

Labelling imaging datasets on the basis of neuroradiology reports: a validation study

David A. Wood¹, Sina Kafiabadi², Aisha Al Busaidi², Emily Guilhem², Jeremy Lynch², Matthew Townend³, Antanas Montvila², Juveria Siddiqui², Naveen Gadapa², Matthew Benger², Gareth Barker⁴, Sebastian Ourselin¹, James H. Cole^{4,5}, and Thomas C. Booth^{1,2}

MR Imaging abnormality Deep learning Identification study (MIDI)
Consortium

¹ School of Biomedical Engineering, King's College London

² King's College Hospital, London, UK

³ Wrightington, Wigan & Leigh NHSFT

⁴ Institute of Psychiatry, Psychology & Neuroscience, King's College London

⁵ Centre for Medical Image Computing, Dementia Research, University College London

Abstract. Natural language processing (NLP) shows promise as a means to automate the labelling of hospital-scale neuroradiology magnetic resonance imaging (MRI) datasets for computer vision applications. To date, however, there has been no thorough investigation into the validity of this approach, including determining the accuracy of report labels compared to image labels as well as examining the performance of non-specialist labellers. In this work, we draw on the experience of a team of neuroradiologists who labelled over 5000 MRI neuroradiology reports as part of a project to build a dedicated deep learning-based neuroradiology report classifier. We show that, in our experience, assigning binary labels (i.e. normal vs abnormal) to images from reports alone is highly accurate. In contrast to the binary labels, however, the accuracy of more granular labelling is dependent on the category, and we highlight reasons for this discrepancy. We also show that downstream model performance is reduced when labelling of training reports is performed by a non-specialist. To allow other researchers to accelerate their research, we make our refined abnormality definitions and labelling rules available, as well as our easy-to-use radiology report labelling tool which helps streamline this process.

Keywords: Natural language processing · Deep learning · Labelling.

1 Introduction

Deep learning-based computer vision systems hold promise for a variety of applications in neuroradiology. However, a rate-limiting step to clinical adoption is the labelling of large datasets for model training, a laborious task requiring

considerable domain knowledge and experience. Following recent breakthroughs in natural language processing (NLP), it is becoming feasible to automate this task by training text classification models to derive labels from radiology reports and to assign these labels to the corresponding images [12][13][14][7]. To date, however, there has been no investigation into the general validity of this approach, including determining the accuracy of report labels compared to image labels as well as assessing the performance of non-specialist labellers.

In this work we draw on the experience of a team of neuroradiologists who labelled over 5000 magnetic resonance imaging (MRI) neuroradiology reports as part of a project to build a dedicated deep learning-based neuroradiology report classifier. In particular, we examine several aspects of this process which have hitherto been neglected, namely (i) the degree to which radiology reports faithfully reflect image findings (ii) whether the labelling of reports for model training can be reliably outsourced to clinicians who are not specialists (here we examined whether the performance of a neurologist or radiology trainee (UK registrar grade; US resident equivalent) is similar to that of a neuroradiologist) (iii) the difficulty of creating an exhaustive and consistent set of labelling rules, and (iv) the extent to which abnormalities labelled on the basis of examination-level reports are detectable on MRI sequences likely to be available to a computer vision model.

Overall, our findings support the validity of deriving image labels from neuroradiology reports, but with several important caveats. We find that, contrary to basic assumptions often made for this methodology, radiological reports are often less accurate than image findings. Indeed, certain categories of neuroradiological abnormality are inaccurately reported. We conclude that, in our experience assigning binary labels (i.e. normal vs abnormal) to images from reports alone is very accurate. The accuracy of more granular labelling, however, is dependent on the category, and we highlight reasons for this discrepancy.

We also find that several aspects of model training are more challenging than is suggested by a review of the literature. For example, designing a complete set of clinically relevant abnormalities for report labelling, and the rules by which these were applied, took our team of four neuroradiologists more than six months to complete with multiple iterations, and involved the preliminary inspection of over 1,000 radiology reports. To allow other researchers to bypass this step and accelerate their research, we make our refined abnormality definitions and labelling rules available. We also make our radiology report labelling tool available which helps streamline this manual annotation process. Importantly, we found that even when enabled with the labelling tool and set of abnormalities and rules, report annotation for model training must be performed by experienced neuroradiologists, because a considerable reduction in model performance was seen when labelling was performed by a neurologist or a radiology trainee.

