Automated Domain-Specific Healthcare Knowledge Graph Curation Framework: Subarachnoid Hemorrhage as Phenotype

Article *in* Expert Systems with Applications · December 2019

DOI: 10.1016/j.eswa.2019.113120

Khalid Malik

University of Michigan-Flint

123 PUBLICATIONS **1,303** CITATIONS

SEE PROFILE

Madan Krishnamurthy

Elanco

13 PUBLICATIONS **86** CITATIONS

SEE PROFILE

Maqbool Hussain

Sejong University

77 PUBLICATIONS **1,016** CITATIONS

SEE PROFILE

Fakhare Alam

Oakland University

11 PUBLICATIONS **74** CITATIONS

SEE PROFILE

Automated Domain-Specific Healthcare Knowledge Graph Curation Framework: Subarachnoid Hemorrhage as Phenotype

Khalid Mahmood Malik[a], Madan Krishnamurthy[a], Mazen Alobaidi[a] , Maqbool Hussain[b, *], Fakhare Alam[a] , Ghaus Malik[c]

[a] *Department of Computer Science & Engineering, Oakland University, 115 Library Drive, Rochester, MI, 48309, USA*
{mahmood, malobaid, mkrishna, fakharealam }@oakland.edu
[b] *Department of Software, Sejong University, South Korea, {maqbool.hussain}@sejong.ac.kr*
[c] *Department of Neurosurgery, Henry Ford Hospital, 2799 West Grand Boulevard, Detroit, MI 48202, USA, {gmalik1}@hfhs.org*

Corresponding Authors *: Khalid Mahmood Malik and Maqbool Hussain {email: *mahmood @oakland.edu* and *maqbool.hussain@sejong.ac.kr*}

Abstract: To derive meaningful insights from voluminous healthcare data, it is essential to convert it into machine understandable knowledge. Currently, machine understandable domain specific healthcare knowledge curation framework does not exist for complex neurological diseases such as subarachnoid hemorrhage stroke. We envisage futuristic clinical decision support systems and tools backed with such knowledge will aide in complex neurological disease prognosis, diagnosis, and treatment. Existing knowledge graphs (KGs) only contain concepts and relationships between them and offer this knowledge to information extraction and knowledge management applications. However, the proposed domain-specific automated KG curation framework enables extraction of concepts, relationships, individual and cohort graphs, and predictive knowledge. By employing ontology-based information extraction, ensemble learning and word embedding based on skip-gram techniques on structured and unstructured data from electronic health records of 1025 patients with an intracranial aneurysm, this paper proposes a novel fully automated framework to curate knowledge graph, consisting of concepts, different hierarchical and non-hierarchical relationships, and predictive rules for prediction of subarachnoid hemorrhage. The evaluation shows that proposed framework achieves 78% precision and 71% recall respectively, for concept extraction from clinical text. Taxonomic relationships evaluation had precision and recall of 68%, and 95%, respectively. Evaluation of knowledge to predict unruptured status using validation dataset shows accuracy, precision, recall, of 73%, 76%, and 90% respectively.

Keywords: Knowledge Graph, Ontology, Electronic Health Records, Intracranial Aneurysm, Association Rules, Ensemble Learning, Subarachnoid Hemorrhage Stroke

1. Introduction

Biomedical data has the characteristics of big data including volume, variety, velocity, and veracity (Bresnick, 2017; Lee & Yoon, 2017; Ross, Wei, & Ohno-Machado, 2014). The exponential growth of clinical data, that is mostly heterogenous and unstructured, requires semantic, statistical, and predictive analysis to convert this multidimensional data into structured and machine-understandable knowledge, known as knowledge graph (KG). These KGs should follow FAIR (findable-accessible-interoperable-reusable) principle (Wilkinson et al., 2016), so that actionable information provided by KGs, in form of cognitive services, could be consumed by applications to improve their cognitive 'abilities' (Sheth, Yip, Iyengar, & Tepper, 2019). Recently, the role of knowledge has significantly increased due to the emergence of voice assistants such as Google Home, Amazon, Siri, etc. and other AI services such as Google's Semantic Search, and IBM's Watson. These applications consume formally represented domain specific knowledge to perform contextualization, personalization, and abstraction to transform their current simple, scripted conversations to the highly intelligent one (Sheth et al., 2019). Recently, many applications have employed KGs to solve complex AI tasks such as deep understanding of clinical text in the absence of colossal quantity of training data (Sabra, Mahmood Malik, & Alobaidi, 2018). Knowledge graphs could be quite useful for entity recognition particularly when the objects, such as implicit entities, in the clinical text are complex (Costa, 2015). Examples of KGs include Google's Knowledge Graph (Steiner, Verborgh, Troncy, Gabarro, & Van De Walle, 2012), and Linked Open Data (Bizer, 2011) based sources such as DBpedia (Lehmann et al., 2015) ,Yago (Hoffart, Suchanek, Berberich, & Weikum, 2013), Wikidata (Vrandečić & Krötzsch, 2014), Baidu Knowledge Graph , and Sogou Knowledge Cube (Yu et al., 2017).

Currently, healthcare KGs are mainly used by researchers to do basic clinical tasks such as semantic based document retrieval (Zhao, Kang, Li, & Wang, 2018), information extraction systems, and knowledge management systems. However, in near future, healthcare KGs are expected to pave the informatics path toward realization of the evidence-based and precision medicine by integrating and linking diverse biological data, clinical text, and imaging data to extract knowledge for understanding unknown risk factors and their association with the disease, assisting clinical decision making, allowing personalized treatment recommendations, and making preventive strategies for comorbid conditions (Ping, Watson, Han, & Bui, 2017) . The knowledge represented in knowledge graphs that is derived from diverse knowledge sources, will enable the development of new biomedicine applications. For example, clinicians will be having artificial intelligence assistants to help prescribe or formulate plan for individualized treatment of complex diseases. These assistants will communicate with high performance computing clusters to get an integrated view of synthesized knowledge about prognosis, diagnosis, and effective treatment plans.

Currently, knowledge graph creation and curation are mostly manual or somewhat semi-automated (Rotmensch, Halpern, Tlimat, Horng, & Sontag, 2017), and thus it is a labor-intensive process. Furthermore, automatically extracting reliable and consistent knowledge particularly from structured and unstructured sources at scale has proven to be a formidable challenge. Very few attempts have been made on automated construction of health knowledge graphs, and their focus was limited to creation of triplets with having only one type of relationship. Additionally, to develop semantic, correlative, and causal relationships among domain's concepts in knowledge graphs using an automated approach is challenging, as traditional models do not consider semantic inferences and contextualization among data elements acquired from disparate sources. None of these approaches have focused on building hierarchical relationships among extracted concepts. Additionally, concept extraction using either word embedding, or ontology-based information extraction does not give reliable accuracy, and this also affects accuracy of relationship extraction. Lastly, efforts have not been made to develop predictive knowledge which should be interpretable to both machines and humans to enable true symbiotic human-machine and machine-machine interactions.

This paper attempts to solve above-mentioned challenges by proposing an automated domain specific knowledge graph construction by making use of structured and unstructured data of patients diagnosed with cerebral aneurysms. Stroke, the 5th biggest cause of death and the leading cause of disability in the US, has three different types ("Stroke Information | cdc.gov," n.d.). One type of stroke is subarachnoid hemorrhage (SAH) stroke that is caused due to rupture of intracranial aneurysm (a.k.a. cerebral or brain aneurysm). Statistics show that every 18 minutes a brain aneurysm ruptures in the US, and SAH has a 40% fatality rate. Of those who survive, about 66% suffer some permanent neurological deficit ("Statistics and Facts - Brain Aneurysm Foundation," n.d.). Since neither decision support systems nor large training corpus/knowledge for complex neurological diseases such as SAH exist, there is a need to have KG for such diseases to synergistically enhance translational research and clinical decision making. This paper attempts towards extraction of knowledge required to build CDSS for SAH, which will be pivotal to develop proposed hybrid machine learning and knowledge based CDSS: NeuroAssist ("NeuroAssist - BA Foundation.," n.d.). Figure 1 represents the conceptual view of the proposed ASKG curation framework which uses semantic, statistical, and predictive analysis on structured and unstructured clinical data to acquire semantic, statistical and predictive knowledge respectively. Semantic analysis aims to recognize entities and relationships among them by understanding the context and meaning of each term of given unstructured clinical text; while statistical analysis characterizes the inferential knowledge of properties (e.g. aneurysmal rupture risk ration of Caucasian sub-cohort etc.) in semantic knowledge of anterior communicating artery cohort, or properties of predictive knowledge (e.g. accuracy of model used to generate the knowledge). Lastly, by using machine learning, predictive analysis generates predictive knowledge (on cohort level KG or multiple KGs) by exploiting values of features (concepts) extracted in semantic analysis process.

As shown in Figure 1, semantic knowledge acquisition gives triplets in form of concepts and relationships. Additionally, predictive knowledge acquisition by taking the concepts from semantic knowledge, their corresponding values from clinical notes, and deriving production rules using hybrid approach that uses association mining and machine learning models. Statistical knowledge acquisition is updated by both semantic and predictive knowledge layers. Figure 1 also highlights various processes that are employed by each layer.

