Development and External Validation of a Deep Learning Algorithm to Identify and Localize Subarachnoid Hemorrhage on CT Scans

Thanellas, Antonios ; Peura, Heikki ; Lavinto, Mikko ; Ruokola, Tomi ; Vieli, Moira ; Staartjes, Victor E ; Winklhofer, Sebastian ; Serra, Carlo ; Regli, Luca ; Korja, Miikka

DOI: https://doi.org/10.1212/WNL.0000000000201710

Published Ahead of Print on January 13, 2023 as 10.1212/WNL.0000000000201710

OPEN

Neurology®
The most widely read and highly cited peer-reviewed neurology journal
The Official Journal of the American Academy of Neurology

AMERICAN ACADEMY OF NEUROLOGY®

Neurology Publish Ahead of Print
DOI: 10.1212/WNL.0000000000201710

Development and External Validation of a Deep Learning Algorithm to Identify and Localize Subarachnoid Hemorrhage on CT Scans

Author(s):
Antonios Thanellas[1]; Heikki Peura, MD[2]; Mikko Lavinto[3]; Tomi Ruokola[3]; Moira Vieli, BM[4]; Victor E Staartjes, MD[4]; Sebastian Winklhofer, MD[4]; Carlo Serra, MD[5]; Luca Regli, MD[4]; Miikka Korja[2]

Corresponding Author:
Miikka Korja, miikka.korja@hus.fi

Affiliation Information for All Authors: 1. Department of Information Management, Helsinki University Hospital, Helsinki, Finland; 2. Department of Neurosurgery, University of Helsinki and Helsinki University Hospital, Helsinki, Finland; 3. CGI, Helsinki, Finland; 4. Machine Intelligence in Clinical Neuroscience (MICN) Laboratory, Department of Neurosurgery, Clinical Neuroscience Center, University Hospital Zurich, University of Zurich, Zurich, Switzerland; 5. Department of Neuroradiology, Clinical Neuroscience Center, University Hospital Zurich, University of Zurich, Zurich, Switzerland

Equal Author Contribution:
1 &2: These authors contributed equally to this work.

Contributions:
Antonios Thanellas: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of data
Heikki Peura: Major role in the acquisition of data; Analysis or interpretation of data
Mikko Lavinto: Analysis or interpretation of data
Tomi Ruokola: Analysis or interpretation of data
Moira Vieli: Analysis or interpretation of data
Victor E Staartjes: Analysis or interpretation of data
Sebastian Winklhofer: Analysis or interpretation of data
Carlo Serra: Analysis or interpretation of data
Luca Regli: Analysis or interpretation of data
Miikka Korja: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Study concept or design; Analysis or interpretation of data

Figure Count:
2

Table Count:
5

Search Terms:
[7] Intracerebral hemorrhage, [8] Subarachnoid hemorrhage, [119] CT, artificial intelligence, deep learning, [323] Class III

Acknowledgment:
This work is part of the AI Head Analysis project of the CleverHealth Network ecosystem (e6), and we thank the ecosystem partners for supporting the project.

Study Funding:
The authors report no targeted funding

Disclosures:
The authors report no relevant disclosures.

Preprint DOI:

Received Date:
2022-06-09

Accepted Date:
2022-11-07

Handling Editor Statement:
Submitted and externally peer reviewed. The handling editor was José Merino, MD, MPhil, FAAN.

ABSTRACT

Objective

In medical imaging, a limited number of trained deep learning algorithms have been externally validated and released publicly. We hypothesized that a deep learning algorithm can be trained to identify and localize subarachnoid haemorrhage (SAH) on head computed tomography (CT) scans, and that the trained model performs satisfactorily when tested using external and real-world data.

Methods

We used non-contrast head CT images of patients admitted Helsinki University Hospital between 2012 and 2017. We manually segmented (i.e. delineated) SAH on 90 head CT scans, and used the segmented CT scans together with 22 negative (no SAH) control CT scans in training an open-source convolutional neural network (U-Net) to identify and localize SAH. We then tested the performance of the trained algorithm by using external datasets (137 SAH and 1242 control cases) collected in two foreign countries, and also by creating a dataset of consecutive emergency head CT scans (8 SAH and 511 control cases) performed during on call hours in 5 different domestic hospitals in September 2021. We assessed the algorithm's capability to identify SAH by calculating patient- and slice-level performance metrics, such as sensitivity and specificity.

Results

In the external validation set of 1379 cases, the algorithm identified 136 out of 137 SAH cases correctly (sensitivity 99.3%, specificity 63.2%). Of the 49064 axial head CT slices, the algorithm identified and localized SAH in 1845 out of 2110 slices with SAH (sensitivity 87.4%, specificity 95.3%). Of 519 consecutive emergency head CT scans imaged in

September 2021, the algorithm identified all 8 SAH cases correctly (sensitivity 100.0%, specificity 75.3%). The slice-level (27167 axial slices in total) sensitivity and specificity were 87.3% and 98.8%, as the algorithm identified and localized SAH in 58 out of 77 slices with SAH. The performance of the algorithm can be tested on through a webservice.

