Research and Applications

Using UMLS for electronic health data standardization and database design

Andrew P. Reimer^{1,2} and Alex Milinovich³

¹Frances Payne Bolton School of Nursing, Case Western Reserve University, Cleveland, Ohio, USA, ²Critical Care Transport, Cleveland Clinic, Cleveland, Ohio, USA, and ³Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, Ohio, USA

Corresponding Author: Andrew P. Reimer, PhD, RN, Cfrn, Frances Payne Bolton School of Nursing, Case Western Reserve University, 2120 Cornell Dr, Cleveland, OH 44106, USA (axr62@cwru.edu)

Received 10 February 2020; Revised 8 July 2020; Editorial Decision 10 July 2020; Accepted 21 July 2020

ABSTRACT

Objective: Patients that undergo medical transfer represent 1 patient population that remains infrequently studied due to challenges in aggregating data across multiple domains and sources that are necessary to capture the entire episode of patient care. To facilitate access to and secondary use of transport patient data, we developed the Transport Data Repository that combines data from 3 separate domains and many sources within our health system.

Methods: The repository is a relational database anchored by the Unified Medical Language System unique concept identifiers to integrate, map, and standardize the data into a common data model. Primary data domains included sending and receiving hospital encounters, medical transport record, and custom hospital transport log data. A 4-step mapping process was developed: 1) automatic source code match, 2) exact text match, 3) fuzzy matching, and 4) manual matching.

Results: 431 090 total mappings were generated in the Transport Data Repository, consisting of 69 010 unique concepts with 77% of the data being mapped automatically. Transport Source Data yielded significantly lower mapping results with only 8% of data entities automatically mapped and a significant amount (43%) remaining unmapped.

Discussion: The multistep mapping process resulted in a majority of data been automatically mapped. Poor matching of transport medical record data is due to the third-party vendor data being generated and stored in a nonstandardized format.

Conclusion: The multistep mapping process developed and implemented is necessary to normalize electronic health data from multiple domains and sources into a common data model to support secondary use of data.

Key words: data management, data curation, electronic data processing, data warehousing, transportation of patients

INTRODUCTION

The large-scale adoption and widespread availability of electronic health record (EHR) data has accelerated the use of patient data beyond clinical and administrative purposes. In particular, EHR data are now able to be rapidly ingested, analyzed, and repurposed to support clinical

practice via applications such as clinical decision support systems, supporting progress toward realizing a learning health system that uses data in near-real time.¹ While extracting, transforming, and using data within a single institution or health care system with 1 EHR system can be challenging, incorporating data from multiple information systems that can include third-party EHRs, presents even more challenges.²

Health systems are continuing to consolidate specialty services at urban-based large academic medical centers, increasing the necessity for patients that require comprehensive or specialty services to undergo transfer via ambulance or helicopter.^{3,4} Supporting transferred patients via real-time clinical decision support or secondarily via research is complicated due to the challenges of accessing data from multiple domains and sources that can include in-house developed information gathering systems, commercial EHRs, and third-party vendor transport EHRs.

The primary challenge in extracting and normalizing EHR data outside of the original data collection environment is the nonstandard data coding approach, sometimes proprietary, that each system uses. Further, installations of the same EHR at different institutions can result in differing coding structures for data entity storage and retrieval. Thus, significant effort has been focused on developing ontologies and common data models capable of identifying and normalizing discrete data entities into a common form that enables aggregating data into 1 unified structure that supports large-scale use.

Ontology development ranges from specific use case applications to broad use common data models. An example of individualized ontology development is the Ontology for Cancer Research Variables⁵ that was specifically developed to link data from heterogeneous datasets to support cancer research projects. Additional approaches include leveraging existing semantic terminologies, such as SNOMED CT,⁶ to facilitate modelling data by enhancing the terminology by adding additional data standards, such as LOINC and RxNORM, to enable searching additional relationships of associated patient data.⁷ One limitation of these approaches is the amount of work required to work within 1 or merge several standardized terminologies into a common data model.

