Entity Event Knowledge Graph for Powerful Health Informatics

Ravi Bajracharya
Franz Inc.
Kathmandu, Nepal
ravi@datum.md

Richard Wallace
Franz Inc.
Portland, ME, USA
rsw@franz.com

Jans Aasman
Franz Inc.
Lafayette, CA, USA
jans.aasman@franz.com

Parsa Mirhaji
Montefiore Medical Center
Bronx, NY, USA
pmirhaji@montefiore.org

***Abstract*—This paper introduces the Entity-Event Knowledge Graph (EEKG) model for clinical data stored in graph databases. We describe how the EEKG model dramatically simplifies the representation of patient data, facilitates temporal queries, enables a 360 view of patients and promotes scalability by partitioning patient data into shards. We solved the practical problem that not all clinical data and life science knowledge can be sharded. The solution is to federate each individual shard with common shared data in a knowledge graph. One such shared data source is the UMLS (Unified Medical Language System) knowledge base, which contains genetic, drug clinical trials and Metathesaurus data that we link to individual patient records. We report on several use cases including EMR patient retrieval, matching patients with clinical trials, patient control group selection, and care quality measures.**

Keywords—entity-event model, knowledge graph, distributed graph database, umls skos knowledge graph, clinical trials knowledge base

I. Introduction

We describe an EMR and Analytics data system based on the Entity-Event Knowledge Graph (EEKG) model where patient data is sharded into a distributed graph database and linked to knowledge bases that include facility, provider, payer, coding, as well as medical and scientific knowledge. The approach is primarily based on the earlier work of P. Mirhaji at Einstein Medical College and Montefiore Health System [1]. To provide as realistic a demonstration as possible without relying on confidential patient information, we utilized an open-source synthetic patient generator called Synthea to generate demographic, clinical and claims records for 1 million patients [2]. Using the automated sharding feature of the Allegrograph[1] Resource Description Framework (RDF) triple store, we sharded the patient data across a number of servers. We link the sharded patient data to a common knowledge base that also includes UMLS (Unified Medical Language System) and an NIH/NLM project that integrates multiple sources of biomedical knowledge, vocabulary and standards [3]. We report on several use cases including EMR patient retrieval, matching patients with clinical trials, patient control group selection, and care quality measures.

II. Method

In this section, we will describe our approach to building an EEKG model based on patient health data and linking it to a biomedical knowledge graph constructed from multiple sources of biomedical knowledge bases such as UMLS and the clinicaltrials.gov dataset [6]. For the purpose of this demonstration we will use synthetic patient data generated using an open-source project called Synthea [2]. Patient health data concepts will be linked to a knowledge graph by normalising concepts to standard vocabulary which in this case will be the UMLS Metathesaurus.

A. Entity-Event Model

The Entity-Event (EE) model is a method of data organization for information stored in a graph database. By "Entity" we mean an element of the most common or, subjectively, the primary or core class of objects in the graph.

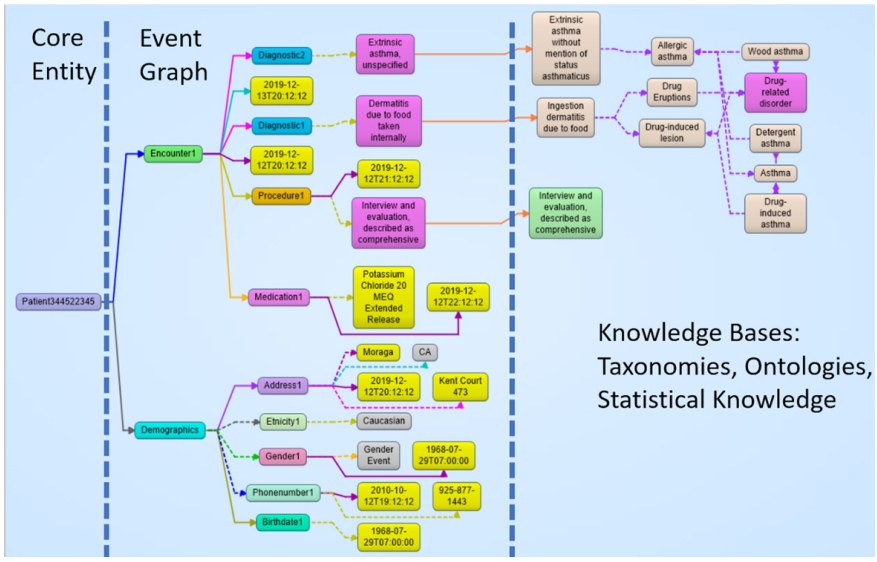

Figure 1 Patient health data and events represented as EE model linked to shared knowledge base

