Automated extraction of dementia related volumetric assessments from radiology reports using natural language processing

Dr Adam Mayers, Dr Angus Roberts, Prof Dag Aarsland, Dr Ashwin Venkataraman & Prof. Robert Stewart, King's College London

Introduction:

Brain MRI scans are performed on most patients coming to NHS memory clinic. We demonstrate the use of a machine learning algorithm to extract and label text relating to brain volume from a large real-world database of brain MRI reports performed on patients referred from multiple memory clinics in South London (SLAM Image Bank).

Materials and Methods:

Due to the presence of nested/overlapping terms of interest and long phrases a span categorisation was used. This involves tagging specific contiguous groups of tokens as referring to a particular category, and six target categories were identified;

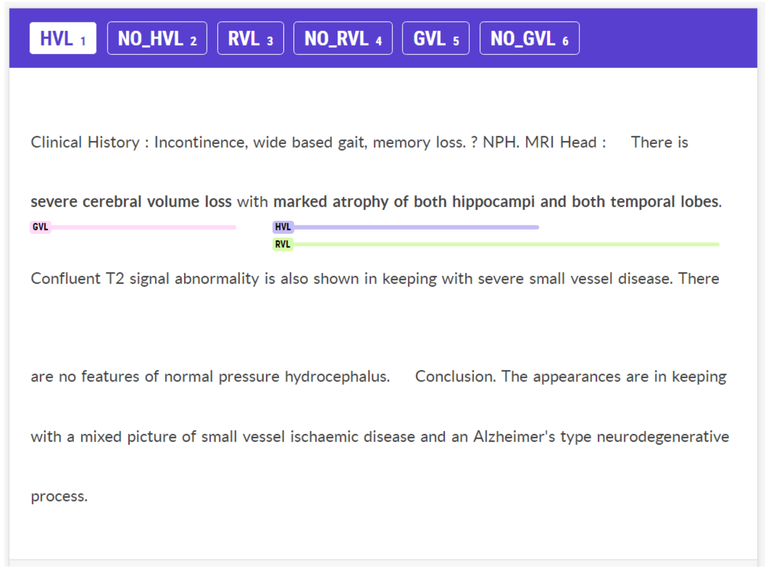
1) presence of global volume loss
2) absence of global volume loss
3) presence of regional volume loss
4) absence of regional volume loss
5) presence of hippocampal / medial temporal lobe volume loss
6) absence of hippocampal /MTL volume loss

4025 brain MRI reports were extracted from the radiology information system from February 2008 to September 2021. These plain text reports were then annotated using the prodigy annotation software.

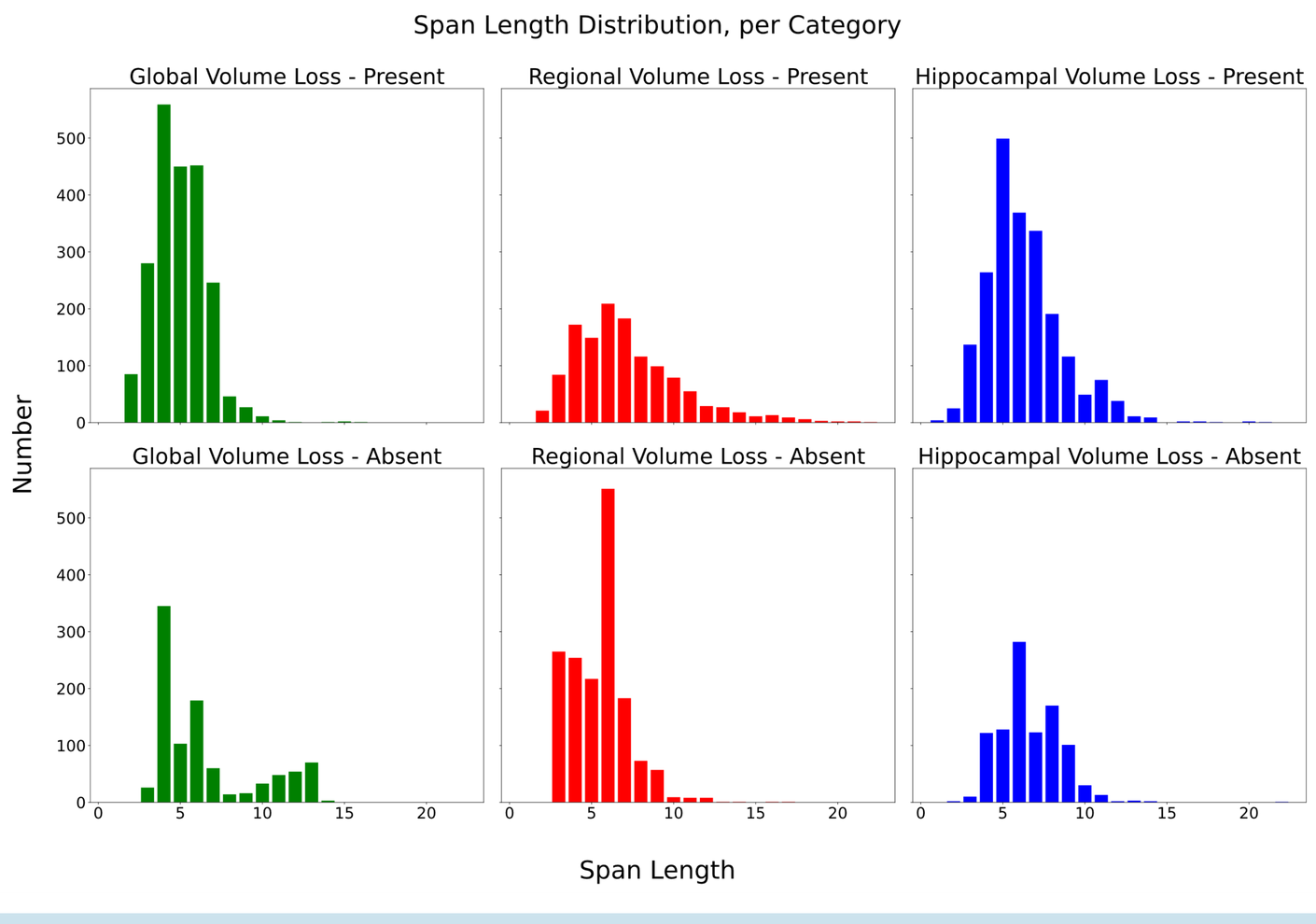
Annotation and Statistics:

