Deep learning facilitates rapid cohort identification using human and veterinary clinical narratives

Arturo Lopez Pineda, PhD[1,§], Oliver J. Bear Don't Walk IV, MS[1,2,§], Guhan R. Venkataraman, BS[1,§], Ashley M. Zehnder, DVM, PhD, Dipl. ABVP(Avian)[1,3,§], Sandeep Ayyar, MS[1], Rodney L. Page, DVM, MS[4], Carlos D. Bustamante, PhD[1,5], Manuel A. Rivas, PhD[1,*]

1. Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA
2. Department of Biomedical Informatics. Columbia University, New York, NY 10032, USA
3. Fauna Bio, San Francisco, CA 94103, USA
4. Department of Clinical Sciences, Colorado State University, Fort Collins, CO 80523, USA
5. Chan Zuckerberg Biohub, San Francisco, CA 94158, USA

§ Authors contributed equally

* Corresponding author

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

Abstract

Background: In public health research, there is currently a need to close the gap between care delivery and cohort identification. We need dedicated tagging staff to allocate a considerable amount of effort to assigning clinical codes after reading patient summaries. Machine learning automation can facilitate the classification of these clinical narratives, but sufficient availability of electronic medical records is still a bottleneck. Veterinary medical records represent a largely untapped data source that could be used to benefit both human and non-human patients. Very few approaches utilizing veterinary data sources currently exist.

Methods: In this retrospective cross-sectional and chart review study, we trained separate long short-term memory (LSTM) Recurrent Neural Networks (RNNs) on 52,722 human records and 89,591 veterinary records, tested the models' efficacy in a standard train-test split setup, and probed the portability of these models across species domains. We trained versions of our models using first the free-text clinical narratives, and then only using extracted clinically relevant terms from MetaMap Lite, a natural language processing tool intended for this purpose.

Findings: We show that our LSTM approach correctly classifies across top-level codes in the veterinary records (F_1 score =0·83), and identifies top-level neoplasia records in veterinary records (F_1 score = 0·93). The model trained with veterinary data can be ported over to identify neoplasia records in the human records (F_1 score = 0·70).

Interpretation: Our findings suggest that free-text clinical narratives can be used to learn classification models that allow the rapid identification of patient cohorts. Ultimately, this effort can lead to new insights that can address emerging public health concerns. Digitization of health information will continue to be a reality in both human and veterinary data; our approach serves as first proof-of-concept regarding how these two domains can learn from, and inform, one another.

Funding: Stanford University & The Chan Zuckerberg Biohub Investigator Award

Keywords: Clinical Coding; Electronic Health Records; Machine Learning; Neural Networks (Computer); One Health; Public Health Informatics

Research in context

Evidence before this study. We systematically reviewed PubMed using the Mesh terms "Clinical Coding" and "Electronic Health Records", finding 50 publications in the last five years. The topics that arise from this body of literature include accuracy of clinical coding, challenges to cohort identification, estimation of disease incidence, evaluation of quality of care, and decision support systems. Furthermore, filtering our query to include the Mesh term "Machine Learning" revealed that only five studies have attempted to use automatic tagging of clinical codes from clinical narratives. However, these studies are still limited by the availability of training data, and are heavily relying on human input for data curation and harmonization.

Added value of this study. In addition to rule-based and natural language processing strategies, the use of deep learning approaches to automatically classify clinical narratives could be a promising tool in accelerating public health research. Our analysis of non-traditional data sources (e.g. veterinary medical records) suggests that it is possible to grasp models circumventing the need for data preprocessing and harmonization. It is of critical importance to continue studying novel sources of information that can rapidly be used to generate classification models.

Implications of all the available evidence.

The costs for clinical coding could be reduced by implementing systems that automatically classify medical records. These automated systems have the benefit of accelerating public health research by quickly identifying cohorts of interest. The use of veterinary data offers a promising way to facilitate the identification of human cohorts, thus exponentially increasing the availability of research data.