Broader scale approaches to EHR data management and normalization have developed comprehensive common data models that are used nationally to support large-scale data aggregation projects and include i2b2,⁸ PCORNet,^{9,10} and Observational Medical Outcomes Partnership (OMOP)^{11,12} (further descriptions provided here¹³). Each common data model provides differing capabilities ranging from PCORNet providing a common data model that requires local data extraction and normalization, to i2b2 that normalizes data ingested into the platform with querying capabilities, to OMOP that provides additional analytic capabilities including patient level prediction and population estimation. However, depending on the model used, there may be limited-to-no ability to create custom entries or incorporate data mappings from other standardized terminologies or new data sources into the common data model that is exportable outside the local application.

Another approach to normalizing data is the National Library of Medicine's Unified Medical Language System (UMLS) that provides a robust index of terminologies and standards in 1 framework that enables incorporating data regardless of how the data are originally structured and stored. For example, structured data stored as ICD-10 codes will map to a unique concept identifier (CUI) in UMLS. That same UMLS concept can also be matched to free text extracted from an unstructured clinical note that exactly or closely matches the text contained in the concept description. UMLS resources have been applied in various applications to enable data and/or concept identification, annotation, and phenotype extraction that simultaneously leverages multiple terminologies.^{14–17} A strength of UMLS is the ability to locally customize tables and update or add new concepts when needed.

This report describes the application and use of the UMLS as the primary coding standard to identify and arrange structured and unstructured data within a Transport Data Repository. To address the

paucity of research regarding transfer patients, we developed the Transport Data Repository using a simple relational database design. In this article, we describe the development and application of a multistep mapping process based on the UMLS Metathesaurus¹⁸ to identify and standardize data from each data domain and source, using the unique UMLS concept identifier as the primary data organizer. We describe the extracting and normalizing data workflow implemented via a multistep mapping approach, describe data handling procedures including developing custom data-mapping user interfaces to handle unmapped data, and present initial data-mapping results.

MATERIALS AND METHODS

Repository design

The Transport Data Repository is a simple relational database. The repository combines data from 3 primary domains: 1) critical care transport log, an in-house developed program that contains every request for transfer and related data (ie, referring and receiving destinations, transfer mode, times, etc); 2) EHRs (ie, Epic) and administrative data sources from the health system Research Data Warehouse that contains UMLS-mapped EHR data of patients for referring and receiving hospital encounters; and 3) transport EHR—in this case, Golden Hour/emsCharts.

Patient matching

Data extraction is initiated via the Critical Care Transport Log data table (Figure 1, #1). Each patient transport results in referring and receiving hospital encounters, unless the transport originated in the prehospital setting that is the case for trauma or mobile stroke unit transfers. When the Critical Care Transport Log identifies a patient that was transferred exclusively within the health system, a patient-matching algorithm attempts to match patients in 2 phases. The first phase consists of an exact match based on patient last name, first initial, date of birth, and medical record number (MRN). Because the Transport Log is an in-house designed system, we included the ability to search and input existing patient MRNs from the health system admission/discharge/transfer system while generating a new patient record within the Transport Data Repository, resulting in approximately 90% of patient encounters and transports automatically matching on first attempt. A typical cause of the patient MRN not matching was when the patient was new to the health system and a new MRN was generated at the time of the encounter but later identified as already having an existing MRN that was reconciled administratively after the initial transport log entry was created.

Unmatched patient transfers are sent to a manual matching cue (Figure 1, #2) and are manually mapped using the unmatched encounters user program that presents the unmatched transfer and suggested encounters as potential matches on an individual patient level. Suggested encounters are generated by implementing a broader matching algorithm that compares the transport date/time to all existing patient encounters and transfer date/time (± 24 hours from recorded transport times to enable for differences in admission/discharge transfer/data entries between discharge and admission). The wide calipers around the transport date/time are necessary to handle cases where conflicts between the sending encounter discharge date/time conflict with the transport date/times and/or the receiving hospital encounter admission date/time. Typical conflicts include failure of the sending hospital to discharge the patient upon transport team arrival that results in a discharge date/time after the transport date/time—thus precluding automatic