9154 span examples were generated following annotation. Each category contained between 951 and 2165 examples, and span length ranged from 1 to 22 tokens with a median span length of 6.

Average span distinctiveness and boundary distinctiveness were 1.84 and 1.42 respectively.



Example of prodigy annotation interface:

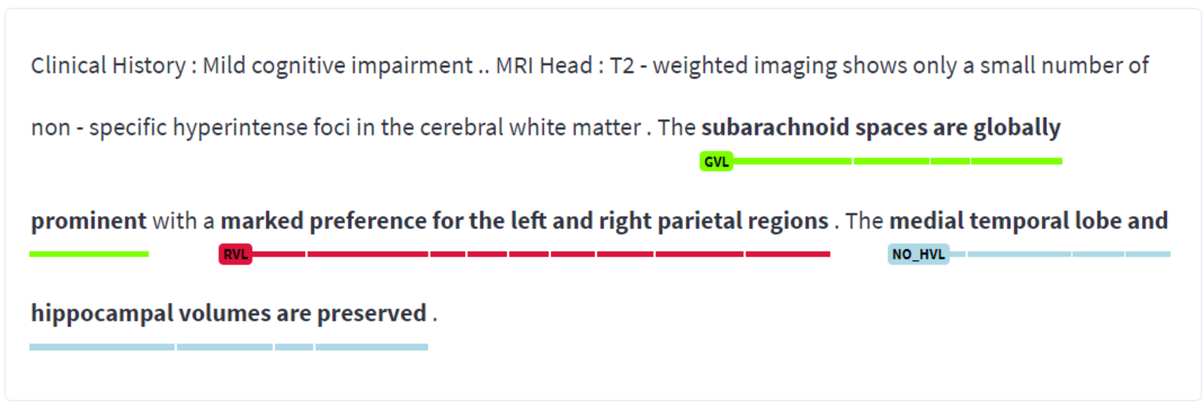


Results:

3206 of the reports (80% of the total) formed the corpus for the training of a neural network model using the open-source spaCy python library, with a further 801 reports (20%) kept as a separate test set.

5-fold cross validation was used to assess model performance during hyperparameter tuning. Against the test set over all six categories the model's F-score was 0.88 (precision = 0.92, recall = 0.84), with variability between the different categories:

Category	Precision	Recall	F-Score
Global Volume Loss – Present	0.93	0.94	0.94
Global Volume Loss – Absent	0.94	0.76	0.84
Regional Volume Loss – Present	0.80	0.60	0.68
Regional Volume Loss – Absent	0.95	0.81	0.88
Hippocampal Volume Loss – Present	0.91	0.89	0.90
Hippocampal Volume Loss – Present	0.98	0.90	0.94



	text	label_	start	end	start_char	end_char	scores
0	subarachnoid spaces are globally prominent	GVL	33	38	174	216	0.9994
1	medial temporal lobe and hippocampal volumes are preserved	NO_HVL	51	59	287	345	0.9993
2	marked preference for the left and right parietal regions	RVL	40	49	224	281	0.7827

Example of model output:

Discussion:

The relative underperformance of the 'regional volume loss' category is expected given the underlying problem; if regional volume loss is present, then any portion of the brain volume may be mentioned making the category very heterogenous. This is reflected in the data, with the RVL category having the broadest range of span length (2-22) and the lowest span distinctiveness (1.38).

Conclusions:

These results support the feasibility and accuracy of natural language processing applied to routine radiology reports and the potential for automated generation of novel meta-data from dementia assessments in electronic health records. This approach is planned to be implemented as part of routine data processing.

KING'S College LONDON
NHS South London and Maudsley NHS Foundation Trust
NHS Guy's and St Thomas' NHS Foundation Trust
NHS Digital Academy
AI CENTRE for Value Based Healthcare