An event is anything associated with the entity that has a starting time and optionally, an ending time. The authors have used this approach in call centers, where the entities are customers and the events include calls and outcomes; in

[1] AllegroGraph is a Horizontally Distributed, Multi-model (Document and Graph), Entity-Event Knowledge Graph technology solution from Franz Inc. : https://allegrograph.com/products/allegrograph/

Submitted to the 10th IEEE INTERNATIONAL CONFERENCE ON HEALTHCARE INFORMATICS
Rochester, Minnesota, USA
June 11th-14th, 2022

aviation where the entities are aircraft and the events are flights, accidents, and maintenance steps; and in many other use cases where the data may be partitioned by entities that have temporal events. This article focuses on the health care use case, where the entities are patients.

Patient data is typically represented as a hierarchical tree as shown in figure 1. At the top of the tree is the patient as the core entity. The next level of the tree consists of events and sub-events. Usually the top level events are inpatient and outpatient encounters with sub-events such as diagnostics, observations, medication orders, procedures and vital signs. The model deals equally well with social determinant and insurance claims data. Each of these event objects have a similar shape because they all have a main type and an associated start and, optionally, end time. Their other properties distinguish them from other events: Encounters have providers, medications have dosages, observations have values, and so on, but they all have start times. The graph database representation of patient events is unified and simple: everything that happens to a patient has the same simple event structure.

The EEKG design is uniquely future proof. If we decide to add yet another type of event, we don't have to make any changes to earlier data, we just define a new event type and we can start adding new events of that type.

Importing data from a relational database via an ETL becomes much easier with the EEKG model. Instead of transforming input data from many different silos into a new relational database model, this approach requires one only to map the data into simple event objects represented as a graph.

The EE Knowledge Graph needs to be HIPAA compliant. The triple attributes feature of Allegrograph permits the implementation of various levels of permissive access. Attributes are name/value pairs that can specify which users can access which triples. Users can also be given attributes, which can be compared to the triple attributes. Suppose for example there is an attribute name "Sensitivity" with values "HIPAA Protected", "MHSA Data" and "HIV Data". Then a clinician accessing the data may have to "break the glass" to obtain permission to view each sensitive category, by acknowledging a treating relationship or special circumstance.

The EE model also makes it straightforward to partition the data by the primary entity or object represented in the data. In a graph database, a patient event is represented by a collection of subject-verb-object triples. In an RDF triple store like Allegrograph, the triple also has an associated "fourth element" typically used to represent a subgraph of data. Even though the patient is represented as a tree, the fourth element of every triple that comprises this tree is the patient identifier. By making this fourth element the patient ID, the distributed graph database can automatically partition the data over shards. In our application each shard will contain several hundred thousand patients.

One practical problem that we had to solve was that not all data can be partitioned or sharded based on entity. If we choose the patient as the primary entity in healthcare, the patient may also be linked to information about facilities, providers, payers, devices, coding systems, ontologies and other resources that cut across all patients. In the EEKG model, this information that cannot be partitioned becomes part of the "Knowledge Graph", which is federated with each of the shards containing the unique patient information. The largest part of this 'unshardable' knowledge graph consists of taxonomies and ontologies from the medical domain and life sciences.

B. UMLS-SKOS Knowledge Graph

Unified Medical Language System (UMLS) is a long-term project of the NLM (National Library of Medicine) which involves mapping biomedical concepts across multiple sources using a concept unique identifier (CUI) to normalize concepts and organize concepts across 54 broad semantic type categories called the semantic network [3]. In addition, UMLS Metathesaurus also contains relationships between concepts imported from component sources. E.g. Hierarchical relationships include PAR/CHD and narrower or broader relationships for example. We will be using SKOS ontology to represent UMLS concepts in our knowledge graph as shown in figure 2.

