIH-funded Northern Manhattan Study (NOMAS) of stroke risk and cognitive decline, Rundek and colleagues conducted a reproducibility analysis for carotid plaque burden measured by two sonologists trained in performing and reading research ultrasound studies. Among 105 randomly selected subjects, both sonologists identified 101 plaques and were in complete in agreement in terms of the number of plaques and plaque locations within an individual (kappa=1). The intraclass correlation coefficient (ICC) was 0.94 for the measurements of total carotid plaque area between the two readers (unpublished data; with permission from the corresponding author). The NOMAS team also reported excellent reproducibility of maximum carotid plaque thickness (interrater intraclass correlation coefficient: 0.87 and intra-rater intraclass correlation coefficient: 0.94).[8]

In the Tromsø study, reproducibility of TPA measurements was reported in detail by Fosse et al.[9] , and was also summarized in another paper [10]. The inter-observer mean arithmetic difference (MAD) (SD) of plaque area measurements was –1.0 mm^2 (4.3) and the limits of agreement ± 8.4 mm^2. The intra-observer MAD (SD) for sonographer 1 was 0.2 mm^2 (3.0) and the limits of agreement ± 6.0 mm^2. The corresponding values for sonographer 2 was 0.01 mm2 (3.6) and ± 7.1 mm^2. These differences are very small in comparison to the average plaque burden of vascular patients. In the database of the Stroke Prevention & Atherosclerosis Research Centre, London, Canada, among 9,864 patients referred for vascular prevention, the mean $\pm$ SD of age was 62.4 $\pm$ 14 years; the range of TPA was from 0 to 1289 mm^2; mean 119 $\pm$

135 mm^2. Among patients with TPA > 0 mm^2, mean age was 64.56 ± 13.07 years, and mean TPA was 134.46 ± 137.42 mm^2.

TPA can be also used in multicenter studies assessing atherosclerosis. For the purpose of demonstrating generalizability of our results to other ultrasound laboratories and clinics, we carried out a study of interobserver reliability in which plaque area measurements in 25 patients were repeated a week apart by 2 technicians using 2 different machines. The senior technologist, who had been performing these measurements for 8 years, used a new, high-resolution ATL HDI 5000 scanner; the junior technologist, who had been doing such measurements for 1 year, used an ATL Mark 9 duplex scanner. The reliability (intraclass correlation) was 0.85, with the senior technician using the higher-resolution machine systematically measuring more plaque."[11] Romanens et al. also reported intra-observer reproducibility of TPA in a study of 2 middle-aged cohorts from Switzerland and Germany. For the cut-offs of TPA, The linear weighted kappa value was excellent at 0.97 (95% CI 0.92–1.00) according to the TPA cut-offs (0–9, 10–49, 50–99 and ≥100 mm^2).[12]

Plaque thickness

Plaque thickness (plaque height) and area are both highly reproducible. In the Northern Manhattan Study (NOMAS), the intra-class correlation coefficient for inter-observer and intra-observer reliability of plaque thickness were 0.77 and 0.94, respectively. The intraclass correlation coefficients for plaque area were 0.78 and 0.94.[13]

Both plaque thickness and area showed similar predictive value In models fully adjusted for risk factors, HR for cPTmax (plaque thickness) was 3.13 (95% CI 1.80–5.51, $P < 0.001$) for all MACE, similar to that of cPB (plaque area). IMT did not improve risk prediction significantly. Non-categorical net reclassification index (NRI) for cPTmax was 0.173 (95% CI 0.109–0.243, $P < 0.001$) for all MACE, similar to cPB. IMT assessment did not result in significant NRI.[14]

References:

[1] Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Journal of clinical epidemiology. 2011;64:96-106.
[2] Mirhaghi A, Heydari A, Mazlom R, Ebrahimi M. The Reliability of the Canadian Triage and Acuity Scale: Meta-analysis. N Am J Med Sci. 2015;7:299-305.
[3] Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical therapy. 2005;85:257-68.
[4] Rettew DC, Lynch AD, Achenbach TM, Dumenci L, Ivanova MY. Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. International journal of methods in psychiatric research. 2009;18:169-84.
[5] Lopez-Melgar B, Fernandez-Friera L, Sanchez-Gonzalez J, Vilchez JP, Cecconi A, Mateo J, et al. Accurate quantification of atherosclerotic plaque volume by 3D vascular ultrasound using the volumetric linear array method. Atherosclerosis. 2016;248:230-7.
[6] Johri AM, Nambi V, Naqvi TZ, Feinstein SB, Kim ESH, Park MM, et al. Recommendations for the Assessment of Carotid Arterial Plaque by Ultrasound for the Characterization of Atherosclerosis and Evaluation of Cardiovascular Risk: From the American Society of Echocardiography. J Am Soc Echocardiogr. 2020.
[7] Johri AM, Nambi V, Naqvi TZ, Feinstein SB, Kim ESH, Park MM, et al. Recommendations for the Assessment of Carotid Arterial Plaque by Ultrasound for the Characterization of Atherosclerosis and Evaluation of Cardiovascular Risk: From the American Society of Echocardiography. J Am Soc Echocardiogr. 2020;33:917-33.
[8] Rundek T, Arif H, Boden-Albala B, Elkind M, Paik M, Sacco RLJN. Carotid plaque, a subclinical precursor of vascular events: the Northern Manhattan Study. 2008;70:1200-7.
[9] Fosse E, Johnsen SH, Stensland-Bugge E, Joakimsen O, Mathiesen EB, Arnesen E, et al. Repeated visual and computer-assisted carotid plaque characterization in a longitudinal population-based ultrasound study: the Tromso study. Ultrasound Med Biol. 2006;32:3-11.
[10] Johnsen SH, Mathiesen EB, Fosse E, Joakimsen O, Stensland-Bugge E, Njolstad I, et al. Elevated high-density lipoprotein cholesterol levels are protective against plaque progression: a follow-up study of 1952 persons with carotid atherosclerosis the Tromso study. Circulation. 2005;112:498-504.
[11] Spence JD, Eliasziw M, DiCicco M, Hackam DG, Galil R, Lohmann T. Carotid plaque area: a tool for targeting and evaluating vascular preventive therapy. Stroke. 2002;33:2916-22.
[12] Romanens M, Sudano I, Adams A, Warmuth W. Advanced carotid atherosclerosis in middle-aged subjects: comparison with PROCAM and SCORE risk categories, the potential for reclassification and cost-efficiency of carotid ultrasound in the setting of primary care. Swiss medical weekly. 2019;149:w20006.
[13] Rundek T, Arif H, Boden-Albala B, Elkind MS, Paik MC, Sacco RL. Carotid plaque, a subclinical precursorof vascular events: The Northern Manhattan Study. Neurology. 2008;70:1200-7.
[14] Sillesen H, Sartori S, Sandholt B, Baber U, Mehran R, Fuster V. Carotid plaque thickness and carotid plaque burden predict future cardiovascular events in asymptomatic adult Americans. Eur Heart J Cardiovasc Imaging. 2018;19:1042-50.
[15] Barnett PA, Spence JD, Manuck SB, Jennings JR. Psychological stress and the progression of carotid artery disease. Journal of hypertension. 1997;15:49-55.
[16] Rundek T, Arif H, Boden-Albala B, Elkind MS, Paik MC, Sacco RL. Carotid plaque, a subclinical precursor of vascular events: the Northern Manhattan Study. Neurology. 2008;70:1200-7.

[17] Mitchell C, Korcarz CE, Gepner AD, Kaufman JD, Post W, Tracy R, et al. Ultrasound carotid plaque features, cardiovascular disease risk factors and events: The Multi-Ethnic Study of Atherosclerosis. Atherosclerosis. 2018;276:195-202.
[18] Lopez-Melgar B, Fernandez-Friera L, Oliva B, Garcia-Ruiz JM, Penalvo JL, Gomez-Talavera S, et al. Subclinical Atherosclerosis Burden by 3D Ultrasound in Mid-Life: The PESA Study. J Am Coll Cardiol. 2017;70:301-13.
[19] Sandholt BV, Collet-Billon A, Entrekin R, Sillesen HH. Inter-Scan Reproducibility of Carotid Plaque Volume Measurements by 3-D Ultrasound. Ultrasound in medicine & biology. 2018;44:670-6.
[20] López-Melgar B, Fernández-Friera L, Oliva B, García-Ruiz JM, Peñalvo JL, Gómez-Talavera S, et al. Subclinical Atherosclerosis Burden by 3D Ultrasound in Mid-Life: The PESA Study. Journal of the American College of Cardiology. 2017;70:301-13.