Conclusions

We show that the shared algorithm identifies SAH cases with a high sensitivity, and that the slice-level specificity is high. In addition to openly sharing a high-performing deep learning algorithm, our work presents infrequently used approaches in designing, training, testing and reporting deep learning algorithms developed for medical imaging diagnostics.

Classification of Evidence

This study provides Class III evidence a deep learning algorithm correctly identifies the presence of subarachnoid hemorrhage on CT scan.

INTRODUCTION

The use of head CT imaging has continued to increase among adults during the 21st century[1]. Moreover, in keeping with the increasing trend in favouring health care system integrations and consolidations, many countries have centralised radiology services during on-call hours. This leads to significantly higher volumes and complexity of on-call imaging cases, which in turn place increasing pressure on on-call radiologists. In fact, the overall on-call workload for radiologists has quadrupled in the past 15 years[2].

Head computed tomography (CT) scans are among the most frequently requested after-hours imaging studies in hospitals. Head CT scans outside normal working hours are mostly requested by emergency departments, where findings in an urgent head CT scan can change

the patient's medical care. Perhaps the two most common patient groups who are imaged with an urgent head CT scan are headache and stroke patients, for whom any delays in ruling out issues, like intracranial bleedings, may be tragic. Of the types of intracranial bleedings, undiagnosed subarachnoid haemorrhage (SAH) is among the most alarming ones, because if the frequent cause, *i.e.* a ruptured intracranial aneurysm, is left untreated, at least 75% of today's SAH patients die within a year[3]. In middle-aged people, SAH deaths surpass the number of ischemic stroke deaths, and SAH deaths are in fact the most common type of stroke deaths in particularly middle-aged women[4].

Even though the rate of missed or misdiagnosed head CT findings is low especially at academic centres, misinterpretations do happen particularly during after-hours, which are often covered by somewhat less experienced clinicians. It has been found that after-hours head CT reports provided by radiology residents at an academic large centre were inaccurate in 4.6% of the cases[5]. Fortunately, however, only 0.62% of the cases that were not identified or were inaccurately reported were intracranial haemorrhages (one-third of these were SAHs)[5]. These facts considered, the primary research question being addressed in this study was as follows: can a deep learning algorithm correctly identify and localize the presence of SAH on head CT scans.

METHODS

Head CT images for deep learning training. We extracted non-contrast head CT images from the Helsinki University Hospital (HUH) Picture Archiving and Communication Systems (PACS) archive. First, using the HUH electronic medical records, we identified [based on the 10th version of the International Classification of Disease (ICD-10) category code I60] SAH patients treated at HUH between 2012 and 2017 (Table 1). Similarly, we

created a negative control group (no SAH on a head CT scan) by searching for subjects who were admitted to the HUH emergency rooms between 2011 and 2018 (Table 1), imaged with a head CT scan, and discharged home on the same admission day with a discharge diagnosis of headache (ICD-10 codes R51 and G44.2). Since the head CT studies were performed with various multislice CT scanners, reconstructed slice thicknesses varied between 2 and 5 mm (Table 1). Similarly, the used imaging protocols varied by year, scanner and hospital. Secondly, of the tens of thousands of identified cases and controls, we extracted up to 1237 non-contrast head CT studies of the identified SAH patients and 2353 control subjects from the PACS archive, which contains more than 21 million digitally stored Digital Imaging and Communications in Medicine (DICOM) imaging studies. The extracted DICOM image series of SAH patients consisted of axially reconstructed multiplanar reformatted volumes (MPR) imaged with four different CT scanners at HUH hospitals (Table 1). A similar image dataset of control subjects originated from five different CT scanners (Table 1). In 2021, the HUH had altogether 19 different CT scanners. Thirdly, after a slice-wise review of the extracted DICOM image series, two study authors (AT, MK) selected 98 MPR volumes corresponding to 96 SAH patients with one patient having two follow-up CT scans, and 985 MPR volumes corresponding to 949 control people with headache (no SAH detected on head CT scans), as 18 people were imaged at least twice. Apart from SAH, no other inclusion criteria were applied [such as demographics, findings of medical interventions (e.g. aneurysm clips, aneurysm coils, ventricular catheters, etc.), image artefacts, image quality, image reconstruction methods, or image resolution] for the selected MPR volumes of SAH patients.

Segmentation of SAH on head CT images. Figure 1 and our previous publication[6] present the concepts of annotation and segmentation. In brief, using the open-source utility dcm2niix, we converted the selected DICOM images to the Neuroimaging Informatics Technology Initiative (NIfTI) open file format for further processing. A trained medical image analyst

(AT) performed a manual segmentation task (*i.e.* delineated SAH evident on head CT scans) using the open source software ITK-SNAP (e1) and 3D Slicer (e2). Following this, the study neurosurgeon (MK) reviewed and adjusted the segmentations for the training set, but not for a dataset segmented in order to assess a pixel-level algorithm performance. We carried out adjustments to the segmented data only when a mutual (AT, MK) agreement was achieved. These segmentations (*i.e.* ground truths) were drawn only onto the axial MPR planes, since the axial MRPs are also used in clinical diagnostics, and since the resolution of reconstructed coronal and sagittal MPRs was low and rarely informative.

