FasTag: automatic text classification of unstructured medical narratives

Arturo Lopez Pineda[1,*], Oliver J. Bear Don't Walk IV[1,2,*], Guhan R. Venkataraman[1,*], Ashley M. Zehnder[1,3,*], Sandeep Ayyar[1], Rodney L. Page[4], Carlos D. Bustamante[1,5], Manuel A. Rivas[1]

1. Department of Biomedical Data Science, School of Medicine, Stanford University, 1265 Welch Road, Stanford, CA 94305, USA
2. Department of Biomedical Informatics, Vagelos College of Physicians and Surgeons, Columbia University, 622 West 168th Street, New York, NY 10032, USA
3. Fauna Bio, 181 2nd Street, San Francisco, CA 94105, USA
4. Department of Clinical Sciences, College of Veterinary Medicine and Biomedical Sciences, Colorado State University, 1678 Campus Delivery, Fort Collins, CO 80523, USA
5. Chan Zuckerberg Biohub, San Francisco, 499 Illinois St, San Francisco, CA 94158, USA
* These authors contributed equally to this work

Address correspondence to:
Manuel A. Rivas, PhD
Department of Biomedical Data Science
Stanford University
1265 Welch Road, MSOB X321, 94305
Stanford, California, United States
Email: mrivas@stanford.edu
Telephone: 650-724-6077

Keywords: Clinical Coding; Electronic Health Records; Machine Learning; Neural Networks (Computer); One Health; Public Health Informatics
Word count: 3,932 (excluding title page, abstract, references, figures and tables)

ABSTRACT

Objective: Unstructured clinical narratives are continuously being recorded as part of delivery of care in electronic health records, and dedicated tagging staff spend considerable effort manually assigning clinical codes for billing purposes; despite these efforts, label availability and accuracy are both suboptimal.

Materials and Methods: In this retrospective study, we trained long short-term memory (LSTM) recurrent neural networks (RNNs) on 52,722 human and 89,591 veterinary records. We investigated the accuracy of both separate-domain and combined-domain models and probed model portability. We established relevant baselines by training Decision Trees (DT) and Random Forests (RF), and using MetaMap Lite, a clinical natural language processing tool.

Results: We show that the LSTM-RNNs accurately classify veterinary and human text narratives into top-level categories with an average weighted macro F_1 score of 0.74 and 0.68 respectively. In the "neoplasia" category, the model built with veterinary data has a high accuracy in veterinary data, and moderate accuracy in human data, with F_1 scores of 0.91 and 0.70 respectively. Our LSTM method scored slightly higher than that of the DT and RF models.

Discussion: The use of LSTM-RNN models represents a scalable structure that could prove useful in cohort identification for comparative oncology studies.

Conclusion: Digitization of human and veterinary health information will continue to be a reality, particularly in the form of unstructured narratives. Our approach is a step forward for these two domains to learn from, and inform, one another.

1. OBJECTIVE

The increasing worldwide adoption of electronic health records (EHRs) has created numerous clinical narratives that are now stored in clinical databases. However, given the nature of the medical enterprise, a big portion of the data being recorded is in the form of unstructured clinical notes. Cohort studies, a form of cross-sectional studies that sample a group of patients with similar clinical characteristics, require quality phenotype labels, oftentimes not readily available alongside these notes.

In place of such labeling, diagnostic codes are the most common surrogates to true phenotypes. In clinical practice, dedicated tagging staff assign clinical codes to diagnoses either from the International Classification of Diseases (ICD) [1] or the Systematized Nomenclature of Medicine (SNOMED) after reading patient summaries. However, this time-consuming, error-prone task leads to only 60–80% of the assigned codes reflecting actual patient diagnoses [2], misjudgment of severity of conditions, and/or omission of codes altogether. For example, the relative inaccuracy of oncological medical coding [3-6] affects the quality of cancer registries [7] and cancer prevalence calculations [8-10]. Poorly-defined cancer types and poorly-trained coding staff who overuse the “not otherwise specified” code when classifying text exacerbate the problem.

Challenges in clinical coding also exist in veterinary medicine in the United States, where neither clinicians nor medical coders regularly apply diagnosis codes to veterinary visits. There are few incentives for veterinary clinicians to annotate their records; a lack of 1) a substantial veterinary third-party payer system and 2) legislation enforcing higher standards of veterinary EHRs (the U.S. Health Information Technology for Economic and Clinical Health Act of 2009 sets standards for human EHRs) compound the problem. Billing codes are thus rarely applicable across veterinary institutions unless hospitals share the same management structure and records system; even then, hospital-specific modifications exist. Less than five academic veterinary centers of a total of thirty veterinary schools in the United States have dedicated medical coding staff to annotate records using SNOMED-CT-Vet [11], a veterinary extension of SNOMED-CT constructed by the American Animal Hospital Association (AAHA) and maintained by the Veterinary Terminology Services Laboratory at the Virginia-Maryland Regional College of Veterinary Medicine [12].

The vast majority of veterinary clinical data is stored as free-text fields with very low rates of formal data curation, making data characterization a tall order. Further increasing variance in the data, veterinary patients come from many different environments, including hospitals [13], practices [14], zoos [15], wildlife reserves [16], army facilities [17], research facilities [18], breeders, dealers, exhibitors [19], livestock farms, and ranches [20]. It is thus important that a general method, agnostic of patient environment, is able to categorize veterinary EHRs for cohort identification solely based on free-text.

Automatic text classification is an emerging field that uses a combination of tools such as human medical coding, rule-based systems queries [21], natural language processing (NLP), statistical analyses, data mining, and machine learning (ML) [22]. In a previous study [23], we have shown the feasibility of automatic annotation of veterinary clinical narratives across a broad range of diagnoses with minimal preprocessing, but further exploration is needed to probe what we can learn from human-veterinary comparisons. Automatically adding meaningful disease-related tags to human and veterinary clinical notes using the same machinery would be a huge step forward in that exploration and could facilitate cross-species findings downstream.