2 Related work

NLP models have previously been employed to assign image labels in the context of training computer vision models for neuroradiology applications using radiology reports from both computed tomography (CT) [12][14][9] and MRI [13] examinations. In all cases, classification performance was reported for the primary objective of labelling reports. However, there was no comparison of either the predicted or annotated labels with the images. The closest published work to our paper is therefore a conference abstract highlighting discrepancies between the findings detailed in chest radiograph reports and the corresponding images when labelling a limited set of abnormalities [8]. To the best of our knowledge no such investigation has been performed in the context of neuroradiology, nor have the challenges of creating an NLP labelling tool for neuroradiology applications been described.

Previous work has investigated the accuracy of using crowdsourcing to label images in the context of general [5] as well as medical [4] computer vision tasks. However, we know of no work in the context of neuroradiology which investigates the level of expertise required for accurate manual annotation of reports. Although it might seem obvious that experienced neuroradiologists are required for this task, previous works have instead employed post-graduate radiology and neurosurgery residents [14] or attending physicians [12][9], without providing any insight into the possible reduction in labelling accuracy that such delegation may invite.

Automated brain abnormality detection using either T₂-weighted or diffusion-weighted images (DWI) and employing supervised [11][10] and unsupervised [2] deep learning models has previously been reported. However, in each case only a limited set of abnormalities were available during training and testing, and there was no investigation into the range of abnormalities likely to be detected by the computer vision system using only these sequences. In fact, to the best of our knowledge no investigation has determined what fraction of abnormalities are visible to expert neuroradiologists inspecting only a limited number of sequences. Resolving this point could help narrow the architecture search space for future deep learning-based abnormality detection systems.

3 Data and methods

The UK's National Health Research Authority and Research Ethics Committee approved this study. 126,556 radiology reports produced by expert neuroradiologists (UK consultant grade; US attending equivalent), consisting of all adult (> 18 years old) MRI head examinations performed at Kings College Hospital NHS Foundation Trust, London, UK (KCH) between 2008 and 2019, were included in this study. The reports were extracted from the Computerised Radiology Information System (CRIS) (Healthcare Software Systems, Mansfield, UK) and all

data was de-identified. Over the course of more than twelve months, 5000 reports were annotated by a team of neuroradiologists to generate reference standard report labels to train the neuroradiology report classifier described in [13] (ALARM classifier). Briefly, each unstructured report was typically composed of 5-10 sentences of image interpretation, and sometimes included information from the scan protocol, comments regarding the patient’s clinical history, and recommended actions for the referring doctor. In the current paper, we refer to these reference standard labels generated on the basis of manual inspection of radiology reports as ‘silver reference standard labels’. Prior to manual labelling, a complete set of clinically relevant categories of neuroradiological abnormality, as well as the rules by which reports were labelled, were generated following six months of iterative experiments involving the inspection of over 1000 radiology reports. The complete set of abnormalities, grouped by category, are presented in the supplemental material.

Three thousand reports were independently labelled by two neuroradiologists for the presence or absence of any of these abnormalities. We refer to this as the ‘coarse dataset’ (i.e. normal vs. abnormal). Agreement between these two labellers was 94.9%, with a consensus classification decision made with a third neuroradiologist where there was disagreement. Separately, 2000 reports were labelled by a team of three neuroradiologists for the presence or absence of each of 12 more specialised categories of abnormality (mass e.g. tumour; acute stroke; white matter inflammation; vascular abnormality e.g. aneurysm; damage e.g. previous brain injury; Fazekas small vessel disease score [6]; supratentorial atrophy; infratentorial atrophy; foreign body; haemorrhage; hydrocephalus; extra-cranial abnormality). We refer to this as the ‘granular dataset’. There was unanimous agreement between these three labellers across each category for 95.3% of reports, with a consensus classification decision made with all three neuroradiologists where there was disagreement.