Simple Knowledge Organization System or SKOS is a framework for semantic representation of thesauri, classification scheme, subject heading systems, controlled vocabularies and taxonomies [4]. SKOS provides a standard way to represent knowledge organization systems using the Resource Description Framework (RDF). As such, it is a perfect system for semantic representation of UMLS which is a concept-oriented knowledge organization system.

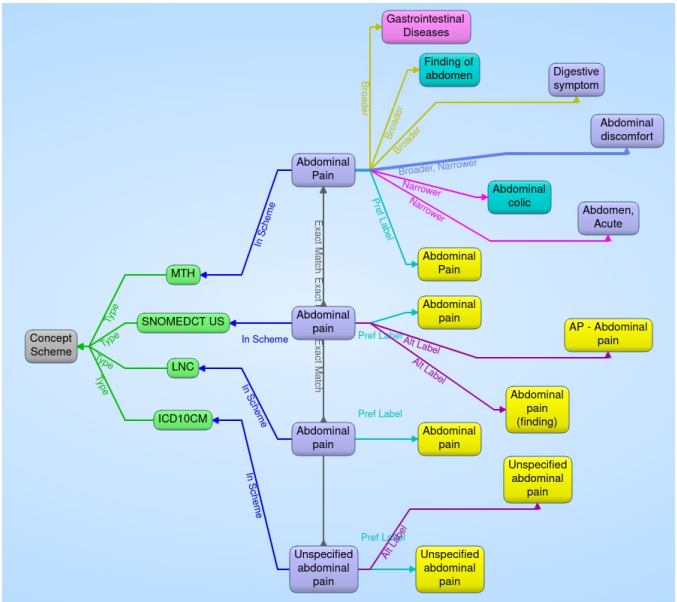

Figure 2 SKOS representation for the UMLS concept of "Abdominal Pain" across multiple concept schemes along with prefLabel, altLabel and broader/narrower relationship for the concept in different schemes.

Following bullets highlight the features of a UMLS SKOS knowledge graph:

- UMLS concepts become SKOS concept with skos:prefLabel set to Metathesaurus preferred term of the concept.

- PAR/CHD and "RB"/"RN" relationships from UMLS Metathesaurus are interpreted as SKOS:broader and SKOS:narrower relationships respectively.
- UMLS semantic network types are used as a rdfs:type for concepts.

C. *Knowledge Graph Augmentation with Clinical Trials Dataset*

SKOS representation of UMLS makes the UMLS knowledgebase far more accessible, interoperable and machine-readable than it's original form particularly in the context of linked data and semantic web. Moreover, UMLS SKOS knowledge graph can be further augmented by adding content from other biomedical sources to further enrich the baseline knowledge graph. We will be augmenting the knowledge graph with clinicaltrials.gov database, a collection of privately and publicly funded clinical studies conducted around the world [5]. The dataset is available as a set of XML files containing detailed information about the studies including recruiting information, eligibility criteria and clinical outcomes where available or reported. The condition and drug intervention fields in the dataset can be resolved to UMLS concepts and the structured data within the dataset can be linked to a UMLS SKOS knowledge graph as shown in figure 3.