Said integration has the potential to improve both veterinary and human coding accuracy as well as comparative analyses across species. Comparative oncology, for example, has accelerated the development of novel human anti-cancer therapies through the study of companion animals [24], especially dogs [25-28]. The National Institute of Health recently funded a multi-center institution called the Arizona Cancer Evolution Center (ACE) that aims to integrate data from a broad array of species to understand the evolutionarily conserved basis for oncology. As this group utilizes animal clinical and pathology data to identify helpful traits like species-specific cancer resistance, they would greatly benefit from improved cohort discovery through automated record tagging.

Veterinary schools across the United States (15 out of 30) have formed partnerships with their respective medical schools in order to perform cross-species translational research within the Clinical and Translational Science Award One Health Alliance (COHA, [29]). Of these schools, only two have active programs to assign disease codes to their medical records. The data for the rest represents the very use case of automatic text classification.

2. BACKGROUND AND SIGNIFICANCE

Automatic medical text classification aims to reduce the human burden of handling unstructured narratives. These computational natural language processing (NLP) methods can be divided into two groups: a) semantic processing and subsequent ML; and b) deep learning.

Semantic processing and subsequent ML. These methods range from simple dictionary-based keyword-matching techniques and/or direct database queries to tools capable of interpreting the semantics of human language through lemmatization (removal of inflectional word endings), part-of-speech tagging, parsing, sentence breaking, word segmentation, and entity recognition [30]. Building the underlying dictionaries and manually crafting the rules that capture these diverse lexical elements both require time and domain expertise.

There is a growing interest in medical concept classification for clinical text; as such, many domain-specific semantic NLP tools (with various objectives, frameworks, licensing conditions, source code availabilities, language supports, and learning curves) have been developed for the medical setting. Such tools include MedLEE [31], MPLUS [32], MetaMap [33], KMCI [34], SPIN [35], HITEX [36], MCVS [37], ONYX [38], MedEx [39], cTAKES [40], pyConTextNLP [41], Topaz [42], TextHunter [43], NOBLE [44], and CLAMP [45]. However, there is no single NLP tool that can handle the broad problem of general medical classification. Instead, each method solves specific problems and applies its unique set of constraints.

ML downstream of the methods above requires featurization (concept extraction into columns and subsequent feature selection) in order to characterize text narratives in a machine-readable way. This can be done via term frequency-inverse document frequency (tf-idf), other vectorization techniques like Word2Vec [46], or manually curated rules. Semantic processing and downstream ML models have been shown to achieve high classification accuracy in human [47,48] and veterinary [49] free-text narratives for diseases well represented in training datasets (e.g. diabetes, influenza, and diarrhea). Additional success has been achieved in overall classification of clinical narratives with Decision Trees (DTs), Random Forests (RFs), and Support Vector Machines (SVMs) [50].

Deep learning. Deep learning (DL) methods eliminate the need of feature engineering, harmonization, or rule creation. They learn hierarchical feature representations from raw data in an end-to-end fashion, requiring significantly less domain expertise than traditional machine-learning approaches [51].

DL is quickly emerging in the literature as a viable alternative method to traditional ML for the classification of clinical narratives, even in situations where limited labeled data is available [50]. The technique can help in the recognition of a limited number of categories from biomedical text [52,53]; identify psychiatric conditions of patients based on short clinical histories [54]; and accurately classify whether or not radiology reports indicate pulmonary embolism [55,56] whilst outperforming baseline methods (e.g. RFs or DTs). Previous studies have shown the possibility of using DL to label clinical narratives with medical subspecialties [57] (e.g. cardiology or neurology) or medical conditions [58] (e.g. advanced cancer or chronic pain), outperforming concept-extraction based methods. Furthermore, the use of DL to analyze clinical narratives has also facilitated the prediction of relevant patient attributes, such as in-hospital mortality, 30-day unplanned readmission, prolonged length of stay, and final discharge diagnosis [59].

Significance: Text classification of human and veterinary medical records

Traditional NLP methods boast interpretability and flexibility but come at the steep cost of data quality control, formatting, normalization, domain knowledge, and time needed to generate meaningful heuristics (which oftentimes are not even generalizable to other datasets). Automatic text classification using deep learning is thus a logical choice to bypass these steps, classifying medical narratives from EHRs by

leveraging supervised deep learning on big data. We expect that our efforts could facilitate rapid triaging of documents and cohort identification for biosurveillance.

3. METHODS

Study Design

This retrospective cross-sectional chart review study uses medical records collected routinely as part of clinical care from two clinical settings: the veterinary teaching hospital at Colorado State University (CSU) and the Medical Information Mart for Intensive Care (MIMIC-III) from the Beth Israel Deaconess Medical Center in Boston, Massachusetts [60]. Both datasets were divided in two smaller datasets - training datasets containing 70% of the original datasets (used to build TensorFlow [61] deep learning models), and validation datasets containing 30% of the original datasets. We measured the accuracy of the models, calculating the F_1 score of each top-level disease category.

For comparison, we investigated the effect of using MetaMap [33], a NLP tool that extracts clinically-relevant terms, on the accuracy of our models. We also explored the possibility of out-of-domain generalization, testing the MIMIC-trained model on the CSU validation data and vice versa (and ran separate tests for MetaMapped versions, as well). Finally, we investigated the effect of merging the MIMIC and CSU training datasets to test the efficacy of data augmentation. Figure 1 shows a diagram of our study design. Our code to run all models can be found in a public repository (https://github.com/rivas-lab/FasTag).

Clinical Settings

Veterinary Medical Hospital at Colorado State University (CSU). This is a tertiary care referral teaching hospital with inpatient and outpatient facilities, serving all specialties of veterinary medicine. After consultation, veterinarians enter patient information into a custom-built veterinary EHR, including structured fields such as entry and discharge dates, patient signalment (species, breed, age, sex, etc.), and SNOMED-CT-Vet codes. There are also options to input free-text clinical narratives with various sections including history, assessment, diagnosis, prognosis, and medications. These records are subsequently coded; the final diagnostic codes represent single or multiple specific diagnoses or post-coordinated expressions (a combination of two or more concepts).