Supplemental Table 1. Summary of basic characteristics of the included studies to identify reproducibility of carotid plaque measurements.

Authors/	Year	Name of study/ Country	Type of measurement	Index assessment	Number of images/participants	Outcomes	Effect size
Barnett et al. [15]	1997	SPARC/Canada	ICC	TPA	50 scans - 25 participants	Inter-rater reliability/ intra-rater reliability	0.99/ 0.94
Spence et al.[11]	2002	SPARC/ Canada	ICC	TPA	25 participants	Inter-rater reliability	0.85
Rundek et al.[16]	2008	NOMAS/ USA	ICC	TPA	110 participants	Inter-rater reliability/ intra-rater reliability	0.87/ 0.94
Romanens et al.[12]	2019	Switzerland	ICC	TPA	57 patients for both carotid arteries	Intra-rater reliability	0.964
Romanens et al.[12]	2019	Germany	ICC	TPA	56 patients for both carotid arteries	Intra-rater reliability	0.976
Mitchell et al.[17]	2018	MESA/ USA	ICC	TPA	24 plaque images	Inter-rater reliability	0.98
López-Melgar et al.[18]	2017	PESA/ Spain	ICC	TPV*	69 participants	Inter-rater reliability/ Intra-rater reliability	0.81 / 0.91
Sandholt et al.[19]	2018	---/ Denmark	Pearson's correlation coefficient	TPV*	38 plaques from 26 patients	Inter-rater reliability	0.96

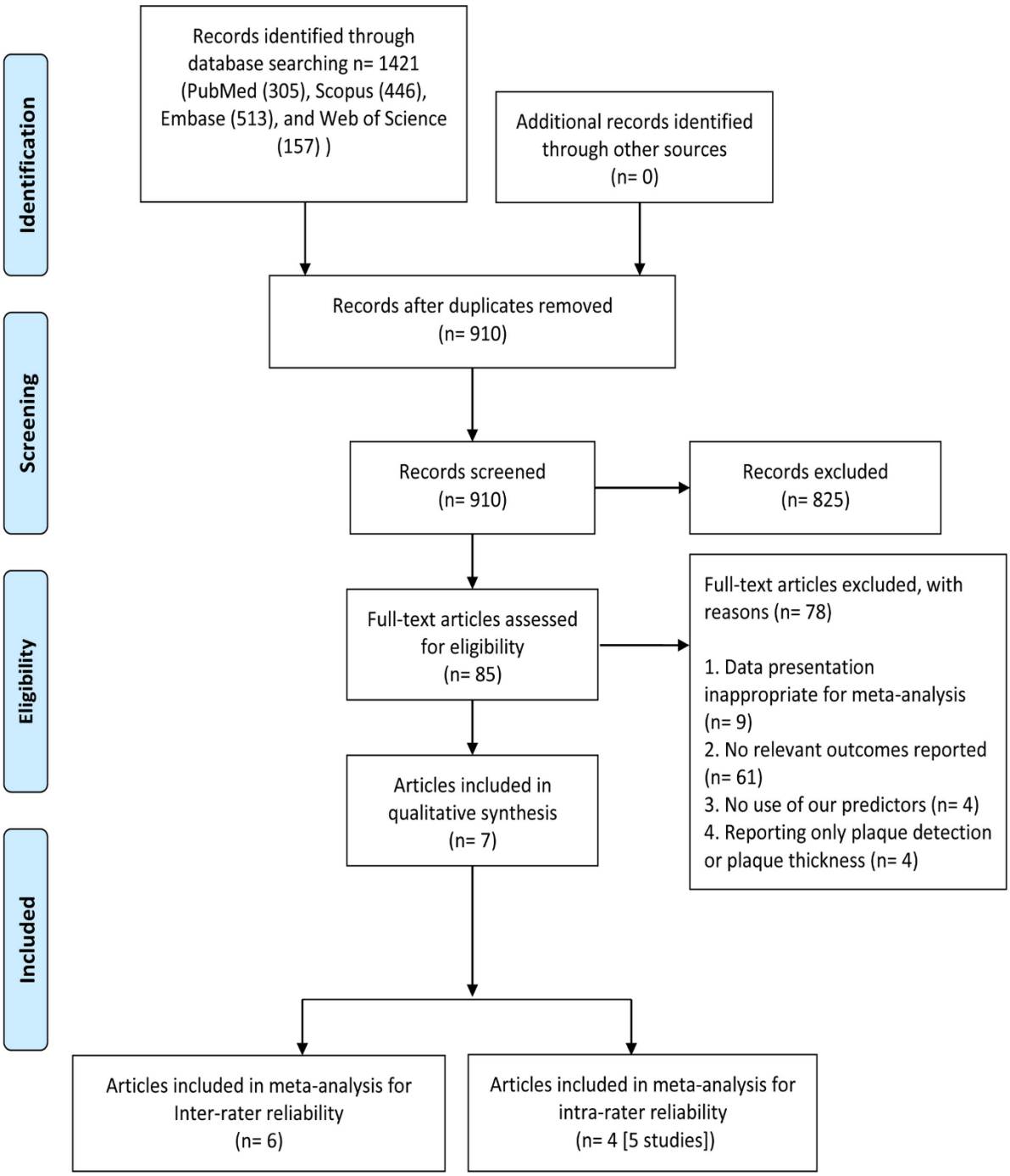
*Three-dimensional ultrasound.