the text “foot pain” but not be stored with an associated ICD-10 code. An exact text match would result in a match to “foot pain” UMLS concept C0016512. Remaining unmatched data progress on to a fuzzy-matching algorithm (Step 3) that recursively searches and compares the data entity with data entities contained in the UMLS atoms table to find potential matches. The UMLS generated atoms table contains each UMLS concept and all associated mappings to that concept from other standardized terminologies. Fuzzy matching identifies and automatically assigns candidate matches that yield a similarity threshold $\geq 90\%$ and constrained within an appropriate semantic type. A more detailed explanation of the similarity threshold and additional matching metrics are provided here.¹⁹ For example, a diagnosis may appear in the raw data as “foot pain left” with no exact match in UMLS. However, “pain in left foot” is a concept in UMLS (C2141243) with 9 associated atoms for potential mapping and achieves a similarity index score of 0.9749, meeting the prespecified threshold thus resulting in a successful match and addition to the diagnosis table. Thus, the diagnosis, “foot pain left” is mapped to UMLS CUI C2141243 with the associated concept description of “pain in left foot” and added to the diagnosis table. Data entities that receive a similarity index score $< 90\%$ are reviewed for approval or passed on to the fuzzy-matching step.

Remaining unmatched data are fed into a cue (Figure 1, #5) for manual matching/assignment via 1 of 2 approaches: 1) additional ad hoc programming or 2) manual-matching user interface. First, reviewing unmatched data often results in identifying certain patterns that precluded the data from being successfully mapped via the automated processes. Sometimes the solution is a simple addition or deletion of unnecessary data or recoding of the source data that results in data entities being automatically mapped via the coercive code. Remaining unmapped data entities are cued to a concept-matching user interface. The manual-mapping user interface presents unmatched data entities (Figure 2). The user interface enables clinicians and researchers to view the unmapped data entity, associated sample values, and locations of the data entity within the EHR to provide contextual understanding when necessary. Mapping the data entity to a UMLS concept is accomplished by either clicking on an unmapped concept from the dashboard (Figure 2A, #1) or searching for a specific concept via text entry in a search box (Figure 2A, #2). Once either option is selected the next screen displays the concept of interest with several options to aid exploring the concept in context (Figure 2B, #3). Searching for a specific concept to map to is accomplished by selecting the “search UMLS” option icon (Figure 2B, #4) that presents several search options and associated search results of UMLS concepts via a pop-up menu (Figure 2C). Viewing possible matches is achieved by inputting the data entity, in this example “foot pain left” (Figure 2C, #5), constraining the search via a specific semantic type if necessary (Figure 2C, #6), and reviewing and selecting a match (Figure 2D, #7). Once selected and confirmed, the new mapping is entered into the table with the newly associated UMLS CUI. Alternatively, concepts may remain unmapped if they are deemed to not provide value such as mapping a particular hospital name to a high-level concept such as Hospitals (C0019994) that has relationships to specific types of hospital encounters. In these cases, mapping the concept substantially diminishes the information contained in the data entity and that concept would remain an unmapped concept.

Lastly, a custom concepts table is created to handle all other concepts not contained within UMLS. Custom entities are unique UMLS-like IDs created within our local installation of the UMLS Metathesaurus. Examples include creating unique codes for pro-

viders and service locations that do not exist natively in UMLS. This allows us to standardize naming conventions as well as to query relationships between locations/providers and their specialties in the same fashion as querying UMLS relationships.

RESULTS

Table 1 presents results of data entities progressing through the data-mapping process for Transport Data Sources, Transport Data Repository, and the health system Research Data Warehouse. The percentages in each step represent the number of data entities effectively processed in each step. The results for source code and same source exact text match indicate the number of data entities matched in that particular step. Results for fuzzy-matching and manual-mapping steps include data entities that can remain unmatched after either not receiving a high enough similarity score, as is the case for a large percentage of Transport Source Data, or remain as an unmapped data entity after manual review. For example, a majority of the suggested fuzzy-matching results are below the 90% similarity threshold for Transport Source Data, resulting in those data entities progressing to the manual-mapping cue.