Medical Information Mart for Intensive Care (MIMIC-III). The Beth Israel Deaconess Medical Center is a tertiary care teaching hospital at Harvard Medical School in Boston, Massachusetts. The MIMIC-III database, a publicly available dataset which we utilize in this study, contains information on patients admitted to the critical care unit at the hospital [60]. We were interested in the free-text hospital discharge summaries in this database. These records are coded for billing purposes and have complete diagnoses

per patient (the database is publicly available, and thus represents the best possible medical coding annotation scenario for a hospital). Free-text fields in this database contain no protected health information.

Top level disease categories

Mapping between ICD and SNOMED codes is a challenging task but can promote semantic interoperability between our two domains. We used ICD top-level groups of diseases as the labels for the records that we aimed to extract. Table 1 shows the mapping between codes in ICD (versions 9 and 10) and SNOMED-CT (including the Veterinary extension), which was manually curated by two board-certified veterinarians trained in clinical coding (co-authors AZM, and RLP).

Table 1. Top-level coding mapping between ICD 9, 10, and SNOMED-CT

Top-level category	Description	ICD 9	ICD 10	SNOMED-CT
1	Infectious and parasitic diseases	001-139	A00-B99	105714009, 68843000, 78885002, 344431000009103, 338591000009108, 40733004, 17322007
2	Neoplasms	140-239	C00-D49	723976005, 399981008
3	Endocrine, nutritional and metabolic diseases, and immunity disorders	240-279	E00-E90	85828009, 414029004, 473010000, 75934005, 363246002, 2492009, 414916001, 363247006, 420134006, 362969004
4	Diseases of blood and blood-forming organs	280-289	D50-D89	271737000, 414022008, 414026006, 362970003, 11888009, 212373009, 262938004, 405538007
5	Mental disorders	290-319	F00-F99	74732009
6	Diseases of the nervous system	320-359	G00-G99	118940003, 313891000009106
7	Diseases of sense organs	360-389	H00-H59, H60-H95	50611000119105, 87118001, 362966006, 128127008, 85972008
8	Diseases of the circulatory system	390-459	I00-I99	49601007
9	Diseases of the respiratory system	460-519	J00-J99	50043002
10	Diseases of the digestive system	520-579	K00-K93	370514003, 422400008, 53619000
11	Diseases of the genitourinary system	580-629	N00-N99	42030000

12	Complications of pregnancy, childbirth, and the puerperium	630-679	O00-O99	362972006, 173300003, 362973001
13	Diseases of the skin and subcutaneous tissue	680-709	L00-L99	404177007, 414032001, 128598002
14	Diseases of the musculoskeletal system and connective tissue	710-739	M00-M99	105969002, 928000
15	Congenital anomalies	740-759	Q00-Q99	111941005, 32895009, 66091009
16	Certain conditions originating in the perinatal period	760-779	P00-P96	414025005
17	Injury and poisoning	800-899	S00-T98	85983004, 75478009, 77434001, 417163006

Mapping of top-level categories was manually curated by two board-certified veterinarians trained in clinical coding.

Deep learning

We chose a long short-term memory (LSTM) recurrent neural network (RNN) architecture (which is able to handle variable-length sequences while using previous inputs to inform current time steps) for this multi-label text classification task [62]. The LSTM shares parameters across time steps as it unrolls, which allows it to handle sequences of variable length. In this case, these sequences are a series of word “embeddings” (created by mapping specific words to corresponding numeric vectors) from clinical narratives. Words are represented densely (rather than sparsely, as in Bag-of-Words or tf-idf models) using the Global Vectors for Word Representation (GloVe) [63] word embeddings. These embeddings learn a vector space representation of words such that words with similar contexts appear in a similar vector space, and also capture global statistical features of the training corpus.

LSTMs have proven to be flexible enough to be used in many different tasks, such as machine translation, image captioning, medication prescription, and forecasting disease diagnosis using structured data [62]. The RNN can efficiently capture sequential information and theoretically model long-range dependencies, but empirical evidence has shown this is difficult to do in practice [64]. The sequential nature of text lends itself well to LSTMs, which have memory cells that can maintain information for over multiple time steps (words) and consist of a set of gates that control when information enters and exits memory, making them an ideal candidate architecture.

We first trained the model over a variety of hyperparameters for the model trained on MIMIC data and calculated the model’s validation accuracy over all combinations of them, finding the set of [learning rate = 0.001, dropout rate = 0.5, batch size = 256, training epochs = 100, hidden layer size = 200, LSTM layers =

1] to be the optimal setting. We proceeded to use this hyperparameter set for all of the models trained, assuming that this set would be amenable to the task at hand regardless of training dataset. We then proceeded to train a set of six models on the MIMIC, CSU, and MIMIC+CSU data; one each in which MetaMap was used to map terms back to UMLS terms, and one each in which MetaMap was not. We finally determined F_1 scores on the corresponding validation sets for each of these models.

Baseline classifier comparisons

A combination of several NLP and ML models have similarly aimed to classify clinical narratives [50,65]. We selected two of these classifiers: DTs and RFs. DTs are ML models constructed around a branching boolean logic [66]. Each node in the tree can take a decision that leads to other nodes in a tree structure; there are no cycles allowed. The RF classifier is an ensemble of multiple decision trees created by randomly selecting samples of the training data. The final prediction is done via a consensus voting mechanism of the trees in the forest.

We featurized the narratives using tf-idf, a statistic that reflects word importance in the context of other documents in a corpus and a standard ML modeling strategy for representing text, to convert the narratives into a tabular format [50]. The hyperparameters of both baseline models, like the LSTM, were tuned on the validation set.

We used MetaMap Lite [67], a NLP tool which leverages the Unified Medical Language System (UMLS) Metathesaurus to identify SNOMED [68] or ICD [69] codes from clinical narratives. MetaMap's algorithm includes five steps: 1) parsing of text into simple noun phrases; 2) variant generation of phrases to include all derivations of words (i.e. synonyms, acronyms, meaningful spelling variants, combinations, etc.); 3) candidate retrieval of all UMLS strings that contains at least one variant from the previous step; 4) evaluation and ranking of each candidate, mapping between matched term and the Metathesaurus concept using metrics of centrality, variation, coverage, and cohesiveness; 5) construction of complete mappings to include those mappings that are involved in disjointed parts of the phrase (e.g. 'ocular' and 'complication' can together be mapped to a single term, 'ocular complication'). MetaMap incorporates the use of ConText [70], an algorithm for the identification of negation in clinical narratives. For additional information on how we used and evaluated MetaMap, please refer to Supplementary Material 1.

