Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head CT using deep learning: an algorithm development and multicentre validation study

Miguel Monteiro, Virginia F J Newcombe*, Francois Mathieu, Krishma Adatia, Konstantinos Kamnitsas, Enzo Ferrante, Tilak Das, Daniel Whitehouse, Daniel Rueckert, David K Menon†, Ben Glocker†*

Summary

Background CT is the most common imaging modality in traumatic brain injury (TBI). However, its conventional use requires expert clinical interpretation and does not provide detailed quantitative outputs, which may have prognostic importance. We aimed to use deep learning to reliably and efficiently quantify and detect different lesion types.

Methods Patients were recruited between Dec 9, 2014, and Dec 17, 2017, in 60 centres across Europe. We trained and validated an initial convolutional neural network (CNN) on expert manual segmentations (dataset 1). This CNN was used to automatically segment a new dataset of scans, which we then corrected manually (dataset 2). From this dataset, we used a subset of scans to train a final CNN for multiclass, voxel-wise segmentation of lesion types. The performance of this CNN was evaluated on a test subset. Performance was measured for lesion volume quantification, lesion progression, and lesion detection and lesion volume classification. For lesion detection, external validation was done on an independent set of 500 patients from India.

Findings 98 scans from one centre were included in dataset 1. Dataset 2 comprised 839 scans from 38 centres: 184 scans were used in the training subset and 655 in the test subset. Compared with manual reference, CNN-derived lesion volumes showed a mean difference of 0·86 mL (95% CI –5·23 to 6·94) for intraparenchymal haemorrhage, 1·83 mL (–12·01 to 15·66) for extra-axial haemorrhage, 2·09 mL (–9·38 to 13·56) for perilesional oedema, and 0·07 mL (–1·00 to 1·13) for intraventricular haemorrhage.

Interpretation We show the ability of a CNN to separately segment, quantify, and detect multiclass haemorrhagic lesions and perilesional oedema. These volumetric lesion estimates allow clinically relevant quantification of lesion burden and progression, with potential applications for personalised treatment strategies and clinical research in TBI.

Funding European Union 7th Framework Programme, Hannelore Kohl Stiftung, OneMind, NeuroTrauma Sciences, Integra Neurosciences, European Research Council Horizon 2020

Lancet Digital Health 2020

Published Online
May 14, 2020
https://doi.org/10.1016/S2589-7500(20)30085-6

See Online/Comment
https://doi.org/10.1016/S2589-7500(20)30106-0

*Equal first authors

†Equal senior authors

Biomedical Image Analysis Group, Department of Computing, Imperial College London, London, UK (M Monteiro MSc, K Kamnitsas PhD, Prof D Rueckert PhD, B Glocker PhD); **Division of Anaesthesia, Department of Medicine, University of Cambridge, Cambridge, UK** (V F J Newcombe PhD, F Mathieu MD, K Adatia MD, T Das PhD, D Whitehouse MD, Prof D K Menon PhD); **and sinc(i), FICH-Universidad Nacional del Litoral, CONICET, Santa Fe, Argentina** (E Ferrante PhD)

Correspondence to:
Mr Miguel Monteiro, Department of Computing, Imperial College London, London SW7 2AZ, UK
miguel.monteiro@imperial.ac.uk

Introduction

With an estimated global incidence of more than 60 million cases per year, traumatic brain injury (TBI) is the leading cause of mortality in young adults and a major cause of morbidity worldwide.[1,2] CT is the imaging modality of choice to assess the extent and distribution of injury, provide input to prognostic models, and assess the requirement for surgery.[3] Being able to automatically and accurately quantify lesion load and its progression would provide a more objective basis than qualitative assessment by visual inspection for medical and surgical treatment decision making.

A substantial focus of TBI research has been to refine the current classification schemes into more therapeutically meaningful categories by incorporating information on a patient's genetic, blood, and cerebrospinal fluid biomarkers along with clinical and neuroimaging data.[1,4] Hence, being able to reliably and efficiently differentiate lesion types and compute their spatial distribution, number, and volumes would enable optimised and more individualised treatment strategies. Such automated assessment would also facilitate the analysis of large imaging datasets, which are emerging as an essential research resource. Finally, by far the greatest burden of TBI is in low-income and middle-income countries,[2] where radiological expertise is likely to be be less easily available. Having automatic CT analysis algorithms would be of particular benefit in such contexts.[2]

Substantial inter-centre variability and discordance by radiologists exists when reporting CT scan results from patients with TBI.[5] Automating such quantitative measurements would, in theory, circumvent inter-observer

Research in context

Evidence before this study

We searched PubMed for machine learning or deep learning studies focusing on automated lesion quantification of traumatic brain injury (TBI) in head CT published before Jan 31, 2020, with the terms: ("traumatic brain injury" OR "TBI") AND ("computed tomography" OR "CT" OR "neuroimaging") AND ("deep learning" OR "convolutional neural network" OR "artificial intelligence" OR "machine learning"). This search was not restricted to any language. We supplemented this list with manuscripts from past knowledge of the literature and discussions with colleagues. We used these, as well as those identified in the initial PubMed search, as the basis for a further literature search. This process identified several publications addressing the use of machine learning for TBI in head CT. However, previous approaches to automated assessment of CT images after TBI have been largely limited to the undifferentiated detection of haemorrhagic lesions, with no routine volumetric analysis. Although such binary image-level detection of abnormalities can prove useful for triaging patients in need of urgent medical attention, it has little value for analysis of lesion progression and predictive modelling.

Added value of this study

In this study, we report quantitative multiclass segmentation results using a convolutional neural network (CNN) for intraparenchymal haemorrhage, extra-axial haemorrhage, intraventricular haemorrhage, and perilesional oedema. We show that these lesion types can be detected and measured with high accuracy. These attributes are relevant for image-based diagnosis, assessment of injury type, quantification of injury burden, and measurement of lesion progression, both for clinical care and research. We have made the algorithm freely available to facilitate future research.

Implications of all the available evidence

CNN-based processing of CT images in TBI can be used to quickly and accurately detect the type, distribution, and extent of injury after TBI. Such algorithms are likely to be of use in research studies, facilitate clinical radiology workflows by flagging scans that require urgent attention, aid reporting in resource-constrained environments, and help to detect pathoanatomically relevant features for prognostication and characterisation of lesion progression.

variability and allow for analysis of large-scale imaging datasets. Until recently, attempts to automate acute intracranial haemorrhage segmentation on CT have relied on techniques such as intensity thresholding and active contouring, which still require some degree of manual input, and have only been applied to small datasets, raising concerns about the robustness and generalisability of these models.[6–9] Little past success in this context probably reflects two challenges in working with this patient population. First, the heterogeneity of radiographic phenotypes in TBI makes the development of accurate segmentation rules challenging. Second, the diffuse nature of the injury in a large proportion of patients with TBI renders the manual annotations required to establish a ground truth reference dataset difficult and time consuming.

