Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study

Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, Prashant Warier

Lancet 2018; 392: 2388–96

Published Online
October 11, 2018
http://dx.doi.org/10.1016/S0140-6736(18)31645-3

See Comment page 2331

Qure.ai, Goregaon East, Mumbai, India (S Chilamkurthy BTech, R Ghosh BTech, S Tanamala MTech, P Rao PhD, P Warier PhD); CT & MRI Center, Dhantoli, Nagpur, India (M Biviji DNB); Department of Radiology, Mayo Clinic, Rochester, MN, USA (N G Campeau MD); and Centre for Advanced Research in Imaging, Neurosciences and Genomics, New Delhi, India (V K Venugopal MD, V Mahajan MBA)

Correspondence to:
Mr Sasank Chilamkurthy, Qure.ai, Goregaon East, Mumbai 400063, India
sasank.chilamkurthy@qure.ai

Summary

Background Non-contrast head CT scan is the current standard for initial imaging of patients with head trauma or stroke symptoms. We aimed to develop and validate a set of deep learning algorithms for automated detection of the following key findings from these scans: intracranial haemorrhage and its types (ie, intraparenchymal, intraventricular, subdural, extradural, and subarachnoid); calvarial fractures; midline shift; and mass effect.

Methods We retrospectively collected a dataset containing 313 318 head CT scans together with their clinical reports from around 20 centres in India between Jan 1, 2011, and June 1, 2017. A randomly selected part of this dataset (Qure25k dataset) was used for validation and the rest was used to develop algorithms. An additional validation dataset (CQ500 dataset) was collected in two batches from centres that were different from those used for the development and Qure25k datasets. We excluded postoperative scans and scans of patients younger than 7 years. The original clinical radiology report and consensus of three independent radiologists were considered as gold standard for the Qure25k and CQ500 datasets, respectively. Areas under the receiver operating characteristic curves (AUCs) were primarily used to assess the algorithms.

Findings The Qure25k dataset contained 21 095 scans (mean age 43 years; 9030 [43%] female patients), and the CQ500 dataset consisted of 214 scans in the first batch (mean age 43 years; 94 [44%] female patients) and 277 scans in the second batch (mean age 52 years; 84 [30%] female patients). On the Qure25k dataset, the algorithms achieved an AUC of 0·92 (95% CI 0·91–0·93) for detecting intracranial haemorrhage (0·90 [0·89–0·91] for intraparenchymal, 0·96 [0·94–0·97] for intraventricular, 0·92 [0·90–0·93] for subdural, 0·93 [0·91–0·95] for extradural, and 0·90 [0·89–0·92] for subarachnoid). On the CQ500 dataset, AUC was 0·94 (0·92–0·97) for intracranial haemorrhage (0·95 [0·93–0·98], 0·93 [0·87–1·00], 0·95 [0·91–0·99], 0·97 [0·91–1·00], and 0·96 [0·92–0·99], respectively). AUCs on the Qure25k dataset were 0·92 (0·91–0·94) for calvarial fractures, 0·93 (0·91–0·94) for midline shift, and 0·86 (0·85–0·87) for mass effect, while AUCs on the CQ500 dataset were 0·96 (0·92–1·00), 0·97 (0·94–1·00), and 0·92 (0·89–0·95), respectively.

Interpretation Our results show that deep learning algorithms can accurately identify head CT scan abnormalities requiring urgent attention, opening up the possibility to use these algorithms to automate the triage process.

Funding Qure.ai.

Introduction

Non-contrast head CT scans are among the most commonly used emergency room diagnostic tools for patients with head injury or for those with symptoms suggesting a stroke or rise in intracranial pressure. The wide availability and low acquisition time of these scans make them a commonly used first-line diagnostic method.[1] The percentage of annual US emergency room visits that involve a CT scan has been increasing for the past few decades[2] and the use of head CT to exclude the need for neurosurgical intervention is on the rise.[3]

The most critical, time-sensitive abnormalities that can be readily detected on CT scan include intracranial haemorrhages, raised intracranial pressure, and cranial fractures. A key assessment goal in patients with stroke is exclusion of an intracranial haemorrhage, which depends on CT imaging and its swift interpretation.[4] Similarly, immediate CT scan interpretation is crucial in patients with a suspected acute intracranial haemorrhage to assess the need for neurosurgical treatment. Cranial fractures, if open or depressed, will usually require urgent neurosurgical intervention. Cranial fractures are also the most commonly missed major abnormality on head CT scans,[5] especially if coursing in an axial plane.

Although these abnormalities are found on only a small proportion of CT scans, streamlining the head CT scan interpretation workflow by automating the initial triage process has the potential to substantially decrease time to diagnosis and expedite treatment, which might in turn decrease morbidity and mortality consequent to stroke and head injury. An automated head CT scan triage system might also be valuable for queue management in a busy trauma care setting, or could facilitate decision making in remote locations without availability of an immediate radiologist.

Research in context

Evidence before this study

We searched for machine learning or deep learning studies focusing on computer-aided diagnosis of head CT. We searched Google Scholar for articles published before Feb 15, 2018, with the terms "deep learning" OR "machine learning" AND "head CT" AND "hemorrhage" OR "midline shift" OR "skull fracture". We also reviewed reference lists of eligible texts. We identified several studies on the development and validation of computer-aided diagnosis algorithms that used small numbers of head CT scans. Deep learning has previously been used to detect intracranial haemorrhages. Traditional computer vision techniques were more common for detection of fractures and midline shift. In most studies, training and validation datasets had fewer than 200 head CT scans, raising concerns about the robustness of these algorithms. We identified no standard public head CT datasets to allow direct comparison with our algorithms' performance.

Added value of this study

We developed deep learning algorithms to separately detect as many as nine critical findings on head CT scans. We described the use of deep learning for detection of calvarial fractures and midline shift. We validated all the algorithms with a large dataset versus clinical radiology reports. We also validated the algorithms versus consensus of three radiologists using a dataset acquired from a completely different source than that of the development dataset.

Implications of all the available evidence

The strong performance of our deep learning algorithms suggests that they can potentially be used for triaging or notification of patients with critical findings as soon as a head CT scan is acquired. A clinical trial is required to determine if such triage or notification improves radiologist efficiency and patient care.

