Interleaved Text/Image Deep Mining on a Large-Scale Radiology Database for Automated Image Interpretation

Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, Ronald M. Summers
Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, National Institutes of Health Clinical Center

Various patient data of a large population are available on the patient archive and communication system (PACS) of many hospitals or clinical institutions. However such data are not widely studied, due to the challenges encountered in analyzing a large clinical dataset. Nonetheless, efficient analysis of large data can lead us to gain useful, possibly unprecedented insights in the area under study. With the big-data analysis on large collection of radiology images, we aim to achieve 'predictive medicine' – detecting diseases using large patient population image screening.

This work is the continuation of the study performed in [9]. In [9], about 780,000 radiology reports were collected from the PACS of the National Institutes of Health Clinical Center, comprising about 1 billion words. Manually examining and annotating such a large collection of radiology text and images is not only challenging, but also requires an expertise in radiology. To alleviate this challenge, in [9]: *(i)* non-parametric topic modeling algorithm (Latent Dirichlet Allocation (LDA) [1]) was employed to analyze the large collection of reports and to divide them into a number of categories with semantic levels; *(ii)* convolutional neural networks (CNNs) were trained to classify the images into the report categories; and *(iii)* recurrent neural networks (RNNs) and CNNs were trained to predict the "keywords" associated with the images, e.g. to predict "adenopathy", "masses", "lung", given a CT image with lung cancer. The rate of predicted disease-related words matching the actual words in the report sentences (recall-at-K, K=1 (R@1 score)) was 0.56.

While the keywords generation in [9] can aid the interpretation of a patient scan, the generated key-words, e.g. "spine", "lung", are not very specific to a disease in an image. Nonetheless, one of the ultimate goals for large-scale radiology image/text analysis would be to automatically diagnose disease from a patient scan. In order to achieve the goal of automated disease detection, in this work we added an additional pipeline of mining disease words rather than disease-related words using radiology semantics, and predicting these in an image using CNNs with softmax cost-function.

The Unified Medical Language System (UMLS) [5, 7] integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and inter-operable biomedical information systems and services, including electronic health records. The Metathesaurus [8] forms the base of the UMLS, where it is organized by concept. Each concept has specific attributes defining its meaning and is linked to the corresponding concept names. It has 133 semantic types that provide a consistent categorization of all concepts represented in it, and we chose to focus on "T033: finding" and "T047: disease or syndrome" semantic types, as they seemed most likely to be disease specific. RadLex [6] is a unified language to organize and retrieve radiology imaging reports and medical records. While the Metathesaurus has a vast resource of biomedical concepts, we also used RadLex to confine our disease-term-mining more specifically to radiology related terms. The mined words are one word terms appearing in the "T033: finding" and "T047: disease or syndrome" of the UMLS Metathesaurus appearing also in RadLex (RadLex is not a subset of Metathesaurus).

We are interested not only in disease terms associated with an image, but also whether the disease mentioned is present or absent. After detecting semantic terms of "T033: finding" and "T047: disease or syndrome", we used the assertion/negation detection algorithm of [2, 3] to detect presence and absence of disease terms. The number of occurrences "T033: finding" and "T047: disease or syndrome" detected as assertion or negations in radiology reports are shown in Figure 1. While the assertion/negation detection of "T047: disease or syndrome" seemed specific enough, the detection of "T033: finding" was not. For example, it seemed difficult to derive any specific disease information from 43,219 occurrences of possible "unchanged" and 422 occurrences of negated "unchanged". We therefore decided to focus on "T047: disease or syndrome" terms only, and further ignored the terms which occurred less then 10 times in the whole radiology reports. The total number of "T047: disease or syndrome" terms for detecting their presence are 59, and the total number of the terms for detecting their absence are 18.

