Automated Segmentation and Volume Measurement of Intracranial Internal Carotid Artery Calcification at Noncontrast CT

Gerda Bortsova, MSc • Daniel Bos, MD, PhD • Florian Dubost, PhD • Meike W. Vernooij, MD, PhD • M. Kamran Ikeram, MD, PhD • Gijs van Tulder, MSc • Marleen de Bruijne, PhD

From the Biomedical Imaging Group Rotterdam, Department of Radiology and Nuclear Medicine (G.B., M.d.B.), Department of Epidemiology (D.B., M.W.V., M.K.I.), and Department of Radiology and Nuclear Medicine (M.W.V.), Erasmus MC, PO Box 2040, 3000 CA Rotterdam, the Netherlands; Department of Biomedical Data Science, Stanford University, Stanford, Calif (F.D.); Faculty of Science, Radboud University, Nijmegen, the Netherlands (G.v.T.); and Machine Learning Section, Department of Computer Science, University of Copenhagen, Copenhagen, Denmark (M.d.B.). Received September 18, 2020; revision requested November 24; revision received May 31, 2021; accepted June 7. Address correspondence to G.B. (e-mail: g.bortsova@erasmusmc.nl)

This research is part of the research project Deep Learning for Medical Image Analysis (project no. P15-26), funded by the Dutch Technology Foundation STW, which is part of the Netherlands Organisation for Scientific Research, and which is partly funded by the Ministry of Economic Affairs.

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2021; 3(5):e200226 • <https://doi.org/10.1148/ryai.2021200226> • Content codes: **AI** **CA** **HN**

Purpose: To develop and evaluate a fully-automated deep learning–based method for assessment of intracranial internal carotid artery calcification (ICAC).

Materials and Methods: This was a secondary analysis of prospectively collected data from the Rotterdam study (2003–2006) to develop and validate a deep learning–based method for automated ICAC delineation and volume measurement. Two observers manually delineated ICAC on noncontrast CT scans of 2319 participants (mean age, 69 years $\pm$ 7 [standard deviation]; 1154 women [53.2%]), and a deep learning model was trained to segment ICAC and quantify its volume. Model performance was assessed by comparing manual and automated segmentations and volume measurements to those produced by an independent observer (available on 47 scans), comparing the segmentation accuracy in a blinded qualitative visual comparison by an expert observer, and comparing the association with first stroke incidence from the scan date until 2016. All method performance metrics were computed using 10-fold cross-validation.

Results: The automated delineation of ICAC reached a sensitivity of 83.8% and positive predictive value (PPV) of 88%. The intraclass correlation between automatic and manual ICAC volume measures was 0.98 (95% CI: 0.97, 0.98; computed in the entire dataset). Measured between the assessments of independent observers, sensitivity was 73.9%, PPV was 89.5%, and intraclass correlation coefficient was 0.91 (95% CI: 0.84, 0.95; computed in the 47-scan subset). In the blinded visual comparisons of 294 regions, automated delineations were judged as more accurate than manual delineations in 131 regions, less accurate in 94 regions, and equally accurate in the rest of the regions (131 of 225, 58.2%; $P = .01$). The association of ICAC volume with incident stroke was similarly strong for both automated (hazard ratio, 1.38 [95% CI: 1.12, 1.75]) and manually measured volumes (hazard ratio, 1.48 [95% CI: 1.20, 1.87]).

Conclusion: The developed model was capable of automated segmentation and volume quantification of ICAC with accuracy comparable to human experts.

Supplemental material is available for this article.

© RSNA, 2021

Intracranial arteriosclerosis is a major risk factor for stroke (1,2) and has been linked to an increased risk of dementia (3). An important indicator of intracranial arteriosclerosis is intracranial internal carotid artery calcification (ICAC), which can be visualized using CT (4,5). ICAC is thus a promising imaging marker for assessing the risk of cerebrovascular diseases. However, the quantitative measurements of ICAC currently rely on time-consuming and error-prone manual annotations (6–8). Automating ICAC assessment could therefore facilitate research on the causes and clinical consequences of intracranial arteriosclerosis and may ultimately enable the use of ICAC assessment in clinical practice.

Automating ICAC detection at CT is challenging because of its proximity to bony structures with similar attenuation, similarity in appearance to other structures (eg, dural calcifications), and image artifacts.