The past year has seen several advances in application of deep learning[6–9] for medical imaging interpretation tasks, with robust evidence that deep learning can perform specific medical imaging tasks including identifying and grading diabetic retinopathy[10] and classifying skin lesions as benign or malignant[11] with accuracy equivalent to specialist physicians. Deep learning algorithms have also been trained to detect abnormalities on radiological images such as chest radiographs,[6,7] chest CT,[12,13] and head CT[8,9] through classification algorithms, as well as to localise and quantify disease patterns or anatomical volumes[14–16] through segmentation algorithms.

The development of an accurate deep learning algorithm for radiology requires—in addition to appropriate model architectures—a large number of accurately labelled scans that will be used to train the algorithm.[17] The chances that the algorithm generalises well to new settings increase when the training dataset is large and includes scans from diverse sources.[18]

We describe the development and validation of fully automated deep learning algorithms that are trained to detect abnormalities requiring urgent attention on non-contrast head CT scans. The trained algorithms detect five types of intracranial haemorrhage (namely, intraparenchymal, intraventricular, subdural, extradural, and subarachnoid) and calvarial (cranial vault) fractures. The algorithms also detect mass effect and midline shift, both used as indicators of severity of the brain injury.

Methods

Datasets

We retrospectively collected 313 318 anonymous head CT scans from around 20 centres in India between Jan 1, 2011, and June 1, 2017. These centres, which included both in-hospital and outpatient radiology centres, use a range of CT scanner models (listed in the appendix, p 4) with slices per rotation ranging from 2 to 128. Each of the scans had an electronic clinical report associated with it, which we used as the gold standard during the algorithm development process.

See Online for appendix

Of the 313 318 scans, we selected scans of 23 263 randomly chosen patients (Qure25k dataset) for validation and used the scans of the remaining patients (development dataset) to train and develop the algorithms. We removed postoperative scans and scans of patients younger than 7 years from the Qure25k dataset. This dataset was not used during the algorithm development process.

An additional validation dataset (CQ500 dataset) was provided by the Centre for Advanced Research in Imaging, Neurosciences and Genomics, New Delhi, India. This dataset was a subset of head CT scans taken at six radiology centres in New Delhi between Jan 1, 2012, and Feb 1, 2018. Half the centres are stand-alone outpatient centres and the other half are radiology departments embedded in large hospitals. There was no overlap between these centres and those used to obtain the development dataset or Qure25k dataset. CT scanners used at these centres had slices per rotation varying from 16 to 128 (see appendix p 4 for list of models). Data were pulled from local picture archiving and communication system (PACS) servers and anonymised in compliance with internally defined Health Insurance Portability and Accountability Act (HIPAA) guidelines. Because both datasets were retrospectively obtained and fully anonymised, the study was exempt from institutional review board approval.

Similar to the development and Qure25k datasets, clinical radiology reports associated with scans in the CQ500 dataset were available. Although we did not use them as gold standards in this study, we used them for the dataset selection.

We collected the CQ500 dataset in two batches. The first batch was collected by selecting all head CT scans taken at the centres for 30 days starting from Nov 20, 2017. The second batch was selected from the remaining scans. First, a natural language processing (NLP) algorithm was used to detect intraparenchymal, intraventricular, subdural, extradural, and subarachnoid haemorrhages, and calvarial fractures from clinical radiology reports. Second, reports were randomly selected so that there were around 80 scans with each of intraparenchymal, subdural, extradural, and subarachnoid haemorrhages, and calvarial fractures. Each of the selected scans were then screened for the following exclusion criteria: postoperative defect; absence of non-contrast (plain) axial series covering complete brain; and patient was younger than 7 years (estimated from cranial sutures[19] if data were unavailable).

Follow-up scans for a patient were not excluded in the selection process. We removed any duplicate scans found in the dataset.

Reading the scans

Three senior radiologists (including NGC) served as independent raters for the CT scans in the CQ500 dataset. They had corresponding experience of 8, 12, and 20 years in cranial CT interpretation. None of the three raters was involved in the clinical care or assessment of the enrolled patients, nor did they have access to clinical history of any of the patients. Each of the radiologists independently evaluated the scans in the CQ500 dataset with the instructions for recording the findings and query resolution as shown in the appendix (pp 10–12). The order of presentation of the scans was randomised so as to minimise recall of the patients' follow-up scans.

Each of the raters recorded the following findings for each scan: (1) the presence or absence of an intracranial haemorrhage and if present, its types (intraparenchymal, intraventricular, extradural, subdural, and subarachnoid); (2) the presence or absence of midline shift and mass effect; (3) the presence or absence of fractures, and if present, if the fracture was (partly) a calvarial fracture.

