nature research

Corresponding author(s): Eric K Oermann, M.D.

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Sample sizes were not pre-determined, but represented the largest samples that we could pull from our PACs system utilizing the tools available to use when the query was made. We elected to obtain as large a dataset as possible under the general premise that deep learning systems perform better with more data.

The sample of non-contrast CT head studies utilized for training and testing of the algorithms described in the present study represents the largest labeled collection of such studies compiled to date. The studies were acquired following search for relevant study descriptions of the picture archiving and communication system at our institution within the IRB-approved dates.

For the natural language processing (NLP) algorithm verification, 1004 reports were sampled randomly from each year of the 10 year period approved for inclusion in the study and were annotated by three physicians with majority vote deciding discrepancies. Details regarding the NLP pipeline have been published in Zech et al, Radiology, 2018.

For convolution neural network (CNN) testing, 180 studies were randomly sampled from the 37,236 imaging studies that had been previously labeled by the NLP algorithm until an approximate two-to-one split of critical to non-critical studies were obtained. These cases were subsequently review by physicians to obtain gold-standard labels of the imaging pathology. This review included the entire medical record including note review, prior imaging, and follow-up imaging studies.

2. Data exclusions

Describe any data exclusions.

None

3. Replication

Describe whether the experimental findings were reliably reproduced.

The CNN demonstrated consistent performance in the prediction of gold-standard critical findings with an area under the curve of 0.73. These results were consistent across multiple folds of training. When tested prospectively in a simulated clinical trial, the CNN was able to alert the interpreting radiologist to 50% of critical cases with a false alarm rate of 21%. The clinical simulation also demonstrated that the CNN was able to significantly reduce time to notification from minutes to seconds.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

For natural language processing (NLP) algorithm verification, 1004 reports were sampled randomly from each year of the 10 year period approved for inclusion in the study. This was done to ensure that variations in institutional reporting standards and personnel over the study period were included in the test set.

For convolution neural network (CNN) testing, 180 studies were randomly sampled from the 37,236 imaging studies that had been previously labeled by the NLP algorithm until an approximate two-to-one split of critical to non-critical studies were obtained.

For CNN training and validation, CT studies were allocated to the validation and training data sets randomly. We repeated this randomization five times for each of the folds and subsequently integrated them into the ensemble model.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

When comparing neural network performance to physicians, blinding was performed by restricting the physicians ability to study the clinical scenario associated with a study and by restricting image manipulation capabilities normally available to physicians working in practice. A description of blinding procedures is included in the supplemental methods section.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	☐ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	☐ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

The NLP algorithms utilized for silver-standard label generation are described in detail in Zech et al, Radiology, 2018. In brief, we tested multiple NLP approaches to CT head report label generation and utilized the best performing model (BOW with unigrams, bigrams, and trigrams plus average word embeddings vector) to generate the silver-standard labels utilized for subsequent CNN training and validation.

We developed a novel 3D CNN intended to identify critical findings on noncontrast CT head studies. The architectures tested and the ensemble of CNNs trained are described in detail in the supplemental methods section. All deep learning preprocessing, training, and inference were performed in C++ and Python. All models were trained on a pair of nVidia GTX 1080s using Tensorflow 1.0.2 and Keras 2.0.1.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	The imaging studies used to train and test the neural network described in the manuscript are subject to privacy regulations and cannot be made available in totality. Subsets of the imaging studies can be made available following verification of appropriate deidentification. If accepted for publication, our group would make the code discussed in the manuscript publicly available.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used.
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No eukaryotic cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animal were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Demographics related to the 37,236 included patients are as follows: mean (standard deviation) age 61.5 ± 21.0, female sex 52%, male sex 48%. The noncontrast CT Head studies included normal studies as well as a broad range of pathologies including intracranial hemorrhage, stroke, hydrocephalus, mass lesions, fractures, and post-operative changes. The revised manuscript contains detailed tables documenting the schema for critical finding identification as well as demographic data and clinical presentation descriptions for the patients included in the gold-standard set of studies.