Topics in Companion An Med 37 (2019) 100363

Research Article

TRANSLATOR Database—A Vision for a Multi-Institutional Research Network

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

[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, North Grafton, MA, USA*
[d]*Animal Bioscience Inc, Boston, MA, USA*

A B S T R A C T

The formation of the CTSI One Health Alliance (COHA) network has generated the infrastructure necessary to support "Big Data" collaborative comparative and translational research in veterinary medicine. We describe the first step in the design, implementation, and dissemination of a collaborative information technology infrastructure that will serve the public and clinicians (COHA public/member based web site at https://ctsaonehealthalliance.org/) and its research focused COHA Research Workbench application. The core research infrastructure, TRANSLATOR (**TR**anslational **AN**imal **S**hared Col**LA**bora**T**ive Observational Research), represents the foundation of a federated research-capable network to enable pooling large datasets from both electronic health records and publications. The public facing COHA website is a mechanism for both the dissemination of knowledge to the public and to foster collaborations amongst veterinary clinician scientists and the greater medical research community.

Background—One Health Dataset Workshops

Two workshops were convened in January, 2018 and February, 2019 focused on the development and implementation of an OMOP data platform. The steering committee for the "COHA Advancing One Health Datasets" project 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, and 2 veterinary databases. The session was supported by a COHA grant and sponsorship by Morris Animal Foundation and Mars Petcare.

The interests of the participants in the development of a data platform were identified as translational research and data sharing. The first workshop provided a synopsis of an implementation of a common data model (CDM) in human medicine and early pilot results of OMOP implementation at a single site, Tufts Cummings Veterinary Medical Center. Background on existing veterinary medical datasets, for example, the longstanding Veterinary Medical Database (VMDB) and American College Veterinary Emergency and Critical Care-Veterinary Committee on Trauma (ACVECC-VetCOT) was presented. Future opportunities in data science related to veterinary medicine were brainstormed. Three areas of further development were recognized: challenges and opportunities relating to veterinary data harmonization; development of use cases that could be realized from veterinary data harmonization efforts, for example, obesity; and finally, sustainability of the data platform—including governance, engagement, and financial/resource needs. Technical and leadership subcommittees were formed to examine 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.

Several examples of successful human data platforms have come into existence over the past 20 years, for example, PORTAL (Patient Outcomes Research to Advance Learning), Vaccine Data Link, Cancer Research Network, SUPREME-DM (Surveillance, Prevention, and Management of Diabetes Mellitus). Case studies in area of comparative effectiveness research, medical product safety surveillance and public health monitoring have shown the utility of these health data research networks.

The formation of TRANSLATOR therefore creates the opportunity to utilize observational data found in EHR systems among the COHA network members. The 2 short-term challenges in realizing the rewards from this opportunity include optimizing the use of the data locally and collaboratively and its governance.

COHA Research Infrastructure

Many if not all Clinical Translational Science Awardees have embarked on building the capacity to support "big data" translational research (86% in 2010 survey of Clinical Translational Science Awardees),[1] leveraging large volumes of patient visit data for observational/retrospective research to complement prospective clinical trial-based research.[2,3] Observational research takes advantage of data readily available within electronic health record (EHR) systems representing data captured in "real-world" clinical settings. Large and small clinical trials typically have a narrow focus and operate

* Corresponding author. Manlik Kwong, Tufts Medical Center, 800 Washington St., Boston, MA 02111
E-mail address: mkwong@tuftsmedicalcenter.org (M. Kwong).

Abbreviations: ACVECC-VetCOT, American College Veterinary Emergency and Critical Care-Veterinary Committee on Trauma; CDM, Common Data Model; CLR, Case Literature Review; COHA, CTSI One Health Alliance; COHAWB, COHA Research Workbench; CTSA, Clinical and Translational Science Award; CTSI, Clinical and Translational Science Institute; EHR, Electronic Health Record; FHIR, HL7 Fast Healthcare Interoperability Resource; ETL, Extract/Transform/Load; i2b2, Informatics for Integrating Biology & Bedside; IT, Information Technology; LOINC, Logical Observation Identifiers Names and Codes; MeSH, Medical Subject Headings; OHDSI, Observational Health Data Sciences; OMOP, Observational Medical Outcomes Partnership; PCORNet, The National Patient-Centered Clinical Research Network; PMID, PubMed ID; SNOMED, Systematized Nomenclature of Medicine; PORTAL, Patient Outcomes Research to Advance Learning; VMDB, Veterinary Medical Database