Statistical analysis

Evaluation metric. For all models we trained (LSTM, DT, and RF), we used the same evaluation metrics previously reported for MetaMap Lite [67]: a) precision, defined as the proportion of documents which were assigned the correct category; b) recall, defined as the proportion of documents from a given category that were correctly identified; and c) F_1 score, defined as the harmonic average of precision and recall. Formulas for these metrics are provided below:

$$Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive} \quad \text{Eq. 1}$$

$$Recall = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative} \quad \text{Eq. 2}$$

$$F_1 = 2 \cdot \frac{precision\ \cdot\ recall}{precision\ +\ recall} \quad \text{Eq. 3}$$

Our task is framed as a multi-label classification problem, where each approach predicts multiple top-level categories for each observation using a single model. In order to combine all class-specific F_1 scores, we averaged the F_1 score for each label, weighting the labels by their supports (the number of true instances for each label, to account for label imbalance).

Domain adaptation. The portability of trained algorithms on independent domains has previously been used as a metric of model robustness in systems that leverage NLP and machine learning [71]. We evaluated the ability of our trained LSTM models to be used in a cross-species context. We utilized the MIMIC-trained model to classify the medical records in the CSU database, and vice versa, assessing performance as before. We also assess the classifier trained on the combined training sets.

4. RESULTS

We investigated the application of deep learning to free-text unstructured clinical narratives on two cohorts: veterinary medical records from CSU, and human medical records in the MIMIC-III database. We show the evaluation of the deep learning models built using human and veterinary records, as well as the portability between them.

Patients

The CSU dataset contains medical records from 33,124 patients and 89,591 hospital visits between February 2007 and July 2017. Patients encompassed seven mammalian species, including dogs (*Canis Lupus*, 80.8%), cats (*Felis Silvestris*, 11.4%), horses (*Equus Caballus*, 6.5%), cattle (*Bos Taurus*, 0.7%), pigs (*Sus Scrofa,* 0.3%), goats (*Capra hircus*, 0.2%), sheep (*Ovis Aries*, 0.1%), and other unspecified mammals (0.1%). In contrast, the MIMIC-III database contains medical records from 38,597 distinct human adult patients (aged 16 years or above) and 7,870 neonates admitted between 2001 and 2008, encompassing 52,722 unique hospital admissions to the critical care unit between 2001 and 2012. Table 2

summarizes the category breakdowns of both databases. Only those patients with a diagnosis in their record were considered.

Table 2. Database statistics of patients, records, and species (records with diagnosis).

	CSU	**MIMIC**
Medical Records	N = 89,591	N = 52,722
Patients	33,124	41,126
Hospital Visits	89,591	49,785
Species		
Humans (*Homo Sapiens*)	n.a.	52,722
Dogs (*Canis Lupus*)	72,420	n.a.
Cats (*Felis Silvestris*)	10,205	n.a.
Horses (*Equus Caballus*)	5,819	n.a.
Other mammals	1,147	n.a.
Category		
1. Infectious	11,454	10,074
2. Neoplasia	36,108	6,223
3. Endo-Immune	17,295	24,762
4. Blood	10,171	13,481
5. Mental	511	10,989

6. Nervous	7,488	9,168
7. Sense organs	15,085	2,688
8. Circulatory	8,733	30,054
9. Respiratory	11,322	17,667
10. Digestive	22,776	14,646
11. Genitourinary	8,892	14,932
12. Pregnancy	136	133
13. Skin	21,147	4,241
14. Musculoskeletal	22,921	6,739
15. Congenital	3,347	2,334
16. Perinatal	54	3,661
17. Injury	9,873	16,121

The mappings in Table 1 were used to generate the categories and numbers presented here in Table 2. The seventeen categories represent the text classification labels.

Evaluation of Deep Learning models

We trained deep-learning models (as well as DT/RF baselines) for the human, veterinary, and merged (human and veterinary) datasets and tested each on their own domain, as well as the other domains. We built our LSTM model using the Python programming language (version 2.7), TensorFlow [61] (version 1.9), and the 'scikit-learn' library (version 0.19.2) [72]. The training was performed on an Amazon® Deep Learning AMI, a cloud-based platform running the Ubuntu operating system with pre-installed CUDA dependencies. Average weighted macro F_1 scores for models across all categories are shown in Table 3; a full list of F_1 scores by category can be found in Supplementary Material 1. The "neoplasia" category results, which we found interesting, are shown in Table 4.

Table 3. Average F_1 scores using various training and validation dataset combinations for all categories.

Configuration			Model evaluation (Weighted F_1 score)		
Training	**Validation**	**MetaMap**	**DT**	**RF**	**LSTM**
MIMIC	MIMIC	No	0.60	0.64	**0.65**
		Yes	0.60	0.63	**0.70**
CSU	CSU	No	0.55	0.61	**0.72**
		Yes	0.54	0.60	**0.75**
MIMIC	CSU	No	0.22	0.24	**0.28**
		Yes	0.23	0.20	**0.31**
CSU	MIMIC	No	**0.31**	0.20	0.23
		Yes	0.28	0.19	**0.36**
MIMIC+CSU	CSU	No	0.57	0.62	**0.67**
		Yes	0.57	0.62	**0.76**
MIMIC+CSU	MIMIC	No	0.60	**0.63**	0.58
		Yes	0.60	**0.63**	0.60
MIMIC+CSU	MIMIC+CSU	No	0.59	0.64	**0.68**
		Yes	0.59	0.63	**0.71**
Average			0.489	0.506	**0.571**

Evaluation metrics for Decision Tree (DT), Random Forest (RF), and Long Short Term Memory Recurrent Neural Network (LSTM-RNN) on validation datasets with and without MetaMap term extraction. Bolded and underlined numbers represent the best scores for the specific configuration of training data, validation data, and MetaMap toggle.