Pre-processing of training images. We down-sampled the 512x512 image resolution to 256x256, which in other words downscaled the NiFTI image slices by a factor of 2 both in horizontal and vertical directions. In down-sampling, we kept the original slice numbers of every scan. We clipped the intensities of the head CT scans using the window range of [0, 150] Hounsfield units. Following this, we divided the segmented and pre-processed NiFTI MPR volumes into training and test sets.

Training of the deep learning algorithm. In training, we used an open-sourced and standard 2-dimensional 5-level U-Net-type architecture[7,8], in which each level consisted of two convolutional layers followed by max-pooling on the downscaling path and up-sampling on the upscaling sides. The number of feature maps per each level was 30, 60, 120, 240 and 480. Simplified, U-Net is a convolutional neural network that has been designed particularly for medical image segmentation. The U-Net architecture is based on fully convolutional layers, and therefore it may be trained with fewer images yet yielding accurate segmentations. In training, the network learns to classify pixels either positive or negative, based on segmented (*i.e.* every pixel including the lesion of interest outlined positive) training images. When the fed input image travels through each convolutional layer, so-called feature maps are generated by superimposing different filters (*i.e.* mathematical functions) on the input image,

and the output value of the filter function is called a feature map. The feature map size changes at each convolutional layer, and the network learns to identify lesion-specific image features. Following training, the network is fed with raw images (no segmentations), and the trained U-Net creates a segmentation mask (*i.e.* outlines identified lesions) as a visual output.

We used the training set simply in training of the algorithm to segment SAH, whereas the small test dataset (8 head CT scans) was reserved for testing the trained model along the training process. Of the 98 head CT scans with SAH, we picked randomly 90 for training. Of these 90 MPR volumes with SAH, 23 were head CT scans taken on admission prior to any invasive treatments. The remaining 67 MPR volumes were postoperative, of which 40 included aneurysm clips and clip-related artefacts, 22 included aneurysm coils and coil-related artefacts, and 5 were volumes showing ventricular catheters. We used the remaining 8 out of 98 MPR volumes of SAH patients as the small test dataset during training in order to continuously evaluate performance of the model. Of the negative (no SAH) control group of 985 head CT scans, we used 22 for training.

External validation. Table 2. For the external validation, we used two different datasets, namely Zurich and CQ500 datasets. These datasets were not used in any training or testing phases. We assessed the algorithm's capability to identify SAH, and reported the results based on patient- and slice-level annotations. We calculated the patient- and slice-level[9] performance metrics of the Zurich dataset. Since the CQ500 dataset included only case-level (not slice-level) annotations, we calculated only patient-level metrics for the CQ500 dataset.

Zurich external dataset. The co-authors from Zurich, Switzerland, selected and extracted head CT images of 100 consecutive SAH patients and 1000 consecutive control subjects (without SAH) from the PACS system of the University Hospital Zurich. In order to retrieve authentic real-world clinical data from another large hospital, we provided no other advice for the case and control selection process. Furthermore, we suggested no limitations apply for

the CT scanners, imaging parameters or imaging dates. We provided the co-authors with the trained algorithm, and the whole external validation process was conducted independently in Zurich. DICOM files were converted to the NiFTI files using the dcm2niix software. Pre-processing was carried out with the scripts provided by the HUH research team, and these operations were run on an offline machine (Windows 10, AMD 1950X 32-Thread, 64 GB RAM, GTX 980 Ti). The algorithm's segmentations were visually checked using ITK-SNAP, and two raters (MV, VS) calculated the slice-level and patient-level performance metrics.

Open source external dataset CQ500. A subset of open source dataset CQ500[9] and its patient-level annotations from three raters as a ground truth was used as the third external dataset for validation. We rated the head CTs of patients as SAH cases when all three raters had annotated accordingly. Similarly, the head CT scan was considered negative (a control) if none of three rates found an intracranial bleeding in the scan. The final set consisted of 37 head CT scans with SAH and 242 head CT scans with no intracranial bleedings.

Simulated real-world validation. Since the external validation set from Zurich originated from a large neurosurgery unit of a tertiary university hospital, which provides emergency care mostly for unconscious patients and patients already diagnosed with emergence lesions on head CT scans, we collected all consecutive emergency head CT scans imaged in September 2021 in five HUH hospitals, which have no neurosurgical services. These five hospitals and their case-mix may therefore better resemble smaller on-call hospitals with head CT imaging facilities but no neurosurgical services. All collected CT scans were anonymized (no radiological reports available), and annotated (slice-level) followingly for SAH by three co-authors (MK, HP, AT). Similar to the CQ500 dataset, an agreement of all three raters was considered as a ground truth. After annotation, we analyzed all head CT scans using the algorithm.

Pixel-level accuracy. Since neither the external validation datasets nor the real-world validation dataset were segmented, *i.e.* they did not include pixel-level information about the true positives and negatives, one co-author (AT) segmented additional 49 SAH cases as described earlier to test the model's pixel-level performance. The co-author (AT) randomly selected 46 SAH out of 1237 non-contrast head CT studies of the identified SAH patients (eTable 1), and included additional three SAH cases in which the diagnosis of SAH was originally missed, despite of positive head CT imaging findings.[10] After segmentation, we analyzed all head CT scans using the algorithm.