http://dx.doi.org/10.1016/j.tcam.2019.100363
1938-9736/

within a defined clinical context, which may not be generalizable across a broader patient population (e.g., teaching medical center vs. a community clinic). The formation of the CTSA One Health Alliance (COHA https://ctsaonehealthalliance.org) in 2014 with its initial 15 member veterinary institutions and partners laid the foundation to apply an informatics-based "big data" [4] research strategy to comparative veterinary research. Employment of informatics approaches to analyze large data sets creates opportunities to leverage available information to ask numerous hypothesis-driven questions, ranging from quality improvement, research, patient engagement, public health surveillance, clinical decision support, among others. This database is the first national effort of its kind across veterinary medical colleges. In addition to supporting research collaboration teams, the COHA research infrastructure facilitates dataset analysis across broad geographic areas, permitting more comprehensive analysis of datasets to answer complex questions. With integrated feedback of derived knowledge, this infrastructure also supports the future goal of building a continuously learning healthcare infrastructure.[5-7]

A CDM that supports multi-institutional collaboration through sharing of data, tools, and research findings across a virtual network is critical for the continued growth of translational research. Importantly, data can be shared while maintaining security and relative independency of data sources and local EHR documentation practices. While there are many CDMs utilized across human research networks[8] (i2b2—https://www.i2b2.org/, PCORNet—https://pcornet.org/, FHIR—https://www.hl7.org/fhir/overview.html, etc.), in line with our respective human medical center and CTSI counterparts, we have adopted the Observational Health Data Sciences and Informatics' (OHDSI—https://www.ohdsi.org/) Observational Medical Outcomes Partnership (OMOP) common data model (CDM) (Fig 1) at Tufts Cummings Veterinary School (TCSVM), Colorado State University Veterinary School (CSU), and UC Davis Veterinary School (UC Davis). By maintaining the same OMOP CDM across both human and veterinary domains, queries initiated in human or veterinary patient records will utilize the same standard vocabularies and concepts (e.g., SNOMED [Systematized Nomenclature of Medicine], RxNORM, LOINC [Logical Observation Identifiers Names and Codes]), and where necessary, custom vocabularies and concepts that are veterinary specific.

A primary characteristic of the OMOP CDM is the requirement that all inbound data be normalized and mapped (translated from its native EHR descriptive form) to a common language (e.g., defined vocabularies and concepts). While this requirement represents significant up-front resources and effort in terms of designing Extraction/Transform/Load (ETL) data handling software and processes, the benefits become readily apparent in the context of the multi-institutional network of OMOP CDMs. For example, a query designed at one institution can readily be distributed to other OMOP CDM sites for execution with defined cohort specifications (clinical observations, drug exposures, device exposures, conditions, measurements, etc.) will have the same meaning across all OMOP CDM databases. Constructing a data research warehouse using a shared vocabulary of concepts also makes the data more portable. Portability in this sense means the operational research database can be put in various network environments as dictated by resource and technology governance.

Each institutional member of the COHA network has available a number of implementation options to create and manage its own OMOP CDM on: (i) local servers; (ii) private cloud hosted OMOP CDM servers; or to (iii) contribute to a centralized OMOP CDM registry. OMOP CDM registry contribution is based on an individual institutions information technology capabilities, resources, and security policies and practices. For example, choosing a local OMOP CDM may make sense for an institution that is starting to build its ETL processes and want direct and efficient movement of data from its source EHR system into the new OMOP CDM warehouse. This approach is focused on data mapping and learning the ropes of managing a research warehouse. A cloud deployment may be suitable for those institutions that have worked out the ETL processes locally and have a fully loaded OMOP CDM ready to participate in multisite clinical trials. With data use agreements and operational procedures in place, snapshots of the institutions OMOP CDM can then be uploaded to the cloud service provider at defined intervals (quarterly, monthly, etc.). For institutions that have been managing a pre-existing veterinary registry, they may choose to transform the registry data into a centralized OMOP CDM. Those institutions that do not have the IT infrastructure and resources to manage a local OMOP CDM may also choose to upload their normalized/mapped data to a centralized