Table 4. F_1 scores using various training and validation dataset combinations for the "neoplasia" category.

Configuration			Model evaluation (Weighted F_1 score)		
Training	**Validation**	**MetaMap**	**DT**	**RF**	**LSTM**
MIMIC	MIMIC	No	0.39	0.45	**0.66**
		Yes	0.4	0.45	**0.76**
CSU	CSU	No	0.81	0.86	**0.91**
		Yes	0.8	0.86	**0.91**
MIMIC	CSU	No	0.3	0.53	**0.69**
		Yes	0.45	0.37	**0.75**
CSU	MIMIC	No	0.46	0.58	**0.70**
		Yes	0.5	**0.58**	0.54
MIMIC+CSU	CSU	No	0.74	0.8	**0.87**
		Yes	0.74	0.8	**0.87**
MIMIC+CSU	MIMIC	No	0.4	0.47	**0.67**
		Yes	0.42	0.45	**0.72**
MIMIC+CSU	MIMIC+CSU	No	0.81	**0.86**	0.85
		Yes	0.81	0.86	**0.90**
Average					

Evaluation metrics for the "neoplasia" category Decision Tree (DT), Random Forest (RF), and Long Short Term Memory Recurrent Neural Network (LSTM-RNN) on validation datasets with and without MetaMap term extraction. Bolded and underlined numbers represent the best scores for the specific configuration of training data, validation data, and MetaMap toggle.

5. DISCUSSION

Applying deep learning to unstructured free-text clinical narratives in electronic health records offers a relatively simple, low-effort means to bypass the traditional bottlenecks in medical coding. Circumventing the need for data harmonization was very important for the datasets, which contained a plethora of domain- and setting-specific misspellings, abbreviations, and jargon (these issues would have greatly impacted the performance of the LSTM and the NLP's entity recognition). MetaMap was useful in this regard given its

ability to parse clinical data, but much work is still needed to improve recognition of terms in veterinary and human domains.

There is moderate evidence of domain adaptation in the “neoplasia” category, with F_1 scores of 0.69-0.70 (Table 4). This process involved training a model on the data in one database and testing on the data in the other, without fine-tuning. It is evident that the high classification accuracy (F_1 score = 0.91) obtained by the CSU model in the neoplasia category is decreased when testing the same model on the MIMIC data. One possible explanation is the difference in clinical settings; CSU is a tertiary care veterinary hospital specializing in oncological care, and the clinical narratives that arise in a critical care unit like the MIMIC dataset do not necessarily compare. Moreover, the records were not coded in the same way, the clinicians did not receive the same training, and the documents apply to different species altogether. Despite these differences, however, our LSTM model was general enough to be able to accurately classify medical narratives at the top level of depth independently in both datasets. The achieved cross-domain accuracy is nonetheless encouraging. Given enough training data and similar-enough clinical narratives, one could conceivably imagine a general model that is highly effective across domains.

Models performed usually better on their respective validation datasets in those categories with more training samples. For example, the CSU-trained model (25,276 samples) had significantly better performance in the “neoplasia” category than the MIMIC-trained model (4,356 samples), while the MIMIC-trained model (21,038 samples) had better performance in the diseases of the circulatory system category than the CSU-trained model (6,133 samples).

The usefulness of even top-level characterizations in the veterinary setting cannot be understated; usually, a veterinarian must read the full, unstructured text in order to get any information about the patient they are treating. Rapid selection of documents with specific types of clinical narratives (such as oncological cases, which our model performed well on) could lead to better cohort studies for comparative research. The repeated use of a series of such LSTM models for subsequent, increasingly-specific classifications thus represents a scalable, hierarchical tagging structure that could prove extremely useful in stratifying patients by specific diseases, severities, and protocols.

6. CONCLUSION

In this era of increasing deployment of EHRs, it is important to provide tools that facilitate cohort identification. Our deep learning approach (LSTM model) was able to automatically classify medical narratives with minimal human preprocessing. In a future with enough training data, it is possible to foresee a scenario in which these models can accurately tag every clinical concept, regardless of data input. The expansion of veterinary data availability and the subsequently enormous potential of domain adaptation like we saw in the neoplasia category could prove to be exciting chapters in reducing bottlenecks in public

health research at large; it is thus of critical importance to continue studying novel sources of data that can rapidly be used to augment classification models.

A reliable addition to existing rule-based and natural language processing strategies, deep learning is a promising tool for accelerating public health research.

DECLARATIONS

Data availability

Veterinary data presented here belongs to the Colorado State University, which may grant access to this data on a case-by-case basis to researchers who obtain the necessary Data Use Agreement (DUA) and IRB approvals.

Human data presented here belongs to the Beth Israel Deaconess Medical Center in Boston, Massachusetts, which can be accessed after signing a data usage agreement in the MIT Lab for Computational Physiology at https://mimic.physionet.org/

Acknowledgments

The authors wish to acknowledge Dr. Katie M. Kanagawa for her valuable support in editing this manuscript. Also, Devin Johnson, DVM, MS, for her contribution to clinical coding and comparison with coding from the MetaMap tool.

Contributions

ALP, OBDW, GRV, and AMZ designed the study. RLP provided access to the veterinary data. OJBDW, GRV, ALP, AMZ, and SA extracted, formatted, and performed analysis of the data. AMZ, RLP, CDB and MAR provided interpretation of the results. ALP drafted the manuscript, and all authors contributed critically, read, revised and approved the final version.

Competing interests

CDB is Principal and Chairman of CDB Consulting LTD. He has advised Fauna Bio, Imprimed, Embark Vet and Etalon DX as a member of their respective Scientific Advisory Boards, and is a Director of Etalon DX AMZ is the CEO of Fauna Bio, LLC.
The remaining authors declare no conflicts of interest.