Post-processing of segmentations. As a sensitivity analysis, we applied simple post-processing steps to the patient-level segmentations in order to reduce the number of false positive cases. For the Zurich dataset, we visually thresholded the number of cases where only one slice with a single pixel cluster was segmented positive. This single cluster in only one positive slice was considered negative (no SAH detected). For the CQ500 and HUH September 2021 datasets, we computed a Python script to evaluate the thresholding similarly, *i.e.* if the case had only one slice with one segmented SAH cluster, the case was considered negative.

Statistical analyses. Patient-level metrics for the CQ500 and patient-, slice- and pixel-level metrics for the HUH datasets were calculated automatically using Python scripts computed for these tasks. These metrics include sensitivity, specificity, false positive rate, false negative rate, and accuracy. We performed all statistical analyses with the Python package numpy, and generated statistical plots with matplotlib.

Ethical considerations. The local institutional review board of HUH approved the retrospective data collection and study design, and granted a waiver for acquiring an informed consent (HUS/365/2017; HUS/163/2019; HUS/190/2021). According to Finnish legislation, no separate ethics committee approval is needed for retrospective studies that

involve a secondary use of registry or archive data. We gathered all imaging data for algorithm training from the HUH, which consists of 23 separate hospitals and has a catchment area of approximately 2.2 million inhabitants. All five Finnish university hospitals, including the HUH, are publicly funded non-profit organizations that provide tertiary health care services for all people living in Finland, regardless of socioeconomic status, insurance status, or race/ethnicity. Therefore, we believe that the HUH imaging data for algorithm training is not inherently biased or deliberately discriminative. We conducted the study in line with the Declaration of Helsinki[11].

In Switzerland, the study was approved by the Zurich Cantonal Ethics Board (KEK Nr. 2020-02725) and the Data Governance Board of the University Hospital Zurich (Nr. DUP-66).

Data availability. Finnish healthcare data for secondary use can be obtained through FINDATA (Social and Health Data Permit Authority according to the Secondary Data Act). The used Finnish and Swiss healthcare data cannot be shared openly. Access to the CQ500 image set can be obtained through a website (e3). In order to share the algorithm code with others, we uploaded the code to the GitHub repository (e4). For the sake of reliability and transparency, we launched a website (e5), where anyone can test the algorithm performance by uploading head CT scans for analysis.

RESULTS

External validation. The external validation dataset consisted of 1379 head CT scans (137 SAH cases) (Table 2). Few head CT scans from the external validation set were imaged with the same CT scanner (GE Discovery CT750 HD) that was used in imaging the training dataset (Table 1 and 2). The confusion matrices show the patient-level (Table 3) and slice-level (Table 4) results. Figure 2 shows four examples of how the algorithm identified and

localized (*i.e.* segmented) SAH. The overall patient-level sensitivity and specificity were 0.99 and 0.63 for SAH, respectively (Table 3). The 1379 head CT scans were composed of 49064 reconstructed axial slices, of which 2110 included SAH (Table 4). The slice-level sensitivity and specificity were 0.87 and 0.95, respectively (Table 4).

The algorithm incorrectly classified one (0.7%) out of 137 SAH cases as negative (Table 3, eFigure 1). At the slice-level, the false negative misclassification rate was 12.6% (Table 4). In terms of false positives, results of the external validation showed a false positive rate of 36.8% at the patient level (Table 3). Some of the false positive cases were other abnormal findings than SAH. For example, of the 34 false positive cases in the CQ500 dataset, the algorithm falsely segmented one tumor, one artefact, 8 cases with calcifications and 23 cases with no abnormal findings. Similarly, of the 423 false positive cases in the Zurich dataset, 138 (32.6%) were postoperative hematomas/hemostatic sealants, 54 (12.8%) ischemic lesions, 23 (5.4%) chronic subdural hematomas, and 21 (5.0%) tumors. At the slice-level, the false positive rate was 4.7% (Table 4).

Simulated real-world validation. Of the 519 consecutive emergency head CT scans imaged during on-call hours in September 2021 in five smaller HUH hospitals without neurosurgical services, the algorithm identified all 8 SAH cases (Table 5). All CT scanners in the five smaller hospitals were newer and differed from those used in imaging the training dataset. The patient-level sensitivity and specificity were 1.00 and 0.87, respectively (Table 5). The slice-level sensitivity and specificity were 0.75 and 0.99, respectively (Table 5). At the slice-level, the false positive rate was 1.2% (Table 5). Patient- and slice-level IRRs for 519 consecutive head CT scans were high (eTable 2).

Pixel-level accuracy. Since neither the external validation dataset nor the simulated real-world validation dataset included segmented images, we segmented and analyzed additional 49 SAH cases to test the model's pixel-level performance (eTable 1). The slice-level

sensitivity and specificity were 0.78 and 0.97, respectively (eTable 3). At the slice-level, the false positive rate was 3.3% (eTable 3). The pixel-level sensitivity and specificity were 0.53 and >0.99, respectively (eTable 3). The pixel-level false positive rate was <0.01%. Anecdotally, the algorithm also identified three SAH cases that were originally misdiagnosed in real life (eFigure 2). The CT scanners used in imaging the 49 SAH cases (eTable 1) were mostly the same as the scanners used in imaging the training set (Table 1).