Currently there are just over 10 million total mappings from 242 sources of data covering 1 482 507 unique concepts from health system EHR and administrative data sources. Of the remaining unmapped data entities, approximately 2.5 million originate from Epic and are either system generated data such as classifiers for type of patient admission or fields within the data table entered as a generic “unspecified.” A majority (57%) of health system data is mapped automatically via source code or same source exact matching, while only a small percentage (8%) are mapped via fuzzy-matching or manual approaches.

Transport data, from both the critical care transport log and transport EHR, matched poorly with only 8% of data entities being mapped by source code or same source exact text match. A majority of Transport Source Data (51%) underwent manual mapping, with a large proportion of data (43%) triaged to no mapping. The remaining unmapped Transport Source Data entities from the manual mapping cue consisted of sending hospital names, receiving hospital unit names (eg, G60 medical intensive care unit), and “trip cancelled” categorizations.

A majority of Transport Data Repository data entities were mapped via source code (75%) as it was originally mapped and pulled from the Research Data Warehouse for patient hospital encounter data. Additionally, a large percentage (62%) of data entities were handled by creating custom concepts, as transport-specific concepts not contained in UMLS were normalized into the data repository.

DISCUSSION

Our results indicate that using the UMLS architecture to anchor an EHR research data repository is feasible and results in a majority of data entities being automatically mapped when the data originates from nontransport sources. Specifically, the low yield in automatic mapping in the Transport Source Data was due to normalizing data from a third party EHR that served as the medical transport EHR that used a unique coding structure specific to the vendor and not associated with existing standardized terminologies. This is evident in 46% of the data resulting in potential mappings during the fuzzy-matching step and a majority of data entities having to be handled

Figure 2. (A) Unmapped concepts dashboard. The manual mapping user interface presents a list of unmatched data entities for review. Review is accomplished by either clicking on an unmapped concept from the dashboard (#1) or searching for a specific concept via a search box (#2). (B) Individual concept mapping screen. Screen displaying 1 concept for review and associated selections available to gain contextual knowledge of the concept that include sample values, location of the data, etc (#3, #4). (C) UMLS concept and atoms search pop-up box. Searching for potential mappings is accomplished by entering the concept (#5) and constraining potential matches via exact match or length and selecting an applicable semantic type displayed in the drop-down selection (#6). (D) UMLS concept match selection (#7).

C

Review Data Mapping
Review mapped data for validity

SUCCESS: The concept has been mapped by code

WARNING: The mapping has not been approved

Review Mapping

Pain in left foot
value

240104659
clarity edg | dx id

M79
co

Percent Similarity
0

Percent Confidence
0

CHANGE MAPPING

Sample Values

Search UMLS

Concept : foot pain left #5

Exact

Sort Len

Semantic Types :

Acquired Abnormality #6

Activity

Age Group

Amino Acid Sequence

Amino Acid, Peptide, or Protein

Amphibian

Anatomical Abnormality

Anatomical Structure

Animal

Antibiotic

Concepts returned: 0

D

Review Data Mapping
Review mapped data for validity

SUCCESS: The concept has been mapped by code

WARNING: The mapping has not been approved

Review Mapping

Pain in left foot
value

240104659
clarity edg | dx id

M79
co

Percent Similarity
0

Percent Confidence
0

CHANGE MAPPING

Sample Values

Search UMLS

Concept : foot pain left

Exact

Sort Len

Semantic Types :

Select

medial left foot pain C2128597

pain in the left foot on the inner side C2128597

left foot pain shooting up the inner leg C2128652

left foot pain #7 C2141243

left foot pain (symptom) C2141243

pain in the left foot C2141243

pain of fifth toe of left foot observed on ambulation (physical finding) C2144336

short-term physical therapy goals reduce pain of metatarsophalangeal joints of left foot to specific level C2148239

short-term physical therapy goals reduce pain of metatarsophalangeal joints of left foot to specific level (treatment) C2148239

Concepts returned: 100

Figure 2. (Continued)

manually as fuzzy matching resulted in poor mapping suggestions. A large percentage of data from transport sources consists of transport logistics data that include sending and receiving hospital and provider names, unit designations, and date/time stamps that exist as individual entries within each table—all data entities that do not require an explicit mapping, thus the large number of unmapped data.