Convolutional neural networks (CNNs) have emerged as a powerful tool for image segmentation, with the ability to learn complex non-linear mappings between the input image and segmentation.[10] Previous deep learning studies for segmentation of TBI lesions have focused on the segmentation of undifferentiated haemorrhagic lesions, with no attempts to differentiate pathoanatomical lesion types.[11] Although such binary image-level detection of abnormalities might prove useful for triaging patients in need of urgent medical attention, it has little value in supporting precision medicine, quantifying lesion progression in trials of new therapies, or predictive modelling of clinical outcome. Other studies have focused on lesion detection at an image level with differentiation of intracranial haemorrhage types.[12,13] In addition to detection, one study showed qualitative results for segmentation.[13] However, this study provided no quantitative metrics, did not specifically address TBI, and provided no assessment of oedema. Accurate quantification of lesion volumes can only be achieved when using voxel-wise labels (ie, for segmentation of lesions) as opposed to image-level labels (ie, for classification of images). Voxel-wise labels allow for both volume quantification and localisation of lesions, which may be important for improved understanding of the factors that lead to lesion progression and to more clinically relevant prognostic schemes.

We aimed to develop and validate a new, clinically relevant algorithm based on deep CNNs for multiclass, voxel-wise segmentation, volumetric quantification, and detection of TBI lesion types visible in CT.

For the **study protocol** see https://www.center-tbi.eu/

Methods

Study design and participants

The data used in this study were from the Collaborative European Neuro Trauma Effectiveness Research in TBI study (CENTER-TBI, NCT02210221),[14,15] accessed using the Neurobot platform (RRID/SCR_017004, core data version 2.0, release date May 15, 2019). Patients were recruited between Dec 9, 2014, and Dec 17, 2017, in 60 centres across Europe. Data collection, handling, and storage are described in detail elsewhere.[14,15] CT scans were collected as part of standard clinical practice, using various platforms and imaging parameters.[5]

Ethical approval was obtained in accordance with all relevant laws and regulations for each recruiting site, and informed consent by patients or their legal

representative or next of kin was obtained according to local laws and regulations.[14] A complete ethics statement, which contains a comprehensive list of sites, ethical committees, and approval numbers, is available online.[16]

	Dataset 1 (n=27)	Dataset 2 (n=512)
Age (years)	46 (16–77)	58 (6–89)
Sex		
Female	5 (19%)	163 (32%)
Male	22 (81%)	349 (68%)
Mechanism of injury		
Acceleration or deceleration	7 (26%)	111 (22%)
Blow to head or hit object	4 (15%)	77 (15%)
Fall from height	13 (48%)	208 (41%)
Multi-mechanistic	2 (7%)	99 (19%)
Unknown	1 (4%)	17 (3%)
Injury severity		
Mild (GCS 13–15)	7 (26%)	299 (58%)
Moderate (GCS 9–12)	2 (7%)	57 (11%)
Severe (GCS <9)	18 (67%)	136 (27%)
Missing	0	20 (4%)
Time from injury to first CT scan (h)	2·4 (1·2–8·0)	2·0 (0·2–77·0)
Repeat scan done	26 (96%)	412 (80%)
Time from injury to second CT scan (h)	16·0 (5·0–79·0)	19·0 (0·9–190·0)
Interval between CT scans (h)	14·0 (3·6–77·0)	16·0 (0·1–190·0)
Marshall score		
I	2 (7%)	120 (23%)
II	11 (41%)	234 (46%)
III	2 (7%)	29 (6%)
IV	0	6 (1%)
V	0	2 (<1%)
VI	12 (44%)	121 (24%)
Presence of:		
Epidural haematoma	10 (37%)	54 (11%)
Acute subdural haematoma	13 (48%)	223 (44%)
Traumatic subarachnoid haemorrhage	20 (74%)	313 (61%)
Intraventricular haemorrhage	6 (22%)	88 (17%)
Intraparenchymal haemorrhage	18 (67%)	224 (44%)
Cisternal compression	9 (33%)	99 (19%)
Midline shift >5 mm	8 (30%)	71 (14%)
Glasgow Outcome Score at 6 months		
1	6 (22%)	66 (13%)
2	0	0
3	9 (33%)	84 (16%)
4	7 (26%)	126 (25%)
5	2 (7%)	199 (39%)
Missing	3 (11%)	37 (7%)

Data are median (range) or number (%). Some percentages do not add up to 100 because of rounding. GCS=Glasgow Coma Score.

Table 1: **Cohort details for both datasets**

Procedures

For development and internal validation, we use two datasets from CENTER-TBI: dataset 1 and dataset 2. We used a two-step process to acquire a large number of annotated scans (appendix p 4). The scans in dataset 1 were annotated manually in a bespoke segmentation tool (ImSeg, version 1.9, BioMedIA, London, UK) by trained personnel (FM and KA) and checked by two other experts (VFJN and TD). These segmentations were used to develop the initial segmentation model and then excluded from any subsequent training or evaluation to avoid skewing the analysis of results.

With the model developed on dataset 1, we did automatic lesion segmentation on dataset 2. These automatic segmentations were refined manually by trained personnel (FM and KA) using ITK-SNAP