These challenges complicate the application of simple image processing techniques. Machine learning methods, however, are very suitable for this task because they can “learn” from examples without being explicitly programmed. Machine learning (including deep learning [9,10]) has been extensively applied to assessing arterial calcification in vessel beds other than the intracranial carotid arteries, including the coronary arteries, aorta, cardiac valves, and extracranial carotid arteries (11–14). To our knowledge, so far only our previous study (15) has proposed a method for automating ICAC assessment; however, in that study there was no comparison of model performance with intra- or interobserver agreement or variability, no visual assessment of the quality of the automated segmentations, nor were any associations determined between model outputs and clinically relevant factors.

Abbreviations

FPV = false-positive volume, ICAC = intracranial internal carotid artery calcification, PPV = positive predictive value

Summary

Automated deep learning–based segmentation and volume measurement of intracranial internal carotid artery calcification (ICAC) had similar accuracy to manual assessment by trained observers, and automated ICAC volume was also associated with incident stroke.

Key Points

- Automated delineation of intracranial calcifications reached a sensitivity of 83.8% and positive predictive value (PPV) of 88.0%; this performance compared well to that of manual assessment (73.9% sensitivity and 89.5% PPV).
- In visual examinations of 294 regions by a blinded expert, automated delineations were judged as more accurate than manual delineations in 131 regions, less accurate in 94 regions, and equally accurate in the rest of the regions (131 of 225; $P = .01$).
- Both automated and manual calcium volumes were associated with incident stroke, with adjusted hazard ratios of 1.38 (95% CI: 1.12, 1.75) and 1.48 (95% CI: 1.20, 1.87), respectively.

Keywords

CT, Neural Networks, Carotid Arteries, Calcifications/Calculi, Arteriosclerosis, Segmentation, Vision Application Domain, Stroke

The purpose of this study was to evaluate deep learning–based automated ICAC assessment in terms of its accuracy and clinical value for stroke risk estimation. In the process of validating our model, we assessed our model in three ways: (a) comparison of model segmentation and volume measurement accuracy to an independent observer, using another observer's annotations as the reference standard, (b) comparison of the accuracy of automated segmentations to that of manual segmentations in a qualitative, visual manner (with analysis of error patterns in both manual and automated segmentations), and (c) assessment of ICAC presence and volume in association with stroke for automated and manual ICAC volume measures.

Materials and Methods

Study Participants

The current study focused on a sample of 2319 participants (mean age, 69 years, aged 55 years or older; 1154 women [53.2%]) from the Rotterdam Study (16), a prospective population-based cohort study. The participants underwent a non-contrast CT examination between 2003 and 2006 as part of a study on visualization of arterial calcification. Subsequently, the participants were continuously monitored for incident stroke until January 1, 2016. Further details on the follow-up procedures can be found in Appendix E1 (supplement). The flow of participants and data through our study, with exclusion criteria, is shown in Figure 1.

The Rotterdam Study (16) has been approved by the Medical Ethics Committee (registration number MEC 02.1015) and by the Dutch Ministry of Health, Welfare and Sport (Population Screening Act WBO, license number 1071272–159521-PG).

All participants provided written informed consent to participate in the study and to have their information obtained from treating physicians.

Image Acquisition

Scans used for ICAC assessment were acquired using a 16-section or 64-section multidetector CT scanner (Somatom Sensation 16 or 64; Siemens). Scan parameters were as follows: collimation, 16 mm $\times$ 0.75 mm; 120 kVp; 100 effective mAs; rotation time, 0.5 second; and normalized pitch of 1. Images were reconstructed with in-plane resolution of 0.23 mm $\times$ 0.23 mm, 0.5-mm section spacing, 1-mm section thickness, and 120-mm field of view.

Manual ICAC Assessment

ICAC was segmented bilaterally in the intracranial internal carotid artery from its horizontal petrous segment until the circle of Willis by two physicians (D.B. [3 years of experience] and reader 2 [1 year of experience]). Both readers were blinded to participant data. The observers were trained and supervised exclusively in evaluating ICAC by an expert neuroradiologist with more than 15 years of experience who they could consult in case of doubt while annotating the data. Most scans were segmented by only one observer. A total of 47 randomly selected scans were segmented by both observers independently to assess the interobserver agreement. The segmentation was performed by manually circling calcifications in every section and subsequently selecting pixels with attenuation above 130 HU. This assessment method has been used extensively over the past years with a high intra- and interobserver reliability (2,7,8,17–20).