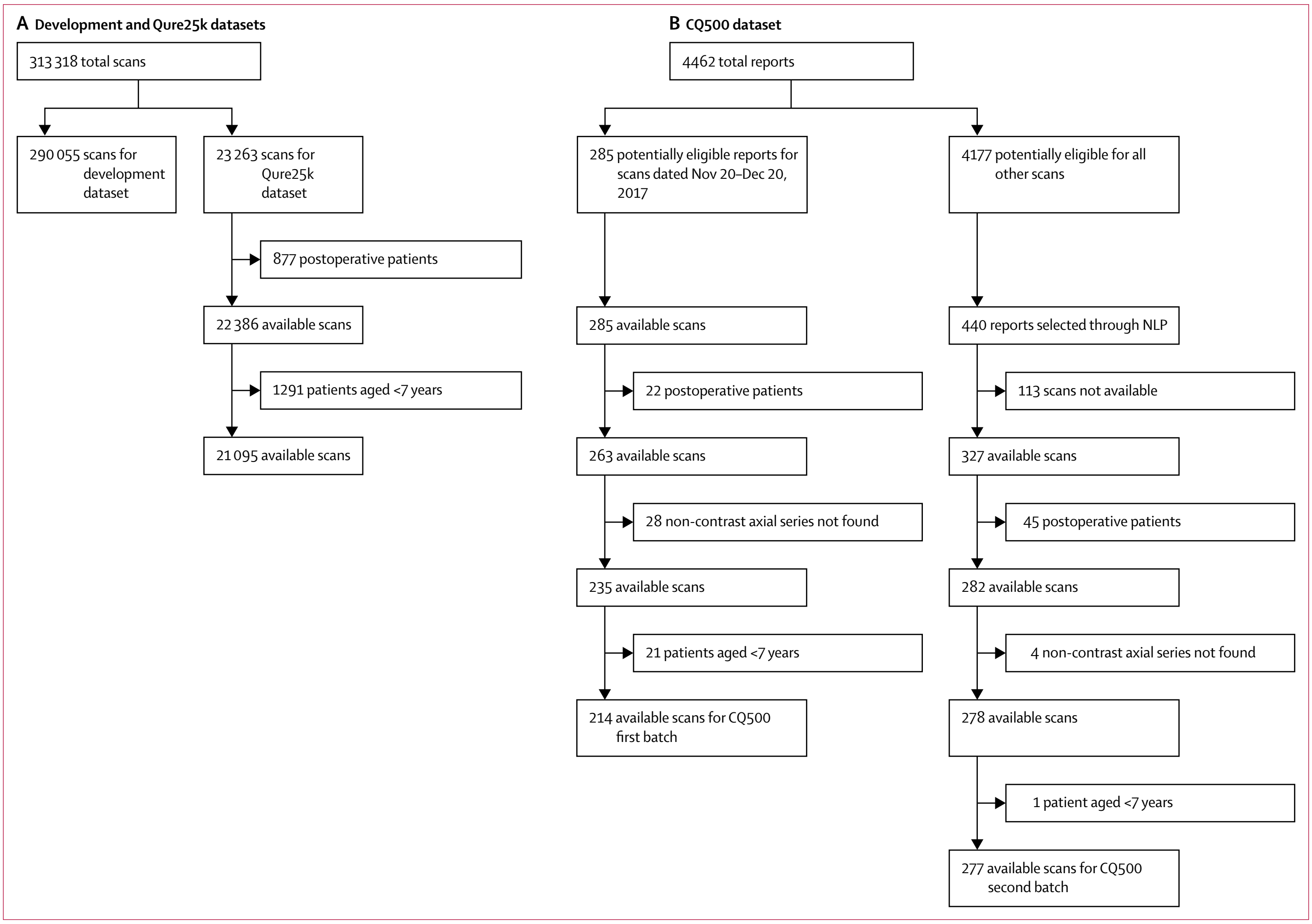

A Development and Qure25k datasets
313 318 total scans
290 055 scans for development dataset
23 263 scans for Qure25k dataset
877 postoperative patients
22 386 available scans
1291 patients aged <7 years
21 095 available scans
B CQ500 dataset
4462 total reports
285 potentially eligible reports for scans dated Nov 20–Dec 20, 2017
4177 potentially eligible for all other scans
285 available scans
22 postoperative patients
263 available scans
28 non-contrast axial series not found
235 available scans
21 patients aged <7 years
214 available scans for CQ500 first batch
440 reports selected through NLP
113 scans not available
327 available scans
45 postoperative patients
282 available scans
4 non-contrast axial series not found
278 available scans
1 patient aged <7 years
277 available scans for CQ500 second batch

Figure 1: **Dataset selection process**
NLP=natural language processing.

Intra-axial presence of blood from any cause (such as haemorrhagic contusion, or tumour or infarct with haemorrhagic component) was included in the definition of intraparenchymal haemorrhage. Chronic haemorrhages were considered positive in this study. Mass effect was defined as any of the following: local mass effect, ventricular effacement, midline shift, or herniation. Midline shift was considered positive if the amount of shift was greater than 5 mm. If there was at least one fracture that extended into the skullcap, the scan was considered to have a calvarial fracture.

If unanimous agreement for each of the findings was not achieved by the three raters, the interpretation of the majority of the raters was used as the final diagnosis.

For the development and Qure25k datasets, we considered clinical reports written by radiologists as the gold standard. However, these were written in free text rather than in a structured format. Therefore, a rule-based NLP algorithm was applied on the radiologists' clinical reports to automatically infer the target findings. We validated this algorithm on a random subset of reports from the Qure25k dataset to ensure that the inferred information was accurate and could be used as gold standard. The validation was achieved by manually labelling reports from this subset and comparing these labels to the NLP algorithm's outputs.

Assessment of the algorithms

We describe the development of the deep learning algorithms in the appendix (pp 1–3). When run on a scan, our algorithms produce a list of nine real valued confidence scores in the range of 0–1 indicating the presence of the following nine findings: intracranial haemorrhage and each of the five types of haemorrhage, midline shift, mass effect, and calvarial fracture. As previously mentioned, the corresponding gold standards were obtained using majority voting for the CQ500 dataset and by NLP algorithm of reports for the Qure25k dataset. Algorithms were assessed independently for each finding.

For both CQ500 and Qure25k datasets, receiver operating characteristic (ROC) curves[20] were obtained for each of the target findings by varying the threshold and plotting the true positive rate (ie, sensitivity) and false positive rate (ie, 1–specificity) at each threshold. Two operating points were chosen on the ROC curve so that sensitivity was approximately 0·9 (high sensitivity point) and specificity approximately 0·9 (high specificity point; see appendix p 5 for algorithm for operating point choice). Areas under the ROC curves (AUCs) and sensitivities and specificities at these two operating points were used to assess the algorithms.

Statistical analysis

Sample sizes for proportions and AUCs were calculated using normal approximation and the method outlined by Hanley and McNeil,[20] respectively. The prevalence of our target abnormalities in a randomly selected sample of CT scans tends to be low; therefore, establishing the algorithms' sensitivity with a reasonably high confidence on an unenriched dataset would require very large sample sizes. For example, to establish a sensitivity with an expected value of 0·7 within a 95% CI of half-length of 0·10, the number of positive scans to be read is approximately 80. Similarly, for a finding with a prevalence of 1%, to establish an AUC within a 95% CI of half-length of 0·05, the number of scans to be read is approximately 20 000.

	Qure25k dataset	CQ500 dataset: first batch	CQ500 dataset: second batch
Number of scans	21 095	214	277
Number of raters per scan	1	3	3
Number of scans for which age was known	21 095	189	251
Mean age, years (SD; range)	43 (22; 7–99)	43 (22; 7–95)	52 (20; 10–95)
Female patients	9030 (43%)*	94 (44%)	84 (30%)
Intracranial haemorrhage	2494 (12%)	35 (16%)	170 (61%)
Intraparenchymal	2013 (10%)	29 (14%)	105 (38%)
Intraventricular	436 (2%)	7 (3%)	21 (8%)
Subdural	554 (3%)	9 (4%)	44 (16%)
Extradural	290 (1%)	2 (1%)	11 (4%)
Subarachnoid	611 (3%)	9 (4%)	51 (18%)
Fracture	1653 (8%)	8 (4%)	31 (11%)
Calvarial fracture	992 (5%)	6 (3%)	28 (10%)
Midline shift	666 (3%)	18 (8%)	47 (17%)
Mass effect	1517 (7%)	28 (13%)	99 (36%)

Data are n (%), unless otherwise stated. *Sex was known for 21 064 scans in the Qure25k dataset.