Online validation portal. We launched a website (e5), where anyone can test the accuracy of the SAH algorithm by uploading (drag and drop) non-contrast head CT scans for analysis. Axial MPR reconstructions should be converted (with any open source DICOM-to-NiFTI converter) to the NiFTI format prior to uploading in order to fully anonymize the image data. The website is deployed and the analysis of one head CT scan with 30-40 axial MPR slices takes around 30 seconds. The segmentation results are presented in color for visual inspection. The website is open for 180 days following online publication.

Classification of Evidence. This study provides Class III evidence a deep learning algorithm correctly identifies the presence of subarachnoid hemorrhage on CT scan.

DISCUSSION

The presented deep learning algorithm identified SAH correctly in 136 (99.3%) out of 137 cases that were imaged with 7 different CT scanners in two countries (India and Switzerland). The only missed SAH was part of the CQ500 dataset (eFigure 1). In terms of specificity, the algorithm incorrectly segmented SAH in 457 (36.8%) out of 1242 controls. The slice-level false positive rate was 2200 (4.7%) per 46954 axial reconstructed head CT slices. A standard reconstructed head CT scan that is used in clinical diagnostics contains usually 30-40 axial MPR slices. If this algorithm was used in a clinical setting, the algorithm would falsely alarm clinicians about SAH in around every third normal (*i.e.* no SAH) head CT scan, and in these

cases, 1-2 incorrectly segmented slices should be carefully inspected to revise the diagnosis. When designing algorithms for life-threatening emergency conditions, the sensitivity should optimally be close to 100% (*i.e.* no missed cases,), even though 100% sensitivity is a challenging goal even for human eyes. If such an algorithm also has a non-zero false positive rate (less than 100% specificity), this obliges clinicians to inspect every positive case (also true positive cases). This may ensure that the algorithm is not replacing clinicians or radiologists, but acts in real-life medical practice more like a collaborative colleague.

Trained imaging algorithms are frequently based on a high number of images. This same applies to algorithms for intracranial haemorrhages, which are often trained with a high number of annotated images[12]. Our approach of using a small number of real-word training images with pixel-level segmentations instead of slice-level annotations may encourage others to adopt a similar strategy in training deep learning algorithms. When training images are segmented, large image datasets are less often needed, and deep learning projects become possible also in smaller medical centres. In addition to high-quality training, a validation process is of paramount importance. Although the sensitivity and specificity of internally validated imaging algorithms for SAH can be very high, their performance metrics when tested with external clinical data are often compromised[12]. Since prior studies reporting deep learning algorithms that localize and identify SAH on head CT scans are scarce, any comparison between our and previous studies is difficult. In a seminal study on which the CQ500 dataset is based and made available for the public, the highest patient-level sensitivity and specificity for identifying (not localizing) SAH were 92% and 90%, respectively[9]. Another deep learning solution, the results of which were validated using an external dataset of a reasonable (>100 positive cases) size, patient-level results showed sensitivity and specificity of 85% and 97%, respectively[13]. In a large external validation study of the world's first and most widely used commercial deep learning solution [which can only interpret thin

0.5-1 mm axial CT images of modern (>64 slices) CT scanners] for identifying intracranial haemorrhages, the patient-level sensitivity for identifying SAH was 93%[14]. Apparently, many previous algorithms have probably been optimized not only for sensitivity but also for specificity, at the expense of sensitivity. In order to avoid a deep learning model surpassing clinicians, our approach was to reach a very high sensitivity and a lower specificity, in which case a clinician-deep learning model collaboration may become more likely. Interestingly, 56% of false positives in our Zurich dataset were in fact other pathological lesions, such as postoperative hematomas. Indeed, the accuracy and particularly the false positive rate of the algorithm can vary depending on "natural confounders" (other blood-containing pathological lesions) and intended use (*e.g.* not intended to be used in postoperative imaging).

One of the study strengths may be that the training dataset included preoperative as well as postoperative artefacts and distortions. The training dataset was imaged using different CT scanners, thus perhaps improving the generalisability of the algorithm. Moreover, since the external validation was conducted by using international datasets, and since the simulated real-world validation dataset consisted of all consecutive head CT scans imaged in September 2021 in five different hospitals with five recently purchased modern CT scanners (none of which were used in imaging any other head CT scans in this study), these results may be generalizable. In addition, benchmarking to our results is feasible with the open-sourced CQ500 dataset. It is generally recommended to use not only open-sourced deep learning tools but also open-sourced datasets when available. We used open source tools for segmentations, file conversions and algorithm development. Despite having no influence on the selection process of images in India and Switzerland, these datasets may still somehow represent optimal cases for our algorithm, and therefore the results can be an overestimate. Since reproducing results based on machine learning algorithms is practically impossible by other researches and hospitals, we also launched a website (e5), where anyone can test the

performance of the algorithm by uploading head CT images in a NiFTI format (*i.e.* fully anonymized data) for validation. Moreover, many deep learning algorithms are incapable of illustrating, visualizing and delineating abnormal imaging findings, whereas the presented algorithm highlights SAH. This visualization may ease and fasten the image interpretation.[15] As a further matter, the used U-Net architecture is small and can therefore be deployed on computers and devices with little computing power. Finally, we shared the algorithm for research purposes and further development in GitHub (e4). Maybe even low-income countries can benefit from this solution.