We manually inspected 500 images (comprising, on average, 6 MRI sequences) to generate reference standard image labels. We refer to labels generated in this way as ‘gold reference standard labels’. 250 images were labelled for the presence or absence of any abnormality, systematically following the same criteria as that used to generate the coarse report dataset. Similarly, 250 images were examined and given 12 binary labels corresponding to the presence or absence of each of the more granular abnormality categories.

Our team designed a complete set of clinically relevant categories capable of accurately capturing the full range of pathologies which present on brain MRI scans. The aim here was to try and emulate the behaviour of a radiologist in the real world, guided by the need for clinical intervention for an abnormal finding. To help other researchers bypass this step, and to encourage standardization across research groups of abnormality definitions, we make our abnormality categories, as well as all clinical rules, available in the supple-

mental material. Our manual labelling campaign was considerably aided by our development of a dedicated labelling app. This tool allows easy visualisation and labelling of reports through a graphical user interface (GUI), and includes functionality for flagging difficult cases for group consensus/review. Two apps were developed - one for binary labelling (Figure 1), and one for more granular labelling (Figure 2) - and we make both available to other researchers at <https://github.com/MIDIconsortium/RadReports>.

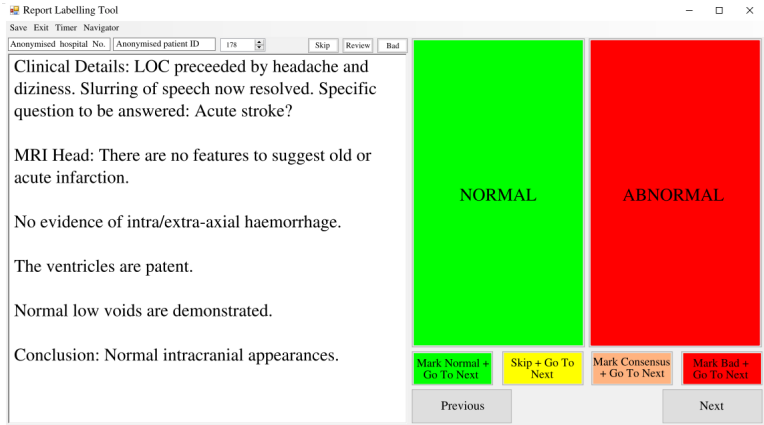


Fig. 1. Binary report labelling tool for the MR Imaging abnormality Deep learning Identification (MIDI) study. The example report should be marked as normal.

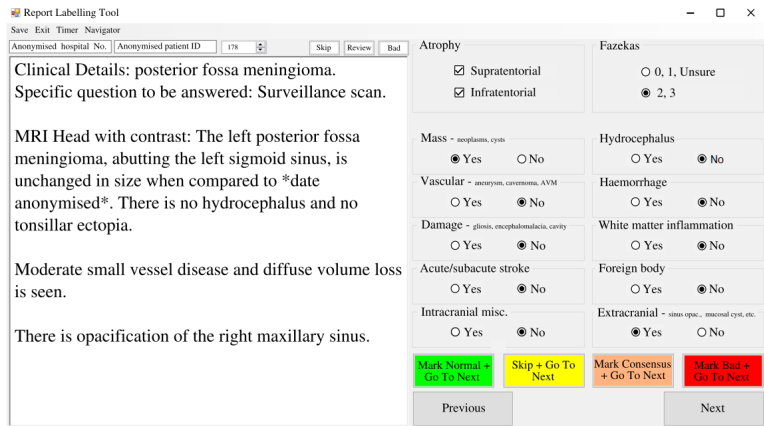


Fig. 2. Granular report labelling tool for the MIDI study. The correct labels for this example report have been selected.

4 Results

4.1 Impact of annotator expertise

To assess the level of expertise required to perform manual annotation of reports for training a text classification model, two experiments were performed.