1. Background

Rapid identification of standardized cohorts, also known as electronic phenotyping, is an emerging field in public health that uses a combination of tools such as rule-based systems queries[1], natural language processing (NLP), statistical analyses, data mining, and machine learning[2]. Currently, significant effort is still required to close the gap between care delivery and medical coding. In clinical practice, dedicated tagging staff assign clinical codes after reading patient summaries, a time-consuming and error-prone task. It is estimated that only 60–80% of the assigned codes reflect actual patient diagnoses[3], resulting in over- and under-coding (and misjudging the severity of conditions, or omitting codes altogether).

Automatic clinical coding technologies aim to reduce human interaction with unstructured narratives while capturing the majority of the critical information captured in data in a structured format. In general, these computational methods can be divided into three groups: a) rule-based and keyword-matching; b) traditional natural language processing; and c) natural language processing with deep learning.

Rule-based and keyword-matching. These methods involve the use of single- or multiple- keyword matching from a dictionary and subsequently direct queries of the database. However, these methods require time and domain expertise to build the underlying dictionaries and manually craft rules that capture diverse lexical elements. In diseases with enough training cases (e.g. diabetes, influenza, and diarrhea), these models have been shown to achieve high classification accuracy in human[4,5] and veterinary[6] free-text narratives.

Natural language processing (NLP). NLP tools are capable of interpreting the semantics of human language through lemmatization, part-of-speech tagging, parsing, sentence breaking, word segmentation, and entity recognition. There are both general-purpose and medical NLP tools. Medical NLP tools are highly heterogeneous, with various frameworks, licensing conditions, source code availability, language support, and learning curves for implementation. These factors generally affect time-to-deployment in clinical settings.

Deep learning (DL). These methods eliminate the need of feature engineering, harmonization, or rule creation. Deep learning approaches are able to learn hierarchical feature representations from raw data in an end-to-end fashion, requiring significantly less domain expertise than traditional machine-learning approaches[7]. Deep learning is slowly emerging in the literature as a viable alternative solution to the analysis of clinical narratives. For example, deep learning has been used to identify patients with chronic conditions[8], achieving a classification accuracy equivalent to, or better than, using keyword matching or NLP approaches. The use of DL to analyze clinical narratives has also facilitated other relevant clinical tasks, such as in-hospital mortality, 30-day unplanned readmission, prolonged length of stay, and final discharge diagnosis[9].

Veterinary records: a novel source of data

Electronic veterinary medical records are being adopted at an increasing rate, with the general goal of emulating the policies of human medical records in terms of recording standardized information for patient visits. However, the vast majority of veterinary clinical data is stored as free-text fields with very low rates of formal data curation. Veterinary patients come from many different places, including hospitals[10], practices[11], zoos[12], wildlife reserves[13], army facilities[14], research facilities[15], breeders, dealers, exhibitors[16], livestock farms, and ranches[17]. It is important to recognize at this time that, with the development of new wearable and sensing technology for animals[18], the amount of data concerning animal health will continue to grow exponentially in coming years.

The integration of these new data streams with those concerning humans has the potential to improve human quality of care, directly addressing emerging public health concerns. Applications of generating these cohorts include biosurveillance for zoonotic diseases (which represent 60–70% of all emerging diseases[19]), chronic disease management, and early detection of environmental pollution factors. Such cohorts could also prove useful in elucidating disease-specific patterns that are consistent across species or in isolating features of diseases specific to human variants, both of which would represent favorable outcomes for downstream translational applications.

Learning on human and veterinary medical records

The breadth and depth of data being generated in the form of clinical narratives largely outperforms our current ability to process them. Traditional NLP methods boast interpretability and flexibility, but come at the steep cost of data quality control, formatting, and normalization, as well as the cost of the domain knowledge and time needed to generate meaningful heuristics (which oftentimes are not even generalizable to other datasets). It is thus a logical choice to bypass these steps, classifying medical narratives from the electronic health record by leveraging supervised deep learning on big data. We expect that our efforts could facilitate cohort identification for biosurveillance, public health research, and quality improvement.

2. Methods

Study Design

This is a retrospective cross-sectional and chart review study, using 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[20]) from the Beth Israel Deaconess Medical Center in Boston, Massachusetts. The STROBE checklist is provided in Supplementary Material 1.

