Developing a Technical Infrastructure for Large-Scale Mining of Veterinary Electronic Health Record Data

Dawson Eliasen, Sooraj Lankala, G. Joseph Strecker

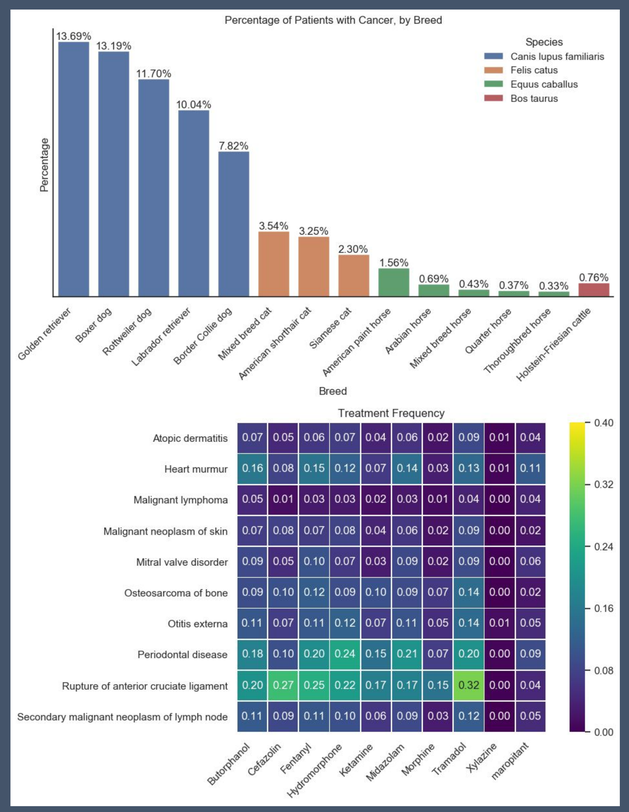
COLLEGE OF VETERINARY MEDICINE AND BIOMEDICAL SCIENCES
COLORADO STATE UNIVERSITY

Introduction

The CSU Veterinary Teaching Hospital has over a million Electronic Health Records (EHRs), each consisting of many fields of free text detailing visits, diagnoses, procedures, histories, and other details. There is potential for enhanced translational discovery by mining these data, especially when multiple human and veterinary EHR data sources are combined. Standardized representation of veterinary EHR data in a Common Data Model (CDM) will make it possible to join these individual data sets. In human medicine, the OMOP structure has been established as a standard EHR data model. As a member of the CTSA One Health Alliance (COHA) Technical Pilot Program (with the University of California-Davis and Tufts University) we aim to pioneer technology best practices and develop tools to facilitate the creation of a nationwide network of databases conforming to the OMOP CDM to support research in veterinary and translational medicine.

The OMOP Common Data Model

The Observational Medical Outcomes Partnership (OMOP) CDM is designed specifically to support health analytics and has significant adoption in human medicine. The model is patient-centric and supports various kinds of healthcare data, including patients, providers, locations, visits, conditions, procedures, measurements, payments, etc. The CDM is structured around the use of "concepts" – named entities associated with a standardized, unique identifier that represent a medical term.



Example analyses supported by the OMOP CDM.
Figure 1. Rates of occurrence of cancer in various species and breeds.
Figure 2. Rates of coincidence of common conditions and drugs.

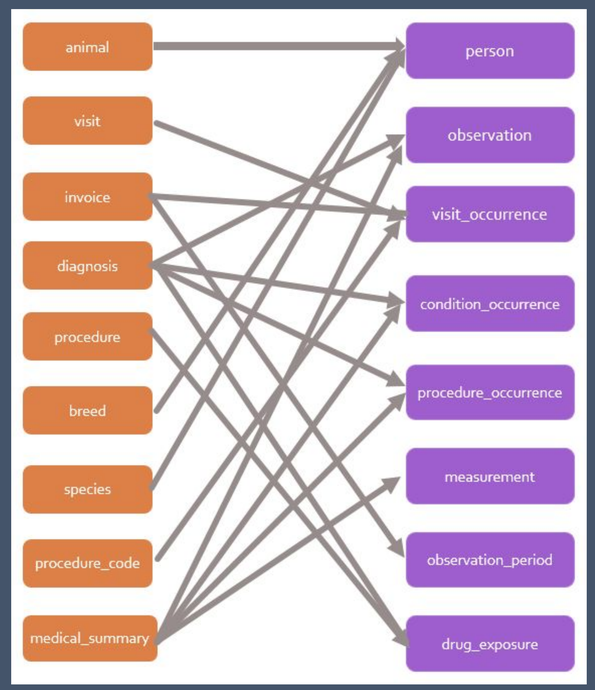


Figure 3. Illustration of data discovery and mapping process.