ASKG automatically generates knowledge graph from both unstructured clinical reports and structured data in formal and machine-understandable format. The proposed framework offers an extensible mechanism to build the concepts incrementally, find relationships among concepts, and add predictive knowledge graph as clinical data grows. The main contribution of this paper is to develop domain specific automated knowledge graph curation framework which can offer domain-specific semantic, statistical and predictive knowledge. The novelties of the proposed framework are as follows:

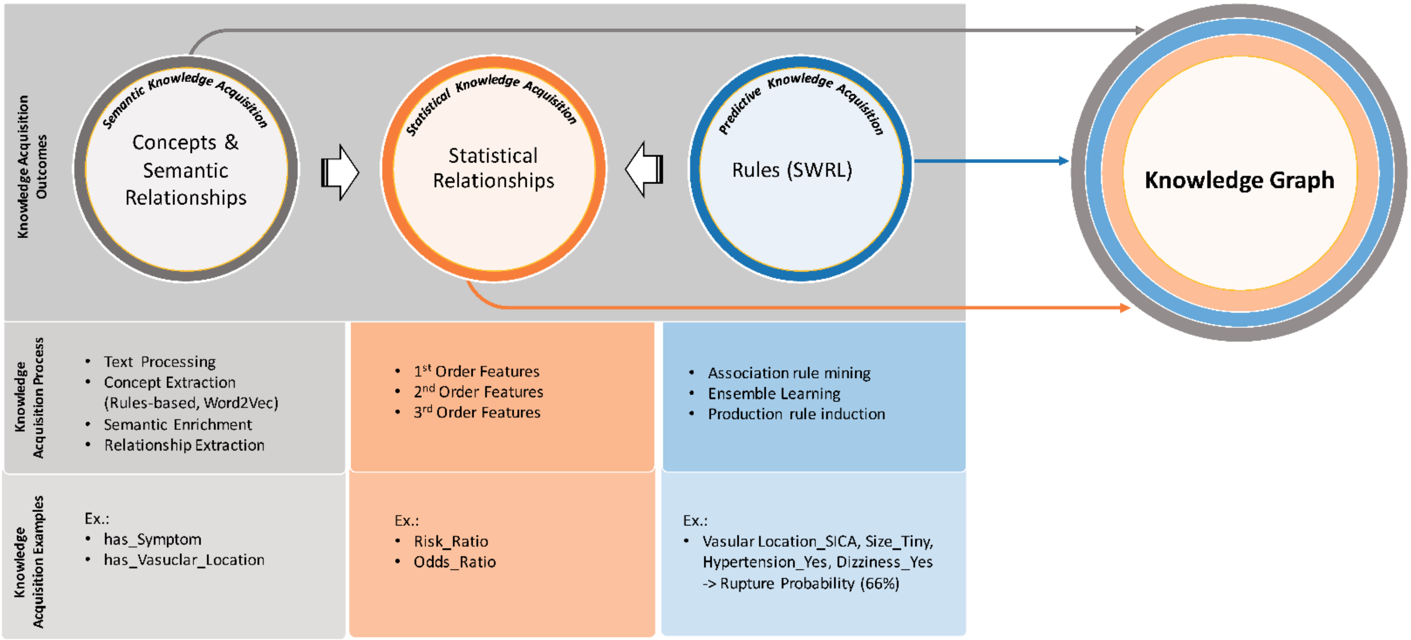

Knowledge Acquisition Outcomes
Semantic Knowledge Acquisition
Concepts & Semantic Relationships
Statistical Knowledge Acquisition
Statistical Relationships
Predictive Knowledge Acquisition
Rules (SWRL)
Knowledge Graph
Knowledge Acquisition Process
• Text Processing
• Concept Extraction (Rules-based, Word2Vec)
• Semantic Enrichment
• Relationship Extraction
• 1st Order Features
• 2nd Order Features
• 3rd Order Features
• Association rule mining
• Ensemble Learning
• Production rule induction
Knowledge Acquisition Examples
Ex.:
• has_Symptom
• has_Vasuclar_Location
Ex.:
• Risk_Ratio
• Odds_Ratio
Ex.:
• Vasular Location_SICA, Size_Tiny, Hypertension_Yes, Dizziness_Yes -> Rupture Probability (66%)

Fig. 1. Conceptual view of processes and outcomes of automated knowledge graph curation framework

- It presents methods to accurately identify clinical concepts and relationships using ontology-based information extraction and word embedding. More specifically, the relationships and concept linking (REL) of ASKG is governed by Linked Biomedical Ontologies (LBO) namely BioPortal ontologies ("Welcome to the NCBO BioPortal | NCBO BioPortal," n.d.), our in-house built extended ICO ontology, and semantic similarity using word embeddings. The hierarchical relationships are extracted from well-defined ontologies of LBO by mapping identified concepts from text to classes of LBO.
- It generates SAH predictive knowledge using hybrid approach which employs association rule and ensemble machine learning model. More specifically, it uses 'Apriori' algorithm for association rules mining and ensemble learning model for building predictive knowledge in the form of production rules having a unique URI. The ensemble model consists of Random Forest (RF), Decision Tree, Support Vector Machine (SVM), Adaptive Boosting (AdaBoost), and Gradient Boosting (GBoost). Additionally, since every intracranial aneurysm leading to SAH poses different characteristics, this knowledge graph employs data taken from structured (taken from images and clinical text) and unstructured clinical data (progress summaries and radiology notes), to develop prediction knowledge in form rupture criticality.

The rest of this paper is organized as follows. We begin by reviewing related works in Section 2. After describing the materials & methods in Section 3, we describe the evaluation in Section 4. Finally, we conclude the paper in Section 5.

2. Related Work

Automated KG construction is relatively new but emerging area of research. Most of KG construction efforts were made in non-healthcare domains by exploiting discourse analysis, semantic frames, machine learning algorithms with existing semantic web data, and open information extraction (Martinez-Rodriguez, Lopez-Arevalo, & Rios-Alvarado, 2018). In healthcare, the focus of KG construction remains on manual construction of domain specific ontology, and then using it to create triplets (concept-relationship-concept) by using semi-automated approaches on biomedical or clinical text. The existing work of knowledge graph construction in healthcare domain could be divided into following four categories: a) KG construction using biomedical literature, b) KG construction using electronic health records (EHRs), c) KG curation using EHRs and Social Media, d) KG construction using databases maintained for traditional Chinese medicine.

Due to unavailability of clinical data to computer scientists, healthcare knowledge graphs are mainly constructed by using published biomedical corpus. For example (Song, Kim, Lee, Heo, & Kang, 2015) and (Yuan et al., 2019) have presented knowledge discovery framework that uses biomedical scientific publications (PubMed) to establish gene-disease, drug-disease, and protein-protein associations. In this work concept extraction involves syntactic parsing using Stanford Core NLP and semantic annotation using ontology referencing (DrugBank, Human Metabolome Database (HMDB), etc.) to generate concepts such as gene, species, cell, anatomical concepts, disease, drugs etc. Furthermore, dependency tree was used for rule-based relationship extraction including grammatical encoding. Likewise, (Yuan et al., 2019) proposed domain specific knowledge graph using PubMed abstracts. This knowledge graph consists of only concepts and non-hierarchical relationships by employing unsupervised entity and relation embedding based on skip-gram, latent relation generation by clustering based on relation embeddings, and a relation refinement on the automatic generated latent (noisy) labels based on a convolutional neural network (CNN) with attention model. For futuristic healthcare decision support system, these KGs could be useful to enable evidence-based treatments, however, they are not quite useful for prognosis or diagnosis of diseases. Thus, it is important to develop KG using clinical sources.

Recently, there have been some attempts on individual aspects of KG construction using semi-automated methods by extracting unstructured clinical notes from electronic health records (EHRs). However, these attempts are limited to have named entities and their relationships extracted from text. For example, (Rotmensch et al., 2017) has constructed KGs constituting disease-symptom relationships using concept extraction from electronic health records (EHRs) linking diseases with symptoms. In their work, string-matching was applied on de-identified data based on common names obtained from Google Healthcare KG and UMLS. Symptom-disease relationships were established using following three probabilistic models: logistic regression, Naive Bayes classification and Bayesian network. Another such attempt was made by (Finlayson, LePendu, & Shah, 2014). In this work, authors have performed analysis on clinical notes by mapping terms into clinical concepts to design a co-occurrence matrix (disease-disease, drug-drug, drug-disease patterns) by grouping patient notes into time-based bins, followed by clinical term identification and extraction built into clinical concept occurrence matrix. The matrix hence obtained is used to evaluate frequency of relation and co-frequency of concepts, quantifying the relatedness among medical concepts. The concept extraction in their proposed work is based on dictionary compiled of 22 clinically relevant ontologies (such as SNOMED-CT, MedDRA).

Likewise, (Shi et al., 2017) presented a framework to convert textual medical knowledge (TMK) into Semantic Health KG. The framework comprises of three models: Medical Knowledge Model (organizes TMK into conceptual graphs), Health Data Model (defines and normalizes structures and relationships of complex and unstructured health data from EHRs), and Terminology Glossary (uses SNOMED CT and International Classification of Diseases (ICD) to provide the instances of both TMK and HDM and their semantic mappings). Overall, the framework organizes TMK into conceptual graphs, defines and normalizes structures, extract relationships from unstructured EHRs, and uses biomedical ontologies to provide the instances of semantic mappings. The KG has defined following parameters: entity (bodily-part, sign/symptom, clinical manifestation, diagnosis, treatment) and relation (located_in, cause, corresponded_to).

It is important to explore possibility of developing context-aware personalized knowledge graphs, by integrating clinical and non-clinical data. Recently, an attempt was also made to construct personalized knowledge graph by using social media data (Reddit), EHR and IoT data (patient and environmental) to define a context-aware personalized health KG. In this work, (Gyrard, Gaur, Shekarpour, Thirunarayan, & Sheth, 2018) provided a case-study on personalized KG for Asthma, obesity and Parkinson's disease. Entity-extraction is achieved by mapping social media text with biomedical ontologies defining concepts. SIDER knowledgebase is used for identification of disorder, treatment, drugs/dosage, side-effects and reactions using the entity-type defined in the context. Furthermore, Kno.e.sis ontology (integration of W3C SOSA ontology, Asthma ontology, FOAF ontology and Weather ontology) was used to enrich and annotate extracted IoT and medical concepts.