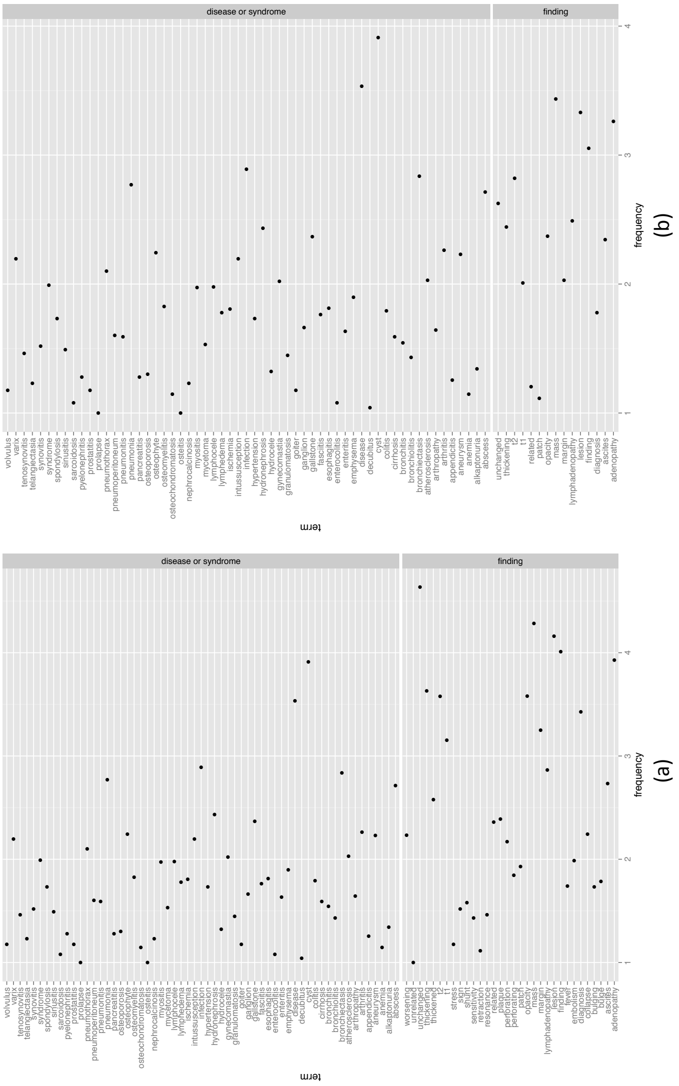
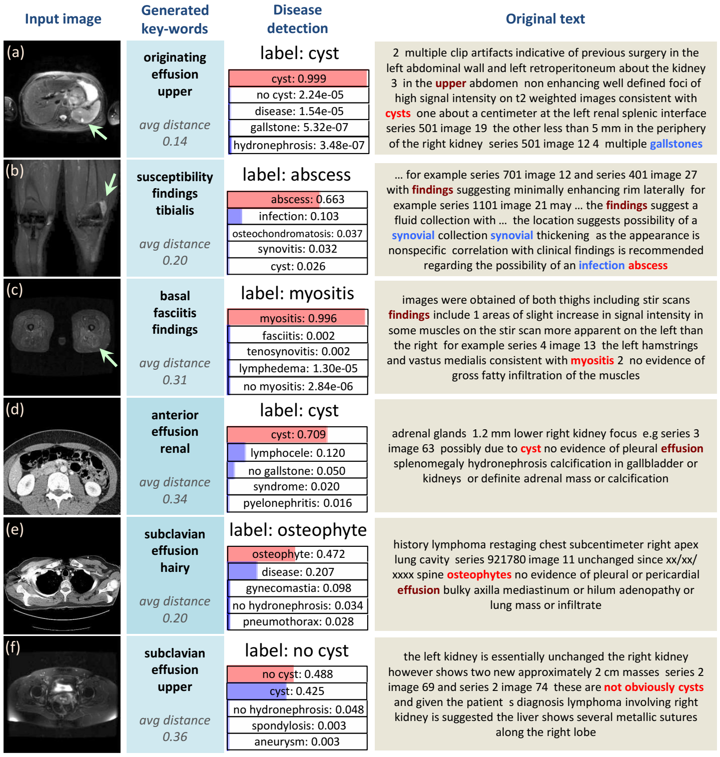


Figure 1: Number of occurrences (frequencies) of semantic terms "T033: finding" and "T047: disease or syndrome" in UMLS Metathesaurus and also appearing in RadLex, detected as (a) assertion and (b) negation in the radiology reports. Frequencies are shown in $\log_{10}$ scale.

Similarly to the object detection task in the ImageNet challenge, we match and detect disease terms found in the sentences of radiology reports referring to the image using CNN and softmax cost function. In addition to assigning disease terms to images, we also assign negated disease terms as absence of the diseases in the images. The total number of labels is 77 (59 present, 18 absent). If more than one disease term is mentioned for an image, we simply assigned the terms multiple times for an image. Some statistics on the number of assertion/negation occurrences per image are shown in Table 1. From the all image-disease-term pairs mined, 85% of image-label pairs were used for training, 5% for cross-validation, and 10% for testing.