A TensorFlow[21] deep learning model of our design classified medical records into 17 categories with a 70-30 train-test split. We built separate models for each database, using the train split of the free-text clinical narratives, and used F_1 score (a measure of a test's accuracy, considering a harmonic mean between precision and recall) per top-level disease category as our evaluation metric on the test set. We also investigated the effect of using MetaMap[22], an NLP tool that extracts only clinically-relevant terms, on the accuracy of our models. To explore the efficacy of model portability, we attempted to also test the MIMIC-trained model on the CSU test data, and vice versa (and ran separate tests for MetaMapped versions, as well). Figure 1 shows a diagram of our training, test, and evaluation design. More information on clinical coding, technologies, and a link to our code can be found in Supplementary Material 2.

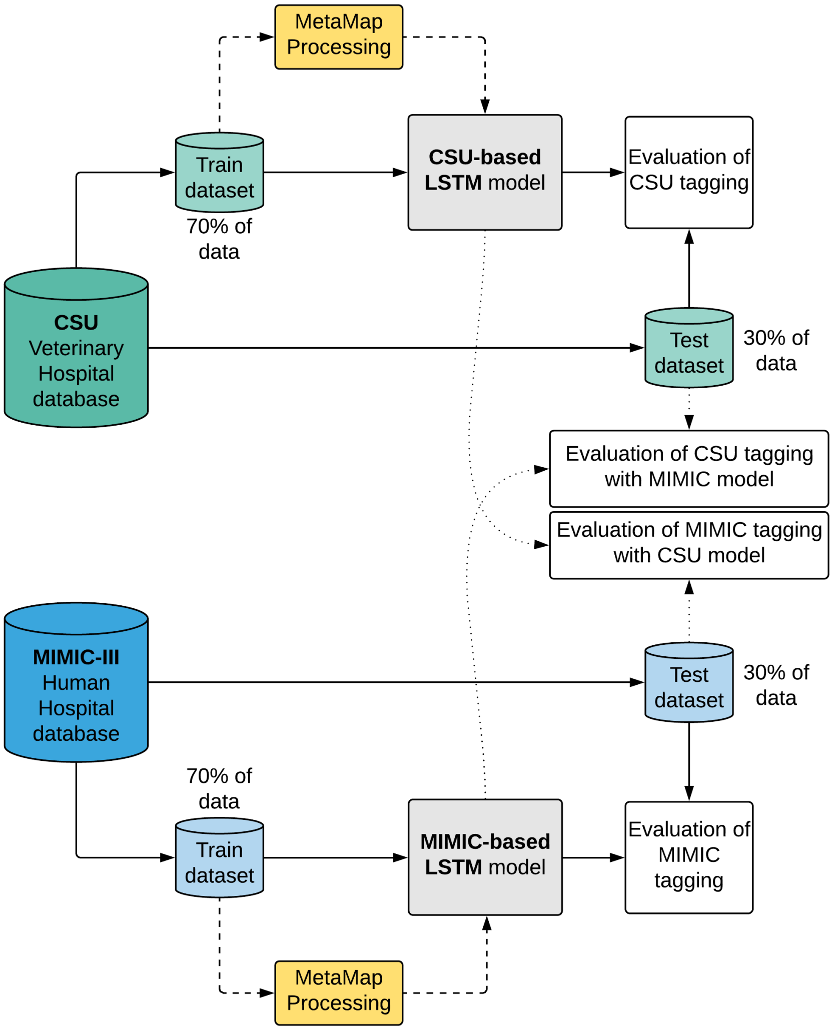

MetaMap Processing
Train dataset
70% of data
CSU-based LSTM model
Evaluation of CSU tagging
CSU Veterinary Hospital database
Test dataset
30% of data
Evaluation of CSU tagging with MIMIC model
Evaluation of MIMIC tagging with CSU model
MIMIC-III Human Hospital database
Test dataset
30% of data
70% of data
Train dataset
MIMIC-based LSTM model
Evaluation of MIMIC tagging
MetaMap Processing

Figure 1. Diagram of the training, test, and evaluation design.

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, a veterinarian or veterinary student enters information about the patient into the custom-built veterinary electronic health record (VEHR), including structured fields, such as open and discharge dates, patient signalment (species, breed, age, sex, reproductive status), and codes applied to the visit. There are also options to input unstructured free-text clinical narratives with various sections, including history, assessment, diagnosis, prognosis, and medications. These records are subsequently coded by trained medical coders and veterinarians, where the final diagnostic codes may consist of a single concept code representing a specific diagnosis or a set of codes referencing multiple diagnoses or post-coordinated expressions.

