Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210192

Using Knowledge Graphs to Plausibly Infer Missing Associations in EMR Data

William VAN WOENSEL[a,1], Chad ARMSTRONG[b], Malavan RAJARATNAM[b], Vaibhav GUPTA[b], and Syed Sibte RAZA ABIDI[a,1]
[a] *NICHE Research Group, Faculty of Computer Science, Dalhousie University, Canada*
[b] *Varian Medical Systems Inc.*

Abstract. Electronic Medical Records (EMRs) are increasingly being deployed at primary points of care and clinics for digital record keeping, increasing productivity and improving communication. In practice, however, there still exists an often incomplete picture of patient profiles, not only because of disconnected EMR systems but also due to incomplete EMR data entry – often caused by clinician time constraints and lack of data entry restrictions. To complete a patient's partial EMR data, we plausibly infer missing causal associations between medical EMR concepts, such as diagnoses and treatments, for situations that lack sufficient raw data to enable machine learning methods. We follow a knowledge-based approach, where we leverage open medical knowledge sources such as SNOMED-CT and ICD, combined with knowledge-based reasoning with explainable inferences, to infer clinical encounter information from incomplete medical records. To bootstrap this process, we apply a semantic Extract-Transform-Load process to convert an EMR database into an enriched domain-specific Knowledge Graph.

Keywords. Knowledge Graphs, Electronic Medical Records, Knowledge-based Reasoning, Medical Taxonomies

1. Introduction

The incompleteness of EMR, due to lack of clinician time and/or cognitive overload combined with a lack of data entry restrictions, has been widely recorded [1]. This reduces the down-stream utility of the data, ranging from proper diagnosis of chronic patients based on long medical histories, to population-wide community health analysis.

As a contribution to solving this problem, this paper proposes a novel EMR information augmentation approach, in the form of a plausible knowledge-based reasoning method. Our method plausibly infers *missing causal associations* between medical EMR concepts, i.e., where a given concept (e.g., treatment) likely occurred due to the associated concept (e.g., diagnosis), based on semantic relations from medical taxonomies (e.g., SNOMED-CT [2], ICD [3]). While plausibly inferred associations do not result from crisp deductive reasoning, but rather a best-effort based on semantic relations, they have been found useful in medical decision making [4]. It may be appealing to consider a machine learning (ML) solution for this problem, but getting a sufficient amount of anonymized and similarly distributed EMR data will often be a

[1] Corresponding authors: Syed Abidi and W. Van Woensel, Faculty of Computer Science, Dalhousie University, B3H 1W5 Halifax, Canada; E-mail: {ssrabidi,william.van.woensel}@dal.ca.

barrier—hence knowledge-driven solutions are more practical and explainable as opposed to state-of-the-art ML methods [5]. Our EMR augmentation approach includes:

- A semantic Extract-Transform-Load (ETL) pipeline to convert semi- and unstructured EMR data into an enriched, domain-specific EMR Knowledge Graph (KG);
- A plausible knowledge-based reasoning method, including a semantic similarity analysis process, to infer missing causal associations within domain-specific EMR KG.

As a use case, we infer missing causal associations between diagnoses and treatments in an installation of the ARIA [6] Oncology Information System (OIS), an oncology-specific EMR (v. 15.1.2.93). While ARIA did allow defining these associations, they were missing from the database as it was not mandatory to enter them.

2. Methods

We previously applied plausible reasoning that computerizes human thought patterns to infer missing knowledge [4]. Whereas probabilistic and fuzzy OWL variants [7] support propositions with a degree of ambiguity, in our case, uncertainty results from missing knowledge. Association rule mining [8] has been used to infer causal relations; KG embedding methods [9] generate latent entity representations, where semantically related entities are closer to each other in space. However, these approaches require large amounts of (in our case, anonymized and similarly distributed) training data, which was missing in our case study. Imputation methods estimate missing EMR values by extrapolating from non-missing values [10], but, as mentioned, no causal associations were present in the EMR. Instead, we present a plausible reasoning method, based on semantic EMR KG, which relies on computerizing existing medical knowledge.

2.1. Domain-Specific EMR Knowledge Graphs using a Semantic ETL Pipeline

We present a generic semantic ETL pipeline to convert semi-structured EMR data into a domain-specific Knowledge Graph (KG) to support subsequent knowledge-based reasoning (Section 2.2). We illustrate this process for an ARIA OIS installation.

Figure 1 shows the final EMR KG for the ARIA OIS (initial EMR data in solid lines, transformed and enriched contents in dashed lines).

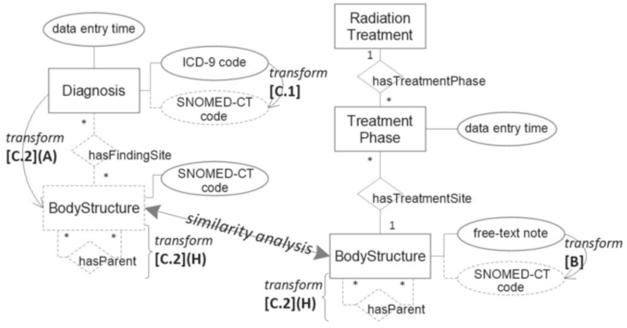

Figure 1. ARIA oncology EMR KG.