First, we compared the coarse labels (i.e. normal vs. abnormal) generated by a hospital doctor with ten years experience as a stroke physician and neurologist, who was trained by our team of neuroradiologists over a six month period, with neuroradiologist-generated labels. The rationale for determining the performance was twofold. Neurologists and stroke physicians frequently interpret reports held on the Electronic Patient Record during patient consultations, therefore it is expected that they would be able to differentiate, and therefore label, normal or abnormal reports accurately. Moreover, given that there are less neuroradiologists than neurologists or stroke physicians, with a ratio of 1:4 in the UK, it is likely to be easier to recruit such physicians to perform such labelling tasks.

We found a reduction in performance of neurologist labelling when compared to the labels created by an expert neuroradiologist (Table 1). Based on classification and evaluation methodology in [13], the state-of-the-art ALARM classifier was trained using these neurologist-derived labels and, for comparison, labels generated by a blinded neuroradiologist (Figure 3). The corresponding reduction in classification performance on a hold-out test set of silver reference-standard labels (i.e. reports with consensus) at an arbitrarily fixed sensitivity of 90% (Table 2) demonstrates the impact of what we have shown to be a sub-optimal labelling strategy. In summary, there is optimal performance when the classifier is trained with reports labelled by an experienced neuroradiologist.

Table 1. Labelling performance of a stroke physician and neurologist.

Accuracy (%)	Sensitivity (%)	Specificity (%)
92.7	77.2	98.9

Table 2. Accuracy, specificity, and F1 score of a neuroradiology report classifier trained using data labelled by either a neurologist or neuroradiologist operating at a fixed sensitivity of 90%. Best performance in bold.

Annotator	Accuracy (%)	Specificity (%)	F1 (%)
Neurologist	89.8	89.5	75.8
Neuroradiologist	96.4	97.7	90.3

As a second experiment, a 3rd year radiology trainee who was also trained by our team over a six month period to label neuroradiology reports, generated

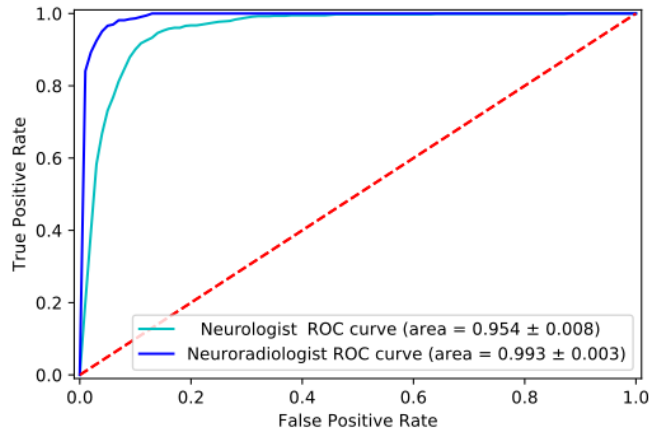


Fig. 3. ROC curve for a neuroradiology report classifier trained on labels generated by a neurologist (cyan) and a neuroradiologist (blue). The area under the curve (AUC) is shown.

labels for our ‘granular dataset’. There was a reduction in radiology trainee performance, averaged across all 12 binary labels, when compared to the silver reference standard labels created by our team of expert neuroradiologists (Table 3). The sensitivity of these labels is clearly too low to be used for model training.

Table 3. Labelling performance of a radiology trainee on the ‘granular dataset’, averaged across all 12 binary labels.

Sensitivity (%)	Specificity (%)	F1 (%)
64.4	98.3	70.8

It is worth highlighting that reliability (inter-rater agreement) and accuracy (performance) should not be conflated for labelling tasks. We demonstrate this in a further experiment where the same neurologist previously described also generated labels for our ‘granular dataset’. The Fleiss κ score for the radiology trainee and the neurologist averaged over all 12 binary categories was 0.64, which is above the threshold previously employed to establish neuroradiology label reliability[14]. Substantial inter-rater agreement (commonly taken as $\kappa > 0.6$), therefore, does not necessarily equate to label accuracy as this experiment has shown.

4.2 Report validation

To determine the validity of assigning image labels on the basis of radiology reports, the granular labels derived from reports (silver reference standard) were

compared to those derived by inspecting the corresponding images (gold reference standard) for 500 cases (Table 4). Although the false positive rate of report labelling is very low for the 12 granular categories of interest, it is clear that the sensitivity of radiology report labelling is category dependent and can be low. On further analysis, we found that insensitive labelling for any given category typically reflects the absence of any reference in the report to that particular category rather than a discrepancy in interpretation. The categories with low sensitivity include hydrocephalus, haemorrhage, extra-cranial abnormalities, and infratentorial atrophy. The reasons for this are discussed below.

