A Resolution-Alignment-Completeness System for Data Imputation over Tabular Clinical Records

Shervin Mehryar

Institute of Data Science, Maastricht University
Paul-Henri Spaaklaan 1, 6229 GT Maastricht, Limburg, NL

The amount of data stored as electronic health records (EHR) in tabular format has grown significantly in recent years, now including an immense quantity of interactions, events and various medical information [1]. As such, data integration will play a transformative role in health information systems for the years to come, bridging the gap between research and applications. Increasingly, clinical knowledge graphs have been adopted for EHR modeling as a means to improving predictive performance and integrating expert knowledge with data-driven insights [2]. To learn their underlying structure, embedding models have been successfully applied to capture hidden hierarchies for downstream clinical tasks (e.g. comorbidity, readmission prediction) [3]. These models however suffer from the inherent incompleteness of knowledge graphs, rooted in missing values and inconsistent codification from legacy systems.

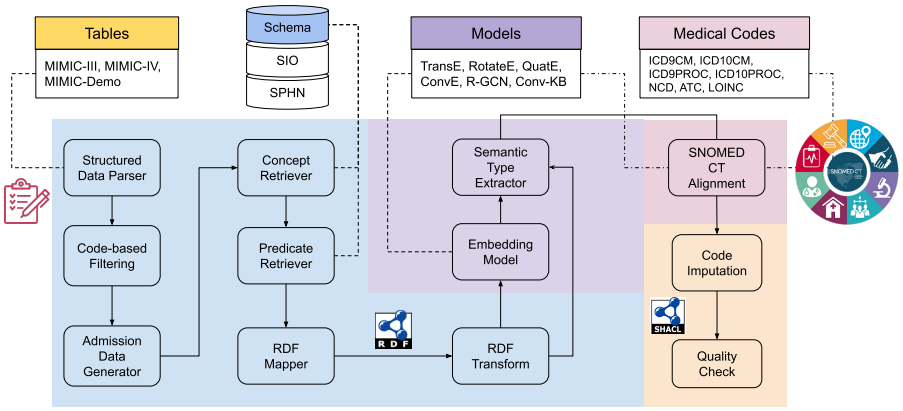

Fig. 1. Resolution (R), Alignment (A), Completeness (C) system for imputing data.

In this work, we propose a robust resolution-alignment-completeness (RAC) system for consolidating tabular EHR into semantically consistent health knowledge graphs, using standard terminologies aligned with medical ontologies. The proposed approach, as shown in Figure 1, follows a modular design:

- **Resolution.** Firstly, important patient entities are extracted from a relational data source and resolved/mapped via semantically equivalent identifiers based on the reference ontology for RDF generation.

- **Alignment.** In the second module, the resulting knowledge base is transformed and vectorized into translational embeddings, excluding class axiomatic and hierarchical information.
- **Completeness.** In the third module, the knowledge graph is aligned with a reference ontology to improve embedded representations before imputation.

In the first module, the groundings of concepts and relations from a reference schema are used to provide type annotations for a given table. Each table is deemed to be a flattened version of a sub-graph rendering from a reference graph. In this vein, we represent relations among entities and columns similar to nodes and concepts in a graph. This view subsequently allows the construction of an embedding space in which the relational entities (i.e. flat mapping of table data) are aligned with the relevant part in the reference schema. Rows are identified with unique URIs as instances of the corresponding RDF class and column attributes are mapped to retrieved RDF predicates. Biomedical codes (e.g. ICD[1], etc) are transformed into unique identifiers to resolve their semantically equivalent instances. After resolving and mapping all entities, the final knowledge graph is generated.

The transformed RDF from above is embedded using multiple translation based algorithms extended to include subsumption and instance checking reasoning. In particular, we apply the algorithm in [4] which introduces a modified loss function to facilitate transitive and hierarchical knowledge. In order to enrich the semantic soundness of predictions, the imputed types are subsequently aligned with a biomedical ontology, namely SNOMED CT[2]. The semantic richness of the latter enables capturing complex clinical relations - for instance “Pain in the left arm" with a type ‘finding’ predicated with attribute “finding site" and value “left arm", and thus facilitates more precise clinical data modeling through the extraction of structured information. Lastly, the semantic types for predictions are extracted and validated against the biomedical ontology (domain and range) constraints in order to add into the knowledge base.

We use the MIMIC III repository for experimentation which contains data associated with 53,423 distinct hospital admissions for adult patients admitted to critical care units [5]. In order to map patient relational records to an RDF schema, we use the Swiss Personal Health Network Schema (SPHN) [3] model (with possible extensions to SIO [4]). The resulting health knowledge graph is embedded thereafter with various models using the Pykeen[5] library. Our experiments demonstrate the contributions of RDF-schema based entity resolution and ontology-based alignment for diagnosis code imputation using several KGE methods. In particular for cardiovascular diagnosis code imputation, our system improves performance by 0.31 to 0.88 points in NDCG@10 over moderately large graph sizes.

[1] `https://icd.who.int/`
[2] `https://www.snomed.org/`
[3] `https://biomedit.ch/rdf/sphn-schema/sphn`
[4] `https://sio.semanticscience.org/`
[5] `https://github.com/pykeen/`

References

1. De Zegher, Isabelle, Kerli Norak, Dominik Steiger, Heimo Mueller, Dipak Kalra, Bart Scheenstra, Isabella Cina, Stefan Shulz, Kanimozhi Uma, Petros Kalendralis, Eno-Martin Lotmam, Martin Benedikt, Michel Dumontier, and Remzi Celebi. "Artificial intelligence based data curation: enabling a patient-centric European health data space." *Frontiers in Medicine*, vol. 11, Frontiers Media S.A., May 15, 2024. `https://doi.org/10.3389/fmed.2024.1365501`
2. Chen, Irene Y., Monica Agrawal, Steven Horng, and David Sontag. "Robustly extracting medical knowledge from EHRs: a case study of learning a health knowledge graph." In *Pacific Symposium on Biocomputing 2020*, pp. 19–30. World Scientific, 2019.
3. Choi, Edward, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. "Learning the graphical structure of electronic health records with graph convolutional transformer." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, pp. 606–613, 2020.
4. Mehryar, Shervin, and Remzi Celebi. "Improving Transitive Embeddings in Neural Reasoning Tasks via Knowledge-Based Policy Networks." In *Joint 2nd Semantic Reasoning Evaluation Challenge and 3rd SeMantic Answer Type, Relation and Entity Prediction Tasks Challenge*, pp. 16–27, 2022.
5. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. "MIMIC-III, a freely accessible critical care database." *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016. Nature Publishing Group.