Transfer Language Space with Similar Domain Adaptation: A Case Study with Hepatocellular Carcinoma

Patricia Balthazar, MD.^a, Scott Jeffery Lee, MD.^a, Terry Desser, MD.^b,
Daniel L. Rubin, MD.^{c,b}, Judy Wawira Gichoya, MD. MS.^{a,d}, Imon Banerjee,
Ph.D.^{d,a}

^aDepartment of Radiology, Emory School of Medicine, Atlanta, USA

^bDepartment of Radiology, Stanford School of Medicine, Stanford, USA

^cDepartment of Biomedical Data Science, Stanford School of Medicine, Stanford, USA

^dDepartment of Biomedical Informatics, Emory School of Medicine, Atlanta, USA

Abstract

Transfer learning is a common practice in image classification with deep learning where the available data is often limited for training a complex model with millions of parameters. However, transferring language models requires special attention since cross-domain vocabularies (e.g. between news articles and radiology reports) do not always overlap as the pixel intensity range overlaps mostly for images. We present a concept of similar domain adaptation where we transfer an inter-institutional language model between two different modalities (ultrasound to MRI) to capture liver abnormalities. Our experiments show that such transfer is more effective for performing shared targeted task than generic language space transfer. We use MRI screening exam reports for hepatocellular carcinoma as the use-case and apply the transfer language space strategy to automatically label thousands of imaging exams.

Introduction

Hepatocellular carcinoma (HCC) is the most common primary liver malignancy and a leading cause of cancer death worldwide¹. Imaging is critical for HCC screening and includes ultrasound and multiphase cross-sectional imaging using CT and MRI². Unlike most cancers, imaging can be used to establish

the definitive diagnosis of HCC noninvasively, and treatment (including major surgical options such as hepatic resection and liver transplant) is usually initiated without confirmatory biopsy³. HCC surveillance programs aim to provide timely diagnosis and detect early preclinical stage of disease where cure is possible with locoregional therapies or liver transplantation, which ultimately reduces mortality⁴.

Although ultrasonography (US) is identified as the most cost-effective tool for HCC screening, MRI is more sensitive and has a specificity of > 90 percent. Thus, in cirrhotic and other high-risk patients (e.g., Asian and African American origin, obese, diabetic, alcohol addicted, hepatitis infection) MRI/CT are frequently used for screening. Patients frequently receive screening studies across multiple modalities based on various institution protocols, resulting in a mix of US, CT and MRI images when reviewing a longitudinal patient record. Several radiology societies have developed structured radiology report templates and standardized coding systems to promote secondary use of the large amounts of unstructured data that exists in radiology departments⁵. One such effort is the Liver Imaging Reporting and Data System (LI-RADS)⁴ developed by the American College of Radiology (ACR) and adopted as the standard coding system for HCC screening in 2018. Despite these efforts, adoption of structured reports remains low⁶, and simple tasks like differentiating between benign and malignant exams from a large number of screening cases, requires hundreds of hours of manual review by the radiologist. In our institution, which is a large multi-site academic center, structured reports coded with LI-RADS were adopted in 2018, representing only a small fraction of available HCC screening data.

Traditional natural language processing (NLP) methods (rule-based, dictionary-based, supervised learning) have been successfully applied to radiology reports for extracting information or performing automatic report classification^{7,8}. NLP techniques have also been explored for information extraction from various liver studies for colorectal adenoma⁹, inflammatory bowel disease¹⁰, fatty liver disease¹¹, aortic aneurysm⁸, and liver steatosis¹². However, a major limitation of traditional NLP techniques is the requirement of large-scale human-labeled data

or explicit hand-written linguistic rules that limit both scalability and generalizability of the system. Moreover, the use of experts to perform required labeling is expensive, and hence medical data annotation is performed on a small cohort of studies. We previously developed a word embedding-based NLP pipeline¹³ for inferring LI-RADS coding by considering only the liver findings written in the narrative US reports where only a minimal amount of hand-curated data is required for training.