Table 4. Accuracy of silver reference standard report labels for granular categories when compared to the corresponding gold standard image labels. Categories with sensitivity > 80% in bold.

Category	Sensitivity (%)	Specificity (%)	F1 (%)
Fazekas	90.5	95.6	93.2
Mass	97.9	93.6	95.9
Vascular	83.3	88.4	86.5
Damage	82.4	92.7	87.8
Acute Stroke	94.4	99.5	94.4
Haemorrhage	69.2	99.6	78.3
Hydrocephalus	70.0	99.6	77.8
White Matter Inflammation	95.6	100	97.7
Foreign Body	100.0	99.6	96.6
Extracranial abnormality	60.0	94.7	54.5
Supratentorial Atrophy	100	94.6	76.9
Infratentorial Atrophy	77.7	94.3	54.5
Macro-average	85.1	96.0	82.8

Importantly, silver standard binary labels indicating the presence or absence of any abnormality in a report (i.e. normal vs. abnormal) were accurate when compared to the image (gold reference standard label) (Table 5).

Table 5. Accuracy of silver reference standard report labels for binary categories (i.e. normal vs abnormal) relative to the corresponding gold standard image labels.

Category	Sensitivity (%)	Specificity (%)	F1 (%)
Normal vs. abnormal	98.7	96.6	98.5

4.3 MRI sequences and abnormality visibility

In another experiment we examined the utility of assigning examination-level labels derived from radiology reports to different MRI sequences. In general,

neuroradiology reports detail findings from multi-modality (i.e. multiple MRI sequences) imaging examinations, with individual sequences providing complementary information to discriminate specific tissues, anatomies and pathologies. For example, the signal characteristics of blood changes over time, the rate of which is sequence dependent. Therefore analysis of images from multiple sequences allows the chronicity of a haemorrhage to be deduced. Assigning the same label to all images in a multi-modality examination can confound computer vision classification if a model isn't optimised to take as its input the individual sequence from which a particular examination-level label was derived. Therefore, we wished to determine whether a minimal number of sequences would be sufficient for use with report-derived labels. At our institution, axial T₂-weighted and DWI images are typically obtained for routine image review, with over 78% of patients receiving both images during an examination. We sought to determine what fraction of abnormalities are visible to a neuroradiologist inspecting only the T₂-weighted and DWI images. Binary labels (i.e. normal vs. abnormal) for 250 examinations were generated by inspecting only these sequences, and compared to labels derived from all available sequences for the same examinations. The agreement between these two labels was 97.8%, showing that these two sequences would be sufficient for use with report-derived labels for most abnormality detection tasks. Examples of the wide range of abnormalities identified on the basis of T₂-weighted and DWI imaging appear in the supplemental material, along with reports describing abnormalities which weren't visible on either of these two sequences.

5 Discussion

In this work we have examined several assumptions which are fundamental to the process of deriving image labels from radiology reports. Overall, our findings support the validity of deriving image labels from neuroradiology reports. In particular, assigning binary labels (i.e. normal vs abnormal) to images from reports alone is highly accurate and therefore acceptable. Until now this has been assumed but has not been thoroughly investigated. The accuracy of more granular labelling, however, is dependent on the category. For example, labelling of acute stroke, mass, neuro-degeneration, and vascular disorders, is shown to be accurate.

The low labelling accuracy seen in some granular labelling categories is a result of low sensitivity. Low sensitivity typically reflects the absence of any reference in the report to that particular category rather than a discrepancy in interpretation. A qualitative analysis by our team of neuroradiologists has determined several reasons for low sensitivity in some categories.

First, in the presence of more clinically important findings, neuroradiologists often omit descriptions of less critical abnormalities which may not necessarily change the overall conclusion or instigate a change in the patient's management.