Table 1: **Dataset characteristics**

	Number of positives	Sensitivity (95% CI)	Specificity (95% CI)
Intracranial haemorrhage	207	0·9807 (0·9513–0·9947)	0·9873 (0·9804–0·9922)
Intraparenchymal	157	0·9809 (0·9452–0·9960)	0·9883 (0·9818–0·9929)
Intraventricular	44	1·0000 (0·9196–1·0000)	1·0000 (0·9979–1·0000)
Subdural	44	0·9318 (0·8134–0·9857)	0·9965 (0·9925–0·9987)
Extradural	27	1·0000 (0·8723–1·0000)	0·9983 (0·9950–0·9996)
Subarachnoid	51	1·0000 (0·9302–1·0000)	0·9971 (0·9933–0·9991)
Fracture	143	1·0000 (0·9745–1·0000)	1·0000 (0·9977–1·0000)
Calvarial fracture	89	0·9888 (0·9390–0·9997)	0·9947 (0·9899–0·9976)
Midline shift	54	0·9815 (0·9011–0·9995)	1·0000 (0·9979–1·0000)
Mass effect	132	0·9773 (0·9350–0·9953)	0·9933 (0·9881–0·9967)

Performance of the natural language processing algorithm in inferring findings from radiologists' reports, measured on 1779 reports from the Qure25k dataset.

Table 2: **Reliability of the gold standards for the Qure25k dataset**

The Qure25k dataset used in our study was randomly sampled from the population distribution and had more than 20 000 scans in accordance with these sample size calculations. However, constraints on the radiologist time necessitated the previously mentioned enrichment strategy for the CQ500 dataset. Manual curation of scans

	Raters 1 and 2		Raters 2 and 3		Raters 1 and 3		All Fleiss' κ
	Agreement, n (%)	Cohen's κ	Agreement, n (%)	Cohen's κ	Agreement, n (%)	Cohen's κ	
Intracranial haemorrhage	437 (89%)	0·78	446 (91%)	0·81	434 (88%)	0·76	0·78
Intraparenchymal	448 (91%)	0·79	445 (91%)	0·77	446 (91%)	0·77	0·77
Intraventricular	472 (96%)	0·70	477 (97%)	0·74	470 (96%)	0·66	0·70
Subdural	432 (88%)	0·49	457 (93%)	0·60	442 (90%)	0·56	0·54
Extradural	478 (97%)	0·51	483 (98%)	0·73	482 (98%)	0·60	0·61
Subarachnoid	457 (93%)	0·68	446 (91%)	0·61	446 (91%)	0·64	0·64
Calvarial fracture	451 (92%)	0·58	452 (92%)	0·37	448 (91%)	0·36	0·45
Midline shift	433 (88%)	0·58	428 (87%)	0·53	460 (94%)	0·70	0·60
Mass effect	424 (86%)	0·65	432 (88%)	0·67	427 (87%)	0·68	0·67

Three radiologists reviewed each of the 491 cases in the CQ500 dataset and majority vote of the raters was used as gold standard. The guidelines of Fleiss and colleagues[24] characterise κ values of more than 0·75 as excellent agreement, 0·40–0·75 as fair to good agreement, and less than 0·40 as poor agreement beyond chance.

Table 3: **Reliability of the gold standards for the CQ500 dataset**

(by referring to the scans themselves) would have had selection bias towards more visually significant positive scans. We mitigated this issue by random selection, in which positive scans were determined from the clinical reports.

We generated confusion matrices for each of the nine critical CT findings at the selected operating points. We then calculated 95% CIs for sensitivity and specificity from these matrices using the exact Clopper-Pearson method[21] based on β distribution. 95% CIs of AUCs were calculated following the distribution-based approach described by Hanley and McNeil.[20] On the CQ500 dataset, we measured the concordance between paired raters on each finding using percentage of agreement and Cohen's κ statistic.[22] We also measured concordance between all three raters on each finding using Fleiss' κ statistic.[23] We did all statistical analyses using scipy, scikit-learn, and statsmodels python packages.

We have also attempted to compare the performance of the algorithms to that of the radiologists. This was only possible on the CQ500 dataset because each rater could be compared with their consensus to obtain their performance metrics (see appendix pp 8–9 for details).

Role of the funding source

The funder of the study was involved in data collection, data interpretation, writing of the report, and the decision to submit for publication. SC, RG, ST, PR, and PW had access to all the data in the study, while NGC, VKV, and VM had access to the CQ500 dataset only. SC, RG, and ST were responsible for the decision to submit for publication.

Results

In the Qure25k dataset, of the 23 263 head CT scans randomly chosen for validation, 21 095 were eligible for inclusion (figure 1). 4462 clinical reports were analysed in the selection process of the CQ500 dataset. Of these, 285 were selected in the first batch and 440 in the second batch. 71 scans in the first batch and 163 scans in the second batch were excluded, resulting in a total of 491 scans. Reasons for exclusion were non-availability of images (n=113), postoperative scans (n=67), scan had no non-contrast axial series (n=32), and patient younger than 7 years (n=22).

Patient demographics and prevalences for each critical finding on head CT scan are summarised in table 1. In the Qure25k dataset, 2494 scans were reported positive for intracranial haemorrhage and 992 were positive for calvarial fracture. The first batch of the CQ500 dataset contained 35 scans reported positive for intracranial haemorrhage and six positive for calvarial fracture. In the second batch, 170 scans were reported positive for intracranial haemorrhage and 28 scans were positive for calvarial fracture.

The NLP algorithm used to infer the target CT findings from clinical reports in the Qure25k dataset was evaluated on a total of 1779 reports. Sensitivity and specificity of the NLP algorithm were fairly high; the least performing finding was subdural haemorrhage with a sensitivity of 0·93 (95% CI 0·81–0·99) and specificity of 1·00 (0·99–1·00), whereas fracture was inferred perfectly with sensitivity of 1·00 (0·97–1·00) and specificity of 1·00 (1·00–1·00; table 2).

Concordance between the three raters on the CQ500 dataset was highest for intracranial haemorrhage (Fleiss' κ=0·78) and intraparenchymal haemorrhage (Fleiss' κ=0·77), representing excellent agreement for these findings (table 3). Calvarial fracture and subdural haemorrhage had the lowest concordance with Fleiss' κ=0·45 and κ=0·54, respectively, indicating fair to moderate agreement.