Our multistep mapping process combines several individual mapping approaches into 1 comprehensive process. The first 2 steps

incorporate matching functions similar to Metamap²⁰ or Sophia.²¹ Because of the multistep process, most data entities are mapped to an existing UMLS concept automatically, overcoming common problems, such as the use of nonstandard concept descriptions,² misspellings, and incomplete or misordered words. Remaining unmapped data entities proceed to the fuzzy-matching step that is similar to LexSynonym²² and relies on recursive partitioning to find and suggest the most appropriate matches based on similarity to

Table 1. Data entity mapping results

	Transport source data		Transport data repository		Research data warehouse	
	n	%	n	%	n	%
Total mappings	14 039 ^a		431 090 ^a		10 025 371 ^a	
Unique data entities (unique concepts)	7577 (2413)		266 934 (69 010)		7 088 651 (1 482 507)	
Unique unmapped entities	5988	43	104 332	24	2 465 157	24
Custom entities	932	7	265 479	62	1 076 033	11
Data-mapping results						
Source code	367	3	322 221	75	4 252 627	42
Same source exact text match	625	5	9049	2	1 547 548	15
Fuzzy matching ^b <i>Similarity mean [IQR]</i>	5817	46	40 941	9	683 645	7
	0.53		0.72		0.66	
	[0.47, 0.76]		[0.59, 0.90]		[0.52, 0.82]	
Manual mapping ^c	7230	51	1268	<1	126 824	1

All percentages are based on comparison to total mappings, thus column percentages may not equal 100% as a data entity can be represented in more than 1 mapping step.

Transport Source Data: critical care transport log and transport electronic health record; Transport data repository: contains all data sources related to transport from transport data source and health system electronic health records related to transports; Research Data Warehouse: all health system electronic health data sources and associated data entities normalized to UMLS structure; n: data entities successfully processed; (%) = percent of data entities successfully processed; IQR: interquartile range.

^aSome mappings are 1-to-many.

^bData entities may remain unmapped after fuzzy matching and progress to the manual-mapping cue.

^cData entities passed to manual mapping can remain unmapped and reflected in the total for unique unmapped entities.

existing UMLS concepts and atoms. In the Transport Data Repository and Research Data Warehouse, only a small percentage of data, 9% and 7% respectively, are identified and mapped during this step. As these data repositories contain data entities from both structured (eg, ICD-10 and current procedural terminology) and unstructured (eg, free-text diagnosis entry in clinical note) sources, further work is necessary to improve performance of the fuzzy-matching step by either supplementing or replacing the approach with additional unsupervised approaches, such as segment convolutional neural networks.²³

Similar to the OMOP common data model, anchoring our data repository using UMLS resources stores the data in a hybrid relational data structure that includes domain tables like PCORNet (eg, diagnoses, encounters, medications, etc) as well as the UMLS-specific concepts, atoms, and relationship tables that enable searching concepts and relations. Unlike the OMOP model, individual data entities are maintained and not transformed into derived values. Due to the relational design, many tables are in long format and still require significant transformation to create wide tables to facilitate patient level analyses. Further, our platform does not provide the analytic capabilities like OMOP to generate individual patient-level or population-level statistics.^{11,12} Future efforts will focus on developing user interfaces that clinicians and researchers can use to explore the data and run queries.

Previous work implementing a multistep concept-matching process for laboratory data identified the need to develop a user program that enables the user to select a match based off of suggested potential matches.²⁴ We developed 2 innovative user interfaces to facilitate manual mapping of unmatched concepts. The primary advantage of the user interfaces is the ability of end users, such as clinicians and researchers that are not informaticians or database programmers, to review individual data entities and select appropriate mappings that are then automatically entered into the data repository as new mappings.