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[20], a publicly available dataset which we utilize in this study, comprises information relating to patients admitted to the critical care unit at the Beth Israel Deaconess Medical Center. We were interested in the unstructured free-text notes in this database, with special attention given to provider progress notes and hospital discharge summaries. These records are coded for billing purposes, and have relatively complete diagnoses per patient (the database is publicly available, and thus represents the best possible medical coding annotation scenario for a hospital). Protected health information was removed from free-text fields.

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 (80·8% *Canis Lupus*), cats (11·4% *Felis Silvestris*), horses (6·5% *Equus Caballus*), cattle (0·7% *Bos Taurus*), pigs (0·3% Sus Scrofa), goats (0·2% *Capra hircus*), sheep (0·1% *Ovis Aries*), and other unspecified mammals (0·1%). In contrast, the MIMIC-III database contains medical records from 38,597 distinct 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 1 summarizes the characteristics and category breakdown of both databases. Only those patients with a diagnosis in their record were considered.

Table 1. 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

Deep learning models

We used 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[23]. 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. 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[23]. Our assumption was that, due to this flexibility, this structure would be ideal for extracting clinically relevant information from across institutions.

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[24]. Because clinical notes average a length of 430 words in the CSU database, and 910 words in the MIMIC database, it is important to use an architecture with an improved capability to store information over many time steps. LSTMs have memory cells that can maintain information for over a longer period of time and that consist of a set of gates that control when information enters and exits memory, making them an ideal candidate architecture.

MetaMap Feature Extraction

We used MetaMap Lite[25], an NLP tool which leverages the Unified Medical Language System (UMLS) Metathesaurus to identify SNOMED[26] and ICD[27] 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[28], an algorithm for the identification of negation in clinical narratives.

Statistical analysis

Evaluation metric. We used the same evaluation metrics previously reported for MetaMap Lite[25]: 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:

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

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

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

Our classification task is a multi-label classification problem, given that clinical narratives can describe multiple top-level categories. We calculated evaluation metrics independently for every top-level category, as if each were a binary classification problem. We obtained global estimates for performance of the models by calculating both the average (macro F_1 score) and the weighted average (micro F_1 score) of the individual F_1 scores, across all classes.

Portability. The portability of trained algorithms on independent datasets has previously been used as a metric of model robustness in systems that leverage NLP and machine learning [29]. We evaluate 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.

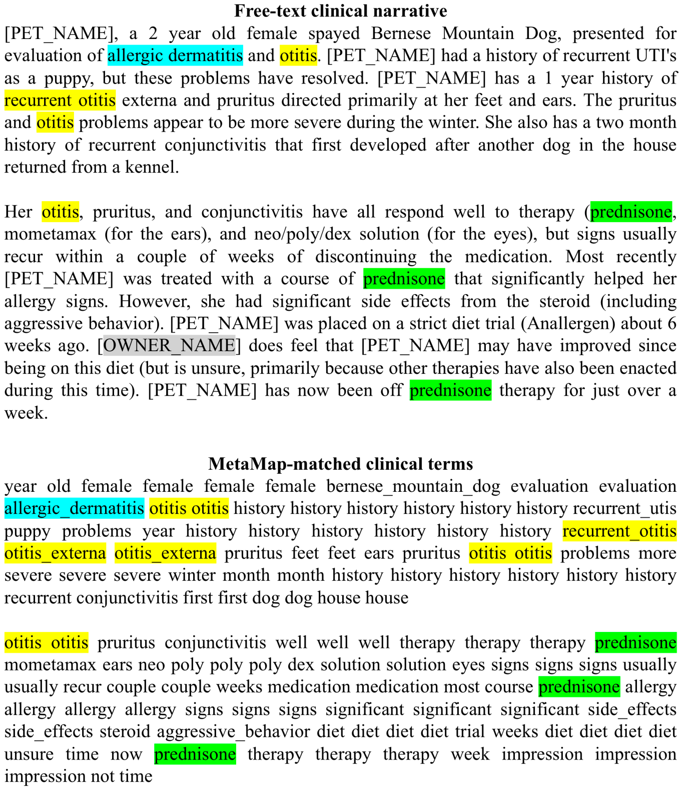
Role of the funding source

The funder had no role in the data analysis, data interpretation, or writing of this paper.