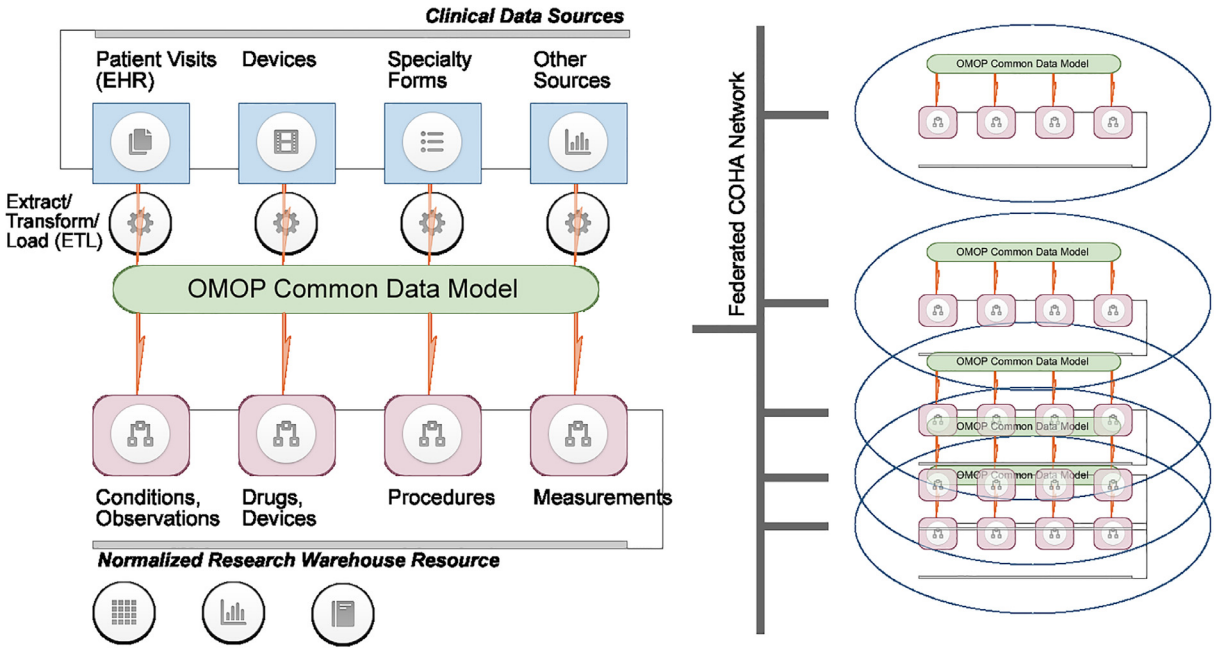

Figure 1. OMOP based research infrastructure.

OMOP CDM in order to participate within the COHA research network without taking on significant local IT infrastructure and support costs. Therefore, the power of this research infrastructure lies in the ability of all tools and queries to operate across all institutional configurations. This provides COHA members flexibility with respect to data use, control, security, management, and IT infrastructure investment needed to participate in multicenter trials.

The next level of technology COHA that is developing in parallel to the OMOP CDM data infrastructure is a resource of research tools encapsulated within an online web-based application called the COHA Research Workbench (COHAWB). The primary goal of the COHAWB is to facilitate comparative and translational research for the benefit of both species. Similar to evidence generation through systematic reviews[9] and data extraction from published literature (i.e., PubMed.gov database and case literature reviews), research databases can also be derived from the complementary observational data generated by clinical EHR systems, including patient visit information, radiology/imaging data, laboratory, and specialty-specific visit data. The COHA network forms the basis for multicenter collaborations, with both human and veterinary data prepared and formatted in a common CDM to readily support multi-institution and multidomain evidence-based research. In addition to providing access to patient clinical data, the COHAWB platform also provides a mechanism to incorporate data from published literature through PubMed.gov. To date, over 320,000 of the 546,826 available literature abstracts representing articles on animal Case Literature Reviews (CLR 23,971), published studies involving cats (133,758), dogs (321,631), and horses (67,466) have been imported from PubMed.gov. The complete 546,826 will be imported by the end of 2019 and made available on the COHAWB application. These uploaded abstracts, MeSH, and Substance terms, as well as EHR-sourced data will automatically be mapped to the OMOP CDM vocabularies and concepts and made available within the COHAWB to provide complementary evidence-based resource to the observational based EHR records. Initial tools available through the COHAWB enable cohort discovery from 2 sources of data: (i) Downloaded PubMed.gov abstracts representing 546,826 veterinary literature reviews (CLRs and clinical trials) parsed and mapped to OMOP vocabulary concepts; and (ii) Local and federated EHR records parsed and mapped to OMOP vocabulary concepts. MeSH and Substance terms from published articles are mapped to OMOP's MeSH and SNOMED vocabularies and managed in an OMOP extension table to hold the publication data (PMID, PMCID, Title, Authors, Abstract, MeSH, Substance, etc.). Similarly, veterinary EHR records are managed within OMOP CDM tables (with extensions to support species, breed, and owner data) utilizing OMOP CDM vocabularies such as SNOMED, LOINC, RxNorm vocabularies. This serves to normalize the patient's condition, observations, drug exposures, device exposures, procedures, and measurements documented within the EHR. Therefore queries based on OMOP vocabularies and concepts (e.g., splenectomy, lyme disease, cancer, etc.) are common across both CLR and EHR OMOP records. CLR can provide evidence of gaps in knowledge, emergent clinical issues, and current clinical trial findings. When applied to EHR records, the same query can provide data on the incidence of a defined patient population across multiple institutions, thereby aiding researchers in the early stages of development of a research project. The aggregation of patient data into a single CDM also enables investigators to view and explore individual patients within a contextual perspective[10,11] in both events (observations, drug/device exposures, procedures, measurements) and time (e.g., adherence to scheduled cancer treatments, progression of measurements and observations).