With the CNN trained to model image to disease presence/absence prediction, the top-1 test accuracy achieved is 0.71, and top-5 accuracy is 0.88.



2 multiple clip artifacts indicative of previous surgery in the left abdominal wall and left retroperitoneum about the kidney 3 in the upper abdomen non enhancing well defined foci of high signal intensity on t2 weighted images consistent with cysts one about a centimeter at the left renal splenic interface series 501 image 19 the other less than 5 mm in the periphery of the right kidney series 501 image 12 4 multiple gallstones

… for example series 701 image 12 and series 401 image 27 with findings suggesting minimally enhancing rim laterally for example series 1101 image 21 may … the findings suggest a fluid collection with … the location suggests possibility of a synovial collection synovial thickening as the appearance is nonspecific correlation with clinical findings is recommended regarding the possibility of an infection abscess

images were obtained of both thighs including stir scans findings include 1 areas of slight increase in signal intensity in some muscles on the stir scan more apparent on the left than the right for example series 4 image 13 the left hamstrings and vastus medialis consistent with myositis 2 no evidence of gross fatty infiltration of the muscles

adrenal glands 1.2 mm lower right kidney focus e.g series 3 image 63 possibly due to cyst no evidence of pleural effusion splenomegaly hydronephrosis calcification in gallbladder or kidneys or definite adrenal mass or calcification

history lymphoma restaging chest subcentimeter right apex lung cavity series 921780 image 11 unchanged since xx/xx/xxxx spine osteophytes no evidence of pleural or pericardial effusion bulky axilla mediastinum or hilum adenopathy or lung mass or infiltrate

the left kidney is essentially unchanged the right kidney however shows two new approximately 2 cm masses series 2 image 69 and series 2 image 74 these are not obviously cysts and given the patient s diagnosis lymphoma involving right kidney is suggested the liver shows several metallic sutures along the right lobe

Figure 2: Some examples of automated image interpretation, where top-1 probability matches the originally assigned label. The probability assigned to the originally assigned label is shown with a red bar, and the other top-5 probabilities are shown with blue bars. Disease region identified in an image is indicated by an arrow.

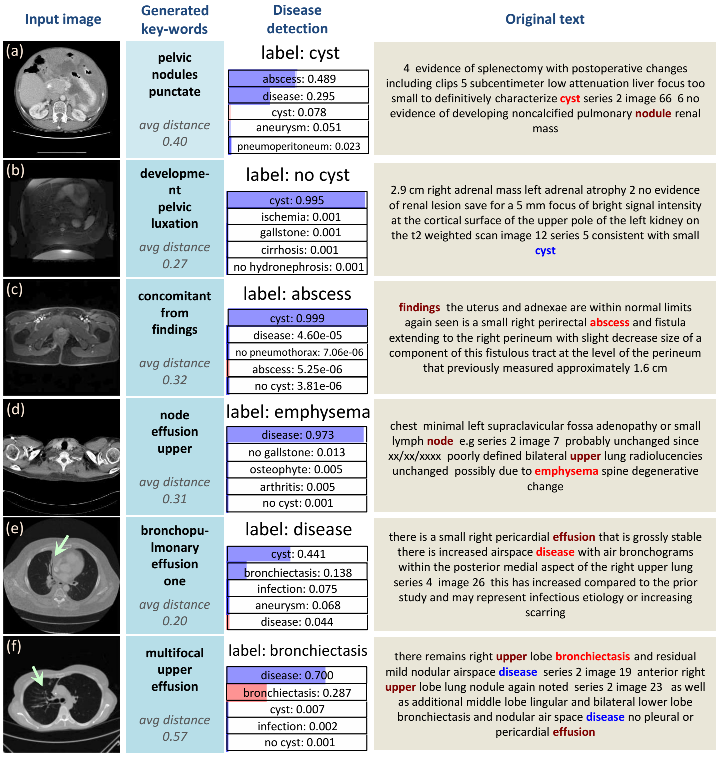
# images		per image mean/std		# assertions per image		# negations per image	
total matching	18291	# assertions mean	1.05	1/image	16133	1/image	1581
total not matching	197495	# negations mean	1.05	2/image	613	2/image	84
with assertions	16827	# assertions std	0.23	3/image	81	3/image	0
with negations	1665	# negations std	0.22	4/image	0	4/image	0

Table 1: Some statistics of images-to-disease presence/absence label matching.