Below, we describe the ETL pipeline and its concrete application:
Extraction. Extract relevant patient, diagnosis, and intervention data from the EMR database using a set of configurable SQL queries.
Oncology EMR KG: the ARIA OIS database comprised a database totaling circa 16GB and 1544 tables; a first step involves extracting a relevant (and more manageable) subset of the data. We used a set of SQL queries to extract radiation treatments, treatment phases (incl. finding sites) and diagnoses for all patients with one or more diagnoses.
Transformation. (**A**) Transform EMR data into semantically structured data, using a domain-specific ontology linked with medical terminologies (ICD-9, SNOMED-CT).
Oncology EMR KG. We developed an OWL ontology to capture relevant oncology EMR concepts, with properties for linking to ICD-9 and SNOMED-CT codes.
(**B**) If needed, transform free-text clinical notes into structured data terms from well-known medical terminologies, e.g., by applying Natural Language Processing (NLP).
Oncology EMR KG: this ARIA OIS system version allowed free-text description of treatment sites. We formulated a complex set of regular expressions to map the free-text notes to SNOMED codes of the described body structures ([B] in Figure 1).
(**C**) Transform semantic EMR data by enriching with relations or properties from medical taxonomies to support knowledge-based reasoning. This includes (*C.1*) if needed, further annotating EMR data with medical terminology codes; and (*C.2*) complementing the EMR data with associational and hierarchical relations where needed.
Oncology EMR KG: (*C.1*) we mapped EMR ICD-9 diagnoses codes to SNOMED-CT codes using their 1-1 mapping [11] ([C.1] in Figure 1). (*C.2*) As per our reasoning needs, using SNOMED-CT, we enriched the semantic data with (a) associations between diagnoses and their finding site ([C.2](A) in Figure 1; e.g., *CancerOfProstate* hasFindingSite *ProstateStructure*); and (b) hierarchical relations between structures ([C.2](H) in Figure 1; e.g., *ThoracicStructure* hasParent *UpperBodyPartStructure*).
Loading. Load the transformed semantic EMR data into a domain-specific KG, which supports expressive querying as needed by knowledge-based reasoning methods.
Oncology EMR KG: we used Apache Jena [12] to create a KG from the extracted and transformed semantic data, which provides expressive querying access.
Applying the semantic ETL pipeline on the ARIA OIS database yielded an oncology EMR KG that supports knowledge-based reasoning methods, such as outlined below.

2.2. Plausible Knowledge-based Reasoning to Infer Missing Associations in EMR

We present a plausible knowledge-based reasoning method that, based on a domain-specific EMR KG, infers missing causal associations between EMR concepts, i.e., where a given concept (e.g., treatment) occurred due to the associated concept (e.g., diagnosis).

Our knowledge-based approach supports the following metrics:

2.2.1. Semantic Similarity Analysis

This process relies on the assumption that *semantic concept similarity* in the EMR KG can plausibly imply a *causal relation between EMR concepts*:

- *Identify suitable EMR concepts for similarity analysis*. While the targeted EMR concepts may not be directly associated in medical taxonomies (e.g., diagnoses and treatments), one may find related, comparable concepts in the EMR KG. In the ARIA use case, the KG relates diagnoses and treatments to their body sites (Figure 1). Hence, more concretely, the assumption here is that *semantically similar finding sites* plausibly imply a *causal relation between associated diagnoses and treatments*. This is illustrated in Figure 1 (similarity-analysis arrow). We revisit this assumption in Section 4.

- *Analyze semantic similarity of the identified EMR concepts*. Semantic similarity metrics typically rely on Conceptual Distance (CD), i.e., cumulative distance between two concepts and their Closest Common Ancestor (CCA) in a hierarchy. The same CD between "deeper", more specific concepts reflects a better semantic similarity than between "higher", more general concepts. We utilize the commonly-used Wu-Palmer similarly metric [13], and calculate the CCA and semantic similarity per pair of relevant EMR concepts using a breadth-first search. One may further rule out overly general CCAs (e.g., Anatomical structure) as these do not reflect conceptual similarity. In doing so, we plausibly infer causal associations between relevant EMR concepts, with the strength of the association depending on the similarity score. One can choose a score threshold for inferring associations per use case (see Section 3). The path between the concepts and CCA serves as an explanation for the similarity.

2.2.2. Temporal Metrics

Here, the assumption is that two EMR concepts, such as diagnoses and treatments, *which were entered soon after one another, are likely causally related.* As before, we revisit the correctness of this assumption for the ARIA use case in Section 4. When the time difference in data entry lies below a given maximum, one can infer a causal association.