The training dataset consisted of Finnish people. Since Finns are genetically considered an independent subpopulation of the European population[16], our algorithm may be biased. Particularly the false positive rate varied between datasets. Whether this depends on the race, remains to be studied. In addition, we lack a Conformité Européenne (CE) mark for the algorithm, which belongs to high risk classes (IIa, IIb, and III) of medical devices. Such accredited assessment and issuing the CE mark are expensive and time-consuming processes, and many university hospitals have little capability to productize medical devices. Moreover, since only one dataset was segmented (*i.e.* every pixel with SAH was delineated), and this dataset came from Finnish hospitals, we were able to calculate pixel-level performance metrics only for this dataset (eTable 3). Since a ground truth segmentation for SAH on head CT scans is a rather impractical measure (*i.e.* it is challenging for experts to agree about true positives and negatives at the pixel level), pixel-level results are clinically less meaningful and seldomly, if ever, reported. However, the pixel-level results were satisfactory (eTable 3), and false positive segmentations consisted of small clusters of incorrectly segmented pixels (results not shown). Inspecting small clusters of false positive pixels (the pixel-level false positive rate <0.01%) in a few slices (the slice-level false positive rate 4.7%) per head CT volume (the patient-level false positive rate 36.8%) puts unlikely a strain on radiologists or

clinicians. However, depending on the intended use, the number of false positive pixels could be decreased with simple postprocessing steps (*e.g.* by ignoring dispersed small pixel clusters) and further development. Finally, we did not test the algorithm prospectively in any emergency room setting. This is an unfortunate but most common shortcoming in developing medical imaging algorithms, as implementing a research algorithm in a hospital PACS system and clinical workflow is legally and technically a cumbersome process, which in addition to financial resources may require close collaboration with the PACS solution provider. However, the simulated real-world validation dataset with all consecutive cases from five hospitals resembled a prospective study setup in this context. On the other hand, the patient-level balance between positive and negative findings varies significantly between every hospital and institution, and therefore even our real-world sensitivity and specificity figures may be imperfectly generalizable.

In conclusion, a similarly trained simple SAH algorithm could serve as a useful tool to assist the diagnosis of SAH in a clinical setting. Since the presented algorithm lacks the CE mark, the algorithm cannot yet be used for a clinical purpose.

http://links.lww.com/WNL/C554

REFERENCES

1. Smith-Bindman R, Kwan ML, Marlow EC, et al. Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016. JAMA : the journal of the American Medical Association. 2019;322:843–856.

2. Bruls RJM, Kwee RM. Workload for radiologists during on-call hours: dramatic increase in the past 15 years. Insights into imaging. 2020;11:121–127.

3. Korja M, Kivisaari R, Jahromi BR, Lehto H. Natural History of Ruptured but Untreated Intracranial Aneurysms. Stroke. 2017;48:1081–1084.

4. Rautalin I, Kaprio J, Korja M. Burden of aneurysmal subarachnoid haemorrhage deaths in middle-aged people is relatively high. J Neurology Neurosurg Psychiatry. 2020;92:563–565.

5. Strub WM, Leach JL, Tomsick T, Vagal A. Overnight Preliminary Head CT Interpretations Provided by Residents: Locations of Misidentified Intracranial Hemorrhage. AJNR American journal of neuroradiology. 2007;28:1679–1682.

6. Thanellas A, Peura H, Wennervirta J, Korja M. Machine Learning in Clinical Neuroscience, Foundations and Applications. Acta Neurochir Suppl. 2021;134:153–159.

7. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Arxiv. Epub 2015.

8. Gibson E, Li W, Sudre C, et al. NiftyNet: a deep-learning platform for medical imaging. Comput Meth Prog Bio. 2018;158:113–122.

9. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet. 2018;392:2388–2396.

10. Vehviläinen J, Niemelä M, Korja M. [Diagnostic challenges of aneurysmal subarachnoid hemorrhage]. Duodecim Lääketieteellinen Aikakauskirja. 2016;132:461–465.

11. (WMA) WMA. Declaration of Helsinki. Ethical Principles for Medical Research Involving Human Subjects. Jahrbuch Für Wissenschaft Und Ethik. 2009;14:233–238.

12. Yeo M, Tahayori B, Kok HK, et al. Review of deep learning algorithms for the automatic detection of intracranial hemorrhages on computed tomography head imaging. J Neurointerv Surg. 2021;13:369–378.

13. Salehinejad H, Kitamura J, Ditkofsky N, et al. A real-world demonstration of machine learning generalizability in the detection of intracranial hemorrhage on head computerized tomography. Sci Rep-uk. 2021;11:17051.

14. Voter AF, Meram E, Garrett JW, Yu J-PJ. Diagnostic Accuracy and Failure Mode Analysis of a Deep Learning Algorithm for the Detection of Intracranial Hemorrhage. J Am Coll Radiol. 2021;18:1143–1152.

15. Watanabe Y, Tanaka T, Nishida A, et al. Improvement of the diagnostic accuracy for intracranial haemorrhage using deep learning–based computer-assisted detection. Neuroradiology. 2021;63:713–720.