Several additional advantages are realized from anchoring data entities to UMLS CUIs. First is the ability to leverage existing platforms, such as cTakes,²⁵ to parse and map large amounts of remaining unmapped data, such as strings of diagnoses from unstructured text, into individual data entities and associated CUIs that can then be integrated into existing tables. Another advantage is the simplification of searching for and identifying specific patient populations through the navigable semantic structure provided via the UMLS concept mappings and relationships across terminologies.²⁶

Using UMLS also offers several benefits when combining data from other domains and sources. The primary advantage is the ability to incorporate data entities from disparate sources and multiple codes/terminologies. Due to the breadth of included terminologies, we are able to incorporate additional data classifications such as coding for Elixhauser Comorbidity Index^{27,28} variables or incorporating the Clinical Classification System^{29,30} coding that maps to ICD codes that are either directly mapped or added by querying the atoms table that contains the applicable codes and concept description for each diagnosis.

A limitation of this work is that we only reported the number of data entities that were effectively processed in each step of the mapping process as it included all data entities regardless of source (ie, medications, laboratory results, diagnoses, etc). Future work will require individual analysis of each data type to assess data-matching accuracy.

The impetus for developing the Transport Data Repository was to facilitate investigating patients undergoing medical transport between hospitals. Until recently, limited research has been conducted on this group of patients due to limited access to data and the ability to harmonize data across sources to support large-scale analyses. To address the limitation of not having primary EHR data for patients prior to being transferred, we leveraged all EHR data from an integrated health system and third-party medical transport EHR. These data are being used by health system administrators, clinical

researchers, and public health researchers to investigate the role medical transfer plays in today's integrated health systems that have consolidated specialty services via a hub-and-spoke approach to specialty care that includes trauma, stroke, cardiac, and complex medical syndromes. For example, the first study we are conducting is to identify electronic phenotypes of patients that benefit from helicopter transfer. Due to the breadth and depth of EHR data contained in the data repository, we are now able to conduct in-depth analyses at patient and population levels by leveraging the latest machine learning approaches to explore what patient and clinical features impact patient needs for care and clinical outcomes and the associated impact on public policy.

CONCLUSION

We developed the Transport Data Repository to combine data from 3 disparate data domains to enable the reuse of clinical data for clinical and secondary research purposes. We developed a multistep mapping process that maps a large amount of data from clinical EHR sources automatically. While a majority of data is mapped automatically, additional programming was necessary to develop end-user interfaces that aid manual mapping of unmapped concepts. Further work is necessary to test the precision of this multistep mapping approach compared to other common data model approaches. Further development is necessary to identify and handle custom data specific to each data domain and source. Postmapped data has improved the interoperability of the data from other domains and sources and increased the use of data to support administrative and research activities.

FUNDING

This work was supported by the National Institute of Nursing Research, National Institutes of Health, (Grant number R15NR017792). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