We combine this with the image-to-topic mapping and key-word generation of [9] to generate the final output for comprehensive image interpretation. Some examples of test cases where top-1 probability output matches the originally assigned disease labels are shown in Figure 2, and some examples of test cases where top-1 probability does not match the originally assigned labels are shown in Figure 3.

Automated mining of disease specific terms enables us to predict disease more specifically with promising results. However, compared to image-to-topic modeling in [9] where image labeling was based on topic modeling and loose coupling of image-to-keyword pairs, by matching the images to more specific disease words we lose about 90% of the images for the analysis due to nonspecific original statements. The proportion of the cases where radiologists indicate a disease as strongly positive or negative is often much less then the cases where they describe a finding rather vaguely. Mining and assigning the semantic label "T033: finding" yields more image to specific-disease-label pairs. However, it is probably less specific to model an image with a generic term as "mass" (which is a more vague indication of a specific disease such as "cyst" or "tumor") and detecting it, than modeling and detecting an image with a more specific term as "cyst" (similarly to "finding" or "unchanged").

Utilizing bigger data will enable us to make a more generalizable model, but labeling will become more challenging as the amount of data gets bigger and becomes more heterogeneous. More advanced natural language processing and comprehensive analysis of hospital discharge summaries, progress notes, and patient histories might address the need to get more specific information relating to an image. We hope that this study will inspire and encourage other institutions interested in mining other large unannotated clinical databases to establish a large-scale resource for medical image research.



4 evidence of splenectomy with postoperative changes including clips 5 subcentimeter low attenuation liver focus too small to definitively characterize cyst series 2 image 66 6 no evidence of developing noncalcified pulmonary nodule renal mass

2.9 cm right adrenal mass left adrenal atrophy 2 no evidence of renal lesion save for a 5 mm focus of bright signal intensity at the cortical surface of the upper pole of the left kidney on the t2 weighted scan image 12 series 5 consistent with small cyst

findings the uterus and adnexae are within normal limits again seen is a small right perirectal abscess and fistula extending to the right perineum with slight decrease size of a component of this fistulous tract at the level of the perineum that previously measured approximately 1.6 cm

chest minimal left supraclavicular fossa adenopathy or small lymph node e.g series 2 image 7 probably unchanged since xx/xx/xxxx poorly defined bilateral upper lung radiolucencies unchanged possibly due to emphysema spine degenerative change

there is a small right pericardial effusion that is grossly stable there is increased airspace disease with air bronchograms within the posterior medial aspect of the right upper lung series 4 image 26 this has increased compared to the prior study and may represent infectious etiology or increasing scarring

there remains right upper lobe bronchiectasis and residual mild nodular airspace disease series 2 image 19 anterior right upper lobe lung nodule again noted series 2 image 23 as well as additional middle lobe lingular and bilateral lower lobe bronchiectasis and nodular air space disease no pleural or pericardial effusion

Figure 3: Some examples of final outputs for automated image interpretation where top-1 probability does not match the originally assigned label. The label assignment of second row example is incorrect, due to the failure in the assertion/negation detection algorithm. Nonetheless, the CNN predicted the "true" label correctly ("cyst").

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

[2] Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

[3] Wendy W Chapman, Dieter Hilert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E Chapman, Michael Conway, Melissa Tharp, Danielle L Mowery, and Louise Deleger. Extending the negex lexicon for multiple languages. *Studies in health technology and informatics*, 192:677, 2013.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[5] Betsy L Humphreys, Donald AB Lindberg, Harold M Schoolman, and G Octo Barnett. The unified medical language system an informatics research collaboration. *Journal of the American Medical Informatics Association*, 5(1):1–11, 1998.

[6] Curtis P Langlotz. Radlex: A new method for indexing online educational materials 1. *Radiographics*, 26(6):1595–1597, 2006.

[7] Donald A Lindberg, Betsy L Humphreys, and Alexa T McCray. The unified medical language system. *Methods of information in Medicine*, 32(4):281–291, 1993.

[8] Peri L Schuyler, William T Hole, Mark S Tuttle, and David D Sherertz. The umls metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2):217, 1993.

[9] Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, and Ronald M. Summers. Interleaved text/image deep mining on a very large-scale radiology database. In *CVPR*, June 2015.