16. Consortium EA, Lek M, Karczewski KJ, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–291.

TABLES

Table 1 The training and control datasets used in the algorithm development. Of the 98 head CT scans with SAH, 90 were randomly picked for training and the remaining 8 were used to continuously evaluate the performance of the deep learning model during training. Of the 985 control head CT scans, 22 and 963 were used for training and testing (during training), respectively. The HUH hospital organisation has 23 hospitals, and head CT scans were performed in different hospitals of the same hospital organisation.

	SAH	Controls
SAH cases	98	985
• Women (%)	68 (69.4)	525 (53.3)
• Men (%)	30 (30.6)	438 (44.5)
• Other (%)	0 (0.0)	5 (0.5)
• Not available	0 (0.0)	17 (1.7)
Number of axial slices		
• SAH	1681	34994
• No SAH	3555	
Slice thickness in mm, mean (SD)	4.2 (0.7)	4.0 (0.3)
Slices per case, mean (SD)	36 (8)	36 (7)
Age in years, mean (SD)	56.7 (12.9)	47.6 (15.6)
SAH cases per scanners	98	985
• Siemens Somatom Definition Edge	1	279
• Siemens Somatom Definition AS+	8	478
• Siemens Somatom Definition Flash	0	45
• GE Lightspeed VCT	84	14
• GE Discovery CT750 HD	5	152
• Not available	0	17

Table 2 The external validation datasets. The five most common imaging diagnoses are presented. Apart from the listed diagnoses, head CT imaging studies were performed for people with for example epileptic seizures, headaches and acute neurological deficits. The dataset also included postoperative images with artefacts. n/a = data not available, NPH = normal pressure hydrocephalus.

	Zurich SAH	Zurich controls	CQ500 SAH	CQ500 controls
Cases ● Women (%)	100 62 (62.0)	1000 442 (44.2)	37 n/a	242 n/a
Number of axial slices	2110	46954	1327[#]	11720[#]
Age in years, mean (SD)	55.2 (13.4)	60.0 (18.8)	n/a	n/a
Axial slices, mean (SD)	44.3 (6.7)	44.6 (15.2)	35.9 (9.3)	48.4 (55.9)[†]
Diagnoses (%) ● Aneurysmal SAH ● Traumatic brain injury ● CSDH* ● Hydrocephalus (NPH) ● Various tumors	 100 (100) 0 (0) 0 (0) 0 (0) 0 (0)	 0 (0) 232 (23.2) 67 (6.7) 62 (6.2) 36 (3.6)	n/a	n/a
CT scanners	● Siemens Somatom Definition Flash		● GE BrightSpeed ● GE Discovery CT750 HD ● GE LightSpeed ● GE Optima CT600 ● Philips MX 16-slice ● Philips Access-32 CT	

*CSDH = chronic subdural hematoma

[#]slices were not annotated

[†]numerous CQ500 head CTs were thin-slice scans without reconstructions

Table 3 Patient-level results of the external validation. CI = confidence interval.

	Zurich SAH	Zurich controls	CQ500 SAH	CQ500 controls	SAH cases in total	Controls in total
Cases	100	1000	37	242	137	1242
Predicted SAH	100	423	36	34	136	457
Sensitivity (95% CIs)	1.00 (0.96-1.00)		0.97 (0.86-1.00)		0.99 (0.96-1.00)	
Specificity (95% CIs)	0.58 (0.55-0.61)		0.86 (0.81-0.90)		0.63 (0.60-0.66)	
False positive rate (95% CIs)	0.42 (0.39-0.45)		0.14 (0.10-0.19)		0.37 (0.34-0.40)	
False negative rate (95% CIs)	0.00 (0.00-0.04)		0.03 (0.00-0.14)		0.01 (0.00-0.04)	
Accuracy (95% CIs)	0.62 (0.59-0.64)		0.87 (0.83-0.91)		0.67 (0.64-0.69)	
CT scanners	• Siemens Somatom Definition Flash		• GE BrightSpeed • GE Discovery CT750 HD • GE LightSpeed • GE Optima CT600 • Philips MX 16-slice • Philips Access-32 CT		• Siemens Somatom Definition Flash • GE BrightSpeed • GE Discovery CT750 HD • GE LightSpeed • GE Optima CT600 • Philips MX 16-slice • Philips Access-32 CT	

Table 4 Head CT slice-level results of the external validation. The CQ500 dataset from India did not have slice-level annotations. Therefore, the dataset was not included in slice-level analyses. CI = confidence interval.

	Zurich SAH	Zurich controls
Slices	2110	46954
Predicted SAH	1845	2200
Sensitivity (95 % CIs)	0.87 (0.86-0.89)	
Specificity (95% CIs)	0.95 (0.95-0.96)	
False positive rate (95% CIs)	0.05 (0.04-0.05)	
False negative rate (95% CIs)	0.13 (0.11-0.14)	
Accuracy (95% CIs)	0.95 (0.95-0.95)	
CT scanners	• Siemens Somatom Definition Flash	

Table 5 Patient- and slice-level results of all 519 emergency head CT scans performed during on call hours in September 2021 in 5 out of 23 hospitals of the study hospital organization (HUH). All 8 SAH cases were traumatic, and all CT scanners were different (as in Table 1) than those used in imaging the training dataset of the model. n/a = not applicable, CI = confidence interval.