Funding

M.A.R. is supported by Stanford University and a National Institute of Health center for Multi and Trans-ethnic Mapping of Mendelian and Complex Diseases grant (5U01 HG009080). This work was supported by National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) under awards R01HG010140. C.D.B. is a Chan Zuckerberg Biohub Investigator. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethics approval

This research was reviewed and approved by Stanford's Institutional Review Board (IRB), which provided a non-human subject determination under eProtocol 46979. Consent was not required.

Bibliography

1 Moriyama IMIM1-2, Loy RM, Robb-Smith AHT, *et al.* History of the statistical classification of diseases and causes of death. 2011.

2 Benesch C, Witter DM, Wilder AL, *et al.* Inaccuracy of the International Classification of Diseases (ICD-9-CM) in identifying the diagnosis of ischemic cerebrovascular disease. *Neurology* 1997;49:660–4.

3 Abraha I, Serraino D, Giovannini G, *et al.* Validity of ICD-9-CM codes for breast, lung and colorectal cancers in three Italian administrative healthcare databases: a diagnostic accuracy study protocol. *BMJ Open* 2016;6:e010547. doi:10.1136/bmjopen-2015-010547

4 Kim SC, Gillet VG, Feldman S, *et al.* Validation of claims-based algorithms for identification of high-grade cervical dysplasia and cervical cancer. *Pharmacoepidemiol Drug Saf* 2013;22:1239–44. doi:10.1002/pds.3520

5 Moar KK, Rogers SN. Impact of coding errors on departmental income: an audit of coding of microvascular free tissue transfer cases using OPCS-4 in UK. *Br J Oral Maxillofac Surg* 2012;50:85–7. doi:10.1016/j.bjoms.2011.01.005

6 Friedlin J, Overhage M, Al-Haddad MA, *et al.* Comparing methods for identifying pancreatic cancer patients using electronic data sources. *AMIA Annu Symp Proc* 2010;2010:237–41.

7 German RR, Wike JM, Bauer KR, *et al.* Quality of cancer registry data: findings from CDC-NPCR's Breast and Prostate Cancer Data Quality and Patterns of Care Study. *J Registry Manag* 2011;38:75–86.

8 Trombert Paviot B, Gomez F, Olive F, *et al.* Identifying prevalent cases of breast cancer in the French case-mix databases. *Methods Inf Med* 2011;50:124–30. doi:10.3414/ME09-01-0064

9 Fisher BT, Harris T, Torp K, *et al.* Establishment of an 11-year cohort of 8733 pediatric patients hospitalized at United States free-standing children's hospitals with de novo acute

lymphoblastic leukemia from health care administrative data. *Med Care* 2014;52:e1–6. doi:10.1097/MLR.0b013e31824deff9

10 Polednak AP, Phillips C. Cancers coded as tongue not otherwise specified: relevance to surveillance of human papillomavirus-related cancers. *J Registry Manag* 2014;41:190–5.

11 Association of American Veterinary Medical Colleges. Association of American Veterinary Medical Colleges. 2019;:1–6.https://www.aavmc.org/

12 Virginia-Maryland Regional College of Veterinary Medicine. Veterinary Terminology Services Laboratory. 2019;:1–1.https://vtsl.vetmed.vt.edu

13 Cummings KJ, Rodriguez-Rivera LD, Mitchell KJ, *et al.* Salmonella enterica serovar Oranienburg outbreak in a veterinary medical teaching hospital with evidence of nosocomial and on-farm transmission. *Vector Borne Zoonotic Dis* 2014;14:496–502. doi:10.1089/vbz.2013.1467

14 Krone LM, Brown CM, Lindenmayer JM. Survey of electronic veterinary medical record adoption and use by independent small animal veterinary medical practices in Massachusetts. *J Am Vet Med Assoc* 2014;245:324–32. doi:10.2460/javma.245.3.324

15 Witte CL, Lamberski N, Rideout BA, *et al.* Development of a case definition for clinical feline herpesvirus infection in cheetahs (Acinonyx jubatus) housed in zoos. *J Zoo Wildl Med* 2013;44:634–44. doi:10.1638/2012-0183R.1

16 Griffith JE, Higgins DP. Diagnosis, treatment and outcomes for koala chlamydiosis at a rehabilitation facility (1995-2005). *Aust Vet J* 2012;90:457–63. doi:10.1111/j.1751-0813.2012.00963.x

17 Poppe JL. The US Army Veterinary Service 2020: knowledge and integrity. *US Army Med Dep J* 2013;:5–10.

18 Field K, Bailey M, Foresman LL, *et al.* Medical records for animals used in research, teaching, and testing: public statement from the American College of Laboratory Animal Medicine. *ILAR J* 2007;48:37–41.

19 Shalev M. USDA to require research facilities, dealers, and exhibitors to keep veterinary medical records. Lab Anim (NY). 2003;32:16. doi:10.1038/laban0603-16a

20 Robinson TP, Wint GRW, Conchedda G, *et al.* Mapping the global distribution of livestock. *PLoS ONE* 2014;9:e96084. doi:10.1371/journal.pone.0096084

21 Gundlapalli AV, Redd D, Gibson BS, *et al.* Maximizing clinical cohort size using free text queries. *Comput Biol Med* 2015;60:1–7. doi:10.1016/j.compbiomed.2015.01.008

22 Shivade C, Raghavan P, Fosler-Lussier E, *et al.* A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* 2014;21:221–30. doi:10.1136/amiajnl-2013-001935

23 Nie A, Zehnder A, Page RL, *et al.* DeepTag: inferring diagnoses from veterinary clinical notes. *npj Digital Medicine 2018 1:1* 2018;1:60. doi:10.1038/s41746-018-0067-8

24 Garden OA, Volk SW, Mason NJ, *et al.* Companion animals in comparative oncology: One Medicine in action. *Vet J* 2018;240:6–13. doi:10.1016/j.tvjl.2018.08.008

25 Saba C, Paoloni M, Mazcko C, *et al.* A Comparative Oncology Study of Iniparib Defines Its Pharmacokinetic Profile and Biological Activity in a Naturally-Occurring Canine Cancer Model. *PLoS ONE* 2016;11:e0149194. doi:10.1371/journal.pone.0149194