For example, on follow-up imaging of previously resected tumours, we have found that the pertinent finding as to whether there is any progressive or recurrent tumour is invariably commented on. In contrast, the presence of white matter changes secondary to previous radiotherapy appears less important within this clinical context. If unchanged from the previous imaging, a statement to the effect of “otherwise stable intracranial appearances” is typical in these cases.

A second source of low sensitivity is the observation that radiology reports are often tailored to specific clinical contexts and the referrer. A report aimed at a neurologist referrer who is specifically enquiring about a neurodegenerative process in a patient with new onset dementia, for example, may make comments about subtle parenchymal atrophy. In contrast, parenchymal volumes may not be scrutinised as closely in the context of someone who has presented with a vascular abnormality, such as an aneurysm, and a report is aimed at a vascular neurosurgeon. Both sources of low sensitivity mentioned above often reflect a “satisfaction of search error” where the radiologist has failed to appreciate the full gamut of abnormalities. After identifying one or two abnormalities the task may appear complete and there is less desire to continue to interrogate the image [1]. It is also noteworthy that abnormalities which are identified by the neuroradiologist by chance may be judiciously omitted from the report on a case by case basis when such “incidentalomas” are thought to be of little consequence. Because of these sources of low sensitivity, labelling categories of abnormality from radiology reports remains challenging for haemorrhage (note that acute haemorrhage is typically detected by CT; MRI reports were often insensitive to those haemorrhages associated with non-critical findings such as micro-haemorrhages), hydrocephalus, extracranial abnormalities and infratentorial atrophy.

In addition to examining the accuracy of radiology reports compared to image findings, we have also demonstrated that most abnormalities typical of a real-world triage environment are picked up using only T_2 -weighted and DWI sequences. This observation may help narrow the architecture search-space for future deep learning-based brain abnormality detection systems, and allow a more accurate comparison of model performance across research groups. However, there are certain abnormalities which may not be visible on these sequences. For example, the presence of microhaemorrhages or blood breakdown products (hemosiderin), are sometimes only visible on gradient echo (T_2^* -weighted) or susceptibility weighted imaging (SWI) [3]. Furthermore, foci of pathological enhancement on post contrast T_1 -weighted imaging can indicate underlying disease which may not be apparent on other sequences. Therefore, whilst we have shown that using T_2 -weighted and DWI sequences alone allows almost all abnormalities to be identified visually, and that plausibly this will translate to efficient computer vision training tasks, it is important to be aware that there are potential limitations.

We briefly discuss several logistical aspects of the report labelling process which were not covered by our more quantitative investigations. Our team designed a complete set of clinically relevant categories capable of accurately capturing the full range of pathologies which present on brain MRI scans. The aim here was to try and emulate the behaviour of a radiologist in the real world, guided by the need for clinical intervention for an abnormal finding. This process, however, was more onerous than is often presented in the literature, requiring the inspection of over 1000 radiology reports by our team of experienced neuroradiologists over the course of more than six months before an exhaustive and consistent set of abnormality categories, as well as the rules by which reports were to be labelled, could be finalised. The rules and definitions constantly evolved during the course of the practice labelling experiments. To allow other researchers to bypass this step and accelerate their research, we make our refined abnormality definitions and labelling rules available as well as our dedicated labelling easy-to-use app.

6 Conclusion

We conclude that in our experience, assigning binary labels (i.e. normal vs abnormal) to images from reports alone is highly accurate. Importantly, we found that even when enabled with the labelling tool and set of abnormalities and rules, annotation of reports for model training must be performed by experienced neuroradiologists, because a considerable reduction in model performance was seen when labelling was performed by a neurologist or a radiology trainee. In contrast to the binary labels, the accuracy of more granular labelling is dependent on the category.