	SAH	No SAH	Slices with SAH	Slices without SAH
Number • Women (%)	8 3 (37.5)	511 280 (54.8)	77 n/a	27090 n/a
Age in years, mean (SD)	76.0 (8.9)	67.6 (20.3)	n/a	n/a
Predicted SAH	8	65	58	329
Sensitivity (95% CIs)	1.00 (0.68-1.00)		0.75 (0.65-0.84)	
Specificity (95% CIs)	0.87 (0.84-0.90)		0.99 (0.99-0.99)	
False positive rate (95% CIs)	0.13 (0.10-0.16)		0.01 (0.01-0.01)	
False negative rate (95% CIs)	0.00 (0.00-0.32)		0.25 (0.16-0.35)	
Accuracy (95% CIs)	0.87 (0.84-0.90)		0.99 (0.99-0.99)	
CT scanners	• Siemens Somatom X.cite • Siemens Somatom go.Top • Toshiba Aquilion Prime 80 • Canon Aquilion Prime 80			

FIGURE LEGENDS

Figure 1 Basic concepts of image annotations and segmentations illustrated. In the patient-level annotation, the whole head CT scan (MPR volume) is classified either positive or negative for SAH (**a**). In slice-level annotations, each of the approximately 30-40 axial slices of the head CT scan (MPR volume) is classified either as positive or negative (**b**). In a pixel-level segmentation, the aim is to delineate every positive pixel in every single slice (**c**). Segmentation of SAH is a time-consuming and laborious procedure, and therefore medical images are mostly annotated (not segmented).

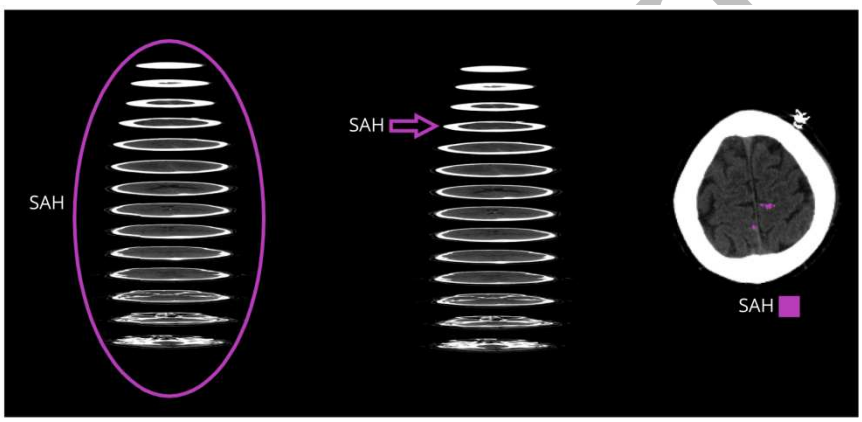
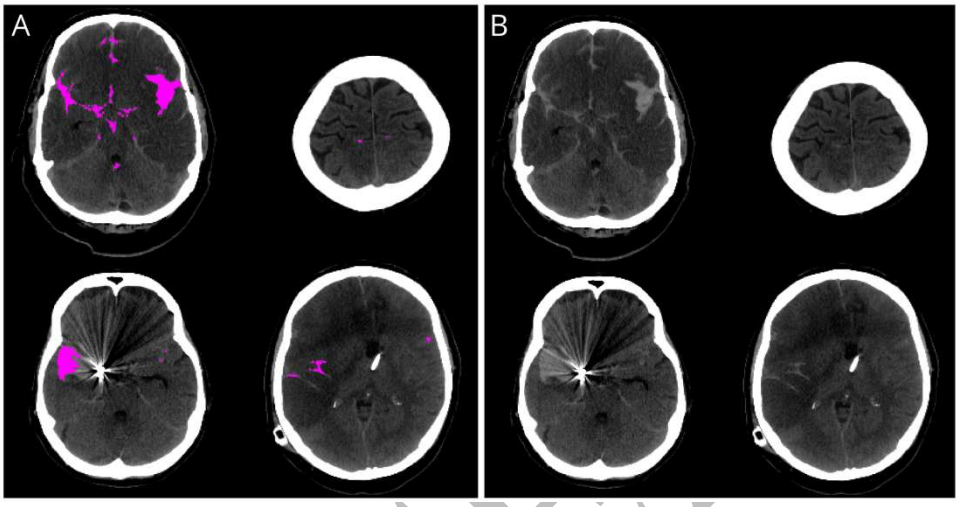


Figure 2 Examples of segmentation results (A) with the trained deep learning algorithm. The overall sensitivity of the algorithm was considered satisfactory, and it identified and localized SAH on axial head CT slices with extensive SAH (upper left image), sulcal SAH (upper right image), streaking clip artefacts (lower left image), and distortions (lower right image). The same images are presented in the panel B without segmentations.



Development and External Validation of a Deep Learning Algorithm to Identify and Localize Subarachnoid Hemorrhage on CT Scans

Antonios Thanellas, Heikki Peura, Mikko Lavinto, et al.

Neurology published online January 13, 2023
DOI 10.1212/WNL.0000000000201710