See Online for appendix

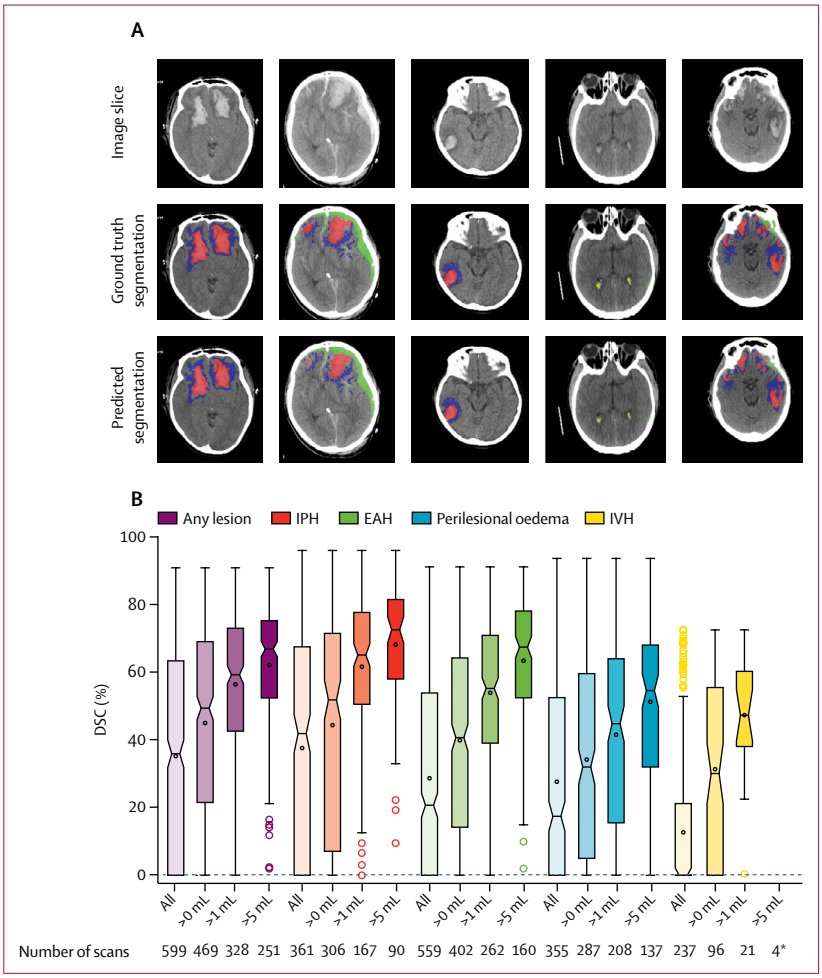

Figure 1: **Qualitative and quantitative multiclass segmentation results**

(A) Qualitative segmentation results. IPH is shown in red, EAH in green, perilesional oedema in blue, and IVH in yellow. (B) Per-class boxplots of DSC progressively including only lesions with volume greater than a threshold. For each individual boxplot, the central line represents the median and the black circle the mean. The box shows the IQR and is indented to indicate the 95% CI of the median. Whiskers adjacent to the boxes represent 1·5 times the IQR. Coloured circles are outliers. The corresponding table is available in the appendix (p 9). DSC=Dice similarity coefficient. EAH=extra-axial haemorrhage. IPH=intraparenchymal haemorrhage. IVH=intraventricular haemorrhage. *Not plotted owing to insufficient data.

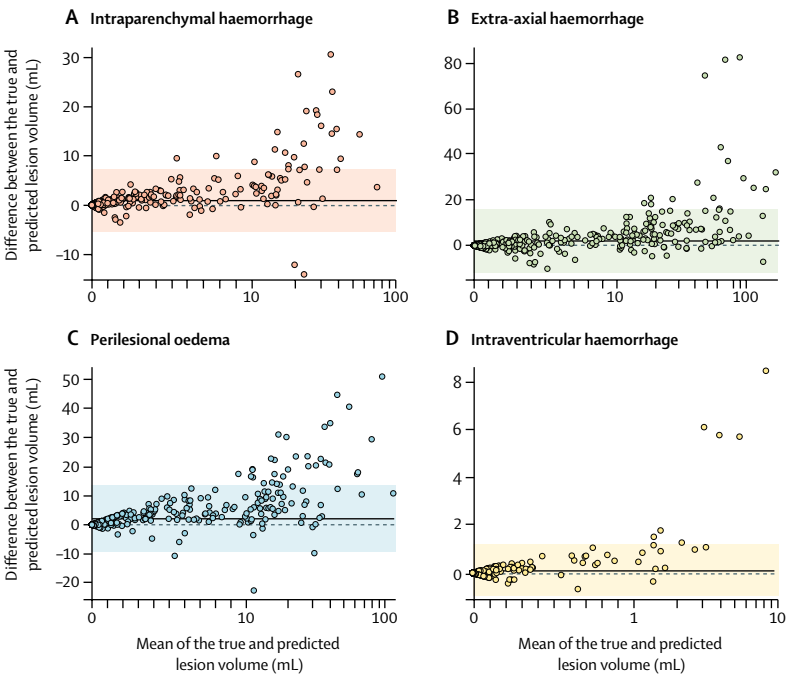

Figure 2: **Bland-Altman plots for lesion volume estimation**
The solid horizontal lines are means and the shaded regions are 95% CIs. The x-axes are on a logarithmic scale to improve visualisation. Axes are plotted on different scales across plots for clarity. Absolute volume errors are shown in the appendix (pp 9–10).

For the **study algorithm** see https://github.com/biomedia-mira/blast-ct/

(version 3.8.0-beta), and the corrections were reviewed by two experts (VFJN and TD) to provide high-quality, accurate ground truth lesion segmentations. The refined segmentations contained four lesion types: intraparenchymal haemorrhage; extra-axial haemorrhage, which includes subdural haematoma, extradural haematoma, and traumatic subarachnoid haemorrhage; perilesional oedema (hereafter referred to as oedema); and intraventricular haemorrhage. Small petechial haemorrhages, which probably arise from diffuse vascular injury and are thought to be a surrogate for accompanying diffuse axonal injury,[17,18] were classified as intraparenchymal haemorrhage.

To establish whether the semi-automatic annotation procedure of dataset 2 provided adequate reproducibility, we did repeat manual segmentation on 20 scans by a single expert (FM) to assess intra-rater reproducibility, and on 25 scans by a second expert (DW) to assess inter-rater variability.

For the subsequent analyses, we split dataset 2 into a training and test set. Different scans from the same patient were placed together in either the training or the test set to avoid the correlation between repeat scans biasing the results. Only scans with more than 1 mL of lesion load were included in the training set to ensure that there was enough training signal for the CNN.

For the segmentation method, we used DeepMedic,[19,20] a three-dimensional CNN with three parallel pathways that process the input at different resolutions. Details on the model and image pre-processing are provided in the appendix (p 2). To facilitate its use in future studies, our algorithm is available online.

For external validation, we used the CQ500 dataset, a publicly available, anonymised, TBI CT dataset provided by the Centre for Advanced Research in Imaging, Neurosciences and Genomics, New Delhi, India.[12,21] This dataset provides image-level labels as opposed to voxel-wise segmentations. However, it is the largest labelled TBI cohort available publicly, and no other dataset provides voxel-wise segmentations.

Outcomes

The primary outcome was the quantification of lesion volume. The secondary outcomes were lesion detection and the assessment of lesion progression.

Statistical analysis

Statistical analysis was done in Python 3.6.8 (appendix p 3). Classic sample size calculation is not directly applicable to CNN-based segmentation. The sample sizes in this work followed the common principle in current deep learning research whereby more data tends to yield better results. Thus, we attempted to maximise the number of scans for training and testing under the constraint of finite resources for expert annotations.

