Topics in Companion An Med 37 (2019) 100364

Research Article

Optimization of Electronic Medical Records for Data Mining Using a Common Data Model

Manlik Kwong, BSEE, BSCS[a,*], Heather L. Gardner, DVM, DACVIM[b], Neil Dieterle, BSc[c], Virginia Rentko, VMD, DACVIM[c,d]

Keywords:
electronic health record
COHA
veterinary
infrastructure
OMOP
research

[a]*Tufts Medical Center, Boston, MA, USA*

[b]*Sackler School of Graduate Biomedical Sciences, Tufts University, Boston, MA, USA*

[c]*Cummings School of Veterinary Medicine, Tufts University, Boston, MA, USA*

[d]*Animal Bioscience Inc., Boston, MA, USA*

A B S T R A C T

The increasing use of electronic health records (EHRs) in veterinary medicine creates an opportunity to utilize the high volume of electronic patient data for mining and data-driven analytics with the goal of improving patient care and outcomes. A central focus of the Clinical and Translational Science Award One Health Alliance (COHA) is to integrate efforts across multiple disciplines to better understand shared diseases in animals and people. The ability to combine veterinary and human medical data provides a unique resource to study the interactions and relationships between animals, humans, and the environment. However, to effectively answer these questions, veterinary EHR data must first be prepared in the same way it is now commonly being done in human medicine to enable data mining and development of analytics to facilitate knowledge formation and solutions that advance our understanding of disease processes, with the ultimate goal of improving outcomes for veterinary patients and their owners. As a first step, COHA member institutions implemented a Common Data Model to standardize EHR data. Herein we present the approach executed within the COHA framework to prepare and optimize veterinary EHRs for data mining and knowledge formation based on the adoption of the Observational Health Data Sciences and Informatics' Observational Medical Outcomes Partnership Common Data Model.

Introduction

Evaluation of medical records for biomedical research is a complex process, historically involving manual evaluation of records and knowledge of the clinical context data with which data is recorded. This effort is confounded by the immense volume of data generated every year in both individual institutional medical records and the published medical literature. The routine application and utilization of data mining is increasingly common in human medicine, driven by improvements in machine learning algorithms, but is still a nascent resource in veterinary biomedical research. The widespread adoption and increasing use of electronic health records (EHR) systems in veterinary medicine provides an unprecedented opportunity to incorporate data mining into veterinary and comparative research. To address this unmet need, the Clinical and Translational Science Institute (CTSI) One Health Alliance (COHA) member institutions came together to create the infrastructure necessary to enable multi-institutional collaborative studies to advance research in diseases shared by animals and people. Specifically, this effort has supported the implementation of a Common Data Model (CDM) based on the Observational Medical Outcomes Partnership (OMOP version 5.2) to standardize EHR data among individual member institutions, enabling collaborative biomedical research employing Data Mining methods.[1,2]

TRANSLATOR - TRanslational ANimal Shared ColLAboraTive Observational Research Network

Two workshops were convened in January, 2018 and February, 2019 with participation from representatives from various COHA member institutions to discuss and debate many topics including the utilization of EHR data to support veterinary and OneHealth research. The discussion regarding utilization of EHR records for veterinary and OneHealth research was guided by the steering committee for the "COHA Advancing One Health Datasets" project which arose from a collaborative, pilot grant held by Colorado State University—University of Colorado Denver, the Ohio State University, Purdue University, Tufts University, and University of Missouri, Veterinarians, physicians, biomedical data scientists, IT and database specialists, and a librarian attended from these institutions and from North Carolina State University, University of California-Davis, and University of Wisconsin-Madison. Other institutions represented were Duke University and Stanford University medical schools, 2 private practice corporations, 2 veterinary nonprofit institutions, 2 veterinary databases.

Technical and leadership subcommittees were formed to examine the technical issues of integrations and establish governance and financial sustainability. The network was dubbed, TRANSLATOR - **TR**anslational **AN**imal **S**hared Col**LA**bora**T**ive **O**bservational **R**esearch. In the 2018 workshop, the interests of the participants in the development of a data platform were identified as translational research and data sharing. This first workshop provided a synopsis of an implementation of a CDM in human medicine and early pilot results of OMOP implementation at a single site, Tufts Cummings Veterinary Medical

*Corresponding author: Manlik Kwong, Tufts Medical Center, 800 Washington Street MA 02111.