Recently some attempts were made to build knowledge graph for the domain of traditional medicine. In this regard, (Yu et al., 2017) has presented a semi-automated approach of constructing triplets for traditional Chinese medicine (TCM) health preservation, by integrating related disconnected knowledge resources of Chinese medicine scientific data. As the base of TCM knowledge graph "TCM health care ontology" was built, which itself consists of two components: "top-level ontology" and "thesaurus". Later this ontology was used to populate the content of health databases into the knowledge graph. Text mining was done to extract entities and relations from text in the databases, and extracted triplets were added into the knowledge graph after expert validation. Likewise, another attempt was made by (Weng et al., 2017) for knowledge graph construction towards traditional Chinese medicine based on semantic analysis of 866 clinical texts of patients diagnosed with hypertension. Their proposed framework consists of a following four modules: medical ontology constructor, knowledge element generator, structured knowledge dataset generator, and a graph model constructor. This framework has employed recurrent neural network and ontology model to develop final knowledge graph.

The limitations of above-mentioned approaches for knowledge graph construction frameworks include a) non-inclusion of statistical and predictive knowledge, b) inclusion of triplets in 'concept-relationship-concept' format without having hierarchical relationships, c) lack of fully automated KG construction framework, and d) use of either only structured or unstructured clinical sources. Additionally, the focus of existing knowledge graphs is limited to offering personalized treatment recommendation. However, ASKG aims to offer knowledge for both treatment, and prognosis. Lastly, none of the existing framework has explored knowledge curation for neurology or neurosurgery domain.

3. Methods

3.1. Dataset

Structured, and unstructured data (progress notes, discharge summaries, and radiology reports) was procured following an approved institutional review board (IRB No 11254). Structured data was prepared manually by analyzing unstructured clinical notes and brain angiograms (i.e. MRA, DSA, and CTA) of patients diagnosed with an intracranial aneurysm from 1997 through 2017. A total of 1025 patient records, 925 from structured data and 100 unstructured de-identified clinical notes from Epic ("Software | Epic," n.d.), were considered.

3.2. Automated Stroke Knowledge Graph (ASKG) Curation Framework

Our methodology of automated knowledge graph curation, by using structured and unstructured data, comprises of following layers: semantic knowledge, statistical knowledge, predictive knowledge, and knowledge factory. Fig 2 depicts four layers of ASKG along with modules within each layer.

The proposed knowledge graph contains relationships to represent following types of knowledge: graph of each cohort (group of patients with same statistical characteristics) showing concepts and relationships at semantic layer, inferential (statistical) knowledge derived from graph of cohorts at statistical layer, and predictive knowledge at predictive layer. The semantic layer of ASKG constructs fourteen (14) cohorts (in form of subgraph) based on aneurysmal arterial location (see Table 01 for arterial locations). In order to use semantic knowledge, the applications consuming knowledge graph would be requiring various statistical measures such as risk ratio and odd ratio of sub-cohort (e.g. Caucasian patients having aneurysm on middle cerebral arterial location). Similarly, for any application to consume predictive knowledge, it is important to see the statistical confidence such as accuracy/precision/recall of model, that was used to generate this predictive knowledge. Therefore, statistical layer encodes inferential knowledge of both semantic and predictive knowledge. Fig 3 graphically illustrates graph having semantic, statistical and predictive knowledge of middle cerebral artery location cohort which is generated from two excerpts of clinical notes. The detail of each layer is given below.

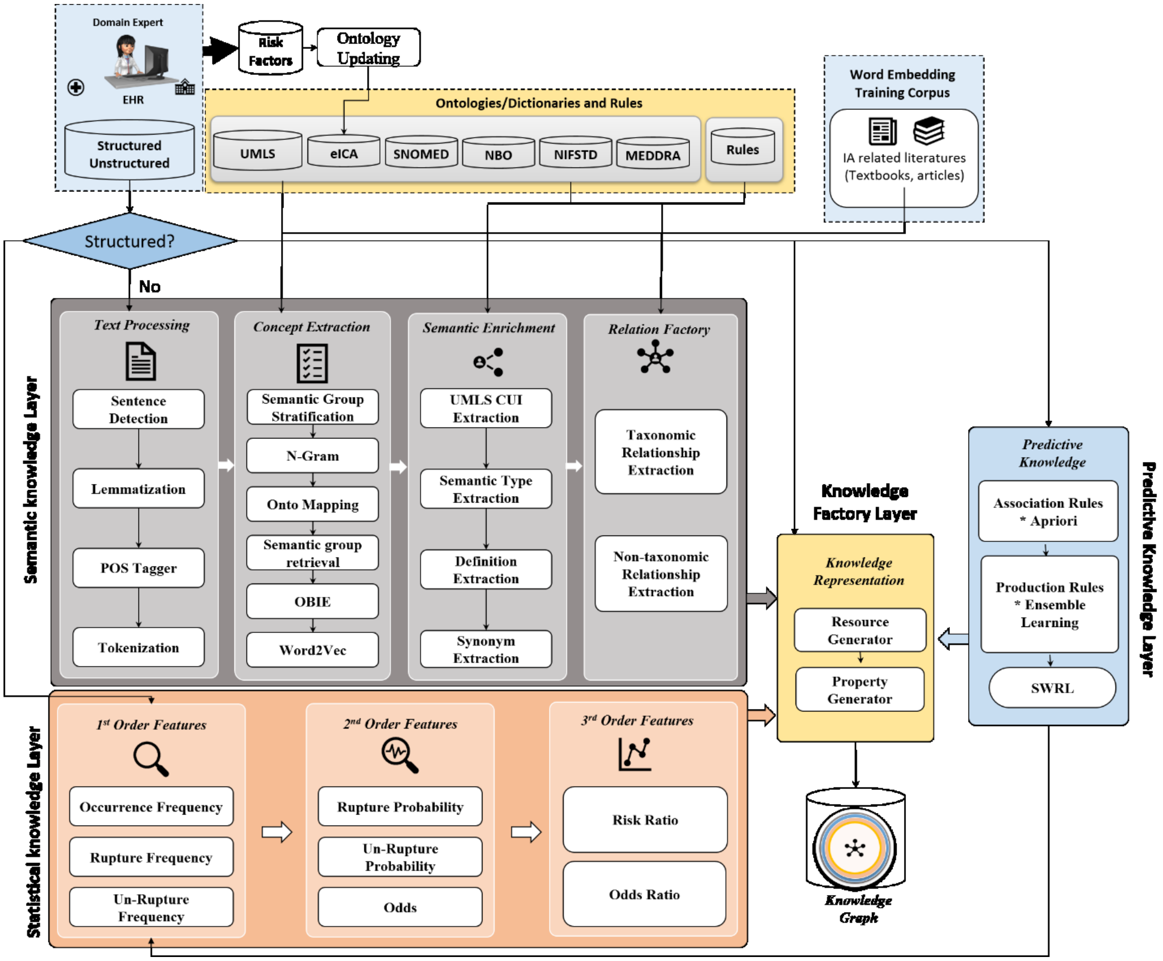


Fig. 2. Functional view of Automated Knowledge Graph Architecture

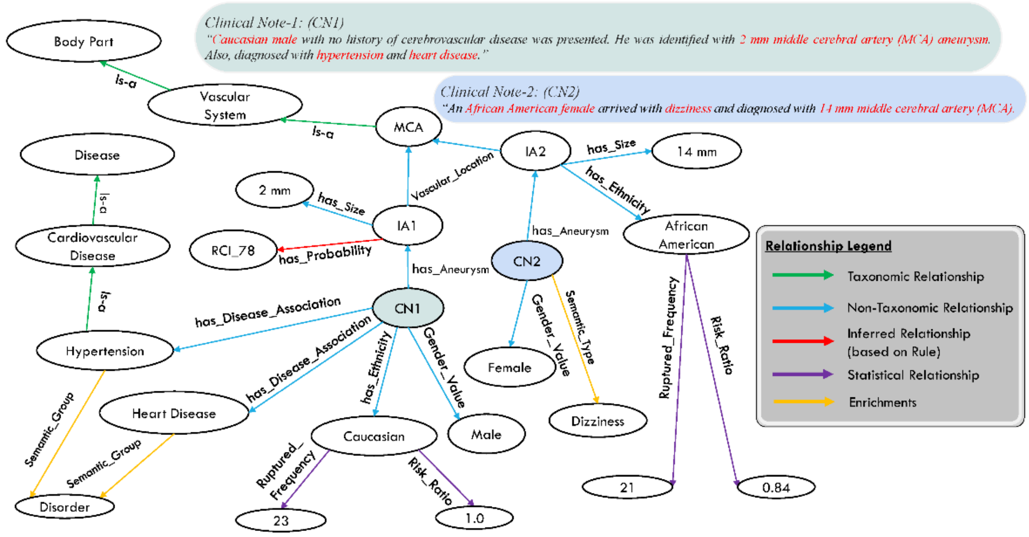


Fig. 3. An excerpt/sub-graph of saccular aneurysm KG of Middle Cerebral Artery defining IA risk factors based on two clinical texts.

3.2.1. Semantic Knowledge Layer

An enormous amount of valuable information relevant to prognosis, diagnosis and treatment is mainly present in clinical text such as progress summaries, radiology reports etc. This voluminous plain text (without any structure or description) could be translated into a structured format and linked to other resources, however, it requires text manipulation tasks provided by techniques such as natural language processing (NLP), information extraction, and information retrieval. By employing these techniques, two main elements are typically extracted and semantically annotated from text: named entities and the semantic relationships between them. The extraction of such elements and their representation on the Semantic Web are the main components of a task known as Relation Extraction and Linking (REL). Broadly speaking, the output of the REL task is a (RDF) graph that currently is known as a Knowledge Graph (KG), wherein nodes represent named entities and edges refer to the semantic relationship between them. In ASKG, the semantic relationships consist of inclusion and the association relationships that exist between intracranial aneurysm concepts. Although the scope of ASKG is not limited to have on extraction of REL, the semantic knowledge layer aims to have REL as a first step towards deducing statistical and predictive knowledge in other layers for futuristic clinical decision support systems. The main challenges of REL is to process the large scale and heterogeneous text by dealing with linguistic problems such as detection of synonymy and ambiguity, entity linking to existing ontologies/KGs, entities selection if entities linked to multiple concepts in ontologies, property linking, and representation (Martinez-Rodriguez et al., 2018) . Additionally, it is important to understand what domain specific REL needs to be extracted to keep high precision and recall in domain specific knowledge graph construction.