Evaluation metrics were computed and stratified by lesion class and volume. A virtual lesion class (any lesion) consisting of the combined lesion map that merged all lesion types into one was created to allow for evaluation in terms of lesion versus non-lesion.

To assess the performance of the algorithm, we used the Dice similarity coefficient (DSC), which measures the agreement between manual and automatic segmentation. Since the mean DSC is sensitive to lesions with small volumes or scans on which lesions are not present, we report DSC scores for lesions above several volume thresholds. DSC is a well accepted metric for assessing accuracy in image segmentation.[22] However, it is not meaningful when assessing performance with respect to clinical utility (appendix pp 2–3). For a clinically relevant assessment, we have provided additional metrics such as lesion volume estimates and receiver operating characteristic (ROC) curves for lesion detection and lesion volume classification.

To assess the accuracy of the algorithm at estimating lesion volume, we extracted lesion volumes from the manual and predicted segmentations to calculate volume error, which we summarised in Bland-Altman plots. We also assessed the accuracy of the algorithm at quantifying lesion progression. To obtain the error in volume change, we calculated the true volume difference and predicted volume difference between repeat scans for patients in the test set who had repeat scans for which both timepoints could be established.

The output of the segmentation algorithm can be used for lesion detection and lesion volume classification. We used the true lesion volume to set a classification target (eg, target is positive if the true volume is greater than 1 mL and negative otherwise). We then used the predicted lesion volume as the score on which a threshold was varied to calculate ROC curves. We addressed three key lesion detection and lesion volume classification problems to assess the clinical applicability of the model: (1) ability to detect lesions, which is equivalent to classifying lesions with a volume greater than 0 mL; (2) classification of lesions with a volume greater than 1 mL, to enable comparison with findings from datasets that did not contain small lesions; and (3) classification of lesions with a volume greater than 25 mL, equivalent to Marshall grade V/VI,[23] which may indicate lesions requiring surgical intervention.

For each curve, we computed the area under the curve (AUC), its 95% CI using the Hanley and McNeil approach,[24] the sensitivity and specificity of the two operating points (sensitivity at a specificity of 0·90 and vice versa), and their 95% CIs using the Clopper-Pearson method.[25]

We used our algorithm to segment the scans in the CQ500 dataset and to calculate lesion volumes. These are used as the classification score to compare with the ground truth image-level labels provided. This dataset was used only at the end for final validation, never during development. This approach validated the lesion detection performance of our algorithm on an external, independent dataset from a different patient population. CQ500 was not annotated for oedema, and so instead of our summated any lesion class we report on intracranial haemorrhage, which includes all haemorrhage classes in our analysis: intraparenchymal haemorrhage, extra-axial haemorrhage, and intraventricular haemorrhage.

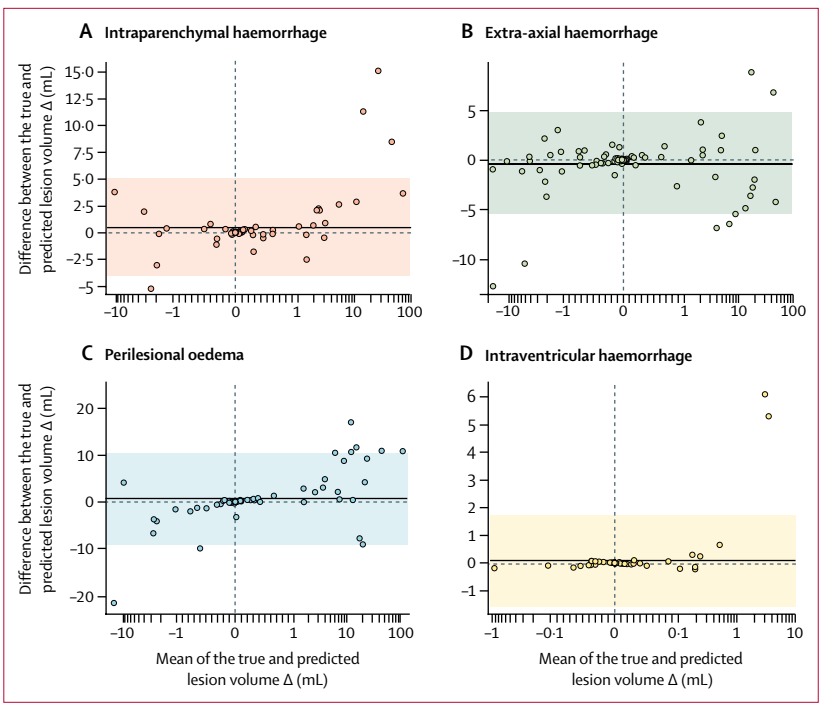

Figure 3: **Bland-Altman plots for lesion progression**
The solid horizontal lines are means and the shaded regions are 95% CIs. The x-axes are on a logarithmic scale to improve visualisation. Axes are plotted on different scales across plots for clarity. Absolute volume change errors are shown in the appendix (pp 9–10).

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Dataset 1 consisted of 98 different CT scanning sessions from 27 patients from one centre (Cambridge University NHS Foundation Trust, Cambridge, UK). Data from this centre were available first as part of a preliminary proof-of-concept study. Dataset 2 consisted of 839 different CT scanning sessions from 512 patients and 38 different centres from which data were available at the time of the study, including Cambridge NHS Foudation Trust. The procedure of semi-automatic segmentation enabled the creation of a much larger dataset (839 *vs* 98 scans) without a commensurate increase in resource requirements. Table 1 shows the cohort characteristics of both datasets, representing the broad spectrum of TBI. From dataset 2, 184 scans were included in the training subset and 655 scans were included in the test subset. Consistent with the known heterogeneity of TBI, 744 (89%) of 839 scans did not contain all four lesion types.[1,15] The distribution of lesions is available in the appendix (p 8).