E-mail address: Mkwong@tuftsmedicalcenter.org (M. Kwong).

Abbreviations: API, Application program interface; CDM, Common Data Model; COHA, CTSI One Health Alliance; CTSA, Clinical and Translational Science Award; CTSI, Clinical and Translational Science Institute; ECG, electrocardiogram; EHR, Electronic Health Record; EMS, emergency medical system; ePCR, electronic patient care record; ETL, Extract/Transform/Load; FTP, File transport protocol; HTML, Hypertext Markup Language; LOINC, Logical Observation Identifiers Names and Codes; OHDSI, Observational Health Data Sciences; OMOP, Observational Medical Outcomes Partnership; SNOMED, Systematized Nomenclature of Medicine; STEMI, ST elevation in myocardial infarction

http://dx.doi.org/10.1016/j.tcam.2019.100364
1938-9736/

Table 1
Veterinary OMOP Database Extensions

New Fields in the table "***omop.person***" – links animal patient to owner table.
New Fields in the table "***omop.person***" – concept to support species and breed using SNOMED codes and custom Veterinary vocabularies.
New Table: ***owner***– Description and management of owner-specific identifiers and demographics.
New Table: ***soap_note***– Natural Language Processing (NLP) table support for standard clinical SOAP (subjective, objective, assessment, plan) narratives.
New Table: ***master_index***– Master Index Table to enable re-identification of OMOP-based veterinary patient records to its true medical record number and owner client number used by the veterinary EHR system.
New Table: ***report_data***– Additional report cache and aggregation tables for active reporting functions and tracking purposes to increase data reporting performance.

Center. The second workshop further refined issues encountered by Tufts Cummings Veterinary Medical Center and early work at Colorado State University's veterinary informatics team regarding designing real-world CDMs on the OMOP platform. The second workshop in 2019 also explored implementation issues across a network of collaborative institutions (local CDM silo, federated CDMs, or single cloud CDM implementations) which is discussed in a companion paper in this issue. In this paper we will explore the topic of optimizing the EHR data resource to implement an adaptable informatics platform that can be implemented locally, within a federated network, or centrally to support more specialized registry type warehouse to support OneHealth focused research (Table 1).

Data Mining of Electronic Health Records

The continued reduction in the overall cost of computation and adoption of EHR systems, reaching 1 billion visits in 2012 (US) in combination with improvements in machine learning algorithms, are driving the use of data mining in veterinary EHR data to complement prospective clinical trial data for knowledge discovery. The widespread adoption and increasing use of EHR systems in veterinary medicine have also reached a point where translation of data mining utilization can be applied to support veterinary research. This is particularly important, given that the retrospective use of patient data generated from medical records, regardless of its origins in human or veterinary medicine, is fraught with limitations, including data collection and lead-time and temporal biases. While data generated from typical patient care may be considered complete for the purposes of future care of the patient, it may not contain all the information necessary for a scientific description.[1] However, the sheer volume and breadth of retrospective information available through the typical care of patients make these data an attractive way to generate questions that can be evaluated in future prospective studies compared with published clinical trial data that are narrowly focused and often only include a small cohort of patients from the larger population of patients being seen. Data mining techniques are increasingly used to generate a more complete dataset from high volume but less controlled clinical care data, which contains much fewer quality controls for data error reduction (e.g., double entry), incomplete data, and data contained in unstructured clinician notes.