Detecting concepts or performing entity linking, by using structured data and unstructured clinical notes from electronic health records (EHRs), the semantic knowledge layer aims to develop an automated schema of cerebral/Intracranial aneurysm and SAH domain. Mainly using clinical and radiology notes, it identifies entities from text that refer to specific concepts of interest, such as disease/disorder, symptom, medication, procedure, risk factor, aneurysm features such as vascular location, aneurysm size, vessel side, etc. and patient demographics such as gender, age, race/ethnicity and automatically enriches concepts with formal semantics by associating them to relevant concepts using LBO which has clean knowledge constructed and verified by domain experts. To further improve the precision and recall of domain specific knowledge graph construction, it applies word embeddings-based similarity between concepts in clinical data with cerebral aneurysm domains. Furthermore, it extracts well-defined relationships that are linked to the concepts occurring within a text automatically. The details of processes used in semantic layers are outlined below.

3.2.1.1. Text processing

The semantic layer performs various text processing tasks namely tokenization, sentence detection, stop-word removal, and lemmatization of input clinical text. Tokenization is the process that splits the artifacts into tokens. Sentence detection segments given clinical text into its constituent sentences, while stop-word removal is the process of removing the most common words, for example, common words such as “an” and “the,” which are not relevant to domain. Lemmatization is the process of grouping modified forms of a word into unique categories so that they can be analyzed as a single item. The details of these processes are available in our earlier work (Alobaidi, Malik, & Sabra, 2018). Algorithm 1 in Figure 4 represents the pseudo code for text processing.

3.2.1.2. Concept Extraction

The main task of the concept extraction module of the semantic knowledge layer is to identify the boundaries of entities, linking entities to related concepts in Linked Biomedical Ontologies (LBO), classifying concepts, pruning concepts, and recognizing aneurysm and subarachnoid hemorrhage concepts in clinical text.

For concept extraction, ASKG uses following two approaches: ontology-based information extraction and semantic similarity using word embeddings. To perform concept extraction using ontology-based information extraction, ASKG uses two sources of background knowledge: extended Intracranial aneurysm ontology (eICO) and LBO as well as clinical text as an input. eICO was extended from existing Intracranial aneurysm ontology (ICO) (Hsu et al., 2015) to formally and explicitly specify the concepts of cerebral aneurysm. Initially ICA had 483 entities only. To allow the flexible growth of ASKG, the following three steps were performed.

a) Following EUPATI (“Risk factors in health and disease - EUPATI,” n.d.) classification, ASKG declares following five classes at root level: behavioral, physiological, demographic, environmental, and genetic classes.
b) Next, we import all the risk factors defined in ICA ontology under one of above-mentioned five classes.

```
(1) ALGORITHM FOR TEXT PROCESSING
Step 1: Read Text
Step 2: Repeat Step 3 For Each Token In Text
Step 3:   If Token In Set {.,?,…,!!}
                    Construct a Sentence
        [END OF LOOP]
Step 4: Repeat Step 5 For Each Token In Sentence
Step 5:   If Token In Stopwords
                     Remove The Token From Sentence
        [END OF LOOP]
Step 6: END
```

```
(2) ALGORITHM FOR CONCEPT EXTRACTION
Step 1: Read Sentence
Step 2: Set WindowSize = 1-4
Step 3: Set Entity = a sequence of N words from
Sentence based on WindowSize
Step 4: Repeat step 5 While Entity != Null
Step 5:   Map Entity to LBO Class
          Set Candidate = Entity
        [END OF LOOP]
Step 6: Repeat Step 7 While Candidate  != Null
Step 7:  Map Candidate To UMLS Semantic group
         If Candidate belong to Semantic group of
interest
            Set Concept = Candidate
        [END OF LOOP]
Step 8: END
```

```
(3) ALGORITHM FOR IS-A RELATION EXTRACTION
Step 1: Read Concepts
Step 2: Repeat Step 3 For each Concept In Concepts
Step 3:   Get hierarchy using rdfs:subOfClass predicate from LBO
          Create graph_model(hierarchy)
        [END OF LOOP]
Step 4: Repeat Step 5 For Each concept pair(Ci,Cj)
Step 5:   Merge the pair(Ci,Cj) graph model
          If(Ci,subClassOf,Cj) or (Cj,subClassOf,Ci)
             Construct subClassOf relation
        [END OF LOOP]
Step 6: END
```

```
(4) ALGORITHM FOR RISKFACTOR RELATION EXTRACTION

Step 1: Read Concepts
Step 2: Repeat Step 3 For each Concept In Concepts
Step 3:   Map Concept To UMLS Semantic group
          If Semantic group In Set {Individual Behavior",
Gender, chemical, Disease),?,…,!!}
              Construct Riskfactor relation
          Else
              If (Concept in ICA Ontology)
                   Construct Riskfactor relation
        [END OF LOOP]
Step 6: END
```

Fig. 4. Semantic knowledge layer algorithms

c) Lastly, we allow a domain expert to add risk factors at run time and update the ASKG. The domain expert updates a predefine CSV file with list of risk factors by specifying risk factors and corresponding type. For example, (risk factors: smoking, risk factor type: behavioral) adds smoking under class 'Behavioral'. The module reads the entered risk factors and updates the ASKG accordingly. A total of 60 risk factors were added by domain expert in this step.

Although ASKG formed so far could be used to perform ontology-based information extraction (OBIE) using input clinical text, it still lacks knowledge of comorbid conditions and associated neurological diseases. Therefore, to further increase the coverage of ASKG, we perform OBIE uses following linked biomedical ontologies (LBO): Neurobehavior Ontology (NBO) , Neuroscience Information Framework (NIF) Standard Ontology (NIFSTD), Unified Medical Language System (UMLS), SNOMED CT (Systematized Nomenclature of Medicine -- Clinical Terms), and Medical Dictionary for Regulatory Activities (MEDDRA).

To identify the boundaries of entities in the given clinical text before performing OBIE, the concept extraction module performs n-gram analysis (Jurafsky & Martin, n.d.) and link entities to related concepts in linked biomedical ontologies based on semantic group and ASKG classes (Table 2), i.e. concepts that are inapt are discarded. This classification of identified concepts based on semantic group and ASKG classes will aid in establishing relationships between the concepts. If semantic group/type poses to be non-granular for concept classification, then ASKG superclass is considered. N-grams are combinations of adjacent words or letters of length n found in the source text, clinical notes in this case. For example, the grouped term "blood pressure", constitutes 2-grams (or "bigrams"). A range of 4-grams were considered for our analysis.

Algorithm 2 in Figure 4 shows pseudo code for concept extraction. The primary aim of this algorithm is to discover the concepts of domain of interest by performing OBIE: mapping/linking entities in clinical to classes of LBO and eICO. The inputs are a) syntactic structure, which is the output of text processing component; b) the window size (number of words) within which to construct ngrams; and c) the list of semantic groups that are targeted.

Table 01: Process of concept Identification

Concept Type	Concepts	Concept Identification process
Location	ICA, MCA, etc.	eICO superclass: Brain region
Size (in mm)	2 mm, 5 millimeters, etc.	Rule-Based
Side	Left, Bilateral, etc.	eICO superclass:Laterality
Type	Saccular, Fusiform, etc.	Rule-Based
Status	Ruptured, Un-ruptured	Rule-Based
Gender	Female, Male, etc.	Rule-Based
Age (in years)	26 yrs., 38-Year-old, etc.	Rule-Based
Ethnicity	Asian, Caucasian, etc.	Rule-Based
Disorder	Hypertension, Diabetes, etc.	Semantic Group: Disorder
Sign or Symptom	Blurred vision, weakness, etc.	Semantic Type: Sign or Symptom
Drugs/ Procedures	Topamax, Clipping, etc.	Semantic Group: Chemicals or Drugs
Risk Factor	Smoking, Family History, etc.	eICO superclass: Aneurysm_Risk_factors
Tests & Procedures	MRI, CT, EEG, etc.	Rule-Based

After performing concept extraction using OBIE, there is possibility of low recall for concept extraction during entity linking which is process of mapping clinical entities in text to existing classes of knowledge graph. Additionally, several entities might appear as the subject or object of the relationship during entity linking. However, existing REL techniques in KGs rely on rule-based technique to restrict certain type of entities, for entity selection, during OBIE. However, such restrictions limit the result to a specific kind of relationships to be extracted. Likewise, existing approaches (Dutta, Meilicke, & Stuckenschmidt, 2015; Martinez-Rodriguez et al., 2018) map relation phrases to KG properties through generated rules and distance-based text similarity measures. However, such mappings may not always occur, and an alternative solution is therefore required. Therefore, as next step, we perform word embedding based similarity using google word2vec ("Software | Epic," n.d.) to filter concepts related to Intracranial Aneurysm. For example, Tylenol is prescribed for headache along with Plavix captured in the same clinical note, here Tylenol cannot be considered to establish the relationship "treated_by" for IA whereas Plavix can be. Tylenol was selected as concept during OBIE using LBO and eICO as it was present in both clinical text and LBO.

We used Word2Vec model implementation in python Gensim library ("Gensim · PyPI," 2018) to get a similarity score between concepts extracted from clinical notes and Intracranial Aneurysm. We used skip gram implementation of word2Vec, since Skip gram shows better performance than CBOW, due to its ability to capture different semantic for same word. Source code for our Word2Vec model implementation can be found in GitHub repository ("GitHub - Word2Vec Implementation," n.d.). The following three-step approach was used for this implementation.

Step- 1 The first step in building the model was to train it on customized corpus. We used large scale (53,378 Kilo Bytes of training corpus) and diverse corpus from published book (Ringer, n.d.), and PubMed abstracts relevant to IA ("Home - PubMed - NCBI," n.d.).