We have expanded our previous work on word embedding for US LI-RADS^{13, 14, 15} to a new institution and a new modality (MRI) coding to build a system that will automatically label a large number of imaging exams to generate a large-scale annotated MRI dataset (benign and malignant). Our approach leverages the benefits of unsupervised learning (word2vec¹⁶) along with expert-knowledge to tackle the major challenges of information extraction from radiology texts, which include ambiguity of free text narrative, lexical variations, use of ungrammatical and telegraphic phrases, arbitrary ordering of words, and frequent appearance of abbreviations and acronyms. This method transforms the free-text radiology reports into dense vectors that help unlock rich source of information for computerized parsing. Distributional semantics models (word2vec) are usually trained on more than millions of similar articles/reports. In comparison, we had a limited number of MRI reports for HCC screening available at our institution. In order to handle the data scarcity, we use *transfer learning with similar domain adaptation* where we fine-tune a language model (LM) which was originally trained on HCC screening US reports, and subsequently applied on the MRI reports. Our hypothesis is that modality shifted similar domain adaptation, even if fine-tuned on a limited data set, is able to capture semantics of radiology better than other LMs trained on completely different domains. This hypothesis is built on the premise that radiologic term representations learned for one task can be directly useful for another task if the domains are analogous. Using this approach allows for curation of large scale datasets for machine learning combining both image/pixel and text information. In this paper, we prove our hypothesis by fine-tuning a language model (LM)

Table 1: Statistics of the cohorts before processing - Stanford US dataset and EUH MRI dataset.

	Stanford US dataset	EUH MRI dataset
Cohort level		
Number of unique words	17194	19828
Common words in two domains	1790	
Report level		
Average number of words (+/- std)	167 (+/- 39)	197 (+/-47)
Average number of sentences (+/- std)	27 (+/-7)	32 ((+/-8)

trained on US reports from Stanford Health Care (SHC) to infer the labels from MR studies ($n = 10,000$, retrospective) performed at Emory University Healthcare (EUH). To demonstrate the benefit of similar domain adaptation, we compare the performance of the fine-tuned LM (from EUH) with the directly transferred US language model from SHC and a second comparison of the EUH LM with the LM initially trained on 3 million Google news articles.

Method

Figure 1 presents our experimentation pipeline where we are using 3 different cohorts to train the LMs- (1) SHC ultrasound reports cohort of HCC screening to initially train a LM; (2) Google news articles to initially train a LM; (3) EUH MRI reports of HCC screening to fine-tuned the LMs. On top of the LM, we leveraged the semi-structured EUH MRI reports (recent) to train the discriminative classifier for identifying the labels and used both semi-structured and unstructured (free-text) EUH MRI reports to validate the performance. In order to explain complexity of the task, we present a sample structured and unstructured MRI report in Figure 2.

Dataset

Stanford US dataset: With the approval of Stanford Institutional Review Board (IRB), we collected all the free-text radiology reports of abdominal ultrasound examinations performed at Stanford Hospital between August 2007 to December

Figure 1: Flow diagram of language model transfer from US to MRI and GoogleNews to MRI

2017. In total, there were 92,848 US reports collected over 10 years, with an average 9,000 US exams performed every year. Among them, 13,860 exams were performed for HCC screening. A total 1,744 abdominal US studies were coded with the US LI-RADS reporting format.

EUH MRI dataset: With the approval of Emory University IRB, we retrieved 10,018 MRI exams performed between 2010 - 2019 at EUH for HCC screening. Among these, only 548 studies were reported using the LI-RADS structured template where a unique LI-RADS score was reported in the impression section (Figure 2.a.). From the LI-RADS coded reports, 99% were malignant cases (LR score >2) since benign cases are often not coded with LI-RADS. 9,470 MRI abdomen exams were documented as free-text narratives where the final diagnosis was recorded without following any structured scoring schema (Figure 2.b.).