One common data mining methodology employed in biomedical research is machine learning, utilizing computational methods to prepare and extract relevant phenotypic features from the data in support of research efforts. Machine learning uses predictive analytics to help inform prospective questions (prediction models) and understand disease patterns (descriptive models) to facilitate a better understanding of behaviors and characteristics within patient cohorts.[2]

To date, machine learning algorithms are routinely implemented in the consumer domain, with online interfaces (e.g. amazon.com, Google search, Facebook, etc.) learning user behavior to make targeted suggestions. This is accomplished by utilization of a combination of artificial intelligence, statistics, probability, and computational complexity theory, among others, to digest and search for patterns in large data sets.[3] While these algorithms can generate results which are directly useful at times, depending on the complexity of the question being asked, the resulting output may be more ambiguous. For example, simple machine learning algorithms that use information from web-based searches to recommend additional websites of interest to the consumer are relatively straightforward. However, predictive algorithms that provide insight into complex multi-factorial diseases (e.g. neurologic diseases, cancer, etc.) are more challenging to accurately implement.[3,4] Therefore, judicious selection of features from large datasets and careful selection of machine learning model (s) will facilitate the utility of these approaches in clinical research settings and underscore the utility of data mining in clinical settings. For example, a collaboration between MIT and Massachusetts General Hospital utilized data mining of mammography data to generate predictive models of breast cancer risk in women of color. When compared with established diagnosis approaches, the incorporation of data mining efforts improved risk assessment modeling in this population of women.[5]

Data Mining in Veterinary Research

Currently, 3 COHA member institutions (Tufts University Cummings School of Veterinary Medicine [Tufts], Colorado State University College of Veterinary Medicine, and Biomedical Sciences [CSU] and University of California-Davis School of Veterinary Medicine [UC Davis]) are implementing an informatics infrastructure to support collaborative EHR based research and form the foundation and resource to enable One Health initiatives within their respective CTSI organizations and as a proof of concept for a wider research network The One Health concept encompasses the interrelationship between animals, humans, and the environment in which they live and the impact on the health of each. One Health initiatives are pilot studies that utilize research warehouses from both humans and animal to explore and tease out their relationships in a number of clinical topics and diseases. In order to facilitate One Health studies, the three CTSA member institutions have implemented an OMOP-based research infrastructure to support human centered cohort discovery and retrospective research activities, including a data warehouse and CTSI services. Each of these veterinary institutions now have the capacity to leverage this human-centered infrastructure and apply it to their respective local veterinary EHR systems. The veterinary data warehouse currently utilizes 2 different veterinary EHR systems among the 3 hospitals, and this platform can accommodate other EHR record systems. All 3 institutions will create a veterinary data warehouse based upon the Observational Health Data Sciences and Informatics' (OHDSI) OMOP CDM (www.ohdsi.org). While each institution will use different development and software tools and strategies based on their own local IT infrastructure, informatics expertise, resources, and underlying EHR data characteristics; All 3 will utilize the same 3 data processing tiers (Fig 1) in their implementation: (1) Messaging; (2) Transformation; and (3) Operational Database.[6]

Messaging

The term "Messaging," also referred to as "data export/integration," consists of connecting the underlying clinical EHR system to automatically request and extract patient data on a regular basis (Fig 1). "Data staging," is the process of storing the native EHR records in a centralized location (e.g. a designated computer folder) to be ready for the next step in this process, data transformation. Data processing is a series of steps (Fig 1 and Tufts' example implementation in Fig 2) in which EHR data moves through the transformation process of being native clinical source data to something that is more structured and normalized that is suitable for data mining and research activities. For example, an Application Programming

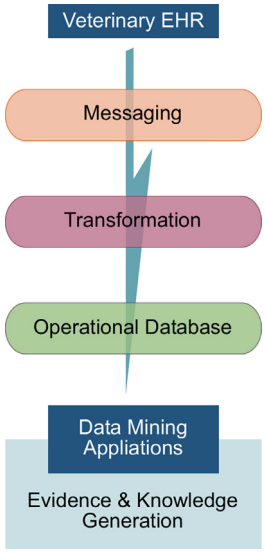

Fig 1. EHR data mining process.