Automated ICAC Assessment

The pipeline of the automated method is shown in Figure 2. The scan was first automatically preprocessed (as described later) and then was processed by an ensemble of four deep learning networks. Each network has the same architecture similar to U-Net (21) and, to increase diversity in the ensemble (22,23), was trained individually using a different loss function: cross-entropy (10), Dice overlap (24), focal loss (25), or a weighted cross-entropy loss upweighting smaller calcifications (descriptions are in Appendix E2 [supplement]). The ensemble produced four probability maps representing network confidence in classifying pixels as ICAC. The four maps, corresponding to the four networks, were averaged to obtain the final probabilistic segmentation map. The ICAC volume was computed as the number of pixels with probability above 0.5 multiplied by pixel dimensions and section spacing.

Data Preprocessing

We aligned all scans to a single arbitrarily chosen reference image using affine image registration with the default similarity metric (AdvancedMattesMutualInformation) using SimpleElastix toolbox (26) and cropped them along the longitudinal axis so that they contained only the intracranial part of the carotid artery. These steps were done completely automatically.

Figure 1: The flow of participants and data through the study. In this study, we first selected participants for whom both head CT scans and manual intracranial internal carotid artery calcification (ICAC) assessments were available ($n = 2319$; only one scan was available per participant). Participants were randomly partitioned into 10 approximately equal-sized nonoverlapping subsets. Tenfold cross-validation was performed by training the model on each of the 10 training and validation sets obtained by exclusion of corresponding test dataset. Finally, the test sets from cross-validation were aggregated into a dataset of 2319 scans with both manual and automated assessments, which were used to evaluate automated assessments against manual ones. In addition to this set, we also performed analyses on its two subsets: a subset for which independent assessments of two observers were available and a subset we used for stroke association analysis and formed using inclusion criteria from a previous study (2). DL = deep learning.

Figure 2: The automated method's processing pipeline. Step 1: preprocessing. Step 2: processing by four trained networks ("deep ensemble"), outputting ICAC probability maps. Step 3: averaging the maps. Step 4: computing the volume corresponding to pixels with the probability above .5. ICAC = intracranial internal carotid artery calcification.

Then, we reduced the resolution of axial sections so that it matched that of the longitudinal axis. The same transformation was performed on the segmentation maps. These three steps were aimed at reducing the size of the input for deep learning networks, which was necessary to cope with the restrictions imposed by limited computational capacities, particularly GPU memory. The resulting image size was $240 \times 240 \times 100$, and each pixel corresponded to 0.5 mm^3 . We used these registered, cropped, and downsized CT volumes and segmentations both to train and evaluate our networks.

Deep Learning

The choice to use multiple networks instead of one was motivated by the fact that ensemble methods have been shown to boost performance and increase stability of predictions (22), especially when the ensemble is diverse (23). To increase the diversity, the ensemble networks were trained using different objective functions, each weighing the importance of different pixel categories differently. Table E1 (supplement) reports the segmentation performance of individual networks in our ensemble (named after objective functions they use) and compares it to the performance of the entire ensemble. The ensemble achieved higher overall performance compared with its individual networks.

The network architecture, shared by all ensemble members, was similar to our previous version of the method (15, fig 1). However, we made several simplifications: We removed auxiliary classifiers, dropout layers, residual connections, and convolutional layers in concatenation blocks. The member of our ensemble trained using cross-entropy loss function is thus the most

similar to our previous version of the method (15), which was a single network that used the same loss.

The rest of the training parameters were as described here. Batches consisted of two large patches of size $178 \times 178 \times 98$ sampled from the scans. The training duration was 75 epochs, with an epoch defined as iterating through 750 training and 1000 validation patches. The lowest validation loss was used to select network parameters. Adadelta algorithm was used for optimization.

Tenfold cross validation was used for model evaluation. Approximately 1590 CT scans were used for training and 500 for validation in every cross-validation fold; the same data split was used for all four models. Each fold had approximately 230 test CT scans.

The method was implemented in Python using deep learning framework Keras 2.3.1 (<https://keras.io>) with TensorFlow 1.14.0 (<https://www.tensorflow.org>) backend.