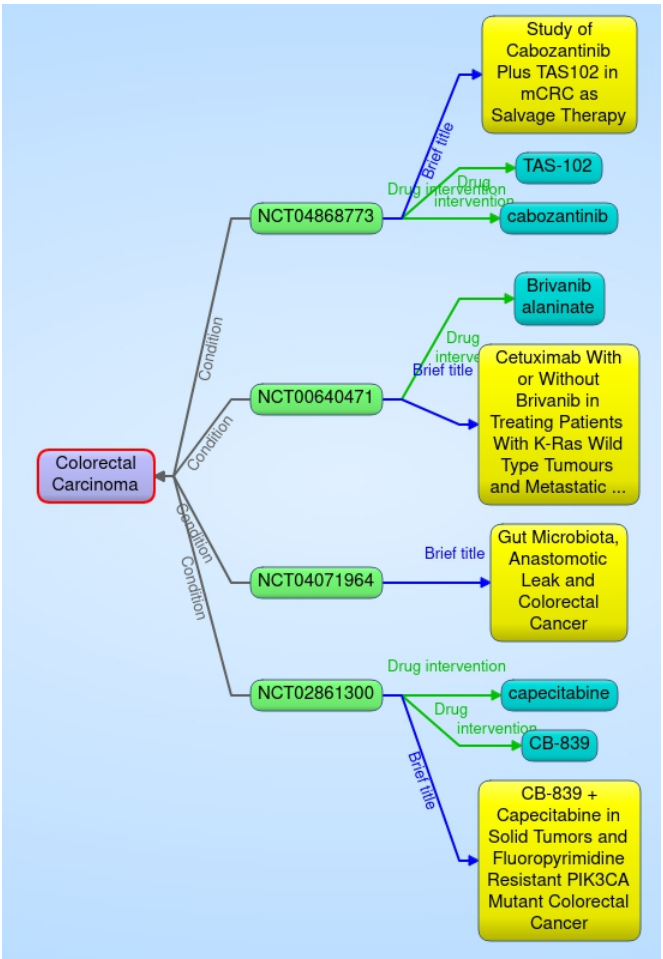

Figure 3 Clinical trials linked to umls-skos concept of colorectal cancer and the drug concept used in the intervention.

Following bullets highlight the features of a clinical trials knowledge graph:

- LinkedCT ontology representation of clinical trial dataset [6].
- Condition and drug intervention fields of clinical trial data linked to corresponding UMLS concept.

III. Use Cases

An EEKG model can have powerful implications in health informatics, population health, temporal analysis of treatment journeys and predictive as well as descriptive analytics of health data. In this section, we will take a look at ability from both perspectives of individual patient journeys and aggregated journeys of a population cohort.

A. *Patient "360 Degree" View*

The EEKG model supports the most basic function of electronic medical records systems (EMRs): looking up an individual patient record. To display a complete patient record, the EMR needs to retrieve all of the patient events, organize these in tables and graphs on the screen, and also retrieve any common knowledge base items related to those events. Sharding the data by patient ID, and storing all patient event data in a graph identified by that ID, permits efficient retrieval of an individual patient record, even in a multi-user environment where many clinicians access many patient records simultaneously.

Whether displaying the data in a UI or packaging it for health data exchange through formats such as a FHIR bundle [9] or a C-CDA document [10], retrieving this "360 Degree" view of the patient becomes problematic when the patient has a huge record [9][10]. In typical EMR systems the majority of patients have small-to-medium sized records. But the few patients who require long-term care or are heavy users of the health care system, may have enormous amounts of data. The EEKG model suits these 'mega-patients' perfectly, because sharding allows large records to be retrieved efficiently, and the required event start date provides a useful parameter for filtering.