Table 4 and figure 2 summarise the performance of the deep learning algorithms. On the Qure25k set, the algorithms achieved AUCs of 0·92 (95% CI 0·91–0·93) for intracranial haemorrhage, 0·92 (0·91–0·94) for calvarial fracture, and 0·93 (0·91–0·94) for midline shift. The algorithms generally performed better on the CQ500 dataset than on the Qure25k dataset. On the CQ500 dataset, AUCs were 0·94 (95% CI 0·92–0·97) for

	AUC (95% CI)	High sensitivity operating point		High specificity operating point	
		Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Qure25k dataset					
Intracranial haemorrhage	0·9194 (0·9119–0·9269)	0·9006 (0·8882–0·9121)	0·7295 (0·7230–0·7358)	0·8349 (0·8197–0·8492)	0·9004 (0·8960–0·9047)
Intra-parenchymal	0·8977 (0·8884–0·9069)	0·9031 (0·8894–0·9157)	0·6046 (0·5976–0·6115)	0·7670 (0·7479–0·7853)	0·9046 (0·9003–0·9087)
Intraventricular	0·9559 (0·9424–0·9694)	0·9358 (0·9085–0·9569)	0·8343 (0·8291–0·8393)	0·9220 (0·8927–0·9454)	0·9267 (0·9231–0·9302)
Subdural	0·9161 (0·9001–0·9321)	0·9152 (0·8888–0·9370)	0·6542 (0·6476–0·6607)	0·7960 (0·7600–0·8288)	0·9041 (0·9000–0·9081)
Extradural	0·9288 (0·9083–0·9494)	0·9034 (0·8635–0·9349)	0·7936 (0·7880–0·7991)	0·8207 (0·7716–0·8631)	0·9068 (0·9027–0·9107)
Subarachnoid	0·9044 (0·8882–0·9205)	0·9100 (0·8844–0·9315)	0·6678 (0·6613–0·6742)	0·7758 (0·7406–0·8083)	0·9012 (0·8971–0·9053)
Calvarial fracture	0·9244 (0·9130–0·9359)	0·9002 (0·8798–0·9181)	0·7749 (0·7691–0·7807)	0·8115 (0·7857–0·8354)	0·9020 (0·8978–0·9061)
Midline shift	0·9276 (0·9139–0·9413)	0·9114 (0·8872–0·9319)	0·8373 (0·8322–0·8424)	0·8754 (0·8479–0·8995)	0·9006 (0·8964–0·9047)
Mass effect	0·8583 (0·8462–0·8703)	0·8622 (0·8439–0·8792)	0·6157 (0·6089–0·6226)	0·7086 (0·6851–0·7314)	0·9068 (0·9026–0·9108)
CQ500 dataset					
Intracranial haemorrhage	0·9419 (0·9187–0·9651)	0·9463 (0·9060–0·9729)	0·7098 (0·6535–0·7617)	0·8195 (0·7599–0·8696)	0·9021 (0·8616–0·9340)
Intra-parenchymal	0·9544 (0·9293–0·9795)	0·9478 (0·8953–0·9787)	0·8123 (0·7679–0·8515)	0·8433 (0·7705–0·9003)	0·9076 (0·8726–0·9355)
Intraventricular	0·9310 (0·8654–0·9965)	0·9286 (0·7650–0·9912)	0·6652 (0·6202–0·7081)	0·8929 (0·7177–0·9773)	0·9028 (0·8721–0·9282)
Subdural	0·9521 (0·9117–0·9925)	0·9434 (0·8434–0·9882)	0·7215 (0·6769–0·7630)	0·8868 (0·7697–0·9573)	0·9041 (0·8726–0·9300)
Extradural	0·9731 (0·9113–1·0000)	0·9231 (0·6397–0·9981)	0·8828 (0·8506–0·9103)	0·8462 (0·5455–0·9808)	0·9477 (0·9238–0·9659)
Subarachnoid	0·9574 (0·9214–0·9934)	0·9167 (0·8161–0·9724)	0·8654 (0·8295–0·8962)	0·8667 (0·7541–0·9406)	0·9049 (0·8732–0·9309)
Calvarial fracture	0·9624 (0·9204–1·0000)	0·9487 (0·8268–0·9937)	0·8606 (0·8252–0·8912)	0·8718 (0·7257–0·9570)	0·9027 (0·8715–0·9284)
Midline shift	0·9697 (0·9403–0·9991)	0·9385 (0·8499–0·9830)	0·8944 (0·8612–0·9219)	0·9077 (0·8098–0·9654)	0·9108 (0·8796–0·9361)
Mass effect	0·9216 (0·8883–0·9548)	0·9055 (0·8408–0·9502)	0·7335 (0·6849–0·7782)	0·8189 (0·7408–0·8816)	0·9038 (0·8688–0·9321)

Neither of the datasets was used during the training process. AUCs are shown for nine critical CT findings in both these datasets. Two operating points were chosen on the ROC curve for high sensitivity and high specificity, respectively. Absolute number used for calculation of sensitivity and specificity are in the appendix (p 7). AUC=area under the receiver operating characteristic curve. ROC=receiver operating characteristic.

Table 4: **Performance of the algorithms on the Qure25k and CQ500 datasets**

intracranial haemorrhage, 0·96 (0·92–1·00) for calvarial fracture, and 0·97 (0·94–1·00) for midline shift.

In a comparison of the performance of the algorithms to that of the radiologists on the CQ500 dataset, at high sensitivity operating point, sensitivities of algorithms and radiologists were not significantly different (p>0·05) but algorithms' specificities were significantly lower (p<0·0001; appendix pp 8–9).