Visual Assessment of Segmentations

We randomly sampled 300 two-dimensional image regions centered at either the left or right intracranial carotid artery and having the difference between manual and automated segmentations of at least 2 mm^2 . An expert reader (8 years of experience in neuroradiology and head and neck radiology) indicated the following for every region: (a) whether the manual or automatic segmentation contour (if any) was more accurate and to what extent (slightly, substantially, or equal; five categories in total); (b) whether the visualization permitted assessment of ICAC; and (c) whether at least one

Table 1: Baseline Characteristics of Study Population

Characteristic	Value
Sample size	2156
No. of women	1154 (53.5%)
Scanner type	
16 section	670 (31.1%)
64 section	1486 (68.9%)
Age (y)	69.3 ± 7
Obesity	510 (23.7%)
Hypertension	1586 (73.6%)
Diabetes mellitus	257 (11.9%)
Hypercholesterolemia	1064 (49.4%)
Low high-density lipoprotein	264 (12.2%)
Past or current smokers	1483 (68.8%)
Presence of ICAC	1748 (81.1%)
ICAC volume* (mm ³)	42.9 (6.3–139.7)

Note.—Unless otherwise noted, values are mean ± standard deviation for continuous variables and number of participants with percentages in parentheses for dichotomous variables. ICAC = intracranial internal carotid artery calcification.

* Median with interquartile range in parentheses.

of the contours was accurate. The visualization technique for presenting segmentations blinded the observer to whether the contours had been generated manually or automatically. See Appendix E3 (supplement) for details regarding the sampling and visualization and Movies 1 and 2 for demonstrations of the visualization.

Statistical Analysis

We evaluated the accuracy of automatic segmentations using the observer segmentations as the reference standard. The performance metrics used were recall (sensitivity), precision (positive predictive value [PPV]), and false-positive volume (FPV; volume corresponding to non-ICAC method–detected pixels). In addition, we computed precision-recall and free-response receiver operating characteristic curves.

To assess the variability between manual and automatic volume measures, we used Spearman and intraclass correlation coefficient ([2, 1] in the Shrout and Fleiss convention [27]), as well as Bland-Altman analysis (28).

We used Cox proportional hazards models to relate manually and automatically assessed ICAC presence and volume to incident stroke, adjusting for age, sex, scanner type, obesity, hypertension, diabetes mellitus, hypercholesterolemia, low high-density lipoprotein cholesterol level, and smoking.

Automatic segmentations and volume measurements used to compute performance metrics and association measures were computed in a 10-fold cross-validation procedure (see Fig 1). All metrics and association measures we report were computed using automated assessments on independent test sets (ie, held out, unseen during training) from the cross-validation procedure; scans from all 10 test sets were combined

into one set (unless otherwise specified) to compute the metrics. All statistics were computed using Python package SciPy version 1.0, IBM SPSS Statistics 24, and R version 3.2.3 (R Foundation for Statistical Computing). Significance threshold was $\alpha = .05$.

Results

Participant Overview

A total of 2319 participants were included for the development of the ICAC segmentation and volume measurement model. The mean age at the scan time was 70 years ± 7, 1154 (53.2%) participants were women, and 1486 (69%) were scanned using the 64-section scanner. In the set of 47 participants for whom ICAC was annotated by two observers independently, the mean age was 67 years ± 5, 21 (45%) participants were women, and 41 (89%) were scanned using the 64-section scanner. Table 1 shows clinical characteristics of a subset of participants included into stroke association analysis (White, no prevalent stroke at the scan time).

Model Training

The time required to train one network (of four) in the ensemble was approximately 1 day. The average time the method took to process one scan was 118 seconds (98 seconds for preprocessing, 20 seconds for applying the ensemble and combining its predictions) using Intel Xeon E5645 processor (six cores, 2.40 GHz) and Nvidia GeForce GTX 1070 graphics card.

Segmentation Performance

Figure 3 shows a comparison of the performance of the automated method with that of the observers using precision-recall curve and receiver operating characteristic curve analysis. All performance metrics for assessing segmentation performance of the method and the observers are summarized in Table 2.