References

1. Berbaum, K., Franken, E., Caldwell, R., Scharz, K., Madsen, M.: Satisfaction of Search in Radiology, p. 121–166. Cambridge University Press, 2 edn. (2018). <https://doi.org/10.1017/9781108163781.010>
2. Chen, X., Konukoglu, E.: Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders (2018), arXiv:1806.04972 [cs.CV]
3. Chiewvit, P., Piyapittayanan, S., Pongvarin, N.: Cerebral venous thrombosis: diagnosis dilemma. *Neurology international* **3**, e13 (11 2011). <https://doi.org/10.4081/ni.2011.e13>
4. Cocos, A., Masino, A., Qian, T., Pavlick, E., Callison-Burch, C.: Effectively crowdsourcing radiology report annotations. In: Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis. pp. 109–114. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015). <https://doi.org/10.18653/v1/W15-2614>, <https://www.aclweb.org/anthology/W15-2614>
5. Crump, M.J.C., McDonnell, J.V., Gureckis, T.M.: Evaluating amazon’s mechanical turk as a tool for experimental behavioral research. *PLOS ONE* **8**(3), 1–18 (03

- 2013). <https://doi.org/10.1371/journal.pone.0057410>, <https://doi.org/10.1371/journal.pone.0057410>
6. Fazekas, F., Chawluk, J., Alavi, A., Hurtig, H., Zimmerman, R.: Mr signal abnormalities at 1.5 t in alzheimer’s dementia and normal aging. *AJR. American journal of roentgenology* **149**, 351–6 (08 1987). <https://doi.org/10.2214/ajr.149.2.351>
 7. Garg, R., Oh, E., Naidech, A., Kording, K., Prabhakaran, S.: Automating ischemic stroke subtype classification using machine learning and natural language processing. *Journal of Stroke and Cerebrovascular Diseases* **28**(7), 2045 – 2051 (2019). <https://doi.org/https://doi.org/10.1016/j.jstrokecerebrovasdis.2019.02.004>, <http://www.sciencedirect.com/science/article/pii/S1052305719300485>
 8. Olatunji, T., Yao, L., Covington, B., Rhodes, A., Upton, A.: Caveats in generating medical imaging labels from radiology reports. *CoRR* (2019), <http://arxiv.org/abs/1905.02283>, arXiv:1905.02283 [cs.CL]
 9. Ong, C.J., Orfanoudaki, A., Zhang, R., Caprasse, F.P.M., Hutch, M., Ma, L., Fard, D., Balogun, O., Miller, M.I., Minnig, M., Saglam, H., Prescott, B., Greer, D.M., Smirnakis, S., Bertsimas, D.: Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. *PLOS ONE* **15**(6), 1–16 (06 2020). <https://doi.org/10.1371/journal.pone.0234908>, <https://doi.org/10.1371/journal.pone.0234908>
 10. Rauschecker, A.M., Rudie, J.D., Xie, L., Wang, J., Duong, M.T., Botzolakis, E.J., Kovalovich, A.M., Egan, J., Cook, T.C., Bryan, R.N., Nasrallah, I.M., Mohan, S., Gee, J.C.: Artificial intelligence system approaching neuroradiologist-level differential diagnosis accuracy at brain mri. *Radiology* **295**(3), 626–637 (2020). <https://doi.org/10.1148/radiol.2020190283>, <https://doi.org/10.1148/radiol.2020190283>, pMID: 32255417
 11. Rezaei, M., Yang, H., Meinel, C.: Brain abnormality detection by deep convolutional neural network (2017), arXiv:1708.05206 [cs.CV]
 12. Shin, B., Chokshi, F.H., Lee, T., Choi, J.D.: Classification of radiology reports using neural attention models. In: 2017 International Joint Conference on Neural Networks (IJCNN). pp. 4363–4370. IEEE (2017)
 13. Wood, D.A., Lynch, J., Kafiabadi, S., Guilhem, E., Busaidi, A.A., Montvila, A., Varsavsky, T., Siddiqui, J., Gadapa, N., Townend, M., Kiik, M., Patel, K., Barker, G., Ourselin, S., Cole, J.H., Booth, T.C.: Automated labelling using an attention model for radiology reports of mri scans (ALARM) (2020), arXiv:2002.06588 [cs.CV]
 14. Zech, J., Pain, M., Titano, J., Badgeley, M., Schefflein, J., Su, A., Costa, A., Bederson, J., Lehar, J., Oermann, E.: Natural language-based machine learning models for the annotation of clinical radiology reports. *Radiology* **287**, 171093 (01 2018). <https://doi.org/10.1148/radiol.2018171093>