Step- 2 To make sure that concepts and keywords in the trained model has same meaning, we customized the model with synonyms and brain aneurysm related phrases. For example, we made sure that 'Anterior Communicating Artery' is treated as one phrase in the model. The other hyper-tuning parameters in the model are set as below:

- Size (number of dimensions of the embedding) - 150
- Window (maximum distance between target words and words around target word) -10
- min_count (minimum count of words to consider while training model) -2
- Workers- (number of partitions during training)-1

Step- 3 We passed each of the extracted concepts to the trained model and received a score by considering its semantic similarity to IA. We obtained the score between the range of -0.27 to 0.71, and analysis was performed to determine the threshold in order to differentiate concepts that are similar/related to IA. Sample concepts and their similarity scores with respect to IA can be found in the supplementary material.

Table 02 shows all the risk factors extracted from structured and unstructured clinical data. We categorize them by aneurysmal features, demographic features, symptomatic features, comorbid/co-occurring disorders, treatments, physical examinations, family history, and smoking history.

Table 02: Features derived from structured and unstructured data

Aneurysmal Features			
Vascular Location	ACoA	Size	Giant (>= 22.6 mm)
	BA SCA		Large (14.6-22.5 mm)
	Basilar Tip		Medium (8.3-14.5 mm)
	Basilar Trunk		Small (4.8-8.2 mm)
	Cavernous Carotid		Tiny (<= 4.7 mm)
	Distal Branch	Side	Bilateral
	MCA		Left
	PICA		Right
	Paraclinoid		Midline
	Pericallosal	Type	Saccular
	SICA		Fusiform
	vertebral artery		Dissecting
	petrous segment	Multiple Aneurysms	Yes
	carotid terminus		No
Demographic Features			
Age	Baby Boomers (56-73 Years)	Ethnicity	Asian/Oriental
	Generation X (38-55 Years)		Black/African American
	Generation Y (<=37 Years)		Native American
	Silent Generation (>=74 Years)		White/Caucasian
Gender	Male		Other
	Female		

Symptomatic Features	Comorbidities	Treatments	Smoking related
Aneurysmal dilatation	COPD	Procedures	Alcohol
Blurred Vision	Anxiety	Bypass	Cigarette
Diplopia	Aphasia	Clip	Current Smoker
Dizziness	Brain damage	Coil	Former Smoker
Face pain	Clot	Craniectomy	Illicit drug
Head pressure	Connective tissue disorder	Embolization	Never Smoker
Headache	Cranial nerve palsy	Endovascular	Smoking
Ischemic	Dementia	External ventricular drain	Tobacco
Motor Deficits	Diabetes	Eyelid weight	Cigar
Nausea	Deconjugate gaze	Flow diversion	Smokeless
Neck remnant	Embolism	Flow reversal	Other features
Neurological symptoms	Facial sensory deficit	Stent	Family related

Nystagmus	Glucophage	Stent Coil	Brother
Obesity	Heart Disease	Trapping	Father
Ptosis	Hydrocephalus	Physical examinations	Mother
Speech Deficits	Hypertension	Angiogram	Sister
Spinning Feeling	Intraventricular hemorrhage	CTA	
Tired eyes	Kidney disease	MRA	
Visual changes	Polycystic Kidney Disease	MRI	
Vomiting	Pronator drift		
Weakness	Pupillary defect		
Medications	Seizure		
Antiplatelet	Sixth nerve palsy		
Aspirin	Stroke		
Minipress	Subarachnoid hemorrhage		
Motrin	Thromboembolic		
Plavix	Thrombosis		
	Vasospasm		

3.2.1.3. Semantic Enrichment

For the purpose of improving semantic interoperability in the knowledge graph, the semantic enrichment module aims to enrich entities (and implicitly the related resources) with formal semantics by associating them to relevant concepts defined in LBOs that are considered in the current research. The semantic enrichment allows the proposed framework to generate one resource for a concept that might be referred in text in different ways. The semantic enrichment module reads all discovered entities and add relevant additional, well-defined metadata to knowledge graph. This addition of metadata in our knowledge graph will help knowledge consuming applications (such as chatbot) to achieve abstraction and contextualization. Abstraction is an inference understandable to users of application which is derived from domain specific raw data (e.g. by using the 'semantic type' of concept), while contextualization is property of knowledge consuming application to deal with variation (e.g. by performing word sense disambiguation using 'definition' property of concept). An example of semantic enrichment output is given in Table 03 and Figure 03.

The enrichment process is summarized as follows:

1) Getting synonyms for each discovered concept by querying the value of label, altlabel properties of resources describing the concept from LBO.
2) Getting the preferred lexical label for each concept by retrieving the prefLabel property of the concept from LBO.
3) Retrieving the definition of the concept from LBO by querying definition and note for the preferable resource.
4) Acquiring the semantic type of a concept by mapping it to semantic type. Since a concept might map to more than one semantic type, the proposed framework considers all of them.
5) Acquiring the UMLS concept unique id (CUI) for the preferable resource.

Table 03: Example of Semantic Enrichment of concepts

Concepts	**Sample Semantic Enrichment of Concepts**	
	Enriched Label	*Remarks*
Brain Aneurysm	Synonyms	Cerebral Aneurysm, Intracranial Aneurysm
	PrefLabel	Brain Aneurysm
	Definition	Brain Aneurysm is an abnormal bulge or "ballooning" in the wall of an artery in the brain
	Semantic type	Disease or Syndrome(T047)
	UMLS CUI	C0751003
Internal Carotid Artery	Synonyms	ICA, arteria carotis interna, cranial carotid artery
	PrefLabel	Internal Carotid Artery
	Definition	A terminal branch of the left or right common carotid artery which supplies oxygenated blood to the brain and eyes.
	Semantic type	Body Part, Organ, or Organ Component (T023)
	UMLS CUI	C0007276

3.2.1.4. Relation Factory

The relation factory module plays a significant role by extracting selected a) taxonomic relationship and b) non-taxonomic/association relationship from LBO and adds to ASKG.

- Taxonomic relationship extraction aims to identify all taxonomic relationships such as (is a, was a, part of, such as, etc.) between two identified concepts within a clinical note if these concepts also exist in LBOs, however, ASKG represents these hierarchical relationship using ‘is-a’ relationship. Algorithm 3 in Figure 4 shows the pseudo-code for is-a relationship extraction. This task uses the results of semantic enrichment as input in line 1. A graph model is built for each concept along with its hierarchy by encoding the semantic enrichment facts and the predefined hierarchy relationships within LBO. A merge graph model is constructed, as shown in Figure 5 below, for each pair of concepts. SPARQL ask query is used to find if is-a relationship exists between the pair of concepts.

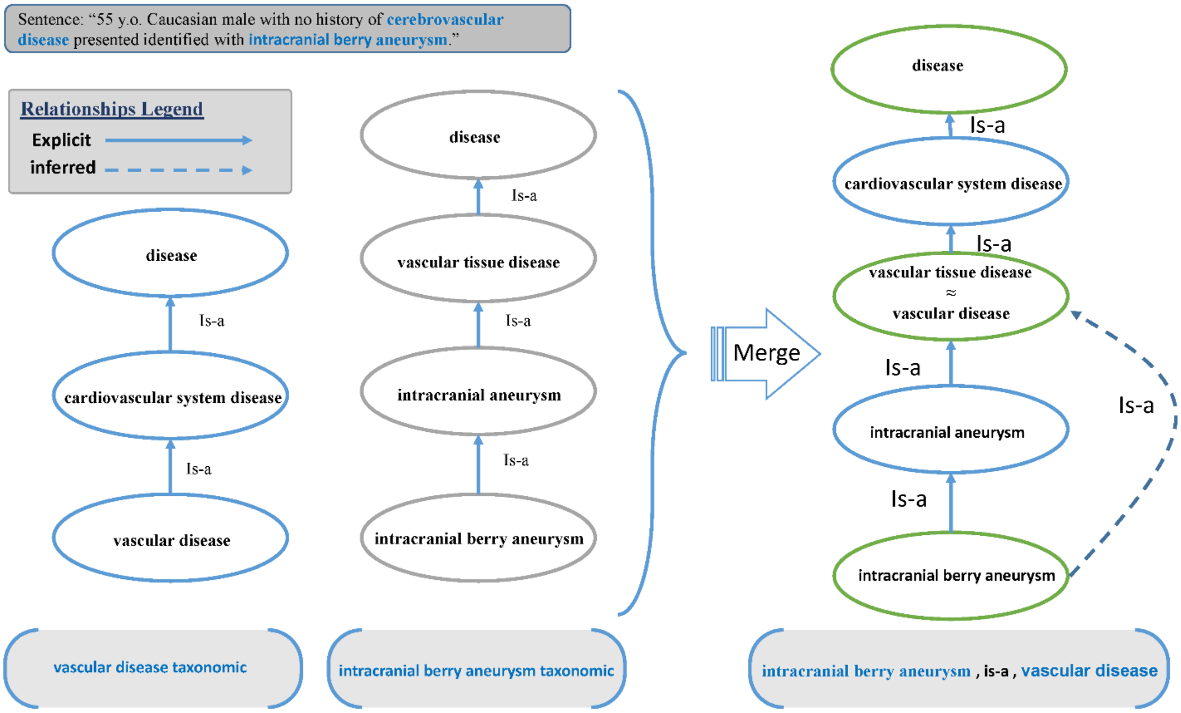

Sentence: "55 y.o. Caucasian male with no history of cerebrovascular disease presented identified with intracranial berry aneurysm."
Relationships Legend
Explicit
inferred
disease
Is-a
cardiovascular system disease
Is-a
vascular disease
vascular disease taxonomic
disease
Is-a
vascular tissue disease
Is-a
intracranial aneurysm
Is-a
intracranial berry aneurysm
intracranial berry aneurysm taxonomic
Merge
disease
Is-a
cardiovascular system disease
Is-a
vascular tissue disease ≈ vascular disease
Is-a
intracranial aneurysm
Is-a
intracranial berry aneurysm
Is-a
intracranial berry aneurysm , is-a , vascular disease

Fig. 5. Sample taxonomic relationship construction illustration

- Non-Taxonomic relationship extraction aims to discover relationship between biomedical concepts, in a clinical note or two items in relational/graph database for structured datasets, based on rule-based approach. The rule-based approach uses semantic relation repository and handcrafted rules. The semantic relation repository, built based on UMLS semantic network, includes 133 semantic types that are mapped to 14 semantic groups. Also, as mentioned earlier, eICO superclass is considered when semantic type or group is not sufficient in establishing the relationships. The proposed approach here is to find the associated relation between Clinical Note (CN) and concepts within CN by using enriched metadata. For example, we create "has_Disease_Association” relationship between "CN" and "diabetes" based on the predefine rules mapping "diabetes" to semantic group, which is a disorder. The examples of the crafted rules for concept relationships are given in Table 4. Algorithm 4 in Figure 4 shows the pseudo code for “has_riskfactor” relations extraction. The primary goal is to classify the concept with the risk factor category and then map the concept to an existing risk factor in the ICA ontology. If it gets mapped, then we consider it as a risk factor and create has_riskfactor relationship. Both taxonomic and non-taxonomic relationships for sample clinical notes can be found in supplementary material in web ontology language (OWL) format.