Discussion

To our knowledge, our study is the first to describe the development of a system that separately identifies critical abnormalities on head CT scans and to conduct a validation with a large number of scans sampled uniformly from the population distribution. We also report the algorithms' accuracy versus a consensus of three radiologists on a second independent dataset, the CQ500 dataset. We have made this dataset and the corresponding reads available for public access so that they can be used to benchmark comparable algorithms in the future. Such publicly available datasets had earlier spurred comparison of the algorithms in other tasks such as lung nodule detection[25] and chest radiograph diagnosis.[6]

Automated and semi-automated detection of findings from head CT scans have been studied by other groups. Grewal and colleagues[9] developed a deep learning approach to automatically detect intracranial haemorrhages. They reported a sensitivity of 0·8864 and a positive predictive value (precision) of 0·8124 on a dataset of 77 brain CT scans read by three radiologists. However, the types of intracranial haemorrhage considered were not mentioned in their report. Traditional computer vision techniques such as morphological processing were used by Zaki and colleagues[26] to detect fractures and by Yamada and colleagues[27] to retrieve scans with fractures. Neither of the two studies measured accuracies on a clinical dataset. Automated midline shift detection was also explored[28–30] using non-deep learning methods. Convolutional neural networks were used by Gao and colleagues[8] to classify head CT scans to help diagnose Alzheimer's disease. More recently, Prevedello and colleagues[31] assessed the performance of a deep learning algorithm on a dataset of 50 scans to detect haemorrhage, mass effect, or hydrocephalus, and suspected acute infarct. The investigators reported AUCs of 0·91 for haemorrhage, mass effect, or hydrocephalus, and 0·81 for suspected acute infarct.

Our work is novel because it is the first large study in which the use of deep learning on head CT scans is used

For more on the **CQ500 dataset and corresponding reads** see http://headctstudy.qure.ai/dataset

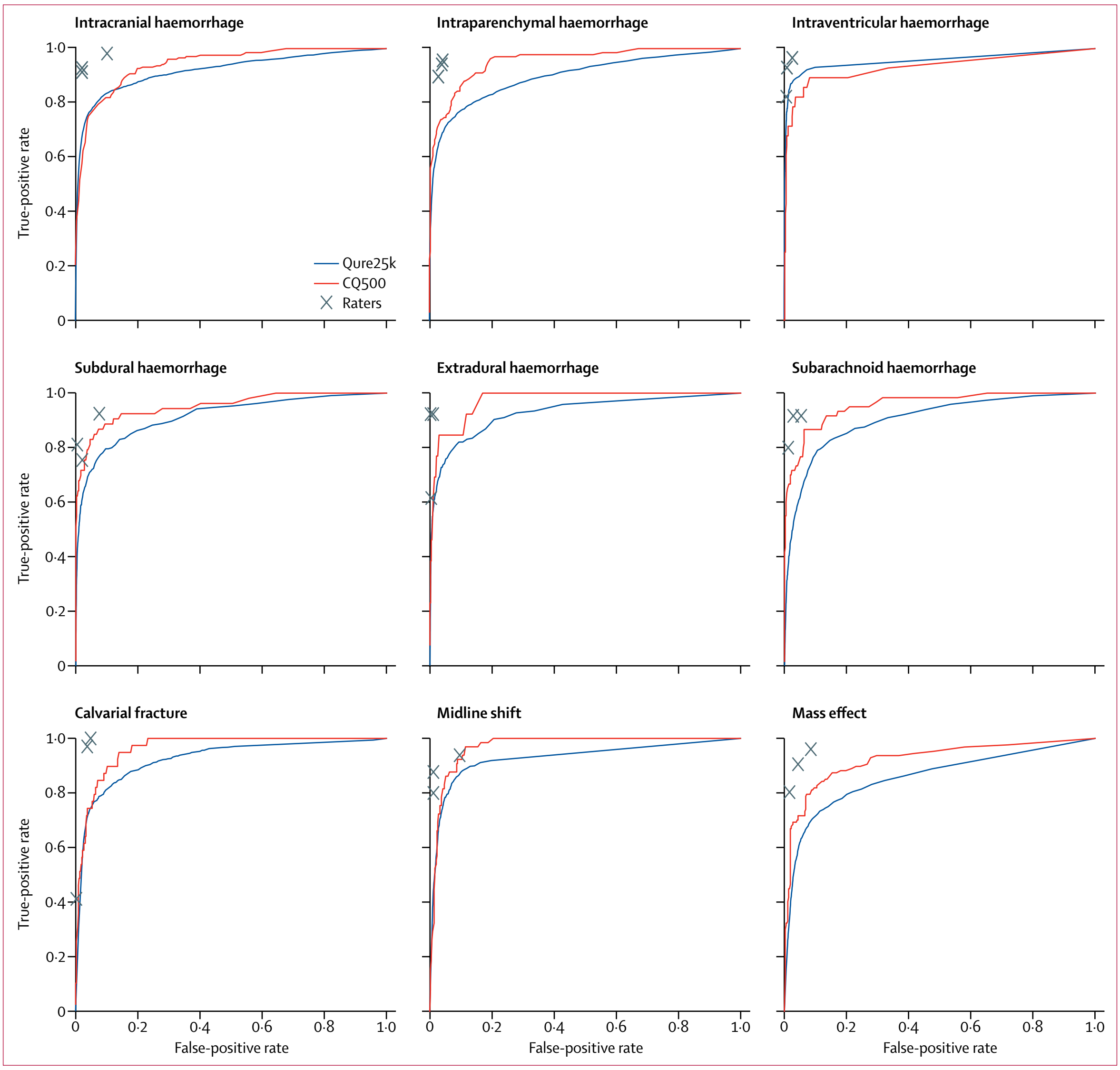


Figure 2: **ROC curves for the algorithms on Qure25k and CQ500 datasets**
Individual raters' true positive and false positive rates measured against their consensus on the CQ500 dataset are also plotted along with the ROCs for comparison. ROC=receiver operating characteristic.

to detect and separately report accuracy for each critical finding, including the five types of intracranial haemorrhage. Furthermore, there is very little literature to date describing the accurate use of deep learning algorithms to detect cranial fractures. We demonstrate that deep learning algorithms are able to perform this task with high accuracy. The validation of algorithms that detect mass effect and midline shift (both used to estimate severity of a range of intracranial conditions and the need for urgent intervention) in such a large number of patients is also unique.

The algorithms produced fairly good results for all the target findings on both the Qure25k and CQ500 datasets. AUCs for all the findings apart from mass effect were greater than or approximately equal to 0·9. AUCs on the CQ500 dataset were better than those on the Qure25k dataset. We hypothesise that this might be because of two reasons. First, because radiologists reading the Qure25k dataset had access to clinical history of the patients, their reads incorporated extra clinical information not available in the scans. The algorithms did not have access to this information and therefore did not perform well. Second, a majority vote of three raters is a better gold standard than that of one rater. Indeed, we observed that AUCs of the algorithms on the CQ500 dataset were lower when a single rater was considered the gold standard instead of the majority vote (appendix p 5).