Expanding the OMOP CDM across all 15 COHA members will generate the necessary volume of patient records to enable informatics approaches employing machine learning to complement ongoing comparative and translational research. The COHAWB tools and resources will also evolve from its current role in the initial steps of research, to include collaborative patient identification, incorporation of unstructured clinical notes, and continuous data reporting. A prototype of human-machine collaborative patient identification within the COHAWB uses machine learning to generate a profile of a study cohort of interest. The profile is then used to assess new patients entering the OMOP CDM to determine whether the new patient is a suitable candidate for review and inclusion into the cohort of study. As each candidate patient is reviewed and accepted to participate in a cohort by the investigator, the machine automatically updates its cohort profile and inclusion rules. Therefore, over time the initial high level of human participation decreases as the machine forms (or learns) a more accurate profile and inclusion criteria, thus reducing the need for the investigator to review every candidate. These profiles are based on the OMOP concepts and therefore can be readily shared with other COHA members who are participating in formation of the same study cohort. This approach may be particularly useful for capturing infrequent clinical events or rare patient visits, which are often relegated to small case series or case reports in the veterinary literature.

COHA Public Website

In addition to providing the COHA research community with a federated research platform to support multisite research initiatives, the public will have access to COHA educational resources through the COHA public website (https://ctsaonehealthalliance.org/). Veterinarians, physicians, academics, and researchers with an interest in studying diseases that affect both animals and people can register for an account through the COHA website. The COHA website provides resources for researchers, including biobanking, ongoing clinical trials, educational opportunities and training fellowships, outreach efforts, and news stories. Notably, training resources to provide standardized training in clinical trial implementation in veterinary medicine are available, in addition to publications and overviews of animal diseases with shared features of the analogous human condition. The website content is generated by the COHA community, with members submitting content and resources of interest to other members.

Data Resource Governance

The expansion and utilization of EHRs in veterinary medicine have supported collection and use of a wide variety of data from clinical, clinical pathologic, pathologic, pharmacy, financial, and environmental. To our knowledge, no research networks include all of these veterinary data at present. At the crux of the successful development and use of such a research platform are the following questions:

1. "If I had access to a database with cross institutional data in the same format, it would advance my research by…"
2. "A multi-institutional veterinary database linked to human, environmental and other data would allow me to…."
3. "What question can't be answered with the database from a single institution and what is clinician researcher willing to give up to use a combined database?

The answers to these questions lie in the confidence of the clinical researcher in the governance of the database. Large population-based studies derived from a research network drive the need for diligent governance policies and procedures.

A second workshop was conducted in February, 2019 with 2 subcommittees (governance and technical) to draft a governance document for TRANSLATOR and to advance the technical infrastructure, respectively. This meeting included veterinary researchers and clinicians, a physician scientist and IT and database specialists. COHA

members represented Colorado State University, University of Colorado Denver, Kansas State University, Ohio State University, Tufts University and University of California.

We view governance as the high-level policies, guidance, and strategies that define how the Network collaborates and makes decisions. By specifying such engagement governance practices proactively, we intend to establish trust and transparency to promote high-quality and efficient scientific collaboration. Governance is never complete, and the Governance document will be a "living" document—a work in progress.

The bedrock principle of TRANSLATOR governance is preservation of the privacy and security of veterinary medical/health information and the legitimate proprietary interests of our contributing partners. The TRANSLATOR Network will honor the willingness of its animal owners, members, and clinical entities and colleagues to share their information and to develop knowledge jointly. We will reward that trust with exemplary stewardship.