To assess the training stability of the method, we computed standard deviations of all metrics across the cross-validation folds. The standard deviation for dataset-wise recall was 83.8% ± 1.8; precision, 88% ± 1; participant-wise recall, 80.6% ± 1.7; FPV among participants with ICAC, 15.7 mm³ ± 2.4; and FPV among ICAC-free participants, 6.2 mm³ ± 2.4, indicating that the method provided consistent performance when trained on different data subsets.

Volume Measurement Performance

The intraclass correlations between the automatic and manual volume measures and between the two observers' measures were 0.98 (95% CI: 0.97, 0.98) and 0.91 (95% CI: 0.84, 0.95; $P = .04$), respectively. The corresponding Spearman correlations were 0.95 (95% CI: 0.84, 0.96) and 0.97 (95% CI: 0.92, 0.99; $P = .38$), respectively. The 95% CIs for Spearman correlations and significance tests for both correlations were computed by bootstrapping with 10 000 replications.

Figure 4 shows Bland-Altman plots of the difference between automatic and manual volume measures and the difference between cubic roots thereof.

Figure 3: Automatic and manual intracranial internal carotid artery calcium (ICAC) segmentation performance. The performance was evaluated on all scans (blue curve) and those annotated by both observers (green curve). The curves were computed by varying the threshold for ICAC probability maps (from 0 to 1); every point thus represents recall and precision (for PRC) or false-positive value (FPV) (for FROC) computed over all pixels in all scans (for PRC) or averaged among participants with ICAC (for FROC) using a specific threshold. Stars represent the interobserver agreement: dataset-wise (ie, across all pixels in all scans, averaged per-participant recall and precision with FPV of reader 2, with reader D.B. used as the reference standard). Dots represent the model performance when α level of .5 was used as the probability cutoff. The interobserver agreement points lie on or under both PRC and FROC, which indicates that the method segmentations agreed with D.B. at least as much as reader 2, when an appropriate threshold was chosen. FP = false positive, ROC = receiver operating characteristic.

Visual Assessment of Segmentations

Six of 300 regions could not be graded, either because the observer could not infer the region orientation from the visualization due to its limited scope or because of the limited spatial resolution. Table 3 presents visual comparison of accuracy of automatic and manual segmentations for the following subsets of regions: (a) all gradable regions, (b) regions in which the method did not miss any ICAC pixels (false-positive findings), and (c) regions in which it did not segment any pixels not indicated in the manual segmentations (false-negative findings). Manual and automatic segmentations were both inaccurate in only six of 294 gradable regions.

ICAC and Incident Clinical Stroke

We assessed associations of presence and volume of ICAC with stroke incidence. For both ICAC presence and volume measures, stroke associations of manual and automated assess-

Table 2: Automatic and Manual ICAC Segmentation Performance

Reference Standard and Performance Metric	Model or Reader		
	Model (<i>n</i> = 2319)	Model (<i>n</i> = 47)	Reader 2 (<i>n</i> = 47)
Reference segmentation	D.B. + reader 2	D.B.	D.B.
Dataset-wise statistics			
Recall	83.8%	75.7%	73.9%
Precision	88.0%	92.2%	89.5%
Participant-wise statistics			
With ICAC (<i>n</i> = 1897)*			
Recall	80.6% $\pm$ 20.8	65.7% $\pm$ 29.7	70.0% $\pm$ 29.8
FPV (mm ³)	15.8 $\pm$ 24.5	7.2 $\pm$ 17.3	10.6 $\pm$ 20.0
Volume (mm ³)	150.3 $\pm$ 204.6	121.9 $\pm$ 191.7	121.9 $\pm$ 191.7
ICAC free (<i>n</i> = 422)*			
FPV (mm ³)	6.2 $\pm$ 12.8	2.1 $\pm$ 2.6	0.0 $\pm$ 0.0

Note.—Participant-wise values are mean $\pm$ standard deviation. D.B. and reader 2 are observers. “Reference segmentation” indicates who performed segmentations that were used as reference. The metrics for the second and third columns were computed using the exact same set of reference segmentations. “Volume” is the intracranial internal carotid artery calcification (ICAC) volume for participants with ICAC as assessed by the observers, reported for comparison with FPV. FPV = false-positive volume.