26 LeBlanc AK, Mazcko CN, Khanna C. Defining the Value of a Comparative Approach to Cancer Drug Development. *Clin Cancer Res* 2016;22:2133–8. doi:10.1158/1078-0432.CCR-15-2347

27 Burton JH, Mazcko C, LeBlanc A, *et al.* NCI Comparative Oncology Program Testing of Non-Camptothecin Indenoisoquinoline Topoisomerase I Inhibitors in Naturally Occurring Canine Lymphoma. *Clin Cancer Res* 2018;24:5830–40. doi:10.1158/1078-0432.CCR-18-1498

28 Paoloni M, Webb C, Mazcko C, *et al.* Prospective molecular profiling of canine cancers provides a clinically relevant comparative model for evaluating personalized medicine (PMed) trials. *PLoS ONE* 2014;9:e90028. doi:10.1371/journal.pone.0090028

29 Clinical Translational Science Award One Health Alliance. Clinical Translational Science Award One Health Alliance (COHA). 2019;:1–5.https://ctsaonehealthalliance.org/

30 Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc* 2011;18:544–51. doi:10.1136/amiajnl-2011-000464

31 Friedman C, Alderson PO, Austin JH, *et al.* A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc* 1994;1:161–74.

32 Christensen LM, Haug PJ, Fiszman M. MPLUS: a probabilistic medical language understanding system. *ACL Workshop on Natural Language Processing in the Biomedical Domain* 2002.

33 Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001;:17–21.

34 Denny JC, Irani PR, Wehbe FH, *et al.* The KnowledgeMap project: development of a concept-based medical school curriculum database. *AMIA Annu Symp Proc* 2003;2003:195–9.

35 Liu K, Mitchell KJ, Chapman WW, *et al.* Automating tissue bank annotation from pathology reports - comparison to a gold standard expert annotation set. *AMIA Annu Symp Proc* 2005;:460–4.

36 Zeng QT, Goryachev S, Weiss S, *et al.* Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med Inform Decis Mak* 2006;6:30. doi:10.1186/1472-6947-6-30

37 Elkin PL, Brown SH, Husser CS, *et al.* Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. *Mayo Clin Proc* 2006;81:741–8. doi:10.4065/81.6.741

38 Christensen LM, Harkema H, Haug PJ, *et al.* ONYX - A System for the Semantic Analysis of Clinical Text. *BioNLP@HLT-NAACL* 2009.

39 Xu H, Stenner SP, Doan S, *et al.* MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc* 2010;17:19–24. doi:10.1197/jamia.M3378

40 Savova GK, Masanz JJ, Ogren PV, *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17:507–13. doi:10.1136/jamia.2009.001560

41 Chapman BE, Lee S, Kang HP, *et al.* Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. *Journal of Biomedical Informatics* 2011;44:728–37. doi:10.1016/j.jbi.2011.03.011

42 Wagner M, Tsui F-C, Cooper GF, *et al.* Probabilistic, Decision-theoretic Disease Surveillance and Control. *Online J Public Health Inform* 2011;3. doi:10.5210/ojphi.v3i3.3798

43 Jackson RG, Ball M, Patel R, *et al.* TextHunter--A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research. *AMIA Annu Symp Proc* 2014;2014:729–38.

44 Tseytlin E, Mitchell K, Legowski E, *et al.* NOBLE - Flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinformatics* 2016;17:32. doi:10.1186/s12859-015-0871-y

45 Lee HJ, 0001 HX, Wang J, *et al.* UTHealth at SemEval-2016 Task 12 - an End-to-End System for Temporal Information Extraction from Clinical Notes. *SemEval@NAACL-HLT* 2016.

46 Mikolov T, Sutskever I, Chen K, *et al.* Distributed Representations of Words and Phrases and their Compositionality. arXiv. 2013;cs.CL:arXiv:1310.4546.

47 Koopman B, Karimi S, Nguyen A, *et al.* Automatic classification of diseases from free-text death certificates for real-time surveillance. *BMC Med Inform Decis Mak* 2015;15:53. doi:10.1186/s12911-015-0174-2

48 Berndorfer S, Henriksson A. Automated Diagnosis Coding with Combined Text Representations. *Stud Health Technol Inform* 2017;235:201–5.

49 Anholt RM, Berezowski J, Jamal I, *et al.* Mining free-text medical records for companion animal enteric syndrome surveillance. *Prev Vet Med* 2014;113:417–22. doi:10.1016/j.prevetmed.2014.01.017

50 Wang Y, Sohn S, Liu S, *et al.* A clinical text classification paradigm using weak supervision and deep representation. *BMC Med Inform Decis Mak* 2019;19:1. doi:10.1186/s12911-018-0723-6

51 Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press 2016.

52 Agibetov A, Blagec K, Xu H, *et al.* Fast and scalable neural embedding models for biomedical sentence classification. *BMC Bioinformatics* 2018;19:541. doi:10.1186/s12859-018-2496-4

53 Du Y, Pan Y, Wang C, *et al.* Biomedical semantic indexing by deep neural network with multi-task learning. *BMC Bioinformatics* 2018;19:502. doi:10.1186/s12859-018-2534-2

54 Tran T, Kavuluru R. Predicting mental conditions based on “history of present illness” in psychiatric notes with deep neural networks. *Journal of Biomedical Informatics* 2017;75S:S138–48. doi:10.1016/j.jbi.2017.06.010

55 Chen MC, Ball RL, Yang L, *et al.* Deep Learning to Classify Radiology Free-Text Reports. *Radiology* 2018;286:845–52. doi:10.1148/radiol.2017171115

56 Banerjee I, Ling Y, Chen MC, *et al.* Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. *Artif Intell Med* Published Online First: 23 November 2018. doi:10.1016/j.artmed.2018.11.004

57 Weng W-H, Wagholikar KB, McCray AT, *et al.* Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Med Inform Decis Mak* 2017;17:155. doi:10.1186/s12911-017-0556-8

58 Gehrmann S, Dernoncourt F, Li Y, *et al.* Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PLoS ONE* 2018;13:e0192360. doi:10.1371/journal.pone.0192360