3. Results

We investigated the application of deep learning to free-text unstructured clinical narratives on two cohorts, veterinary (CSU) and human (MIMIC). First, we present an evaluation of the NLP tool, MetaMap, applied to veterinary records, and then we show the evaluation of the deep learning models built using human and veterinary records, as well as the portability between them.

Evaluation of NLP on veterinary records. MetaMap Lite, a software typically used to extract clinical terms from free-text narratives, has not previously been applied to veterinary data. As such, we endeavored to verify that the software works as expected when applied in this context. Two board-certified veterinarians trained in clinical coding independently evaluated the MetaMap-extracted terms from 19 randomly selected records. Disagreements were resolved via in-depth discussion and consensus. This process resulted in a weighted-average precision of 0·62, recall of 0·82, and F$_1$ score of 0·71 for the CSU data, as compared to a previously reported weighted-average precision of 0·67, recall of 0·53, and F$_1$ score of 0·58 for human clinical narratives[25]. Figure 2 shows an example of one free-text clinical narrative processed with MetaMap.



Free-text clinical narrative

[PET_NAME], a 2 year old female spayed Bernese Mountain Dog, presented for evaluation of allergic dermatitis and otitis. [PET_NAME] had a history of recurrent UTI's as a puppy, but these problems have resolved. [PET_NAME] has a 1 year history of recurrent otitis externa and pruritus directed primarily at her feet and ears. The pruritus and otitis problems appear to be more severe during the winter. She also has a two month history of recurrent conjunctivitis that first developed after another dog in the house returned from a kennel.

Her otitis, pruritus, and conjunctivitis have all respond well to therapy (prednisone, mometamax (for the ears), and neo/poly/dex solution (for the eyes), but signs usually recur within a couple of weeks of discontinuing the medication. Most recently [PET_NAME] was treated with a course of prednisone that significantly helped her allergy signs. However, she had significant side effects from the steroid (including aggressive behavior). [PET_NAME] was placed on a strict diet trial (Anallergen) about 6 weeks ago. [OWNER_NAME] does feel that [PET_NAME] may have improved since being on this diet (but is unsure, primarily because other therapies have also been enacted during this time). [PET_NAME] has now been off prednisone therapy for just over a week.

MetaMap-matched clinical terms

year old female female female female bernese_mountain_dog evaluation evaluation allergic_dermatitis otitis otitis history history history history history history recurrent_utis puppy problems year history history history history history history recurrent_otitis otitis_externa otitis_externa pruritus feet feet ears pruritus otitis otitis problems more severe severe severe winter month month history history history history history history recurrent conjunctivitis first first dog dog house house

otitis otitis pruritus conjunctivitis well well well therapy therapy therapy prednisone mometamax ears neo poly poly poly dex solution solution eyes signs signs signs usually usually recur couple couple weeks medication medication most course prednisone allergy allergy allergy allergy signs signs signs significant significant significant side_effects side_effects steroid aggressive_behavior diet diet diet diet trial weeks diet diet diet diet unsure time now prednisone therapy therapy therapy week impression impression impression not time

Figure 2. Example of free-text and MetaMap-extracted veterinary record. A 2-year old female dog patient with recurrent otitis (marked in yellow), and allergic dermatitis (marked in blue). The narrative shows that the treatment given included prednisone (marked in green). For the purpose of this manuscript, the pet and owner's name (marked in gray) were manually de-identified.

Evaluation of Deep Learning models. We trained one deep-learning model for each of the veterinary and human datasets. Both models were used to test on their own domain (veterinary to veterinary, and human to human), as well as ported over to the other domain (veterinary to human, and human to veterinary). In Figure 3, we provide the classification performance, using the original clinical narratives and the narratives with MetaMap-extracted features.