Figure 1A shows qualitative results for five different cases from our test set, showing the visual agreement between the true and predicted segmentations. Figure 1B shows DSC boxplots. The median DSC for the any lesion class was 36·0% (IQR 0·0–63·4) when including all 599 scans (469 with lesions plus 130 with no lesions but where our model predicted a lesion). In addition to calculating DSCs using all the test scans, we chose the following preplanned thresholds to address different performance levels: 0 mL, 1 mL, and 5 mL (appendix p 9). Limiting the analysis to the 469 scans with lesions increased the median DSC to 49·4% (IQR 21·5–67·1), and the exclusion of lesions of 1 mL or smaller further increased the DSC to 59·3% (42·6–73·1, n=328). A similar relationship between lesion volume and DSC was noted for individual lesion classes (figure 1B). For lesions with a volume greater than 1 mL, the median DSC was 65·2% (IQR 50·6–77·8, n=167) for intraparenchymal haemorrhage, 55·3% (39·1–71·0, n=262) for extra-axial haemorrhage, 44·8% (15·5–64·1, n=208) for oedema, and 47·3% (38·1–60·3, n=21) for intraventricular haemorrhage; for lesion volumes greater than 5 mL, these numbers increased to 72·6% (58·1–81·6, n=90) for

	Number of scans		High-specificity operating point		High-sensitivity operating point		Area under the curve (95% CI)
	Positives	Negatives	Mean sensitivity (95% CI)	Mean specificity (95% CI)	Mean sensitivity (95% CI)	Mean specificity (95% CI)	
>0 mL							
Any lesion	469	186	0·70 (0·66–0·74)	0·90 (0·85–0·94)	0·90 (0·87–0·93)	0·61 (0·54–0·68)	0·89 (0·86–0·91)
IPH	306	349	0·77 (0·72–0·82)	0·90 (0·87–0·93)	0·81 (0·76–0·85)	0·85 (0·80–0·88)	0·87 (0·85–0·90)
EAH	402	253	0·72 (0·67–0·76)	0·90 (0·86–0·94)	0·90 (0·87–0·93)	0·63 (0·57–0·69)	0·89 (0·86–0·91)
Perilesional oedema	287	368	0·80 (0·75–0·85)	0·90 (0·87–0·93)	0·85 (0·80–0·89)	0·82 (0·77–0·85)	0·89 (0·86–0·92)
IVH	96	559	0·70 (0·60–0·79)	0·90 (0·87–0·93)	0·90 (0·82–0·95)	0·75 (0·71–0·78)	0·89 (0·85–0·93)
>1 mL							
Any lesion	328	327	0·89 (0·85–0·92)	0·90 (0·86–0·93)	0·90 (0·87–0·93)	0·87 (0·83–0·91)	0·96 (0·95–0·98)
IPH	167	488	0·96 (0·92–0·98)	0·90 (0·87–0·93)	0·90 (0·85–0·94)	0·97 (0·94–0·98)	0·99 (0·98–1·00)
EAH	262	393	0·89 (0·85–0·93)	0·90 (0·87–0·93)	0·90 (0·86–0·93)	0·89 (0·85–0·92)	0·97 (0·95–0·98)
Perilesional oedema	208	447	0·86 (0·80–0·90)	0·90 (0·87–0·93)	0·90 (0·86–0·94)	0·86 (0·83–0·89)	0·94 (0·92–0·96)
IVH	21	634	0·95 (0·76–1·00)	0·90 (0·87–0·92)	0·90 (0·70–0·99)	0·97 (0·95–0·98)	0·99 (0·95–1·00)
>25 mL							
Any lesion	134	521	0·98 (0·94–1·00)	0·90 (0·87–0·92)	0·90 (0·84–0·95)	0·96 (0·94–0·98)	0·99 (0·98–1·00)
IPH	19	636	1·00 (0·82–1·00)	0·90 (0·88–0·92)	0·95 (0·74–1·00)	0·94 (0·92–0·96)	0·99 (0·97–1·00)
EAH	61	594	0·98 (0·91–1·00)	0·90 (0·87–0·92)	0·90 (0·80–0·96)	0·97 (0·96–0·99)	0·99 (0·98–1·00)
Perilesional oedema	36	619	0·89 (0·74–0·97)	0·90 (0·88–0·92)	0·92 (0·78–0·98)	0·89 (0·87–0·92)	0·98 (0·95–1·00)
External validation set CQ500							
ICH	205	285	0·59 (0·51–0·65)	0·90 (0·86–0·93)	0·90 (0·85–0·94)	0·51 (0·45–0·56)	0·83 (0·79–0·87)
IPH	134	356	0·76 (0·68–0·83)	0·90 (0·87–0·93)	0·89 (0·82–0·94)	0·74 (0·69–0·79)	0·90 (0·86–0·94)
EAH	119	371	0·49 (0·39–0·58)	0·90 (0·87–0·93)	0·91 (0·84–0·95)	0·38 (0·33–0·43)	0·80 (0·75–0·85)
IVH	28	462	0·89 (0·72–0·98)	0·90 (0·87–0·93)	0·93 (0·76–0·99)	0·68 (0·63–0·72)	0·95 (0·89–1·00)

The high specificity and high sensitivity operating points were obtained using a cutoff of 0·90 or the closest possible available. The 0 mL threshold is equivalent to lesion detection. EAH=extra-axial haemorrhage. ICH=intracranial haemorrhage. IPH=intraparenchymal haemorrhage. IVH=intraventricular haemorrhage.

Table 2: **Multiclass detection and classification results for three volume thresholds and detection results for the external validation dataset CQ500**

intraparenchymal haemorrhage, 67·5% (52·5–78·2, n=160) for extra-axial haemorrhage, and 54·6% (32·0–68·1, n=137) for oedema. To compare with previous literature, we combined intraparenchymal haemorrhage and extra-axial haemorrhage and obtained a median DSC of 72·0% (59·2–80·1, n=210) for lesion volume greater than 5 mL.

Figure 2 shows Bland-Altman plots of the agreement between the true and predicted lesion volumes. The mean difference was 0·86 mL (95% CI –5·23 to 6·94) for intraparenchymal haemorrhage, 1·83 mL (–12·01 to 15·66) for extra-axial haemorrhage, 2·09 mL (–9·38 to 13·56) for oedema, and 0·07 mL (–1·00 to 1·13) for intraventricular haemorrhage. For lesions with a volume greater than 5 mL, the median absolute error was 3·57 mL (IQR 1·96 to 7·97, n=90) for intraparenchymal haemorrhage and 4·57 mL (2·18 to 8·88, n=160) for extra-axial haemorrhage. For further discussion regarding absolute volume error see the appendix (p 3). Regarding the reproducibility of the manual annotation procedure, for intra-rater reproducibility (n=20) and inter-rater variability (n=25), we obtained agreements in the range of 0·90–1·00 for all lesion types (appendix p 8).