* As indicated by the observer(s).

ments were similar. Adjusted hazard ratios computed for ICAC presence were 2.51 (95% CI: 1.42, 5.85) for automated and 2.52 (95% CI: 1.44, 5.95) for manual presence assessment ($P =$

Figure 4: The Bland-Altman plots of the difference between manual and automatic intracranial carotid artery calcium volume measures. **(A)** The difference between manual and automatic volumes (V_{GT} and V_{pred} , respectively) and **(B)** the difference between the cubic root thereof. Blue dots represent participants with V_{GT} of zero (constituting 18.2% of the dataset). SD = standard deviation.

Table 3: Blinded Qualitative Visual Comparison of Manual and Automatic Segmentations by an Expert

Characteristic	All (<i>n</i> = 294)	False-Positive Findings (<i>n</i> = 101)	False-Negative Findings (<i>n</i> = 138)
Automatic is substantially more accurate	14 (4.8%)	9 (8.9%)	5 (3.6%)
Automatic is slightly more accurate	117 (39.8%)	69 (68.3%)	33 (23.9%)
Both are equally accurate	69 (23.5%)	7 (6.9%)	33 (23.9%)
Manual is slightly more accurate	86 (29.2%)	16 (15.9%)	59 (42.8%)
Manual is substantially more accurate	8 (2.7%)	0 (0%)	8 (5.8%)
<i>P</i> value	<.001	<.001	.008

Note.—Values are number of regions, with percentage in parentheses. *P* values were computed using Wilcoxon signed rank test and represent the significance of the difference between accuracy of manual and automatic assessments.

.99). Adjusted hazard ratios per 1-standard-deviation increase of measured ICAC volume were 1.38 (95% CI: 1.12, 1.75) for automated and 1.48 (95% CI: 1.20, 1.87) for manual volumes ($P = .12$). The CIs and significance tests were computed by bootstrapping with 10 000 replications. We explore differences in volume and attenuation distributions between automatically and manually identified ICAC lesions in Appendix E5 (supplement) and investigate whether these differences may affect the association of automatically computed ICAC volumes with stroke in Appendix E6 (supplement).

Discussion

We developed a fully automated deep learning–based method for ICAC segmentation and volume measurement at noncontrast CT and evaluated it on a large dataset in a 10-fold cross-validation procedure. Accuracy of automated assessment was comparable to or better than manual assessment by trained observers in several aspects: Segmentation and volume measurement performance measured between automated and manual assessments were comparable to those measured between the observers, and in blinded visual assessment of segmentations, automated segmentations were more accurate than manual (131 of 225 [58.2%] regions with automated and manual de-

lineations judged as not equally accurate, $P = .01$).

Furthermore, the method identified lesions missed by the observers, while detecting few non-ICAC structures. In 77% (78 of 101) of false-positive regions, the automated segmentation was more accurate than manual; the opposite happened in only 16% (16 of 101) of cases (see Table 3). However, the pattern was different in false-negative regions: When the method missed observer-indicated ICAC, it was wrong in 49% (67 of 138) and correct only in 28% (38 of 138) of the cases. This suggests that a union of automated and manual segmentations may yield a more accurate measurement of ICAC volume.

The associations with stroke for the automated presence and volume assessments were similar to those for the corresponding manual assessments. However, our analyses in Appendices E5 and E6 (supplement) suggest that there may be differences in kinds of lesions the method and the observers focused on. More specifically, the method detected more small and low-attenuation lesions than the observers, and that could possibly affect the stroke association for automated volume measurements. Nevertheless, ICAC volumes were found to be associated with stroke and thus could be useful for stroke risk assessment models. For example, automated assessment might be used in primary stroke

Figure 5: Examples of manual (blue) and automated (red) segmentations of Rotterdam Study and trauma dataset scans. The images shown were preprocessed (registered and downsized; details on preprocessing can be found in the corresponding subsection of the Materials and Methods). Trauma dataset images were additionally smoothed with a Gaussian filter to make them more similar to Rotterdam Study scans used for training and thus improve the method's performance. Analysis of the method's performance on the trauma dataset can be found in Appendix E4 (supplement).

prevention in the future, similarly to how coronary artery calcium assessment is currently used to reclassify persons at risk for a first cardiac event (29). We would like to emphasize that the purpose of analyzing stroke associations was to further study potential differences between manual and automated measurements and to assess the potential of the latter for use in stroke risk estimation; the purpose was not to demonstrate the association between ICAC and stroke, which was studied before (2).