We developed simple regex expressions to extract the LI-RADS codes from the impression section of the EUH semi-structured MRI reports. To obtain a representative sample of benign cases (which represented 1% of the LI-RADS coded reports), two radiologists manually annotated 537 benign cases from the

EUH MRI dataset. We selected benign cases from reports performed after 2018 in order to match the report structure of annotated malignant cases.

Figure 2: Sample MR reports: (a) Sample with LI-RADS structured template, (b) Sample free-text report

To generate a large MRI dataset of HCC screening, the aim of the study was to label the 9,470 MRI abdomen free-text reports obtained between 2010 – 2017 without the LI-RADS formatted template. However, ground truth labels were missing for these reports since LI-RADS coding was not adopted before 2018 in our institution. To generate the test set, two radiology residents supervised by a board-certified radiologist independently assigned labels to a set of randomly selected 112 exams by analyzing only the free-text reports.

In Table 1, we present the synopsis of the Stanford US dataset and EUH MRI dataset according to report-level and word-level statistics which reflects a slight diversity between the style of reporting. For instance, the number of words in the reports ranged from 24 to 331 in the US dataset while for MR dataset it varies from 11 to 2116. Same observation holds for the number of sentences in the reports. It is also interesting to note here that 1790 common words between MR and US.

Report pre-processing

Segmentation - We design a python-based liver section segmentation module that works with both MRI/CT and US reports. The model uses a combination of regular expressions and dictionary based sentence retrieval using anatomical vocabularies derived from Foundational Model of Anatomy (FMA)¹⁷ to extract only findings related to liver and its sub-regions from the whole report. The module maintains dependencies between anatomical entities (e.g. all the lesions found within the liver would be extracted even if they are not described in the same paragraph). This segmentation module has been manually validated on randomly selected 100 MRI and 100 US reports and obtained perfect accuracy for segmenting the liver section and liver related statements from both recent (more structured) and historic radiology reports. In order to perform a valid experiment from the LI-RADS formatted US and MRI reports, we excluded the Impression section of the reports since the final LI-RADS assessment category is often reported explicitly in the Impression. The Findings section includes only the imaging characteristics of the liver abnormalities; thus, it does not contain any clear-cut definition of the LI-RADS final assessment classes.

Root term mapping - We used a dictionary-style controlled-term mapping technique, where publicly available CLEVER terminology was used to replace common analogies/synonyms to create more semantically structured texts. As shown in our previous studies^{15,14,13}, this strategy helps to reduce the number of unique words in the vocabulary and optimizes the learning for the distributional semantics model. We focused on the terms that described family, progress, risk, negation, and punctuations, and normalized them using the formal terms derived from the terminology. For instance, ‘mother’, ‘brother’, ‘wife’ .. → ‘FAMILY’, ‘no’, ‘absent’, ‘adequate to rule her out’ .. → ‘NEGEX’, ‘suspicion’, ‘probable’, ‘possible’ → ‘RISK’, ‘increase’, ‘invasive’, ‘diffuse’, .. → ‘QUAL’.

Text cleaning - We normalize the text by converting in to lowercase letters, removing general stop words (e.g. ‘a’, ‘an’, ‘are’, ‘be’, ‘by’, ‘has’, ‘he’, etc.), removing words with very low frequency (< 50). We also removed unwanted

terms/phrases (e.g. medicolegal phrases such as “I have personally reviewed the images for this examination”); these words generally appear in all or very few reports, and are thus of little to no value in document classification. We used the Natural Language Tool Kit (NLTK) library¹⁸ for determining the stop-word list and discarded these words during report indexing. We also discarded radiologist, clinician, and patient-identifying details from the reports. Date and time stamps were replaced by ‘<datetoken>’ phase.

Initial training with Stanford US reports

In our previous study¹³, we leveraged 92,848 US reports collected over 10 years at Stanford hospital to train a word2vec model to learn the similarity between various terms used by different radiologists for reporting the liver US findings across a large time window in an unsupervised way. The primary goal was to successfully map the vocabulary used before (old reports) and after LI-RADS template implementation (new reports). The word2vec model adopts distributional semantics to learn dense vector representations of all words in the pre-processed corpus by analyzing the context. We trained the model on a pre-processed corpus where we extracted the liver section and the model learnt the representation of 2000 unique words. The model demonstrated encouraging performance for learning a LI-RADS vocabulary (semantic dictionary) using word similarity clustering. The ensemble machine-learning model trained on top of the embedding was also able to infer the three LI-RADS final assessment categories from the free-text US reports and achieve performance comparable with human raters¹³.