Interface was implemented within the Tufts EHR software, based on an internal URL that uses a programmatic machine to automatically request daily EHR records (pull model). The data retrieved can be tailored to include information such as patient demographics, visit information (weight, vitals, etc.), invoiced items (prescriptions, medical procedures, consultation, and recheck appointments), and laboratory results. Additionally, "custom forms," such as department and/or disease specific data assessment and documentation forms can be incorporated into the data extraction process using both text and image processing if needed. The daily transaction record is then saved in a designated "inbox" with a time-stamped file name for easy identification and ordered processing of subsequent data analysis. While the "pull model" described above represent 1 method of automatically collected EMR data, other mechanisms can also be employed, such as secure File Transport Protocol (FTP – "push model"), shared folder transfers, etc. with and without encryption depending on whether the data remains within the local information technology network (intranet) or has to transit over the public Internet. Similar approaches and methods will be utilized by other COHA veterinary institutions as depicted in Fig 1, supporting the utility of this approach across many different EHR systems.

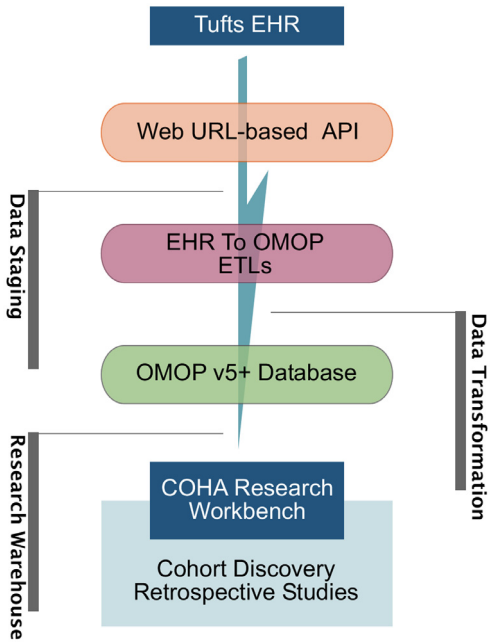

Fig 2. Data warehouse implementation – tufts EHR to OMOP.

Transformation

After data export and integration has been completed, the raw EHR data must be transformed into a useable format that can readily be imported or loaded into the research database. This process is called transformation, or "Extract/Transform/Load" (ETL). As mentioned above, the COHA member institutions are largely adopting the OHDSI OMOP CDM as their target database system. The OMOP CDM utilizes a group of standard vocabularies and concepts to represent patient data elements stored in the database (SNOMED, RxNORM, LOINC, and others), which reduces any unexpected variations in the meaning or semantics of data among OMOP CDM participants and increases the transparency of the resulting information. Notably, this approach is unique among existing CDMs within the medical research community, and bridges existing gaps to support multi-institutional research collaborative efforts. From a practical standpoint, the implementation of a cohort query written at one institution against their local OMOP CDM can be readily shared and executed at another site with their own OMOP CDM and both results can be aggregated without further interpretation or adjustment to local documentation methods and practices. Therefore, ETLs are 1 or more programs that translate local EHR records and data elements into a standardized code. Data "normalization" is the conversion or mapping the original clinical description as documented in the EHR to a standard form or concept code. For example, the EHR might use the abbreviation for breed "AFFN@CANINE" to represent a "normalized" SNOMED description of "Affenpinscher" and OMOP CDM code of "SNOMED.4203003". Another example where a drug is documented in the EHR as "ONSIOR 6MG TABLETS - T@T QTY: 3 GIVE 1 TABLET BY MOUTH EVERY 24 HOURS. **START TOMORROW MORNING**" is "normalized" to the RxNORM description "robenacoxib 6 MG Oral Tablet [Onsior]" and its OMOP CDM code of "RxNorm.42708964". Recognizing normalization of source EHR data to standard codes is a form of data reduction, particularly when transforming veterinary data using largely human based vocabularies and concepts – the OMOP CDM database table design provides "source" data fields for retaining the original source data within its tables. For example, in the above ONSIOR drug exposure, researchers could locate all occurrences where ONSIOR were prescribed with the QTY of 3 tablets using the database source fields in addition to the normalized codes. Retention of the original source values enables traceability for both ETL validation and database queries utilizing local terms and abbreviations familiar to clinicians and researchers.