An aspirational goal of the TRANSLATOR Network is to develop an infrastructure through which animal health data can be integrated with human health data to create multispecies datasets to support One Health initiatives, for example, biomarker discovery, zoonotic and public health diseases, environmental factors in incidence of disease, treatment outcomes etc. The comparative datasets are to be used to generate hypotheses and identify relevant research opportunities. As an effective governance document, it also provides a template for addressing issues as they arise. Based on these goals, we drafted the document incorporating these guiding principles. The governance document provides an overview of the governance process for the TRANSLATOR Network. It describes the organizational structure and committees of TRANSLATOR. It also addresses decision-making within TRANSLATOR; the approach to data governance, data privacy, and confidentiality; conflicts of interest, and scientific misconduct. This governance plan incorporates relevant regulations and policies of the institutions at the levels of the contributing partners and network, building on the rich experience of other networks and evolving national standards for human networks.

We envision a TRANSLATOR Steering Committee to approve and administer the Governance Plan and ongoing revisions. Standard operating procedures and specific activities to carry out the governance policies, guidance, or strategies will be developed and implemented by committees in the network. We foresee a TRANSLATOR Advisory Council to provide input and guidance.

The TRANSLATOR research network will collaborate with clinicians and operational leaders to develop a high-functioning clinical data research network to help support One Health/One Medicine initiatives. TRANSLATOR network will solicit and develop scientifically sound cohorts which can support specific comparative and translational research, for example, observational studies and other veterinary clinical research and trials. The TRANSLATOR Network will emphasize the engagement of clinicians, animal owners and operational leaders in network governance and commits to consistent communication with our stakeholders.

Following this foundational work, the COHA members and other relevant stakeholders will need to review the proposed governance and technical infrastructure and make decisions on the implementation plan for the addition of data (i.e., aggregated vs. line item) from new sources. Two pressing decisions to be made are in relation to open access and financial sustainability. Should the research outputs be distributed free of cost or other barriers, and unrestricted in use or should the research be partially or fully restricted? Considerations for the financial model to date are pay for use, grant or sponsorship.

Conclusions

A unique research infrastructure is being developed to facilitate comparative and translational research. Through implementation of web-based tools to transform clinical EHR data and extract data from published literature, COHA member institutions have an unprecedented opportunity to engage in collaborative multicenter studies and research teams.

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. MacKenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N. Practices and perspectives on building integrated data respositories: results from a 2010 CTSA survey. *J Am Med Inform Assoc* **19:**e119–e124, 2012
2. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc* **20:**117–121, 2013
3. Zhang Y, Guo SL, Han LN, Li TL. Application and exploration of big data mining in clinical medicine. *Chin Med J (Engl)* **129:**731–738, 2016
4. Ross MK, Wei W, Ohno-Machado L. Big data and the electronic heath record. *Yearb Med Inform* **9:**97–104, 2014
5. Daudelin DH, Selker HP, Leslie LK. Applying process improvement methods to clinical and translational research: conceptual framework and case examples. *Clin Transl Sci* **8:**779–786, 2015
6. Daudelin DH, Selker HP. Medical error prevention in ED triage for ACS: use of cardiac care decision support and quality improvement feedback. *Cardiol Clin* **23:**601–614, 2005
7. Daudelin DH, Saya AJ, Kwong M, et al. Improving use of prehospital 12-Lead ECG for early identification and treatment of acute coronary syndrome and ST-elevation myocardial infarction. *Circ Cardiovasc Qual Outcomes* **3:**316–323, 2010
8. Rosenbloom ST, Carroll RJ, Warner JL, Matheny ME, Denny JC. Representing knowledge consistently across health systems. *Yearb Med Inform* **26:**139–147, 2017
9. Ip S, Hadar N, Keefe S, et al. A web-based archive of systemic review data. *Syst Rev* **1:**15, 2012
10. Hsu W, Taira RK, El-Saden S, Kangarloo H, Bui AA. Context-based electronic health record toward patient specific healthcare. *IEEE Trans Inf Technol Biomed* **16:**228–234, 2012
11. Hsueh PY, Ramakrishnan S, Yu K, Akushevich M, Sharma S, Mooiweer P. Development of temporal context-based feature abstractions for enabling monitoring and managing of interventions. *IEEE Stud Health Technol Inform* **205:**471–475, 2014