Apart from automated assessment, another cheaper alternative to manual segmentation-based ICAC assessment is visual

scoring, in which an observer assesses ICAC severity subjectively. Scoring systems categorizing entire scans or arteries into severity grades have substantial (7) to excellent (18) interobserver agreements but show very large variations of ICAC volume within grades. Ahn et al (18), who compared several such systems, concluded that the segmentation-based volume measure is more promising for assessing arteriosclerosis. A section-wise scoring method of Subedi et al (6), shown to be superior to scan-wise scoring, achieved a Spearman correlation of 0.91, whereas for our method this was 0.95. Automated assessment accurately emulates segmentation-based volume measurements

and, unlike visual scoring, does not require any human input, which may make it a better candidate for replacing manual assessment.

The main limitation of the current study was that the method was trained and evaluated on a relatively homogeneous dataset: scans of (mostly White) persons from one country acquired using a standardized protocol on two scanners of one vendor. Evaluating the method on scans of trauma patients reconstructed using a different kernel (see Appendix E4 and Fig E2 [supplement] and Fig 5) showed that, although the agreement with manual annotations was good (intraclass correlation was 0.94 and Spearman correlation 0.95), it was lower than that on the Rotterdam Study data used for training. To generalize our model to data that may be dissimilar from our training data, such as the trauma dataset, the training data could be expanded using scans from the target population of scans (ie, the population we wish to apply the method to) or other type of scans dissimilar to the training data (which would make the training data more heterogeneous and thus may improve generalization). Alternatively, the method itself could be augmented to adapt to new data, or the target data could be transformed to be more similar to the training data.

Another possible limitation was that the quality of manual segmentations used in training and evaluation could have been reduced due to labor intensiveness and tediousness of the annotation process and/or that the annotators did not specialize in neuroradiology. Having a larger number of scans annotated by several observers, including expert neuroradiologists, could allow training a better algorithm and a more comprehensive, accurate comparison between automated and manual assessment.

In this study, we demonstrated an accurate and fast automated method for ICAC assessment. Automated assessment may replace manual assessment in research or clinical settings, facilitating analysis of large amounts of data. Automated assessment could also be used as a starting point for manual annotations, which would speed up annotation and could increase completeness of ICAC burden estimation. The latter use case may also increase completeness of ICAC burden estimation. Automated assessment may thus facilitate studying the causes and clinical consequences of arteriosclerosis, for which ICAC is a proxy. It may thus play an important role in providing a basis for and facilitating the incorporation of ICAC-based imaging markers into clinical practice; for example, ICAC volume may be used in stroke risk assessment.

Acknowledgments: In this study, computations were carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Author contributions: Guarantors of integrity of entire study, G.B., M.d.B.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, G.B., D.B., F.D.; clinical studies, D.B.; statistical analysis, G.B., G.v.T.; and manuscript editing, all authors.

Disclosures of Conflicts of Interest: G.B. disclosed grant to author's institution for this research as part of the research project Deep Learning for Medical Image Analysis (DLMedIA) (project no. P15-26), funded by the Dutch Technology Foun-

dation STW, which is part of the Netherlands Organisation for Scientific Research (NWO) and which is partly funded by the Ministry of Economic Affairs; disclosed money paid to author's institution from Intel for research project of Gerda Bortsova, which is part of the aforementioned DLMedIA project. D.B. disclosed no relevant relationships. F.D. disclosed grant to author's institution from the Netherlands Organisation for Health Research and Development (project no. 104003005). M.W.V. disclosed no relevant relationships. M.K.I. disclosed no relevant relationships. G.v.T. disclosed no relevant relationships. M.d.B. disclosed grants to author's institution from NWO with industrial cofunding from Quantib and Intel; author's institution has patents issued (8811724, EP2240904B1, 7844090, 8126240, 7561727, 7463758).