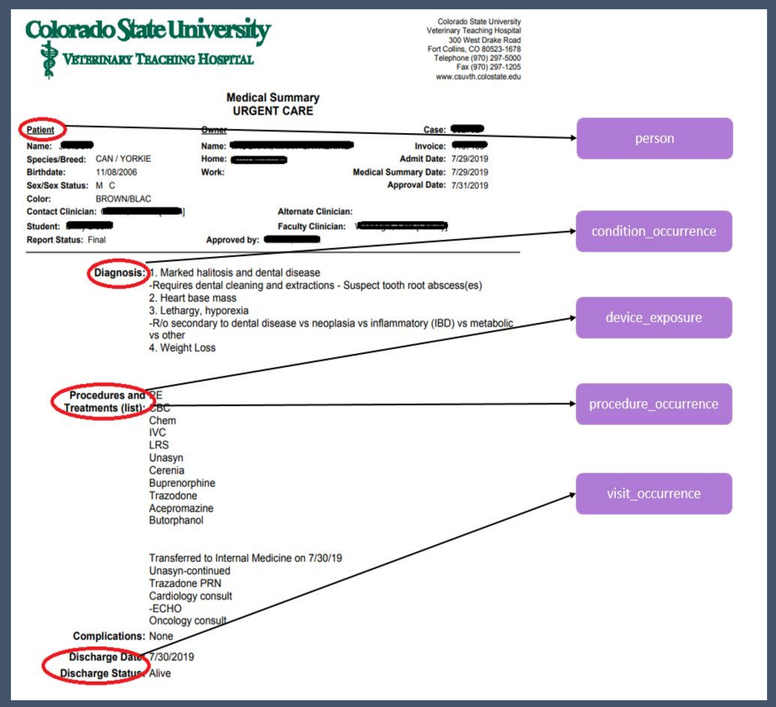


Figure 4. Illustration of how an EHR is represented in the OMOP CDM.

Challenges

Adopting the OMOP CDM involves two notable challenges. The first is the data discovery process, to identify where and how all information that is to be loaded in the OMOP database is currently represented in existing databases. The relationships between records in the source and the locations in the target can be distributed and complex, requiring CDM developers to coordinate with users and administrators of the source data systems. The second challenge is the development of efficient means for converting EHRs into the "standard concepts" that the CDM expects. This often involves processing text having little formal structure, and matching it with the target vocabulary consisting of millions of standard concepts. This can become prohibitively expensive computationally at scale, which has led us to pursue solutions such as the use of existing software such as CLAMP, developing our own efficient algorithms for text matching, as well as experimenting with the use of artificial intelligence (AI) for text processing.

Automated diagnosis coding with artificial intelligence

Following recent research in the area, we are exploring the use of Artificial Intelligence in the form of Deep Learning (DL) to deduce the meaning of EHRs and perform automated standard encoding for loading into the CDM. A DL system can be trained on CSU's large dataset which has been extensively hand-encoded by medical records experts. Once trained, the system may be deployed at other veterinary sites, many of which have little standard-encoded data, to convert raw text EHR data structured concepts for the CDM. We have seen promising results from an experiment training a convolutional neural network (CNN) to identify 146 common diagnoses from free text (precision = ~85%, recall = ~60%). Now, we are attempting to incorporate advances in DL from projects such as the Transformer, DeepTag1, VetTag2, Google's BERT3, and others to improve the effectiveness of the AI.

Original diagnosis note
"Allergic dermatitis - primary differentials include atopic dermatitis or cutaneous adverse food reaction."

Human encoding ("ground truth")	Raw text matching	Simplified multi-labeling with CNN (shallow NN)	Sequence-to-sequence with Transformer (Deep NN)
238575004 - Allergic contact dermatitis 42752001 - Due to 101 - Suspect 370540009 - Adverse reaction to food 101 - Suspect 24079001 - Atopic dermatitis	45955251 - Allergic dermatitis 281647001 - Adverse reaction 24079001 - Atopic dermatitis	238575004 - Allergic contact dermatitis	100 - Recheck 238575004 - Allergic contact dermatitis 42752001 - Due to 100 - Recheck 24079001 - Atopic dermatitis

True positive False positive True positive, but does not match human encoding

Figure 5. Sample diagnosis note, human encoding, and output from different text processing approaches.

Acknowledgements

1. Nie A, Zehnder A, Page RL, Zhang Y, Lopez Pineda A, Rivas MA, Bustamante CD, Zou J. 2018. DeepTag: Inferring Diagnoses from Veterinary Clinical Notes. NJP Digital Medicine 1:60.
2. Zhang Y, Nie A, Zehnder A, Page RL, Zou J. 2019. VetTag: Improving Automated Veterinary Diagnosis Coding via Large-Scale Language Modeling. NJP Digital Medicine 2:35
3. Devlin J, Chang M-W, Kenton L, Toutanova K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805

Supported by NIH/NCATS Colorado CTSA Grant Number UL1 TR002535. Contents are the authors' sole responsibility and do not necessarily represent official NIH views.