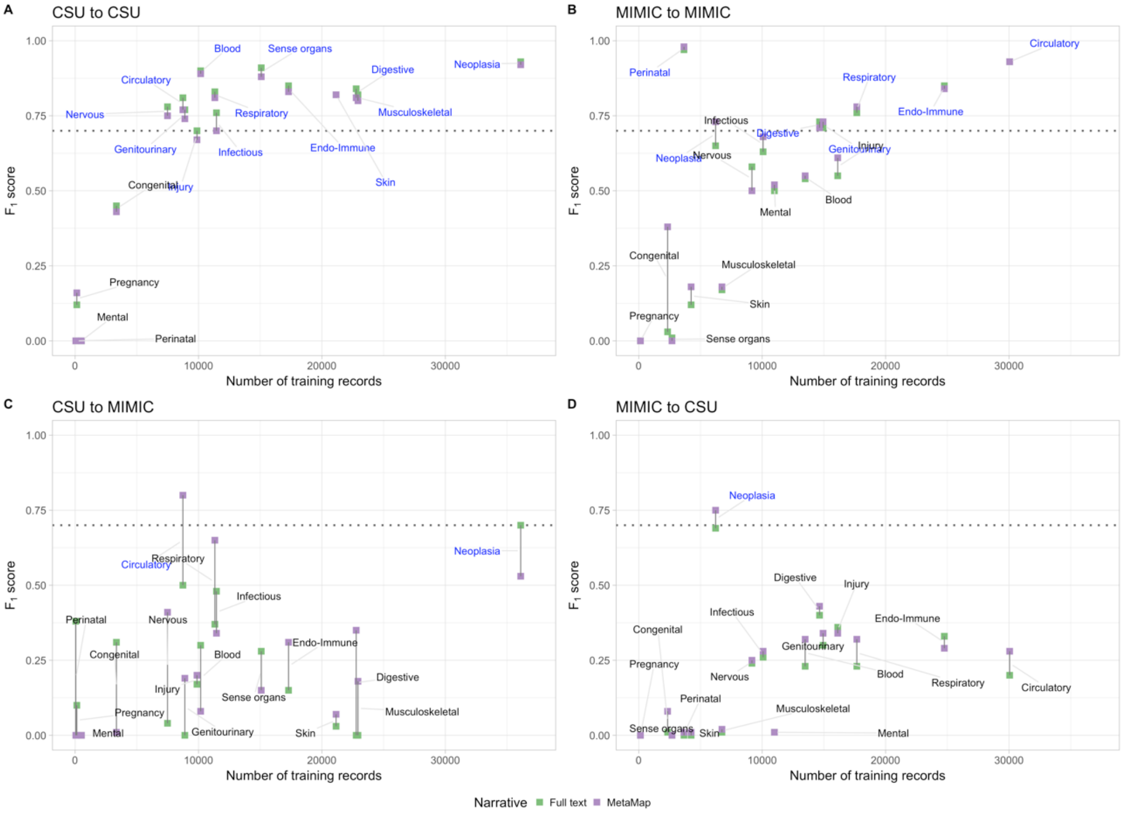


Figure 3. F_1 scores by category in each training-validation model. A) training model with CSU data, validation with CSU data; B) training model with MIMIC data, validation with MIMIC data; C) training model with CSU data, validation with MIMIC data; D) training model with MIMIC data, validation with CSU data. The color of the square represents the type of narrative used, either full free-text (green) or the MetaMap version (purple). The color of the category text is highlighted (blue) if it surpasses the threshold of at least 0.70 F_1 score (dotted horizontal line).

4. 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 electronic phenotyping. Besides assigning ground-truth labels to the CSU veterinary data for the purpose of illustrating the efficacy of our model, none of our efforts involved any manual curation, feature generation, or data harmonization, all of which are time-consuming tasks.

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 greatly impact the performance of the LSTM's vector selection and the NLP's entity recognition. MetaMap was useful in this regard, given its ability to parse clinical data.