Once the automated schema of ASKG has been created, we create instances in graph database from each clinical note based on concepts and relationships in created schema/ontology. Concept CN can have multiple instances i.e. CN1, CN2 etc. that are linked by “type” relationship, each representing individual patient history/record. And within one clinical note there can be multiple aneurysms, IA1, IA1.1, etc., which are instances of concept IA. Similarly, all instances are created only if they have corresponding concept in schema/ontology. For structured data, entities in the relational database were mapped using OBIE as concepts, and the values were taken as instances in KG. The source code of algorithms 1-4 for semantic knowledge and statistical knowledge generation is available on the GitHub repository (“GitHub - TripleIE/ASKG (Source Code),” n.d.).

Table 4. An example of the crafted rules for object relationships

	Concepts		Instance Samples	
Object Relationship	**Domain**	**Range**	**Domain**	**Range**
has_Aneurysm	CN	IA	CN1, CN2, ...	IA1, IA2, …
Vascular_Location	IA	Location	IA1, IA2, …	ICA, MCA, etc.
has_Size	IA	Size	IA1, IA2, …	2 mm, 5 millimeters, etc.
Vessel_Side	IA	Side	IA1, IA2, …	Left, Bilateral, etc.
has_Type	IA	Type	IA1, IA2, …	Saccular, Fusiform, etc.
Rupture_Status	IA	Status	IA1, IA2, …	Ruptured, Un-ruptured
Gender_Value	CN	Gender	CN1, CN2, ...	Female, Male, etc.
has_Age	CN	Age	CN1, CN2, ...	26 yrs., 38-Year-old, etc.
has_Ethnicity	CN	Ethnicity	CN1, CN2, ...	Asian, Caucasian, etc.
has_Disease_Association	CN	Disorder	CN1, CN2, ...	Hypertension, Diabetes, etc.
has_Symptom	CN	Symptom	CN1, CN2, ...	Blurred vision, weakness, etc.
Treated_by	CN	Drugs/Procedure	CN1, CN2, ...	Topamax, Clipping, etc.
has_riskfactor	CN	Smoking/Family History	CN1, CN2, ...	Smoking, Family History, etc.
Diagnosed_by	CN	Tests & Procedures	CN1, CN2, ...	MRI, CT, EEG, etc.
has_Multiple_Aneurysms	CN	Multiple Aneurysms	CN1, CN2, ...	Yes, No

3.2.2. Statistical Knowledge Layer

For applications to use inferential knowledge, ASKG provides various 1st order (e.g. statistical significance), 2nd order (derived from the 1st order), and 3rd order statistics (derived from 2nd order statistics). Both semantic and predictive knowledge layers add various statistical knowledge in the proposed knowledge graph. Figure 6 shows sample statistical relationships of proposed knowledge graph. For example, Risk ratio (RR), represented as 3rd order statistical relationship in ASKG, is well known measure of association in epidemiology. It is the ration of risk of an event in one group (e.g., exposed group) versus the risk of the event in the other group (e.g., nonexposed group). As shown in Figure 03, it is represented as property value of property such Caucasian. An RR of 1.0 indicates that there is no difference in risk between the groups being compared (e.g. Caucasian and African American). An RR more than 1.0 indicates an increase in risk among the exposed compared to the unexposed, whereas a RR <1.0 indicates a decrease in risk in the exposed group. Therefore, we can say Caucasian has 20% less chances of aneurysm rupture compared to African American. As shown in Figure 6 RR is derived from second order statistical relationship 'rupture probability' which itself is calculated from first order relationship 'occurrence frequency'. In order to understand the importance of statistical knowledge, consider an example of futuristic application that aims to educate resident physicians about the criticality of each location w.r.t. aneurysmal rupture. Rather than providing the application only with critical locations, if each location is tagged with statistical facts such as rupture probability and/or risk ratio, etc. indicating risk of rupture when compared to other locations, this would help them to make more informed decision making in their careers. Similarly, if such application also has to use the predictive knowledge, it can verify reliability of predictive knowledge by checking accuracy (along with precision and recall) of employed models, feature importance using p-value, or other measure such as confidence/predictive/tolerance interval, depending upon acquired knowledge, from statistical layer. These measures will assist applications to make informed decisions to whether use the knowledge of any specific model.

3.2.3. Predictive Knowledge Generation

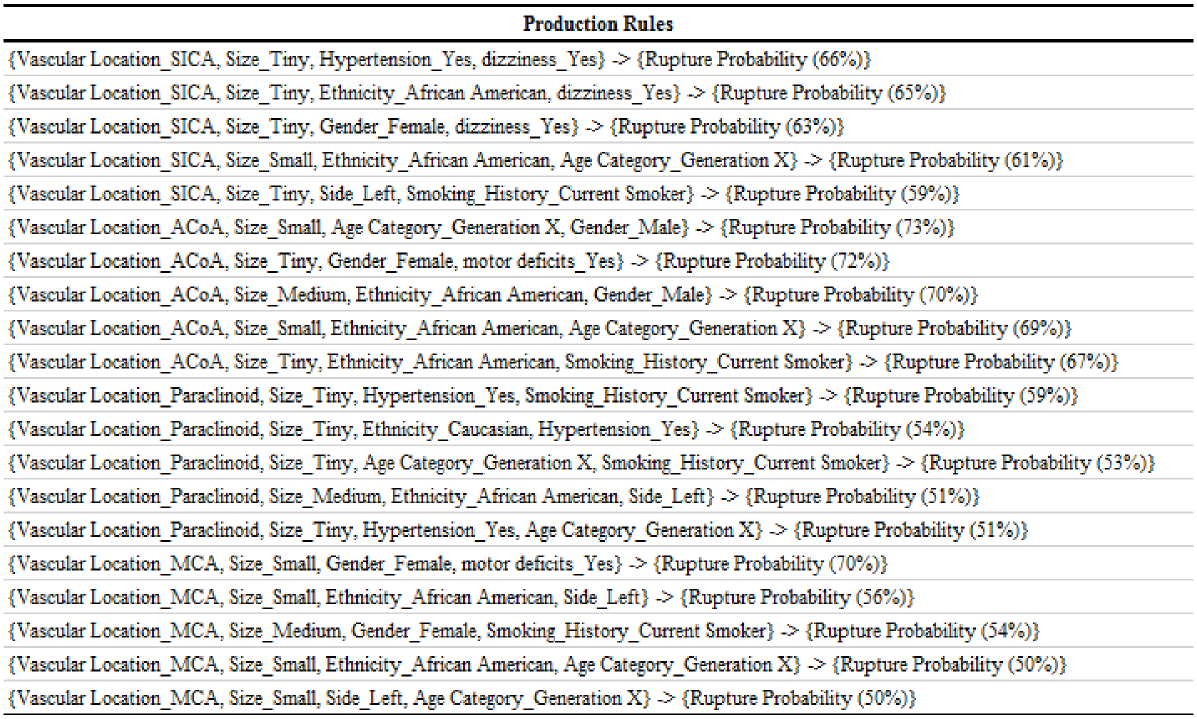
The predictive knowledge generation process creates rule-based knowledge for prediction of the SAH using hybrid approach that involves association rule mining and the ensemble learning. This knowledge enriches the significance of knowledge graph and the assertion knowledge it can provide. This predictive knowledge does not exist in any of existing KGs. The final predictive knowledge is expressed in the form of production rules in knowledge graph, and each rule has unique URI. The final production rules are created using following steps.