References

1. Arenillas JF. Intracranial atherosclerosis: current concepts. *Stroke* 2011;42(1 Suppl):S20–S23.
2. Bos D, Portegies MLP, van der Lugt A, et al. Intracranial carotid artery atherosclerosis and the risk of stroke in whites: the Rotterdam Study. *JAMA Neurol* 2014;71(4):405–411.
3. Bos D, Vernooij MW, de Bruijn RFAG, et al. Atherosclerotic calcification is related to a higher risk of dementia and cognitive decline. *Alzheimers Dement* 2015;11(6):e39–47.e1.
4. Kockelkoren R, De Vis JB, de Jong PA, et al. Intracranial carotid artery calcification from infancy to old age. *J Am Coll Cardiol* 2018;72(5):582–584.
5. Kockelkoren R, Vos A, Van Hecke W, et al. Computed tomographic distinction of intimal and medial calcification in the intracranial internal carotid artery. *PLoS One* 2017;12(1):e0168360.
6. Subedi D, Zishan US, Chappell F, et al. Intracranial carotid calcification on cranial computed tomography: Visual scoring methods, semiautomated scores, and volume measurements in patients with stroke. *Stroke* 2015;46(9):2504–2509.
7. Bleeker L, Marquering HA, van den Berg R, Nederkoorn PJ, Majoie CB. Semi-automatic quantitative measurements of intracranial internal carotid artery stenosis and calcification using CT angiography. *Neuroradiology* 2012;54(9):919–927.
8. de Weert TT, Cakir H, Rozie S, et al. Intracranial internal carotid artery calcifications: association with vascular risk factors and ischemic cerebrovascular disease. *AJNR Am J Neuroradiol* 2009;30(1):177–184.
9. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521(7553):436–444.
10. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal* 2017;42:60–88.
11. Dey D, Slomka PJ, Leeson P, et al. Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. *J Am Coll Cardiol* 2019;73(11):1317–1335.
12. Lessmann N, van Ginneken B, Zreik M, et al. Automatic calcium scoring in low-dose chest CT using deep neural networks with dilated convolutions. *IEEE Trans Med Imaging* 2018;37(2):615–625.
13. Chellamuthu K, Liu J, Yao J, et al. Atherosclerotic vascular calcification detection and segmentation on low dose computed tomography scans using convolutional neural networks. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). Piscataway, NJ: IEEE, 2017; 388–391.
14. van Engelen A, Niessen WJ, Klein S, et al. Atherosclerotic plaque component segmentation in combined carotid MRI and CTA data incorporating class label uncertainty. *PLoS One* 2014;9(4):e94840.
15. Bortsova G, van Tulder G, Dubost F, et al. Segmentation of intracranial arterial calcification with deeply supervised residual dropout networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2017; 356–364.
16. Ikram MA, Brusselle GGO, Murad SD, et al. The Rotterdam Study: 2018 update on objectives, design and main results. *Eur J Epidemiol* 2017;32(9):807–850.
17. Taoka T, Iwasaki S, Nakagawa H, et al. Evaluation of arteriosclerotic changes in the intracranial carotid artery using the calcium score obtained on plain cranial computed tomography scan: Correlation with angiographic changes and clinical outcome. *J Comput Assist Tomogr* 2006;30(4):624–628.
18. Ahn SS, Nam HS, Heo JH, et al. Quantification of intracranial internal carotid artery calcification on brain unenhanced CT: evaluation of its feasibility and assessment of the reliability of visual grading scales. *Eur Radiol* 2013;23(1):20–27.
19. Hernández-Pérez M, Bos D, Dorado L, et al. Intracranial carotid artery calcification relates to recanalization and clinical outcome after mechanical thrombectomy. *Stroke* 2017;48(2):342–347.
20. Compagne KCJ, Clephas PRD, Majoie CBLM, et al. Intracranial carotid artery calcification and effect of endovascular stroke treatment. *Stroke* 2018;49(12):2961–2968.

21. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015; 234–241.
22. Kamnitsas K, Bai W, Ferrante E, et al. Ensembles of multiple models and architectures for robust brain tumour segmentation. In: Crimi A, Bakas S, Kuijff H, et al, eds. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2017. Lecture Notes in Computer Science, vol 10670. Springer, Cham., 2018; 450–462.
23. Kuncheva LI, Whitaker CJ. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach Learn* 2003;51(2):181–207.
24. Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Cham: Springer, 2017; 240–248.
25. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2017; 2999–3007.
26. Marstal K, Berendsen F, Staring M, Klein S. SimpleElastix: A user-friendly, multi-lingual library for medical image registration. 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway, NJ: IEEE, 2016; 574–582.
27. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86(2):420–428.
28. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1(8476):307–310.
29. Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* 2019;140(11):e596–e646.