The databases that we selected, MIMIC and CSU, represent vastly different clinical settings. The clinical narratives that arise in a critical care unit, like those in MIMIC, do not necessarily compare to those from a tertiary referral veterinary care facility, like those in CSU. 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 able to accurately classify medical narratives at the top level of depth in both datasets, without loss of generality in the method. However, the variability in classification performance across categories could be explained by larger number of training cases. There was a direct classification increase in those categories with more training 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. Having any sort of data (e.g. top-level ICDs) on the patient beforehand could be extremely useful in more rapid triage. One can also imagine that more granular characterizations (in any dataset) could arise given sufficient data in each target tag (our models seemed to have high classification accuracies when there were more than approximately 5,000 records in the category of interest). The repeated use of a series of LSTM models for subsequent, increasingly-specific classifications thus represents a scalable, hierarchical tagging structure that could prove extremely useful in bucketing patients into specific departments, severities, and protocols.

Our study has several limitations, including sample size, number of databases investigated, and focus on top-level, rather than downstream, categories. In the future, the increased availability of data from both the human and veterinary domains will facilitate more research in this field. Using a deep learning approach, can facilitate the categorization of unstructured clinical narratives, which are often a bottleneck to the identification of research cohorts, as well as facilitating sharing capabilities across institutions.

Public Health Implications. In this era of rapid digital health and deployment of health records, it is important to provide tools that facilitate cohort identification. Our deep learning approach (LSTM model) was able to automatically classify medical narratives without having any domain knowledge or manual curation of features. The accuracy of classification (F_1 score) was 0·83 for veterinary data, and 0·67 in the human data. With more training data it is possible to foresee a scenario in which these training models can benefit every clinical domain. As an example, in the neoplasia top-level category, the veterinary data had 36,108 clinical notes, which were used to train an LSTM model that correctly identified those clinical narratives in the human clinical notes, with F_1 scores of 0·93 and 0·70, respectively. The expansion of veterinary data availability and the subsequently enormous potential of model portability could prove to be exciting chapters in reducing bottlenecks in biosurveillance and public health research at large.

Declarations

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.

Availability of data

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 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/

Competing interests

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

Funding

Research reported in this publication was partially supported via institutional funds from Stanford University. CDB is a Chan Zuckerberg Biohub Investigator. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

ALP, OJBDW, 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.

Acknowledgments

The authors wish to acknowledge Dr. Katie 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.

Bibliography

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

2 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.

3 Benesch C, Witter DM, Wilder AL, Duncan PW, Samsa GP, Matchar DB. Inaccuracy of the International Classification of Diseases (ICD-9-CM) in identifying the diagnosis of ischemic cerebrovascular disease. *Neurology* 1997; **49**: 660–4.

4 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.

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

6 Anholt RM, Berezowski J, Jamal I, Ribble C, Stephen C. Mining free-text medical records for companion animal enteric syndrome surveillance. *Preventive Veterinary Medicine* 2014; **113**: 417–22.

7 Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning. 2016.

8 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.

9 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**: 18.

10 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.

11 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.

12 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.

13 Griffith JE, Higgins DP. Diagnosis, treatment and outcomes for koala chlamydiosis at a rehabilitation facility (1995-2005). *Aust Vet J* 2012; **90**: 457–63.

14 Poppe JL. The US Army Veterinary Service 2020: knowledge and integrity. *US Army Med Dep J* 2013; : 5–11.

15 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.

16 Shalev M. USDA to require research facilities, dealers, and exhibitors to keep veterinary medical records. Lab Anim (NY). 2003; **32**: 16.

17 Robinson TP, Wint GRW, Conchedda G, *et al.* Mapping the global distribution of livestock. *PLoS One* 2014; **9**: e96084.

18 Smith K, Martinez A, Craddolph R, Erickson H, Andresen D, Warren S. An integrated cattle health monitoring system. *Conf Proc IEEE Eng Med Biol Soc* 2006; **1**: 4659–62.

19 Gates MC, Holmstrom LK, Biggers KE, Beckham TR. Integrating novel data streams to support biosurveillance in commercial livestock production systems in developed countries: challenges and opportunities. *Front Public Health* 2015; **3**: 74.

20 Johnson AEW, Pollard TJ, Shen L, *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; **3**: 160035.

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

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

23 Pham T, Tran T, Phung D, Venkatesh S. DeepCare: A Deep Dynamic Memory Model for Predictive Medicine. In: Advances in Knowledge Discovery and Data Mining. Springer, Cham, 2016: 30–41.

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

25 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.

26 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.

27 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.

28 Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform* 2009; **42**: 839–51.

29 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.