In situations where data cannot be mapped to one of the standard vocabularies used in human medicine (SNOMED, RxNORM, LOINC, etc.), a custom veterinary medicine vocabulary is needed. In our implementation at Tufts Cummings Veterinary School, the source value also contains a unique transaction ID, clinical area representing the data source, and the value to provide complete traceability. Iterative data quality processes and vocabulary normalization processes can therefore use this backwards traceable information to identify documentation errors and shared decision-making regarding proposed new vocabularies and concepts to cover clinical data that were not previously supported in the standard vocabularies. Selecting the OHDSI OMOP CDM was based on 2 drivers: (1) Compatibility with the work already done on human EHR data for research and its use of standard vocabularies and concepts; and (2) Leverage of the various vocabulary standards communities as well as the OHDSI community to further develop its database and accompanying tools and knowledge base in utilizing observational EHR data.

Operational Database

The operational database is the end-product of the process of messaging and transformation. OMOP was originally designed and used in the human research community and has been extended to meet the needs of the veterinary community. The extensions added to the OMOP database schema design include the addition of new data fields to existing OMOP database tables and new tables that are unique to veterinary patients and documentation (Box 1). These extensions were designed and implemented in such a way to be backward compatible with OHDSI OMOP database tools (https://www.ohdsi.org/analytic-tools/). Backwards compatibility allows us to leverage tools and activities generated by the OHDSI community. The veterinary OMOP database is therefore a derivative of the OMOP version 5.2 – with the extension is referred to as "OMOP V5+" indicating its core version base with extension ("+"). The operational database is the locally-owned end-product of the process of Messaging and Transformation meaning each veterinary institution will have full responsibility and therefore control over their local implementation of an OMOP V5+ data warehouse. Access is dictated through data use agreements, COHA data governance policies/guidelines, and deidentified snapshots of the database contributed to a centralized COHA research network warehouse.

The OMOP CDM is a living project that is constantly evolving over time as the OHDSI community gains experience in using this technology and expanding its use cases. We will continue to monitor and evolve our veterinary version of the OMOP CDM as well. In the coming year, we will adapt the current OMOP version 5 to version 6 for veterinary use (OMOP v6+).

Technologies

The CTSI One Health OMOP CDM was designed and implemented using open-source technologies and platforms, which is a viable approach to manage costs and leverage technologies and resources generated by various academic and technology communities. For example, in the Tufts OMOP v5+ data warehouse the Messaging and Transformation stages were implemented using open source software on a virtual server over an intranet connection between the EHR and the local receiving research OMOP CDM server. It is common and recommended practice to physically or virtually separate the production EHR system from any consumer applications and systems to protect and isolate the production system as much as possible to minimize processing and security impact that may arise from secondary use of the EHR data. This approach makes the implementation relatively portable to other operating systems and environments. The transformation of data is managed as a daily scheduled Task within the standard Windows Task Manager. Other institutions can utilize alternate open-source and commercial products based on their individual needs and inhouse engineering expertise. Therefore, the choice of technology across all COHA institutions is a local decision, and as long as the data output is in the form of an OMOP v5+ data warehouse, is compatible with the platform.

Data Mining Applications and Veterinarian Engagement

After the source EHR data are normalized and transferred into an Operational Database, the data can be used to perform a number of tasks, from cohort discovery in preparation for research applications to retrospective observational studies. Enabled by having both the human and animal EHR data transformed and mapped into the same technology CDM such as OMOP within its CTSI organization (Tufts Clinical Translational Science Institute's OMOP and Tufts Cummings Veterinary Medical Center's OMOP v5+), it is anticipated in the near future that we can utilize data mining and machine learning techniques to conduct One Health studies that investigate the connections between owners and their companion animals across a variety of disease states, e.g. Lyme disease, obesity, and zoonotic transmission of disease. The key to efficient execution of single and multi-institution One Health investigations is the normalization of EHR data to common and standardized concepts as described above.

The utilization of a CDM is not limited to data obtained during hospitalization. The OMOP CDM is flexible in its patient-centered design and can support many types of data. For example, veterinary care data from existing registries (Veterinary Medical Database – VMDB https://vmdb.org/, Morris Foundation's Golden Retriever Lifetime Study registry - https://www.morrisanimalfoundation.org/golden-retriever-lifetime-study) can also be adapted and migrated to a common OMOP CDM research infrastructure. The OMOP CDM ***visit_occurrence*** database table readily supports identifying the clinical care context (prehospital, emergency care, hospital, primary care, registry) as well as the start and end of the visit context.