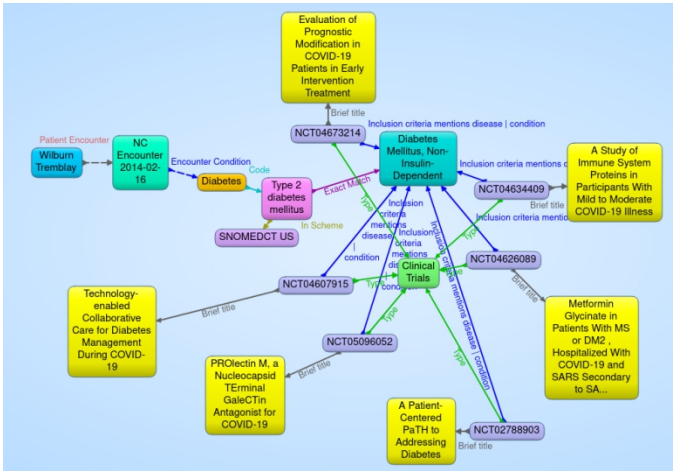

Figure 4 Clinical trials linked to patient "Wilburn Tremblay" based on trials studying patient condition of type 2 diabetes or those that mention the condition in it's inclusion criteria.

```
select ?category (count(?category) as ?count) {
{
  select ?patient (max(?obsDate) as ?obsDate) {
    {
      # sub-query for office visit
      select ?patient (max(?ovDate) as ?ovDate)  {
       # sub-query for diagnosis
       {
          select ?patient (max(?diagnosisDate) as ?diagnosisDate){
               {umls:C0011849 skos:narrower{,4} ?cond .} UNION {BIND(umls:C0011849 AS ?cond)}
               ?snomed skos:exactMatch ?cond; skos:inScheme umls-scheme:SNOMEDCT_US .
               ?condition rdf:type synthea:Condition; synthea:code ?snomed;
                          synthea:startDateTime ?diagnosisDate FILTER (?diagnosisDate > "2018-06-01T00:00:00+00:00"^^xsd:dateTime).
               ?patient rdf:type synthea:Patient; synthea:patientCondition ?condition .
          } group by ?patient
       }

       ?patient synthea:patientEncounter ?officeVisit; synthea:birthdate ?dob .
       ?officeVisit synthea:code/skos:notation "162673000"; synthea:startDateTime ?ovDate .
       bind( year(?ovDate) - year(?dob) - if(month(?ovDate)<month(?dob) || (month(?ovDate)=month(?dob) && day(?ovDate)<day(?dob)),1,0) as ?age
       filter (?ovDate > "2020-06-01T00:00:00+00:00"^^xsd:dateTime && ?age >= 18 && ?age <=75)
      } group by ?patient
    }

    ?patient synthea:patientObservation ?obs .
    ?obs synthea:startDateTime ?obsDate; synthea:code/skos:notation "4548-4" .
    filter (?obsDate > "2020-06-01T00:00:00+00:00"^^xsd:dateTime)

  } group by ?patient
}
?patient synthea:patientObservation ?obs .
?obs synthea:startDateTime ?obsDate; synthea:code/skos:notation "4548-4"; synthea:value ?a1c .
BIND(COALESCE(                                                    # categorize the A1C level
  IF(xsd:float(?a1c) >= 8.0, "CRITICALLY HIGH", 1/0),
  IF(xsd:float(?a1c) >= 7.0, "UNCONTROLLED DIABETES", 1/0),
  IF(xsd:float(?a1c) >= 6.0, "CONTROLLED DIABETES", 1/0),
  IF(xsd:float(?a1c) >= 5.7, "PRE-DIABETES", 1/0),
  IF(xsd:float(?a1c) >= 0.0, "NON-DIABETIC", 1/0),
  "NO MEASUREMENT") as ?category)
} Group by ?category
```

Figure 5 SPARQL query to fetch diabetic patients based on conditions specified for NQF-59 quality measure.

B. Match Patient to Studies

Because the UMLS system is augmented with clinical trial data, it becomes an intriguing possibility to provide the clinician with clinical studies applicable to a specific patient. Clinical trials contain specific inclusion and exclusion criteria. The EMR can match these with the patient's demographic, diagnostic and treatment data to display relevant clinical trials as shown in figure 5.

C. Patient Cohort Selection

Public health organizations and government agencies often want to study the effects of some intervention on a patient population, such as whether diverting high-utilizers into social services has an impact on ED admissions. The time and budget may be unavailable for a proper double-blind scientific study with a control group and a sample population. If the intervention has already been performed on the sample, it may be possible to select a random "control group" retroactively. For example, select a group of patients on Medicaid, of a certain demographic distribution, with a certain range of diagnoses and utilization criteria, matching the sample population, so that their outcomes may be compared with the sample group.

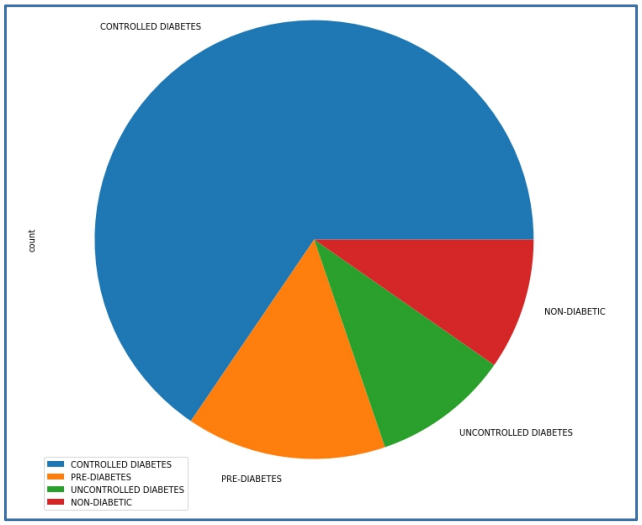

Figure 6 Visualization of NQF-59 query for diabetic patients.