98 patients in the test set who had repeat scans for which both timepoints could be established (196 scans) were included in the calculations of true and predicted volume difference. Figure 3 presents Bland-Altman plots of the agreement between the true and predicted lesion volume change. The mean difference was 0·46 mL (95% CI –4·04 to 4·97) for intraparenchymal haemorrhage, –0·37 mL (–5·42 to 4·69) for extra-axial haemorrhage, 0·68 mL (–9·03 to 10·39) for oedema, and 0·12 mL (–1·48 to 1·71) for intraventricular haemorrhage. In the appendix (p 3), we show that our algorithm enables localisation of lesions (ie, the quantification of lesion volume by brain region).

Table 2 and figure 4 show the results of lesion volume classification and lesion detection for external validation. For image-level detection of lesions, we obtained an AUC of 0·89 (95% CI 0·86–0·91) for the any lesion class, 0·87 (0·85–0·90) for the intraparenchymal haemorrhage class, 0·89 (0·86–0·91) for the extra-axial haemorrhage class, 0·89 (0·86–0·92) for the oedema class, and 0·89 (0·85–0·93) for the intraventricular haemorrhage class. For the 1 mL threshold, the AUCs increased to 0·96 (0·95–0·98), 0·99 (0·98–1·00), 0·97 (0·95–0·98), 0·94 (0·92–0·96), and 0·99 (0·95–1·00), indicating that most of the missed lesions are very small. For the classification of large lesions (>25 mL), the AUCs were 0·99 (0·98–1·00) for any lesion, 0·99 (0·97–1·00) for intraparenchymal haemorrhage, 0·99 (0·98–1·00) for extra-axial haemorrhage, and 0·98 (0·95–1·00) for oedema.

On the external validation set, we reported an AUC of 0·83 (95% CI 0·79–0·87) for the intracranial haemorrhage class, 0·90 (0·86–0·94) for the intraparenchymal haemorrhage class, 0·80 (0·75–0·85) for the extra-axial haemorrhage class, and 0·95 (0·89–1·00) for the intraventricular haemorrhage class.

Discussion

In this study, we found that the voxel-wise segmentation produced by a CNN can be used for volumetric quantification and detection and classification of multiclass TBI lesions in head CT, as well as for the assessment of lesion progression. We were able to accurately quantify and detect lesions on an external, independent dataset. To our knowledge, this is the largest study so far to use a ground truth reference of manually annotated and manually corrected automatic segmentations of CT scans. The size and diversity of this multicentre dataset provide insights into the performance of deep learning in a real-world clinical scenario. We extend findings from previous studies[11–13] by providing quantitative volumetric results separately for intraparenchymal haemorrhage, extra-axial haemorrhage, intraventricular haemorrhage, and perilesional oedema.

The CNN provided a well calibrated prediction of lesion volume since differences between the true and predicted volumes were small when compared with the overall lesion volume. The funnelling observed can be explained by lesions being predicted where there were none and vice versa, which mostly occurs for smaller lesions. For comparison, previous work[11] reported a median absolute error of 8·83 mL (n=39) for intraparenchymal haemorrhage and extra-axial haemorrhage lesions combined while considering only lesions with a volume greater than 5·5 mL. In our analysis, we did fine-grained segmentation of these two classes individually and validated our CNN on a larger dataset. For lesions with a volume greater than 5 mL, our median absolute error was smaller than that reported previously[11] for intraparenchymal haemorrhage and extra-axial haemorrhage.

The potential clinical applicability of the volume estimates is further confirmed by our results on lesion progression. Such progression of intracranial lesions represents a major target for therapies in the acute phase. For example, cerebral contusions are common after TBI, occurring in up to two-thirds of patients admitted to hospital,[26,27] and progression of such lesions is common, occurring in up to half of patients within the first 24–48 h.[28–30] The ability to automatically monitor lesion progression offers key opportunities to improve patient stratification, guide and monitor management, and investigate potential causes and risk factors for lesion progression in large cohort studies such as CENTER-TBI.[15] Until now, the identification of factors that predict or cause contusion progression, or both, has been hampered by the need to estimate lesion volume and change manually, restricting analyses to small sample sizes.[28–30]

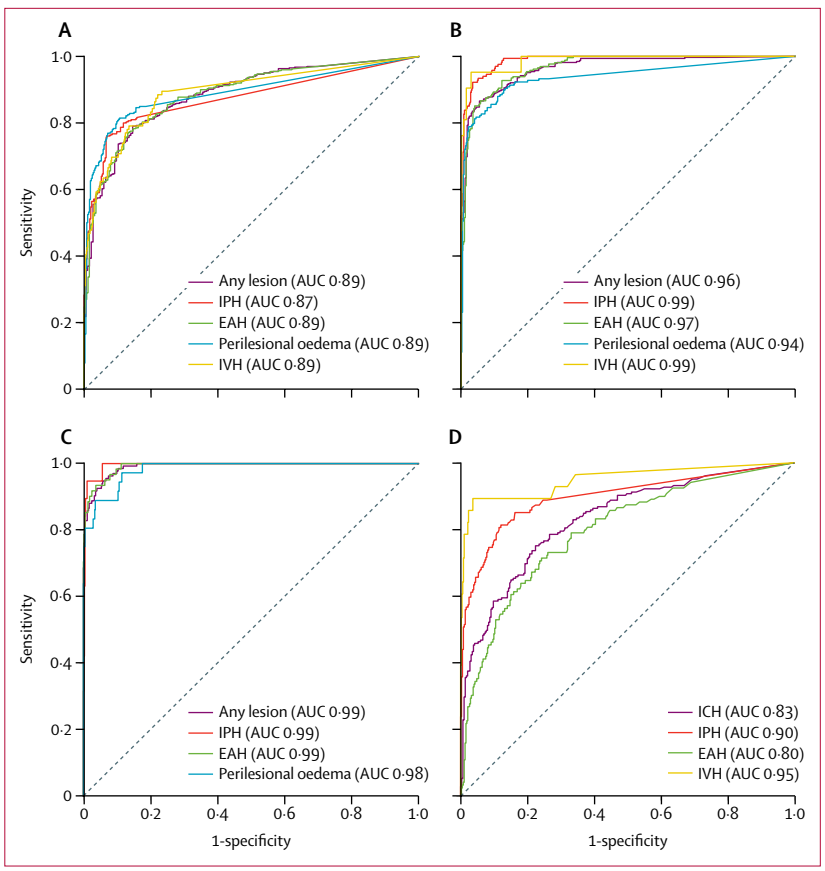

Figure 4: **Receiver operating characteristic curves for lesion detection and classification**
Classification of lesions with a volume greater than 0 mL (A), greater than 1 mL (B), and greater than 25 mL (C) on the internal validation set, and detection of lesions on the external validation set CQ500 (D). AUC=area under the curve. EAH=extra-axial haemorrhage. IPH=intraparenchymal haemorrhage. IVH=intraventricular haemorrhage.