59 Rajkomar A, Oren E, Chen K, *et al.* Scalable and accurate deep learning with electronic health records. *npj Digital Medicine 2018 1:1* 2018;:1–10. doi:10.1038/s41746-018-0029-1

60 Johnson AEW, Pollard TJ, Shen L, *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:160035. doi:10.1038/sdata.2016.35

61 Abadi M, Agarwal A, Barham P, *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv. 2016;cs.DC.

62 Pham T, Tran T, Phung D, *et al.* DeepCare: A Deep Dynamic Memory Model for Predictive Medicine. arXiv. 2016;stat.ML.

63 Pennington J, Socher R, Manning CD. Glove - Global Vectors for Word Representation. *EMNLP* 2014.

64 Pascanu R, Mikolov T, Bengio Y. On the difficulty of training Recurrent Neural Networks. arXiv. 2012;cs.LG.

65 Segura-Bedmar I, Colón-Ruíz C, Tejedor-Alonso MÁ, *et al.* Predicting of anaphylaxis in big data EMR by exploring machine learning approaches. *Journal of Biomedical Informatics* 2018;87:50–9. doi:10.1016/j.jbi.2018.09.012

66 Yu Z, Bernstam E, Cohen T, *et al.* Improving the utility of MeSH® terms using the TopicalMeSH representation. *Journal of Biomedical Informatics* 2016;61:77–86. doi:10.1016/j.jbi.2016.03.013

67 Demner-Fushman D, Rogers WJ, Aronson AR. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. *J Am Med Inform Assoc* 2017;24:841–4. doi:10.1093/jamia/ocw177

68 Barros JM, Duggan J, Rebholz-Schuhmann D. Disease mentions in airport and hospital geolocations expose dominance of news events for disease concerns. *J Biomed Semantics* 2018;9:18. doi:10.1186/s13326-018-0186-9

69 Hanauer DA, Saeed M, Zheng K, *et al.* Applying MetaMap to Medline for identifying novel associations in a large clinical dataset: a feasibility analysis. *J Am Med Inform Assoc* 2014;21:925–37. doi:10.1136/amiajnl-2014-002767

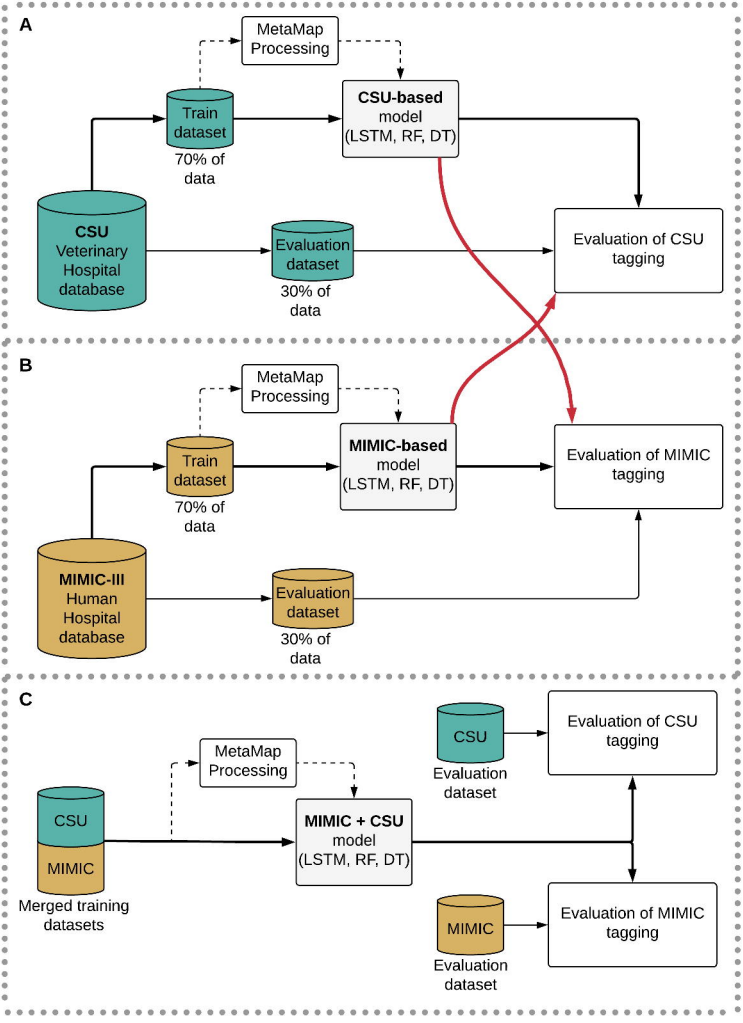
70 Harkema H, Dowling JN, Thornblade T, *et al.* ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics* 2009;42:839–51. doi:10.1016/j.jbi.2009.05.002

71 Ye Y, Wagner MM, Cooper GF, *et al.* A study of the transferability of influenza case detection systems between two large healthcare systems. *PLoS ONE* 2017;12:e0174970. doi:10.1371/journal.pone.0174970

72 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn - Machine Learning in Python. *Journal of Machine Learning Research* 2011.

Figure legends

Figure 1. Diagram of the training and evaluation design. Relevant acronyms: MIMIC: Medical Information Mart for Intensive Care; CSU: Colorado State University; MetaMap, a tool for recognizing medical concepts in text; LSTM: long-short term memory recurrent neural network classifier; RF: Random Forest classifier; DT: Decision Tree classifier


A
MetaMap Processing
Train dataset
70% of data
CSU-based model (LSTM, RF, DT)
CSU Veterinary Hospital database
Evaluation dataset
30% of data
Evaluation of CSU tagging
B
MetaMap Processing
Train dataset
70% of data
MIMIC-based model (LSTM, RF, DT)
Evaluation of MIMIC tagging
MIMIC-III Human Hospital database
Evaluation dataset
30% of data
C
CSU
Evaluation dataset
Evaluation of CSU tagging
MetaMap Processing
CSU
MIMIC
Merged training datasets
MIMIC + CSU model (LSTM, RF, DT)
MIMIC
Evaluation dataset
Evaluation of MIMIC tagging