Abbreviations (alphabetical order):ICC: Intraclass correlation coefficient; ELSA-Brasil: Brazilian Longitudinal Study of Adult Health; MESA: Multi-Ethnic Study of Atherosclerosis; NOMAS: Northern Manhattan Study; PESA: Progression of Early Subclinical Atherosclerosis; SPARC: Stroke Prevention and Atherosclerosis Research Centre; TPA: Total plaque area; TPV: Total plaque volume; VIPVIZA: VIsualiZation of asymptomatic atherosclerotic disease for optimum cardiovascular prevention.

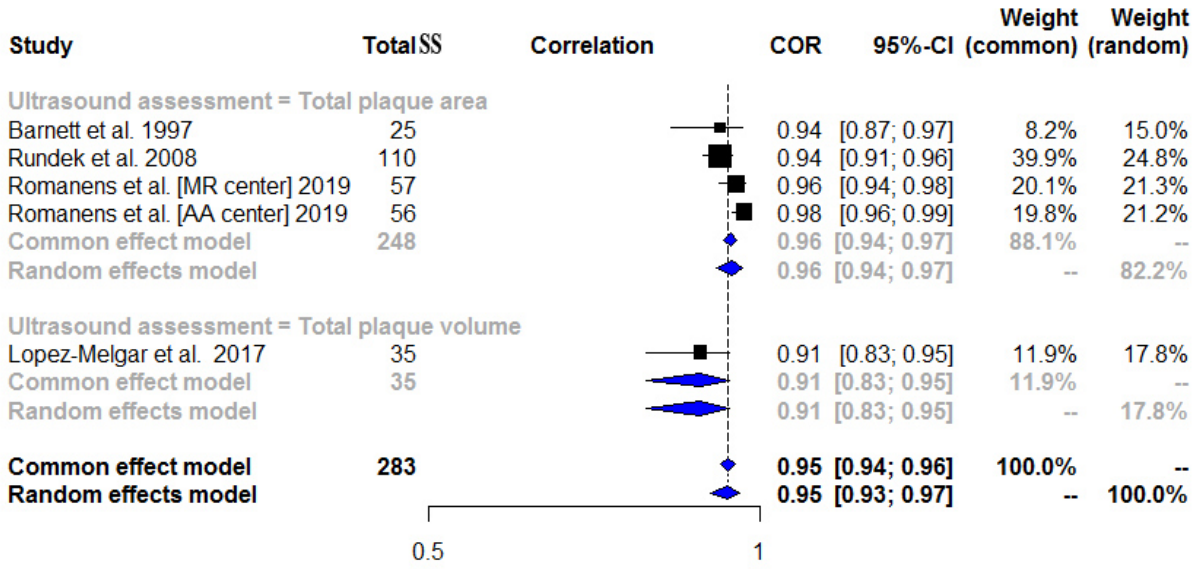
Supplemental Table 2. **Quality assessment results for selected studies to assess the reproducibility of carotid plaque assessments**