Regarding the underlying lesion segmentation, the DSC increased with lesion volume, illustrating that the DSC is sensitive to small or non-existent lesions, which is a limitation of the metric. The median DSC of 73·0% (n=39) reported previously for large intraparenchymal haemorrhage and extra-axial haemorrhage lesions combined (lesion volume >5·5 mL)[11] is similar to that found in our study.

The algorithm performed less well at quantifying perilesional oedema, and by extension mixed density lesions. However, the ability to undertake such quantification has not been reported previously; hence, we are unable to benchmark it against previous work. Although detection and delineation of high-intensity haemorrhagic lesions are straightforward, precise delineation of hypointense oedema can be challenging, even for radiologists. The ability of our algorithm to do this task, in addition to quantifying other lesion types, may be important for prognostication, aid detection and avoidance of secondary injury, the evaluation of neuroprotective measures, and as an intermediate biomarker for clinical

trials aimed at the reduction of cerebral oedema and contusion growth.[31]

The accuracy of our CNN was lower in segmenting small haemorrhagic lesions. From a clinical perspective, however, this reduced accuracy is mitigated by the fact that the volume of these small lesions is less important in terms of prognostication or deciding on therapy. These small lesions are typically microhaemorrhages associated with diffuse vascular injury and are clinically used as a surrogate marker for diffuse axonal injury. Consequently, their clinical significance is dependent on number and distribution, rather than volume of individual lesions.[17]

Although our model was not designed for classification specifically, as a byproduct of the segmentation algorithm, it is able to do so with comparable performance to state-of-the-art methods developed solely for detection.[12,13] On the CQ500 dataset, previous work[12] reported an AUC of 0·94 (95% CI 0·92–0·97) for intracranial haemorrhage, 0·95 (0·93–0·98) for intraparenchymal haemorrhage, 0·95 (0·91–0·99) for subdural haematoma, 0·97 (0·91–1·00) for extradural haematoma, 0·97 (0·92–0·99) for traumatic subarachnoid haemorrhage, and 0·93 (0·87–1·00) for intraventricular haemorrhage.[12] Apart from the intraventricular haemorrhage class, the AUCs we report on the same data are lower. However, our algorithm also has the ability to quantify lesion volume, shape, and location, which can be used to extract other radiological features of potential interest. Additionally, our results are not directly comparable with the previous work by Chilamkurthy and colleagues[12] because they used certain rules to select the optimum scan per patient processed by their algorithm and we were not able to determine those rules for comparison. Instead, we processed all available scans for each patient (up to eight) and calculated the mean predicted volume for subsequent classification. Using a selected set of scans, as done in previous work, is likely to improve our results.

The ability to distinguish between different lesion types is important to aid understanding of pathophysiology and to implement personalised care. The heterogeneity of TBI is well described, encompassing a wide spectrum of pathologies, from axonal injury to focal contusions and extracranial bleeding. The large annotated dataset used in this study is representative of this clinical spectrum. The CENTER-TBI study[14,15] allowed a large variety of vendors and acquisition protocols to be used. Images in this analysis were contributed from 38 centres. Consequently, the performance is not manufacturer or acquisition dependent. The ability to generalise is supported by validation on an external, independent dataset from a different continent, for which the results for lesion detection were comparable with the results obtained on internal data.

Adding the ability to distinguish the different types of extra-axial haemorrhage is important, particularly given that extradural haematomas portend a better prognosis, and the presence of traumatic subarachnoid haemorrhage is a marker for worse outcomes in prognostic models.[26,27,32] Furthermore, expanding on the capability of lesion localisation may help answer key research questions and support clinical reporting of scans.

Future work needs to focus on the optimal incorporation of such algorithms into clinical practice, which must be accompanied by a rigorous assessment of performance, strengths, and weaknesses. Such algorithms will find clear research applications, and, if adequately validated, may be used to help facilitate radiology workflows by flagging scans that require urgent attention, aid reporting in resource-constrained environments, and detect patho-anatomically relevant features for prognostication and a better understanding of lesion progression.

To **submit proposals** see https://www.center-tbi.eu/data

For more on **CENTER-TBI** see https://www.center-tbi.eu/project/overview

For the **CQ500 dataset** see http://headctstudy.qure.ai/dataset

For the **study algorithm** see https://github.com/biomedia-mira/blast-ct/

For the **archived version** see https://doi.org/10.5281/zenodo.3746088

Contributors

MM, VFJN, DKM, and BG conceived and designed the study. MM did implementations, analysed data, and cowrote the manuscript with VFJN. MM, VFJN, DKM, and BG revised and finalised the manuscript. VFJN, FM, KA, and DW did the manual and semi-automatic segmentation of the scans or provided broader clinical input, or both. KK, EF, and BG provided feedback on the development of the model. TD provided specialist neuroradiological oversight of image analysis. VFJN, DR, DKM, and BG secured the funding. All authors read and approved the final manuscript.

Declaration of interests

VFJN reports an Academy of Medical Sciences/The Health Foundation Clinician Scientist Fellowship, during the conduct of this study; and a grant from Roche Pharmaceuticals and honoraria from Neurodiem, outside the submitted work. DR has received grants from EU Horizon 2020, during the conduct of this study; and personal fees from IXICO, Heartflow, and Circle Cardiovascular Imaging, outside the submitted work. DKM reports grants from GlaxoSmithKline and personal fees from NeuroTraumaSciences, Pfizer, Calico, PressuraNeuro, Lantmannen, Integra Neurosciences, Gryphon, and Cortirio, outside the submitted work. BG has received grants from European Commission and UK Research and Innovation Engineering and Physical Sciences Research Council, during the conduct of this study; and is Scientific Advisor for Kheiron Medical Technologies, Advisor and Scientific Lead of the HeartFlow-Imperial Research Team, and Visiting Researcher at Microsoft Research. All other authors declare no competing interests.

Data sharing

The data and algorithm are available at the time of publication. Data access is conditional to an approved study proposal; there are no end dates to the availability. The CENTER-TBI data used in this study is available to researchers who provide a methodologically sound study proposal that is approved by the CENTER-TBI management committee to achieve the aims in the approved proposal. Proposals may be submitted online to CENTER-TBI. A data access agreement is required, and all access must comply with regulatory restrictions imposed on the original study. No patient-identifiable information is made available, and all data have been anonymised. Study protocols and additional information for CENTER-TBI about data collection, recruitment, and participating centres is available online. The anonymised CQ500 data used for the external validation of our algorithm can be accessed and downloaded online. The source code of our algorithm together with pre-trained models and usage instructions are available online. An archive of the version used for the experimental validation in our paper is also available online.