Statistical Features	Statistical Relation	Domain	Range	Sample RDF
1st Order	Occurance_Frequency	All Concepts	No. of CNs the concept was present	<owl:NamedIndividual rdf:about="&untitled-ontology-337;Middle_Cerebral_Artery"> <P-Value rdf:resource="&untitled-ontology-337;.004"/> <Occurance_Frequency rdf:resource="&untitled-ontology-337;1600"/> <Un-Ruptured_Frequency rdf:resource="&untitled-ontology-337;700"/> <Ruptured_Frequency rdf:resource="&untitled-ontology-337;900"/> </owl:NamedIndividual>
	Ruptured_Frequency	All Concepts	No. of CNs the concept was present, and IA was ruptured	
	Un-Ruptured_Frequency	All Concepts	No. of CNs the concept was present, and IA was un-ruptured	
	P-Value	All Concept	Value	
	True Positive	Model Name	Value	<owl:NamedIndividual rdf:about="&untitled-ontology-337;Model_1_(SVM)"> <False_Positive rdf:resource="&untitled-ontology-337;200"/> <False_Negative rdf:resource="&untitled-ontology-337;300"/> <True_Negative rdf:resource="&untitled-ontology-337;500"/> <True_Positive rdf:resource="&untitled-ontology-337;600"/> </owl:NamedIndividual>
	True Negative	Model Name	Value	
	False Positive	Model Name	Value	
	False Negative	Model Name	Value	
2nd Order	Rupture_Probability	All Concepts	Ruptured_Frequency/Occurrence_Frequency	<owl:NamedIndividual rdf:about="&untitled-ontology-337;Middle_Cerebral_Artery"> <Rupture_Probability rdf:resource="&untitled-ontology-337;56%"/> <Un-Rupture_Probability rdf:resource="&untitled-ontology-337;43%"/> <Odds rdf:resource="&untitled-ontology-337;.77"/> </owl:NamedIndividual>
	Un-Rupture_Probability	All Concepts	Un-Ruptured_Frequency/Occurrence_Frequency	
	Odds	All Concepts	Rupture_Probability/Un-Rupture_Probability	
	Accuracy	Model Name	Value	<owl:NamedIndividual rdf:about="&untitled-ontology-337;Model_1_(SVM)"> <Accuracy rdf:resource="&untitled-ontology-337;68%"/> <Precision rdf:resource="&untitled-ontology-337;75%"/> <Recall rdf:resource="&untitled-ontology-337;67%"/> </owl:NamedIndividual>
	Precision	Model Name	Value	
	Recall	Model Name	Value	
3rd Order	Risk_Ratio	All Concepts	Rupture_Probability (current concept)/Rupture_Probability (Reference Concept)	<owl:NamedIndividual rdf:about="&untitled-ontology-337;Middle_Cerebral_Artery"> <Risk_Ratio rdf:resource="&untitled-ontology-337;.70"/> <Odds_Ratio rdf:resource="&untitled-ontology-337;.43"/> </owl:NamedIndividual>
	Odds_Ratio	All Concepts	Odds (current concept)/Odds (Reference Concept)	
	F-Measure	Model Name	Value	<owl:NamedIndividual rdf:about="&untitled-ontology-337;Model_1_(SVM)"> <F1-Score rdf:resource="&untitled-ontology-337;71%"/> </owl:NamedIndividual>

Fig. 6. Sample Statistical Relationships and their encoding in ASKG

- Preprocessing: Randomized undersampling of unruptured data was performed to balance the data set, yielding 371 records each for both ruptured and unruptured cases. More specifically, we under sampled majority class (un-ruptured cases here) by selecting random sample (371) to have equal number of instances for both classes (ruptured and unruptured). This preprocessing was performed to ensure that machine learning model built on top of dataset is not biased towards un-ruptured class prediction as final production rules, which will be used for prediction of unseen cases will be generated based on this model.
- Furthermore, this dataset was segregated into training and validation based on 80:20 ratio. Python inbuilt library scikit-learn ("API Reference — scikit-learn 0.21.3 documentation," n.d.) was used for encoding categorical label values (such as Gender, Location, Size) to numeric for the data to be processed by machine learning algorithms.
- Association Rules: The process of association rule generation was used to acquire production rules. These rules are later used in prediction of aneurysm rupture or subarachnoid hemorrhage. For the formation of the association rules initially, Apriori algorithm (Ng, n.d.)was run separately on ruptured and unruptured dataset to acquire frequently occurring feature combinations. The undersampling performed in preprocessing stage will also guarantee that If the confidence of the same rule is greater in un-ruptured dataset, the rule is rejected and not converted to production rule. This indicates that presence of certain features (rule set) will not always guarantee that case will be ruptured since confidence is more at un-ruptured end. Association rules for locations Supraclinoid Internal Carotid Artery (SICA), Anterior Communicating Artery (ACoA), Paraclinoid and Middle Cerebral Artery (MCA) with tiny, small and medium size were obtained based on 20% support and 50% of confidence rules. The rest of aneurysmal locations were discarded due to lack of statistical significance.
- Production Rules: Random Forest Classifier (RFC), Decision Tree Classifier, Support Vector Machine (SVM), Adaptive Boosting (AdaBoost), and Gradient Boosting (GBoost) models were used to classify between ruptured and un-ruptured aneurysm class with a default boundary probability of 0.5 for ruptured (>=.5) and un-ruptured (<.5) cases. Following this, ensemble learning (Rokach, Schclar, & Itach, 2014) was used to balance out individual weaknesses of the classifiers. Voting Classifier (ensemble technique) and Weighted Average Probabilities (Soft Voting) was used to predict the final label. The validation data was used to

run association rules and find mean probability of rupture against each rule produced through an Apriori algorithm. Furthermore, these rules are converted into a set of consistent Semantic Web Rule Language (SWRL) rules and integrated into the KG. A total of 260 production rules were generated, 27 rules of which belonged to ACOA, 57 to SICA, 84 to MCA, and 92 to Paraclinoid. Figure 7 displays the top 5 ranked rules for each location. A complete list of rules can be found in supplementary materials. The source code of predictive knowledge generation is available on ("GitHub - Predicitiveknowledge Implementation (Source Code)," n.d.).



Production Rules
{Vascular Location_SICA, Size_Tiny, Hypertension_Yes, dizziness_Yes} -> {Rupture Probability (66%)}
{Vascular Location_SICA, Size_Tiny, Ethnicity_African American, dizziness_Yes} -> {Rupture Probability (65%)}
{Vascular Location_SICA, Size_Tiny, Gender_Female, dizziness_Yes} -> {Rupture Probability (63%)}
{Vascular Location_SICA, Size_Small, Ethnicity_African American, Age Category_Generation X} -> {Rupture Probability (61%)}
{Vascular Location_SICA, Size_Tiny, Side_Left, Smoking_History_Current Smoker} -> {Rupture Probability (59%)}
{Vascular Location_ACoA, Size_Small, Age Category_Generation X, Gender_Male} -> {Rupture Probability (73%)}
{Vascular Location_ACoA, Size_Tiny, Gender_Female, motor deficits_Yes} -> {Rupture Probability (72%)}
{Vascular Location_ACoA, Size_Medium, Ethnicity_African American, Gender_Male} -> {Rupture Probability (70%)}
{Vascular Location_ACoA, Size_Small, Ethnicity_African American, Age Category_Generation X} -> {Rupture Probability (69%)}
{Vascular Location_ACoA, Size_Tiny, Ethnicity_African American, Smoking_History_Current Smoker} -> {Rupture Probability (67%)}
{Vascular Location_Paraclinoid, Size_Tiny, Hypertension_Yes, Smoking_History_Current Smoker} -> {Rupture Probability (59%)}
{Vascular Location_Paraclinoid, Size_Tiny, Ethnicity_Caucasian, Hypertension_Yes} -> {Rupture Probability (54%)}
{Vascular Location_Paraclinoid, Size_Tiny, Age Category_Generation X, Smoking_History_Current Smoker} -> {Rupture Probability (53%)}
{Vascular Location_Paraclinoid, Size_Medium, Ethnicity_African American, Side_Left} -> {Rupture Probability (51%)}
{Vascular Location_Paraclinoid, Size_Tiny, Hypertension_Yes, Age Category_Generation X} -> {Rupture Probability (51%)}
{Vascular Location_MCA, Size_Small, Gender_Female, motor deficits_Yes} -> {Rupture Probability (70%)}
{Vascular Location_MCA, Size_Small, Ethnicity_African American, Side_Left} -> {Rupture Probability (56%)}
{Vascular Location_MCA, Size_Medium, Gender_Female, Smoking_History_Current Smoker} -> {Rupture Probability (54%)}
{Vascular Location_MCA, Size_Small, Ethnicity_African American, Age Category_Generation X} -> {Rupture Probability (50%)}
{Vascular Location_MCA, Size_Small, Side_Left, Age Category_Generation X} -> {Rupture Probability (50%)}

Fig. 7: Top 5 ranked rules for each location

3.2.4. Knowledge Factory Layer

This component is a key constituent of our proposed framework. This layer aims to encode semantic, statistical, and predictive knowledge into formal ontological knowledge. It automates the process of encoding the output of concept extraction component (concepts), semantic enrichment output (e.g. synonym, definition), relation factory output (taxonomic and non-taxonomic relationships), and statistical knowledge in OWL (and hence RDF, RDFs), and SKOS format. We selected W3C specifications ontologies over the Open Biomedical Ontologies (OBO) format because the former provides well-defined standards for semantic web and supports the inference of complex properties based on rule-based engines. Additionally, predictive knowledge is added in form of SWRL rules.

Figure 3 shows a sample visual representation of relationships between entities in the knowledge graph. Concepts and instances in the knowledge graph are associated with patient specific relations or IA specific relations. For example, a patient associated with clinical note #10 (CN10) is related to IA10 with "has_Aneurysm" relationship. IA10 is related to instances ACoA, left, small and unruptured with relationships "vascular_location"; "vessel_side";" has_size" and "rupture_status" respectively. Also, CN10 is related to instances specific to patient features such as hypertension, male, cigarette etc. with defined relationships as represented in Table 4.

4. Results and Evaluation

4.1. Concepts and Relationship Extraction

The proposed approach develops and experimentally applies task-based evaluation (Hobson, Dorr, Monz, & Schwartz, 2007) , to evaluate concept and relationships extraction, separately. We employed 100 clinical notes as an input and as gold standard consisting of clinical notes annotated by experts (for samples see supplementary material) for concept extraction evaluation. Furthermore, we built a gold standard for broader taxonomy relationships for all disorder concepts, identified in clinical notes, using Disease Ontology (DO) ("Disease Ontology - Institute for Genome Sciences @ University of Maryland," n.d.).

The performance measure for concept and relationships extraction includes precision, recall, and F-measure. For concept extraction, we compared results of OBIE approach with OBIE + concepts filtering using word embedding approach. For embedding model, we used window size=3. The upper limit of 0.025 and lower limit of 0.15 was used as thresholds to select final concepts from score given by embeddings.

$$\text{Precision} = \frac{\text{Correct retrieved Concepts/relations}}{\text{Total retrieved Concepts/relations}} \quad (1)$$

$$\text{Recall} = \frac{Correct\ retrieved\ Concepts/relations}{Total\ correct\ concepts/relations} \quad (2)$$

$$\text{F-measure} = 2 \text{ x } \frac{Precision\ x\ Recall}{Precision + Recall} \quad (3)$$

Figure 8 (a) shows the that hybrid OBIE and Embedding approach performs better than OBIE alone for concept extraction. Figure 8 (b) shows a comparison between recall and precision of hierarchical relationship extraction. In this evaluation, we use different string-matching methods. The first technique is finding strings that exactly match, where the second technique is Soundex, which converts an alphanumeric string to a four-character code based on how the string sounds when spoken (US1435663A, 1922), and the final technique is Levenshtein distance, which finds strings that match a pattern approximately (rather than exact-match) ("The Levenshtein Distance Algorithm - DZone Big Data," n.d.).