D. NQF-59 - Quality Metric for Diabetes Population

The Centers of Medicare and Medicaid Services provides a collection of MIPS (Merit-Based Incentive Payment Systems) criteria that impact how providers are reimbursed for services [7]. One such measure is NQF-59, measuring diabetes poor control in a patient population [8]. Like many MIPS measures, the NQF-59 has a complex query specification involving patient ages, encounter and diagnostic history, most recent A1c measurement, and exclusion and inclusion criteria. Fortunately these queries are more straightforward in SPARQL than in relational database languages, as shown in figures 5 and 6.

SUMMARY

In this paper, we demonstrated possible use cases of an EEKG based modeling of patient health data in representing individual and aggregate view of patient events and patient journeys in a healthcare system. EEKG model is both effective for looking at 360 view of an individual patient journey and also for looking at journeys of a cohort or patient population in aggregate. We demonstrated these use cases on a collection of synthetic patient data for over 1 million patients, corresponding to 12 billion triples. Moreover, an EEKG model allows for a convenient method to shard patient data and medical knowledge base across a distributed graph database deployment to allow scaling over ten of millions of patients and hundreds of billions of triples.

EEKG model can link patient data to knowledge bases such as UMLS enabling structured query based on UMLS graph and relationships. Similarly, queries can take advantage of the transitive closure of a graph to perform more knowledge based queries on hierarchical data and relationships such as query for all forms of breast carcinoma related trials in a dataset without having to explicitly provide an exhaustive list of all breast carcinoma forms.

We also touched upon some other benefits of the EEKG model, including the ease of ETL from relational databases, the simplicity of adding new event types in the future, and the ability to grant permission for users of different status to access different levels of confidential patient data.

In summary, representing patient health data using EEKG model can enable powerful use cases and analytics in health informatics.

REFERENCES

[1] J. Aasman en P. Mirhaji, "Knowledge graph solutions in healthcare for improved clinical outcomes", CEUR Workshop Proceedings, vol 2180, Jan 2018.

[2] J. Walonoski et al., "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record", Journal of the American Medical Informatics Association, vol 25, no 3, bll 230–238, 08 2017.

[3] O. Bodenreider, "The Unified Medical Language System (UMLS): Integrating Biomedical Terminology", Nucleic acids research, vol 32, bll D267-70, 02 2004.

[4] A. Miles and S. Bechhofer, Skos Simple Knowledge Organization System Reference. [Online]. Available: https://www.w3.org/TR/skos-reference/. [Accessed: 07-Feb-2022].

[5] "Clinicaltrials.gov background," ClinicalTrials.gov. [Online]. Available: https://clinicaltrials.gov/ct2/about-site/background. [Accessed: 07-Feb-2022].

[6] O. Hassanzadeh, A. Kementsietsidis, L. Lim, R. J. Miller, en M. Wang, "LinkedCT: A Linked Data Space for Clinical Trials", arXiv [cs.DB]. 2009.

[7] "Quality measures requirements," CMS. [Online]. Available: https://qpp.cms.gov/mips/quality-measures?py=2021. [Accessed: 07-Feb-2022]

[8] "Quality ID #1 (NQF 0059): Diabetes: Hemoglobin a1c (hba1c ..." [Online]. Available: https://qpp.cms.gov/docs/QPP_quality_measure_specifications/CQM-Measures/2019_Measure_001_MIPSCQM.pdf. [Accessed: 07-Feb-2022].

[9] "Welcome to the HL7 Fhir Foundation," HL7 FHIR Foundation Enabling health interoperability through FHIR. [Online]. Available: http://www.fhir.org/. [Accessed: 14-Feb-2022].

[10] "Consolidated Clinical Document Architecture," Wikipedia, 04-Nov-2020. [Online]. Available: https://en.wikipedia.org/wiki/Consolidated_Clinical_Document_Architecture. [Accessed: 14-Feb-2022].

Submitted to the 10th IEEE INTERNATIONAL CONFERENCE ON HEALTHCARE INFORMATICS
Rochester, Minnesota, USA
June 11th-14th, 2022