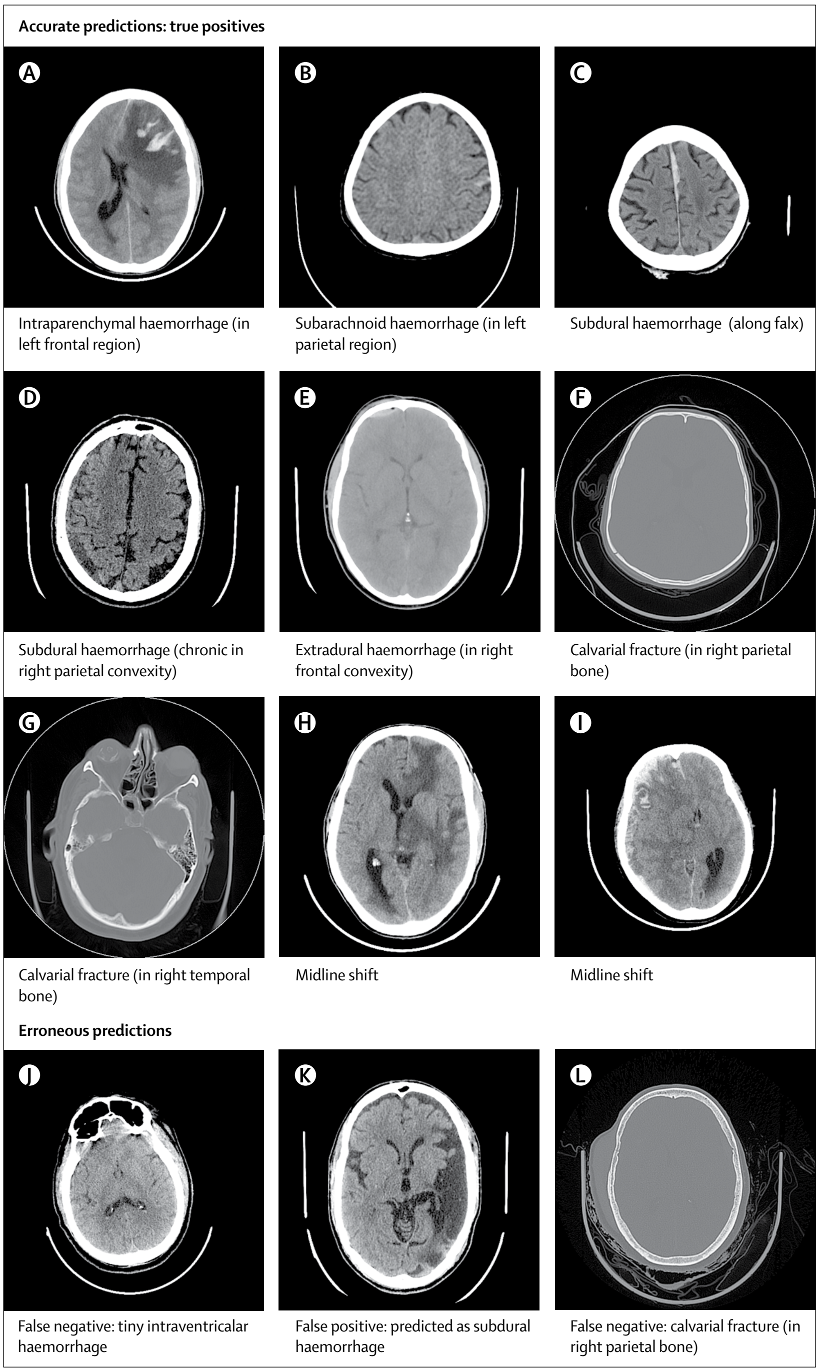
We expect that the Qure25k dataset and the first batch of the CQ500 dataset represent the population distribution of head CT scans. This is because the Qure25k dataset was randomly sampled from a large database of head CT scans, whereas the first batch of the CQ500 dataset consisted of all the head CT scans acquired at the selected centres in a month. The observation that age, sex, and prevalence statistics are similar for both datasets further supports this hypothesis. The CQ500 dataset as a whole, however, is not representative of the population because the second batch was selected for higher incidence of haemorrhages. Despite this difference in prevalence, our performance metrics (ie, AUC, sensitivity, and specificity) should represent the performance on the population because these metrics are prevalence independent.

We did an informal qualitative analysis of the algorithms' outputs on the CQ500 dataset. The algorithms produced good results for normal scans without bleed, scans with medium to large sized intraparenchymal and extra-axial haemorrhages, haemorrhages with fractures, and in predicting midline shift. There was room for improvement for small-sized intraparenchymal, intraventricular haemorrhages and haemorrhages close to the skull base. In this study, we did not separate chronic and acute haemorrhages. This approach resulted in occasional prediction of scans with infarcts and prominent cerebrospinal fluid spaces as intracranial haemorrhages. However, the false positive rates of the algorithms should not impede its usability as a triaging tool. We show some accurate and erroneous predictions of the algorithms in figure 3.

Our study has several limitations. Although the selection strategy ensured that there were a substantial number of positive scans in the CQ500 dataset for most of our target findings, the number of scans with extradural haemorrhage was found only to be 13. This result made the confidence intervals of sensitivities of extradural haemorrhage in this dataset wide. There is also a risk of selection bias in the CQ500 dataset, perhaps because ambiguously worded reports confounded the NLP algorithm and therefore were missed while selecting the second batch. However, this risk is minimal because of the high accuracy of the NLP algorithm when tested on the reports used to select this dataset (appendix p 5).

For the scans in the CQ500 dataset, concordance between the three radiologists was not very high for all findings. In particular, calvarial fracture had low Cohen's κ of 0·58, 0·37, and 0·36 between the pairs of raters. This result might be because of non-availability of clinical history to the raters. We observed that the raters were either very sensitive or very specific to a particular target finding (appendix p 8). For example, two raters were highly sensitive to calvarial fracture whereas the third rater was highly specific.

Another limitation of our study is that we did not exclude follow-up scans of patients from the CQ500 dataset, mainly because very few scans were reported with some of our target abnormalities such as extradural and intraventricular haemorrhages. We could not present the extent of this limitation because of non-availability of unique identifiers of patients in this dataset. Existence of follow-up images in the dataset might mean that the scans are not independent of each other, and therefore presented 95% CIs might be too tight.