Fine-tuning with MRI reports

In the current study, given the limited number of MRI exams (10,018) available to train a language model from scratch, we transferred the language model trained on US reports from Stanford hospital to MRI corpus at EUH. Although both MRI and US LI-RADS exams were targeted to provide a standard diagnostic algorithm and reporting guidelines for HCC, the vocabulary between these

two modalities and scoring criteria differs significantly. For instance, ancillary imaging features (favouring malignancy) are only documented in MRI exams but never appear in US while echogenicity which is an important characterization of lesion in US, cannot be visualized in MRI. We followed three different strategies for the language model transfer - (1) direct transfer - US LM was used to create embedding of the MRI reports. The words which were present in MRI space but not in US, were replaced by zero vector; (2) fine-tuning US LM on the MRI reports - US model fine-tuned with pre-processed text from EUH MRI reports since the language in US and MR reports are not drastically different, even when obtained from different hospitals; (3) fine-tuning Google News LM on the MRI reports - LM trained on 3 millions Google News articles fine-tuned with the pre-processed text from EUH MRI reports.

Mathematically, our objective in fine-tuning is that given the source US domain D_{US} and the corresponding task of vectorization of words T_{US} , learn the probability distribution $P(Y_{MR}|X_{MR})$ in the targeted MR domain D_{MR} with the knowledge learnt from D_{US} and T_{US} where Y_{MR} is the vector representation of words present in the MR vocabulary and X_{MR} is the EUH MR reports. Note that even if both reports are from abdomen and liver disease mostly, the condition between the source D_{US} and targeted D_{MR} domain varies in several ways - (1) words used in the US and MRI domain are not same, i.e. $X_{US} \neq X_{MR}$; (2) marginal probability distribution is also different between US and MRI domain, i.e. $P(X_{MR}) \neq P(X_{US})$.

We hypothesized that fine-tuning an existing US pre-trained language model trained on a large US dataset by continuously training it (i.e. running back-propagation) on the smaller MRI dataset should provide a more semantically meaningful network than training on the smaller dataset from scratch. Since our limited number of MRI reports did not allow us to completely retrain the model, we fine-tuned the US LM by freezing the weights for the words that were present in the US report corpus from Stanford. Using this approach, we only learnt the new words' representation from the EUH MRI report corpus which in theory is a better solution than fine-tuning a LM trained on Google News where the

vocabulary between Google News and MRI have limited overlap. We learnt the vocabulary from the EUH MRI reports by setting a minimum word frequency threshold. Thereafter, we merged an input-hidden weight matrix loaded from the US trained model, where it intersected with the current vocabulary. No extra words were added to the MRI vocabulary, but intersecting words adopted the US weights, and non-intersecting words were left alone to be solely learnt from the MRI corpus.

Train discriminative model

For training the discriminative model to classify the reports, the liver section embedding was created by averaging the word vectors generated via by the models. According to earlier studies with generic text handling¹⁹ and also from our previous experience with radiology report classification tasks^{13, 14, 15}, averaging the embeddings of words in a sentence/paragraph has been proven to be successful to obtain embeddings, where each section vector is computed as: $V_{doc} = \frac{1}{N} \sum_{w \in doc} v_w$, where v_w is vector representation of the word w , N is the total number of word present in the document doc and V_{doc} is the vector representation of the report. The dense and relatively lower dimensional numerical representation of the texts V_{doc} is used as a feature vector to efficiently train a supervised classifier that can leverage the labeled dataset to learn the difference between benign and malignant reports. RandomForest model²⁰, a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and then aggregates voting from the weak learners to improve the predictive accuracy and control over-fitting, was selected as the discriminative classifier. In all cases, we applied RandomForest model with default hyperparameters (number of estimators = 100, minimum sample split = 5).