Acknowledgments

The CENTER-TBI study was supported by the European Union 7th Framework Programme (EC grant 602150), with additional project support from OneMind, NeuroTrauma Sciences, Integra Neurosciences, and European Research Council Horizon 2020 (EC grant 757173). Individual sources of funding were the Engineering and Physical Sciences Research Council (EP/R511547/1, to BG, KK), Academy of Medical Sciences/The Health Foundation (VFJN), and the National Institute for Health Research (DKM).

References

1 Maas AIR, Menon DK, Adelson PD, et al. Traumatic brain injury: integrated approaches to improve prevention, clinical care, and research. *Lancet Neurol* 2017; **16:** 987–1048.

2 Dewan MC, Rattani A, Gupta S, et al. Estimating the global incidence of traumatic brain injury. *J Neurosurg* 2018; **130:** 1080–97.

3 Amyot F, Arciniegas DB, Brazaitis MP, et al. A review of the effectiveness of neuroimaging modalities for the detection of traumatic brain injury. *J Neurotrauma* 2015; **32:** 1693–721.

4 Carney N, Totten AM, O'Reilly C, et al. Guidelines for the management of severe traumatic brain injury, fourth edition. *Neurosurgery* 2017; **80:** 6–15.

5 Vande Vyvere T, Wilms G, Claes L, et al. Central versus local radiological reading of acute computed tomography characteristics in multi-center traumatic brain injury research. *J Neurotrauma* 2019; **36:** 1080–92.

6 Bardera A, Boada I, Feixas M, et al. Semi-automated method for brain hematoma and edema quantification using computed tomography. *Comput Med Imag Grap* 2009; **33:** 304–11.

7 Bhadauria HS, Singh A, Dewal ML. An integrated method for hemorrhage segmentation from brain CT imaging. *Comput Electr Eng* 2013; **39:** 1527–36.

8 Zaki WMDW, Fauzi MFA, Besar R, Ahmad WSHMW. Qualitative and quantitative comparisons of haemorrhage intracranial segmentation in CT brain images. TENCON 2011–2011 IEEE Region 10 Conference 2011; Bali, Indonesia; Nov 21–24, 2011: 369–73.

9 Roy S, Wilkes S, Diaz-Arrastia R, Butman JA, Pham DL. Intraparenchymal hemorrhage segmentation from clinical head CT of patients with traumatic brain injury. *Proc Spie* 2015; **9413**.

10 LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 1998; **86:** 2278–324.

11 Jain S, Vyvere TV, Terzopoulos V, et al. Automatic quantification of computed tomography features in acute traumatic brain injury. *J Neurotrauma* 2019; **36:** 1794–803.

12 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:** 2388–96.

13 Kuo W, Hne C, Mukherjee P, Malik J, Yuh EL. Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. *Proc Natl Acad Sci USA* 2019; **116:** 22737–45.

14 Maas AI, Menon DK, Steyerberg EW, et al. Collaborative European NeuroTrauma Effectiveness Research in Traumatic Brain Injury (CENTER-TBI): a prospective longitudinal observational study. *Neurosurgery* 2015; **76:** 67–80.

15 Steyerberg EW, Wiegers E, Sewalt C, et al. Case-mix, care pathways, and outcomes in patients with traumatic brain injury in CENTER-TBI: a European prospective, multicentre, longitudinal, cohort study. *Lancet Neurol* 2019; **18:** 923–34.

16 CENTER-TBI. Ethical approval. https://www.center-tbi.eu/project/ethical-approval (accessed March 13, 2020).

17 Haacke EM, Duhaime AC, Gean AD, et al. Common data elements in radiologic imaging of traumatic brain injury. *J Magn Reson Imaging* 2010; **32:** 516–43.

18 Figueira Rodrigues Vieira G, Guedes Correa JF. Early computed tomography for acute post-traumatic diffuse axonal injury: a systematic review. *Neuroradiology* 2020; published online March 4. DOI:10.1007/s00234-020-02383-2.

19 Kamnitsas K, Ferrante E, Parisot S, et al. DeepMedic for brain tumor segmentation. In: Crimi A, Menze B, Maier O, Reyes M, Winzeck S, Handels H (eds). Brain lesion: glioma, multiple sclerosis, stroke, and traumatic brain injuries. BrainLes 2016. Lecture notes in computer science, vol 10154. Cham: Springer International Publishing, 2016: 138–49.

20 Kamnitsas K, Ledig C, Newcombe VFJ, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal* 2017; **36:** 61–78.

21 Qure.ai. Download CQ500 Dataset. http://headctstudy.qure.ai/dataset (accessed March 13, 2020).

22 Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging* 2015; **34:** 1993–2024.

23 Marshall LF, Marshall SB, Klauber MR, et al. A new classification of head-injury based on computerized-tomography. *J Neurosurg* 1991; **75:** S14–20.

24 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143:** 29–36.

25 Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934; **26:** 404–13.

26 Edwards P, Arango M, Balica L, et al. Final results of MRC CRASH, a randomised placebo-controlled trial of intravenous corticosteroid in adults with head injury-outcomes at 6 months. *Lancet* 2005; **365:** 1957–59.

27 Roberts I, Yates D, Sandercock P, et al. Effect of intravenous corticosteroids on death within 14 days in 10008 adults with clinically significant head injury (MRC CRASH trial): randomised placebo-controlled trial. *Lancet* 2004; **364:** 1321–28.

28 Narayan RK, Maas AI, Servadei F, et al. Progression of traumatic intracerebral hemorrhage: a prospective observational study. *J Neurotrauma* 2008; **25:** 629–39.

29 Oertel M, Kelly DF, McArthur D, et al. Progressive hemorrhage after head trauma: predictors and consequences of the evolving injury. *J Neurosurg* 2002; **96:** 109–16.

30 Kurland D, Hong C, Aarabi B, Gerzanich V, Simard JM. Hemorrhagic progression of a contusion after traumatic brain injury: a review. *J Neurotrauma* 2012; **29:** 19–31.

31 Mathieu F, Zeiler FA, Whitehouse DP, et al. Relationship between measures of cerebrovascular reactivity and intracranial lesion progression in acute TBI patients: an exploratory analysis. *Neurocrit Care* 2020; **32:** 373–82.

32 Murray GD, Butcher I, McHugh GS, et al. Multivariable prognostic analysis in traumatic brain injury: results from the IMPACT study. *J Neurotrauma* 2007; **24:** 329–37.