Accurate predictions: true positives

Intraparenchymal haemorrhage (in left frontal region)

Subarachnoid haemorrhage (in left parietal region)

Subdural haemorrhage (along falx)

Subdural haemorrhage (chronic in right parietal convexity)

Extradural haemorrhage (in right frontal convexity)

Calvarial fracture (in right parietal bone)

Calvarial fracture (in right temporal bone)

Midline shift

Midline shift

Erroneous predictions

False negative: tiny intraventricalar haemorrhage

False positive: predicted as subdural haemorrhage

False negative: calvarial fracture (in right parietal bone)

Figure 3: **Some accurate and erroneous predictions of the algorithms**

In this study, we have limited our algorithm to the detection of calvarial (cranial vault) fractures. Another missing component is a thoroughly validated algorithm that localises lesions. Both of these are important for a clinical decision support system.

Our results show that deep learning algorithms can be trained to detect critical findings on head CT scans with good accuracy. The strong performance of deep learning algorithms suggests that they could be a helpful adjunct for identification of acute head CT findings in a trauma setting, providing a lower performance bound for quality and consistency of radiological interpretation. We think that it might also be feasible to automate the triage process of head CT scans with these algorithms. This approach might improve radiologist efficiency, but it is also possible that over-reliance on such a triage might lead to automation bias in radiologists whereby false negative scans are overlooked. A prospective clinical trial is necessary to determine the safety and efficacy of such a triage and if it ultimately improves patient care and outcomes.

Contributors

SC, RG, ST, PR, and PW conceived and designed the study. SC, RG, and ST developed the algorithms with the help of clinical input from MB and PR. ST developed and validated the NLP algorithms. VM collected, anonymised, and prepared the data from the Centre for Advanced Research in Imaging, Neurosciences and Genomics, New Delhi, India. NGC contributed to the protocol of the study and rated the scans. SC and RG did the primary statistical analysis while VKV did the secondary analysis. SC and RG wrote the initial draft. All authors subsequently critically edited the report. All authors read and approved the final report.

Declaration of interests

SC, RG, ST, PR, and PW are employees of Qure.ai. All other authors declare no competing interests.

Acknowledgments

This study was funded by Qure.ai.

References

1 Coles JP. Imaging after brain injury. *Br J Anaesth* 2007; **99:** 49–60.

2 Larson DB, Johnson LW, Schnell BM, Salisbury SR, Forman HP. National trends in CT use in the emergency department: 1995–2007. *Radiology* 2011; **258:** 164–73.

3 Papa L, Stiell IG, Clement CM, et al. Performance of the Canadian CT head rule and the New Orleans criteria for predicting any traumatic intracranial injury on computed tomography in a United States level I trauma center. *Acad Emerg Med* 2012; **19:** 2–10.

4 Powers WJ, Rabinstein AA, Ackerson T, et al. 2018 guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 2018; **49:** e46–110.

5 Erly WK, Berger WG, Krupinski E, Seeger JF, Guisto JA. Radiology resident evaluation of head CT scan orders in the emergency department. *AJNR Am J Neuroradiol* 2002; **23:** 103–07.

6 Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. 2017 IEEE Conference on Computer Vision And Pattern Recognition (CVPR); Honolulu, HI; July 21–26, 2017. 3462–71.

7 Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. 2017. https://arxiv.org/abs/1711.05225 (accessed Aug 20, 2018).

8 Gao XW, Hui R, Tian Z. Classification of CT brain images based on deep learning networks. *Comput Methods Programs Biomed* 2017; **138:** 49–56.

9 Grewal M, Srivastava MM, Kumar P, Varadarajan S. RADNET: radiologist level accuracy using deep learning for hemorrhage detection in CT scans. 2017. https://arxiv.org/abs/1710.04934 (accessed Aug 20, 2018).

10 Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; **316:** 2402–10.

11 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542:** 115–18.

12 Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Trans Med Imaging* 2016; **35:** 1207–16.

13 Cheng J-Z, Ni D, Chou Y-H, et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. *Sci Rep* 2016; **6:** 24454.

14 Prasoon A, Petersen K, Igel C, Lauze F, Dam E, Nielsen M. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. *Med Image Comput Comput Assist Interv* 2013; **16:** 246–53.

15 Liao S, Gao Y, Oto A, Shen D. Representation learning: a unified deep learning framework for automatic prostate MR segmentation. *Med Image Comput Comput Assist Interv* 2013; **16:** 254–61.

16 Patravali J, Jain S, Chilamkurthy S. 2D-3D fully convolutional neural networks for cardiac MR segmentation. 2017. https://arxiv.org/abs/1707.09813 (accessed Aug 20, 2018).

17 Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal* 2017; **42:** 60–88.

18 Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 2009; **24:** 8–12.

19 Harth S, Obert M, Ramsthaler F, Reuss C, Traupe H, Verhoff MA. Estimating age by assessing the ossification degree of cranial sutures with the aid of flat-panel-CT. *Leg Med (Tokyo)* 2009; **11:** S186–89.

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

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

22 Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med* 2005; **37:** 360–63.

23 Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971; **76:** 378.

24 Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. Hoboken, NJ, USA: John Wiley & Sons, 2013.

25 Armato SG, McLennan G, Bidaut L, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (LDRI): a completed reference database of lung nodules on CT scans. *Med Phys* 2011; **38:** 915–31.

26 Zaki WMDW, Fauzi MFA, Besar R. A new approach of skull fracture detection in CT brain images. In: Badioze Zaman H, Robinson P, Petrou M, Olivier P, Schröder H, Shih TK, eds. Visual informatics: bridging research and practice. Berlin: Springer, 2009: 156–67.

27 Yamada A, Teramoto A, Otsuka T, Kudo K, Anno H, Fujita H. Preliminary study on the automated skull fracture detection in CT images using black-hat transform. 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; Orlando, FL; Aug 17–20, 2016. 6437–40.

28 Chen W, Belle A, Cockrell C, Ward KR, Najarian K. Automated midline shift and intracranial pressure estimation based on brain CT images. *J Vis Exp* 2013; **74:** 3871.

29 Wang H-C, Ho S-H, Xiao F, Chou J-H. A simple, fast and fully automated approach for midline shift measurement on brain computed tomography. 2017. https://arxiv.org/abs/1703.00797 (accessed Aug 20, 2018).

30 Xiao F, Liao C-C, Huang K-C, Chiang I-J, Wong J-M. Automated assessment of midline shift in head injury patients. *Clin Neurol Neurosurg* 2010; **112:** 785–90.

31 Prevedello LM, Erdal BS, Ryu JL, et al. Automated critical test findings identification and online notification system using artificial intelligence in imaging. *Radiology* 2017; **285:** 923–31.