3. Results

Oncology EMR KG. After applying the semantic ETL pipeline on the ARIA OIS data, the resulting EMR KG included 5999 potential causal associations between diagnoses and treatments. The KG included 154 unique treatment site and 288 diagnosis site codes (SNOMED-CT). We relied on a "gold standard" solution where oncology experts annotated potential associations with YES (likely correct) or NO (likely incorrect).

We utilized the SNOMED-CT "Body Structure" hierarchy for analyzing semantic similarity of diagnosis and treatment body sites. Using a grid search for hyperparameter optimization, based on the gold standard, we found an optimal similarity score threshold of 0.8 and a max. entry time difference of 60 days. Table 1 shows evaluation metrics for plausibly inferring causal associations based on (1) temporal metric (max. of 60 days); (2) semantic similarity (threshold of 0.8); and (3) both temporal and semantic similarity.

Table 1. Evaluation results for the ARIA use case.

	Accuracy	**Sensitivity**	**Specificity**
(1) Temporal metric	0.70	0.99	0.52
(2) Semantic Similarity metric	0.72	0.67	0.75
(1, 2) Both metrics	0.83	0.67	0.93

4. Discussion and Concluding Remarks

Regarding the assumptions that predicate our approach: assuming a causal association between EMR data entered soon after one another, casts a web that is far too wide as it captures very many incorrect associations (specificity of 52%). Assuming a causal association between diagnoses and treatments with semantically similar finding sites, is more often correct (specificity of 75%) but also misses many more associations (sensitivity of 67%). Applying both temporal and semantic similarity metrics yields the best overall results, with higher accuracy (83%) and very good correctness (specificity of 93%) but still misses many associations (sensitivity of 67%). Nevertheless, given the noisy nature of the original EMR data (see below), this seems to indicate the veracity of our assumptions, and hence the utility of these metrics, at least for the ARIA use case.

We presented a knowledge-based reasoning method that plausibly infers missing causal EMR associations, together with a semantic ETL pipeline for preparing an EMR KG. There are clear limitations to our work. Our evaluation suffered from a noisy EMR dataset: even within the gold standard, 11% of treatments were entered *multiple days before their diagnosis*; which may explain the reduced utility of the temporal metric. Circa 20,000 patients had a diagnosis but no intervention; implying that some inferred associations were incorrect because the related EMR concepts simply did not have a relevant counterpart. Secondly, our semantic similarity analysis relies on the *consistency of specificity of taxonomy subhierarchies*. E.g., some sub-hierarchies may be quite fine-grained, differentiating in great detail between body sub-structures; whereas other sub-hierarchies may not be so detailed. We could not find examples of such inconsistencies in SNOMED-CT, but this needs to be separately validated. Future work involves testing other taxonomies (e.g., ICD-10) and evaluating their impact on evaluation metrics.

Acknowledgement: This research was funded by an NSERC Engage Grant. The authors thank Natural Science and Engineering Research Council of Canada.

References

[1] Madden JM, Lakoma MD, Rusinak D, et al. Missing clinical and behavioral health data in a large electronic health record (EHR) system. J. Am. Med. Inform. Assoc. 2016; 23:1143–1149

[2] U.S. National Library of Medicine: SNOMED CT, Available at: https://www.nlm.nih.gov/healthit/snomedct/index.html.

[3] World Health Organization: International Statistical Classification of Diseases and Related Health Problems (ICD). Available at: https://www.who.int/standards/classifications/classification-of-diseases.

[4] Mohammadhassanzadeh H, Van Woensel W, Abidi SR, Abidi SSR. Semantics-based plausible reasoning to extend the knowledge coverage of medical knowledge bases for improved clinical decision support. BioData Min. 2017 Feb 10;10:7.

[5] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H.Causability and explainabilty of artificial intelligence in medicine. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2019;9:e1312.

[6] Varian: ARIA Oncology Information System, Available at: https://www.varian.com/products/software/information-systems/aria-oncology-information-system

[7] Bobillo F, Straccia U. Fuzzy ontology representation using OWL 2. Int. J. Approx. Reason. 2011;52: 1073–94.

[8] Galárraga L, Teflioudi C, Hose K, Suchanek FM. Fast rule mining in ontological knowledge bases with AMIE+. VLDB J. 2015;24:707–30.

[9] Ristoski P, Paulheim H. RDF2Vec: RDF Graph Embeddings for Data Mining. Proceedings of the 5th International Semantic Web Conference; 2016: Springer; p. 498-514.

[10] Hu Z, Melton GB, Arsoniadis EG, Wang Y, Kwaan MR, Simon GJ. Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. J. Biomed. Inform. 2017;68: 112–20.

[11] U.S. National Library of Medicine: ICD-9-CM Diagnostic Codes to SNOMED CT Map, Available at: https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html.

[12] Apache: Apache Jena, Available at: https://jena.apache.org/.

[13] Wu Z, Palmer M. Verbs semantics and lexical selection. Proceedings of the 32nd Annual meeting of the Association for Computational Linguistics; 1994; p. 133–38.