AUTHOR CONTRIBUTIONS

APR and AM conceived and developed the Transport Data Repository, and both drafted and contributed to the content of this manuscript.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Bindman AB. The Agency for Healthcare Research and Quality and the development of a learning health care system. *JAMA Intern Med* 2017; 177 (7): 909–10.
2. Divney AA, Lopez PM, Huang TT, Thorpe LE, Trinh-Shevrin C, Islam NS. Research-grade data in the real world: challenges and opportunities in data quality from a pragmatic trial in community-based practices. *J Am Med Inform Assoc* 2019; 26 (8-9): 847–54.
3. Cutler DM, Scott Morton F. Hospitals, market share, and consolidation. *JAMA* 2013; 310 (18): 1964–70.
4. Frakt A. A sense of alarm as rural hospitals keep closing. *The New York Times*. <https://www.nytimes.com/2018/10/29/upshot/a-sense-of-alarm-as-rural-hospitals-keep-closing.html> Accessed November 2, 2018.
5. Zhang H, Guo Y, Li Q, et al. An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. *BMC Med Inform Decis Mak* 2018; 18 (S2): 41.
6. Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. *AMIA Annu Symp Proc* 2003; 2003: 699–703.
7. Campbell WS, Pedersen J, McClay JC, Rao P, Bastola D, Campbell JR. An alternative database approach for management of SNOMED CT and improved patient data queries. *J Biomed Inform* 2015; 57: 350–7.
8. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. *AMIA Annu Symp Proc* 2006; 2006: 1040.
9. Corley DA, Feigelson HS, Lieu TA, McGlynn EA. Building data infrastructure to evaluate and improve quality: PCORnet. *J Oncol Pract* 2015; 11 (3): 204–6.
10. Qualls LG, Phillips TA, Hammill BG, et al. Evaluating foundational data quality in the national patient-centered clinical research network (PCORnet(R)). *eGEMs* 2018; 6 (1): 3.
11. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc* 2012; 19 (1): 54–60.
12. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med* 2010; 153 (9): 600–6.
13. Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the All of Us Research Program: Transforming i2b2 data into the OMOP common data model. *PLoS One* 2019; 14 (2): e0212463.
14. Becker M, Bockmann B. Semi-automatic mark-up and UMLS annotation of clinical guidelines. *Stud Health Technol Inform* 2017; 245: 294–7.
15. Varghese J, Schulze Sunninghausen S, Dugas M. Standardized cardiovascular quality assurance forms with multilingual support, UMLS coding and medical concept analyses. *Stud Health Technol Inform* 2015; 216: 837–41.
16. Tran LT, Divita G, Carter ME, Judd J, Samore MH, Gundlapalli AV. Exploiting the UMLS Metathesaurus for extracting and categorizing concepts representing signs and symptoms to anatomically related organ systems. *J Biomed Inform* 2015; 58: 19–27.
17. Adamusiak T, Shimoyama N, Shimoyama M. Next generation phenotyping using the unified medical language system. *JMIR Med Inform* 2014; 2 (1): e5.
18. Metathesaurus. Unified Medical Language System; 2020. https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html Accessed January 20, 2020
19. Fuzzy Lookup Transformation; 2017. <https://docs.microsoft.com/en-us/sql/integration-services/data-flow/transformations/fuzzy-lookup-transformation?view=sql-server-ver15> Accessed April 14, 2020
20. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001; 2001: 17–21.
21. Divita G, Zeng QT, Gundlapalli AV, Duvall S, Nebeker J, Samore MH. Sophia: an expedient UMLS concept extraction annotator. *AMIA Annu Symp Proc* 2014; 2014: 467–76.
22. Lu CJ, Tormey D, McCreedy L, Browne AC. Enhanced LexSynonym Acquisition for effective UMLS concept mapping. *Stud Health Technol Inform* 2017; 245: 501–5.
23. Luo Y, Cheng Y, Uzuner O, Szolovits P, Starren J. Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. *J Am Med Inform Assoc* 2018; 25 (1): 93–8.
24. Lee L-H, Groß A, Hartung M, Liou D-M, Rahm E. A multi-part matching strategy for mapping LOINC with laboratory terminologies. *J Am Med Inform Assoc* 2014; 21 (5): 792–800.
25. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17 (5): 507–13.
26. Milinovich A, Kattan MW. Extracting and utilizing electronic health data from Epic for research. *Ann Transl Med* 2018; 6 (3): 42.

27. van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. *Med Care* 2009; 47 (6): 626–33.
28. Epstein RH, Dexter F. Development and validation of a structured query language implementation of the Elixhauser comorbidity index. *J Am Med Inform Assoc* 2017; 24 (4): 845–50.
29. Clinical Classification Software (CCS) for ICD-9-CM; 2017. <https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp> Accessed July 17, 2017
30. Elixhauser A, McCarthy EM. *Clinical Classifications for Health Policy Research, Version 2: Hospital Inpatient Statistics*. Rockville, MD: Agency for Health Care Policy and Research; 1996.