Studies	**Quality assessment questions**								
	Question1	***Question 2***	***Question 3***	***Question 4***	***Question 5***	***Question 6***	***Question 7***	***Question 8***	***Score***
Barnett et al. [15]	*	*	*	-	-	*	*	*	6
Spence et al. [11]	*	-	*	*	*	*	*	-	6
Rundek et al. [16]	*	*	*	-	*	*	*	-	6
Romanens et al. [12]	*	*	*	-	*	*	*	*	7
López-Melgar et al. [20]	*	*	-	-	*	*	*	*	6
Sandholt et al. [19]	*	*	*	-	*	*	*	-	6
Mitchell et al.[17]	*	*	*	-	*	*	*	*	7

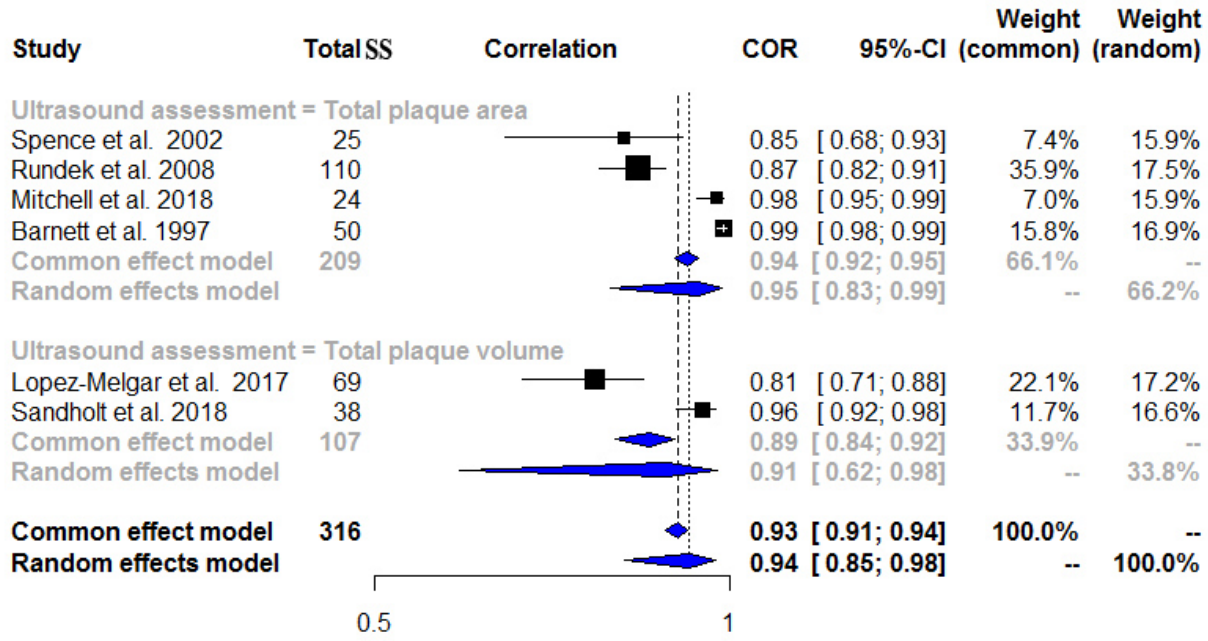
Question #1: Explained how the sample size was chosen. Stated the determined number of raters, subjects/objects, and replicate observations.
Question #2: Described the sampling method.
Question #3: Described the measurement/rating process (e.g., time interval between repeated measurements, availability of clinical information, blinding).
Question #4: Stated whether measurements/ratings were conducted independently. (rating process)
Question #5: Described the statistical analysis.
Question # 6: Stated the actual number of raters and subjects/objects that were included and the number of replicate observations that were conducted.
Question # 7: Described the sample characteristics of raters and subjects (e.g., training, experience).
Question # 8: Reported estimates of reliability and agreement, including measures of statistical uncertainty.



Supplemental Figure 1. Prisma flowchart: Publication selection process.



Supplemental Figure 2. Interobserver reliability for measurement of plaque burden: A comparison between total plaque area and total plaque volume.
Abbreviations: SS: Sample Size; COR: Correlation.



Supplemental Figure 3. Intraobserver reliability for measurement of plaque burden: A comparison between total plaque area and total plaque volume.
Abbreviations: SS: Sample Size; COR: Correlation.