Centralized resources for data mining have generated the infrastructure necessary to support a variety of clinical research efforts. However, engagement of veterinarians and rest of the scientific community is critical to its growth and success. There are 2 broad mechanisms for engagement, including contribution of deidentified patient information to the database through participating COHA member institutions, and proposing projects that utilize the COHA Research Workbench resources.

One of the predominant goals of the COHA Research Workbench is to accelerate multi-institutional research collaborations. The COHA Research Workbench facilitates this process by helping to identify schools with patient populations for multi-center prospective trials and/or cohort discovery to enable multi-site clinical trials. Cohort summaries can be published on the website, to facilitate quick identification of institutional collaborators within the COHA network. This will encourage collaborations across COHA member institutions by connecting researchers and helping to identify where opportunities for collaboration exist.

While retrospective studies can provide information to suggest future prospective studies, they are fraught with limitations, including inconsistent data from medical records, lack of consistency in diagnostic follow-up, and limited medical records data. The Research Workbench also facilitates retrospective studies by identifying patient data with specific cohort criteria across multiple institutions, therefore helping to reduce bias created by regional differences in hospital populations and inconsistent entry of data in the medical record through use of standardized annotations. As patient data continues to be deposited into the database, the repository available for retrospective studies across all COHA member institutions will expand, ultimately increasing the power and relevance of many retrospective evaluations within the literature.

Finally, data mining can be used to enable investigators to conduct large-scale systematic and case reviews, comparing the effectiveness of treatments and performing exploratory surveys.[7,8] Deidentified patient data coded using OMOP V5+ criteria provide standardization across institutional EMRs, providing a rich resource for identifying patient cohorts based on the incidence, type of occurrence, and concurrent treatment regimens received. After generation of criteria for each cohort, the Research Workbench can automatically monitor the database in real-time for patient data matching the generated profile.

Conclusions

As we work towards a research data infrastructure with the capacity to employ CDMs within a network of academic veterinary centers within the CTSA COHA, implementation of the Research Workbench is a first step to facilitate efficient research and large-

scale collaborative work. The data can be pooled into centralized registries or operate as linked, independent, federated silos on which an application layer can operate to form and distribute queries and aggregate and summarize findings across multiple institutions electronically. This is the long-term vision of the COHA Information Technologies infrastructure that is being implemented through its public COHA website and the COHA Research Workbench online web-based application.

Acknowledgments

The contributions of the members of the COHA database leadership and technical committees are gratefully acknowledged: Chris Brandt, Jeff Bryan, Michael Cinkosky, Colleen Duncan Kelly Hall, Elle Holbrook, Majid Jaberi-Douraki, Michael Kahn, Warren Kibbe, Sarah Moore, Wayde Shipman, Joe Strecker, Sue Vandewoude, Alison Zwingenberger.

References

1. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc* **20:**117–121, 2013
2. Bellinger C, Mohomed Jabbar MS, Zaiane O, Osornio-Vargas A. A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health* **17:**907, 2017
3. Meyfroidt G, Guiza F, Ramon J, Bruynooghe M. Machine learning techniques to examine large patient databases. *Best Pract Res Clin Anaesthesiol* **23:**127–143, 2009
4. Zhang Y, Guo SL, Han LN, Li TL. Application and exploration of big data mining in clinical medicine. *Chin Med J* **129:**731–738, 2016
5. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* :182716, 2019
6. Doherty C, Camiña S, White K, Orenstein G. *The Path to Predictive Analytics and Machine Learning*. 1 ed. Sebastopol, CA: O'Reilly Media, Inc.; 2016
7. Law GC, Apfelbacher C, Posadzki PP, Kemp S, Tudor Car L. Choice of outcomes and measurement instruments in randomised trials on eLearning in medical education: a systematic mapping review protocol. *Syst Rev* **7:**75, 2018
8. Bonardi A, Clifford CJ, Hadar N. A Structured Approach Using the Systematic Review Data Repository (SRDR): building the evidence for oral health interventions in the population with intellectual and developmental disability. *Eval Rev* **41:**111–129, 2017