Statistical evaluation

We trained the machine learning models on the EUH structured MRI reports (*Training*) and tested the performance on the EUH historic free-text test set (*Test₁*) as well as on a subset of structured reports (*Test₂*). Among 944

Table 2: Distribution of targeted labels in training and testing set. Number of reports and % within each set has been reported.

	Training - Structured reports (after 2018)	Test ₁ - Structured reports (after 2018)	Test ₂ - Unstructured reports (2010 - 2018)
Malignant	341 (45.16%)	64 (33.86%)	21 (18.75%)
Benign	414 (54.83%)	125 (66.13%)	91 (81.25%)
Total	755	189	112

EUH structured reports with ground truth label (Figure 2.a), we randomly selected 20% of the reports as our test dataset (189 reports) and the remaining reports (755 reports) were used to train our machine-learning model. We also evaluated our model which had been trained only on the structured reports, on a subset of randomly selected free-text reports (112 reports) documented without any structured reporting template annotated manually (Figure 2.b). The distribution of the targeted labels in our *Training*, *Test₁* and *Test₂* is shown in Table 2. In order to present a straight comparison between direct transfer of the US model (US language model) and fine-tuned models on the MRI cohort, we evaluated the models using the same discriminative classifier on both sets of structured (*Test₁*) and free-text reports (*Test₂*).

Result

Interrater reliability was estimated as Cohen’s Kappa Score and the raters were highly consistent (kappa score 0.94) for identifying benign versus malignant cases by reading only the reports. A senior radiologist resolved all conflicting cases manually for preparing the ground truth labels. Table 3 presents the comparative performance in terms of class-wise precision, recall and f1-score along with number of evaluated cases.

Figure 3 presents the visualization of dimension reduced language space (PCA reduced dimension from 300 $\rightarrow$ 2) learnt from only US reports (a), fine-tuned US language space on the MRI report corpus (b), and also the overlapping language spaces (c). Each point in the spaces represents unique word in the vocabulary. As seen from the overlapping language spaces (zoomed representation

Figure 3: Represent distinct language spaces after dimension reduction using TSNE

Figure 3.d), fine-tuning strategy on a small subset of MRI reports was not only able to learn the new words in the MRI domain but was also able to preserve the overall semantics of the US space and learnt correlation between words in two different domains. For instance, ‘hyperintense’ is a word from MRI space but one of its synonyms is ‘bright’ learnt from the US corpus.

The classifier running on report vectors generated by the fine-tuned US LM outperformed the directly transferred LM on all cases. When tested on the LI-RADS structured reports, the fine-tuned US model achieved f1-score of 0.95 for identifying malignant cases while the model with direct transfer scored 0.91. For benign cases, the fine-tuned US model achieved 0.98 f1-score while the model with direct transfer scored 0.96. The performance on structured reports improved with fine-tuning the language space but direct transfer model also

Table 3: Performance of the NLP models on $Test_1$ and $Test_2$: ‘US language model’ - direct transfer of the US model, ‘Fine-Tuned model GoogleNews Language model’ - LM trained on google news and fine-tuned on MRI, and ‘Fine-Tuned model US Language model’ - LM trained on US reports and fine-tuned on MRI,

	Report document with template				Report document without template			
	Precision	Recall	f1-score	No. of cases	Precision	Recall	f1-score	No. of cases
	US language model							
<i>Malignant</i>	0.91	0.92	0.91	64	0.68	0.62	0.65	21
<i>Benign</i>	0.96	0.95	0.96	125	0.91	0.93	0.92	91
	Fine-Tuned GoogleNews language model							
<i>Malignant</i>	0.87	0.91	0.89	64	0.65	0.62	0.63	21
<i>Benign</i>	0.95	0.93	0.94	125	0.91	0.92	0.92	91
	Fine-Tuned US language model							
<i>Malignant</i>	0.98	0.92	0.95	64	0.78	0.67	0.72	21
<i>Benign</i>	0.96	0.99	0.98	125	0.93	0.96	0.94	91

resulted in high accuracy for identifying both benign and malignant cases.

Fine-tuned US LM outperformed fine-tuned Google news LM on the benign and malignant categories. On the structured reports, the fine-tuned Google news model achieved 0.89 f1-score for identifying malignant cases while the model with direct transfer scored 0.95. For benign cases, the fine-tuned google news model achieved 0.94 f1-score while the fine-tuned US model with direct transfer scored 0.98.

The major improvement while fine-tuning the language space is observed during labeling of the free-text dataset which is a more challenging task for a classifier that has only been trained on the structured reports. For example, while labeling malignant cases, the classifier with fine-tuned US LM improved the f1-score from 0.65 to 0.72 by correctly classifying non-trivial cases where the other model failed, e.g. cases with no HCC but other abnormality. The performance on benign cases also improved with fine-tuning the US LM from 0.92 to 0.94. Interestingly, fine-tuning the google news LM only achieved 0.63 f1-score on malignant cases and performance stayed the same as of direct transfer of the US model on the benign cases. Our results demonstrate that fine-tuning LM on similar domain helps the LM to converge fast on a small dataset.

Discussion

The study reports a successful transfer of language models from one domain (US) to a similar domain (MRI) and compares it to not performing adaptation and to adapting a more generic model (trained on Google News). The core novelty of this whole pipeline is that only a limited domain-specific corpus is required to train the word-embedding model while performing the LM transfer. We showed that fine-tuning of the word-embedding model with similar domain adaptation (US $\rightarrow$ MRI) provides more opportunity for semantic knowledge preservation for down-stream classification tasks compared to fine-tuning a completely different domain model (Google news $\rightarrow$ MRI radiology reports).

Our study provides a way to train language space model with limited number of reports by using cross-domain transfer. Our proposed pipeline automatically extracted binary labels (benign/malignant) for MRI imaging dataset for HCC screening with high accuracy. Our proposed framework only need a small subset of labeled MR exams to train the discriminative model on top of the dense vector embedding to generate a large dataset of labeled MR exams. The labeled MR studies may be used for AI development and training, and such automated NLP methods will rapidly reduce the manual workload for creating labeled imaging datasets from hospital databases.

This study lacks experimentation with the sensitivity for word order that limits the ability of learning long term and rotated scope of negex term. However, thanks to the semantic mapping and word occurrence analysis in our study, the adjacent negex terms are concatenated with the targeted entity in the pre-processed text before Word2Vec training, e.g. ‘negex_abnormal’. As a result, the model had the opportunity to learn the representation of the entity and the adjacent negation of the entity without explicitly considering the sentence boundary. In future, we plan to collect a larger multi-institutional radiology reports corpus and perform experimentation with advanced Bert (Bidirectional Encoder Representations from Transformer) sequential embedding model.

Acknowledgement

The study is supported by funding received from Winship Cancer Institute, and partially supported by ‘Stanford NLP collaboration’ funding.

References

- [1] F. X. Bosch, J. Ribes, M. Díaz, R. Cléries, Primary liver cancer: worldwide incidence and trends, *Gastroenterology* 127 (5) (2004) S5–S16.
- [2] M. Sherman, Hepatocellular carcinoma: epidemiology, surveillance, and diagnosis, in: *Seminars in liver disease*, Vol. 30, © Thieme Medical Publishers, 2010, pp. 003–016.
- [3] J. A. Marrero, L. M. Kulik, C. B. Sirlin, A. X. Zhu, R. S. Finn, M. M. Abecassis, L. R. Roberts, J. K. Heimbach, Diagnosis, staging, and management of hepatocellular carcinoma: 2018 practice guidance by the american association for the study of liver diseases, *Clinical Liver Disease* 13 (1) (2019) 1.
- [4] K. M. Elsayes, J. C. Hooker, M. M. Agrons, A. Z. Kielar, A. Tang, K. J. Fowler, V. Chernyak, M. R. Bashir, Y. Kono, R. K. Do, et al., 2017 version of li-rads for ct and mr imaging: an update, *Radiographics* 37 (7) (2017) 1994–2017.
- [5] T. A. Morgan, M. E. Helibrun, C. E. Kahn Jr, Reporting initiative of the radiological society of north america: progress and new directions (2014).
- [6] D. M. Rocha, L. M. Brasil, J. M. Lamas, G. V. Luz, S. S. Bacelar, Evidence of the benefits, advantages and potentialities of the structured radiological report: an integrative review, *Artificial Intelligence in Medicine* (2019) 101770.
- [7] S. Bozkurt, J. A. Lipson, U. Senol, D. L. Rubin, Automatic abstraction of imaging observations with their characteristics from mammography reports, *Journal of the American Medical Informatics Association* 22 (e1) (2015) e81–e92.
- [8] C. Morioka, F. Meng, R. Taira, J. Sayre, P. Zimmerman, D. Ishimitsu, J. Huang, L. Shen, S. El-Saden, Automatic classification of ultrasound screening examinations of the abdominal aorta, *Journal of digital imaging* 29 (6) (2016) 742–748.

- [9] T. D. Imler, J. Morea, C. Kahi, J. Cardwell, C. S. Johnson, H. Xu, D. Ahnen, F. Antaki, C. Ashley, G. Baffy, et al., Multi-center colonoscopy quality measurement utilizing natural language processing, *American Journal of Gastroenterology* 110 (4) (2015) 543–552.
- [10] J. K. Hou, M. Chang, T. Nguyen, J. R. Kramer, P. Richardson, S. Sansgiry, L. W. D’Avolio, H. B. El-Serag, Automated identification of surveillance colonoscopy in inflammatory bowel disease using natural language processing, *Digestive diseases and sciences* 58 (4) (2013) 936–941.
- [11] J. S. Redman, Y. Natarajan, J. K. Hou, J. Wang, M. Hanif, H. Feng, J. R. Kramer, R. Desiderio, H. Xu, H. B. El-Serag, et al., Accurate identification of fatty liver disease in data warehouse utilizing natural language processing, *Digestive diseases and sciences* 62 (10) (2017) 2713–2718.
- [12] E. K. Chang, C. Y. Yu, R. Clarke, A. Hackbarth, T. Sanders, E. Esrailian, D. W. Hommes, B. A. Runyon, Defining a patient population with cirrhosis, *Journal of clinical gastroenterology* 50 (10) (2016) 889–894.
- [13] I. Banerjee, H. H. Choi, T. Desser, D. L. Rubin, A scalable machine learning approach for inferring probabilistic us-li-rads categorization, in: *AMIA Annual Symposium Proceedings*, Vol. 2018, American Medical Informatics Association, 2018, p. 215.
- [14] I. Banerjee, S. Bozkurt, E. Alkim, H. Sagreiya, A. W. Kurian, D. L. Rubin, Automatic inference of bi-rads final assessment categories from narrative mammography report findings, *Journal of biomedical informatics* 92 (2019) 103137.
- [15] I. Banerjee, M. C. Chen, M. P. Lungren, D. L. Rubin, Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest ct cohort, *Journal of biomedical informatics* 77 (2018) 11–20.
- [16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: *Advances in neural information processing systems*, 2013, pp. 3111–3119.
- [17] C. Rosse, J. L. Mejino Jr, A reference ontology for biomedical informatics: the

foundational model of anatomy, Journal of biomedical informatics 36 (6) (2003) 478–500.

- [18] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, ” O’Reilly Media, Inc.”, 2009.
- [19] T. Kenter, A. Borisov, M. De Rijke, Siamese cbow: Optimizing word embeddings for sentence representations, arXiv preprint arXiv:1606.04640 (2016).
- [20] T. K. Ho, The random subspace method for constructing decision forests, IEEE transactions on pattern analysis and machine intelligence 20 (8) (1998) 832–844.