4.2. Ensemble Model Evaluation

Machine learning model was trained on 80% of data and performance was evaluated using 7-fold cross validation (Brownlee, 2018). Figure 8 (c) shows the performance comparison, of individual classifiers and ensembled one, in terms of accuracy, precision, recall, and F1 score. The final evaluation report contains averaged values of all these parameters across 7 folds.

4.3. Rupture Evaluation based on Rules

Unseen 30 patient records were used to evaluate rules of constructed KG. If testcase matches more than 1 rule, average of the probability assigned for each rule is considered as represented in Table 5. If the average probability is >=50%, IA is predicted as "ruptured".

For the 30 test records, the number of actual ruptured and unruptured aneurysms were compared against the predicted number of ruptured and unruptured aneurysms. Results can be interpreted from 2 perspectives. Perspective 1 considering ruptured status as predicted class and perspective 2 considering unruptured status as predicted class. Since majority is un-ruptured in test set, we see precision and recall being higher in second perspective. While the accuracy of 0.73 remains the same in both scenarios. Precision, recall, and F1 scores are shown in the Figure 8 (d).

Unseen 30 patient records were used to evaluate rules of constructed KG. If testcase matches more than 1 rule, the average of the probability assigned for each rule is considered as represented in Table 5. If the average probability >=50%, IA is predicted as "ruptured".

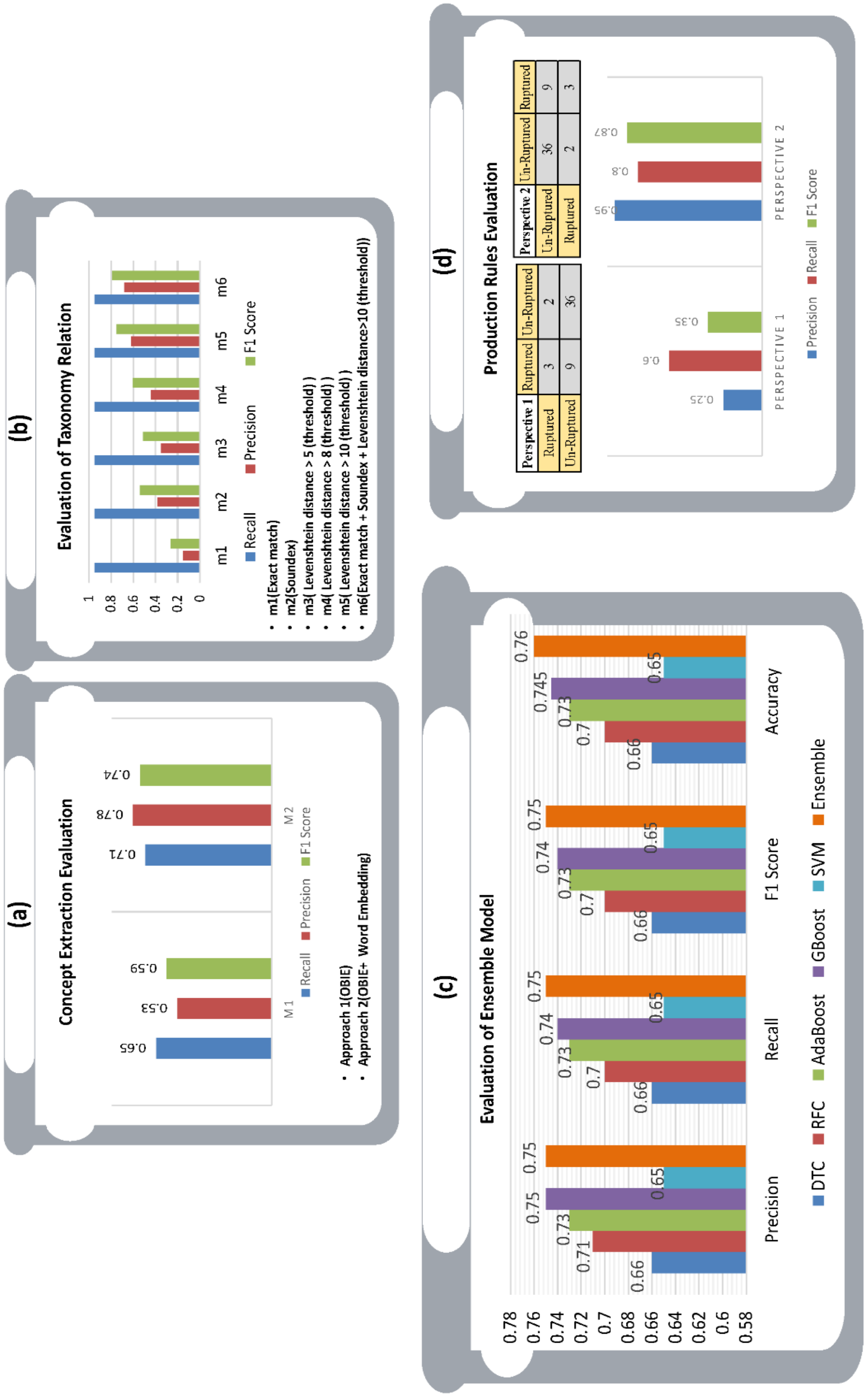
5. Discussion

Most of the existing works (Gyrard et al., 2018; Song et al., 2015; Yuan et al., 2019) have focused on knowledge curation using biomedical scientific publications (e.g. PubMed). The limitations of these works are as follows: a) the extracted knowledge is limited to concepts and relationships only, which lack completeness of knowledge due to lack of detailed clinical entities and relations about specific domain and b) the inability to generate knowledge (facts or rules) by applying data driven approaches that can be used for clinical decision making.

Recently there have been some attempts (Finlayson et al., 2014; Rotmensch et al., 2017; Shi et al., 2017) on individual aspects of automated KG construction such as concept or relationship extraction from EHRs. For example, (Rotmensch et al., 2017) constructs KGs constituting disease-symptom relationship using concept extraction from EHRs by linking diseases with symptoms. However, number of relationships are limited to 1-3 (such as disease-symptom, disease-disease, drug-drug, drug-disease patterns). Additionally, none of these approaches provide consumable knowledge built using data-driven approaches in forms of rules.

To the best of our knowledge, ASKG is the first fully automated attempt towards construction of KG schema and curation of knowledge from structured, unstructured clinical data along with LBO. The schema construction in KGs so far have remained manual, and thus this paper has introduced a novel hybrid approach of OBIE and word embeddings for concept and relationships extraction. The evaluation of final extracted concepts and hierarchical relationships from LBO based on identified concepts in structured and unstructured data sources shows promising results. The use of this hybrid approach also improves the accuracy of relationship extraction.

Unlike some preliminary studies (Finlayson et al., 2014; Gyrard et al., 2018; Rotmensch et al., 2017; Shi et al., 2017; Song et al., 2015), concept extraction in ASKG is not limited to only disease, symptoms, or treatments/drugs, but it also includes risk-factors, patient demographics, medical history, and condition specific features to provide a wholesome and more granular approach for futuristic knowledge-based clinical applications.



Perspective 1	Ruptured	Un-Ruptured
Ruptured	3	2
Un-Ruptured	9	36

Perspective 2	Un-Ruptured	Ruptured
Un-Ruptured	36	9
Ruptured	2	3

Fig. 8. Evaluation and results

Table 5. Predicted and expected rupture statuses for test cases based on rules

	Actual	Rule-based Prediction			
Patient Records	Status	No. of Rules Matched	Probabilities	Probability (Avg)	Status
IA1	Ruptured	1	53	53	Ruptured
IA2	Ruptured	4	41, 43, 49, 50	45.75	Un-Ruptured
IA3	Ruptured	1	69	69	Ruptured
IA4	Ruptured	1	57	57	Ruptured
IA5	Ruptured	7	43, 46, 49, 50, 51, 53, 54	49.43	Un-Ruptured
IA6	Un-Ruptured	5	38, 41, 42, 47, 50	43.6	Un-Ruptured
IA7	Un-Ruptured	6	55, 56, 57, 58, 61, 67	59	Ruptured
IA8	Un-Ruptured	0		0	Un-Ruptured
IA9	Un-Ruptured	1	38	38	Un-Ruptured
IA10	Un-Ruptured	2	64, 66	65	Ruptured
IA11	Un-Ruptured	7	36, 39, 41, 43, 49, 51, 59	45.43	Un-Ruptured
IA12	Un-Ruptured	0		0	Un-Ruptured
IA13	Un-Ruptured	4	43, 44, 49, 50	46.5	Un-Ruptured
IA14	Un-Ruptured	1	57	57	Ruptured
IA15	Un-Ruptured	5	41, 43, 44, 49, 50	45.4	Un-Ruptured
IA16	Un-Ruptured	3	41, 43, 49	44.33	Un-Ruptured
IA17	Un-Ruptured	3	55, 56, 57	56	Ruptured
IA18	Un-Ruptured	2	45, 63	54	Ruptured
IA19	Un-Ruptured	1	38	38	Un-Ruptured
IA20	Un-Ruptured	1	39	39	Un-Ruptured
IA21	Un-Ruptured	3	34,38, 41	37.67	Un-Ruptured
IA22	Un-Ruptured	0		0	Un-Ruptured
IA23	Un-Ruptured	0		0	Un-Ruptured
IA24	Un-Ruptured	7	45, 47, 49, 50, 52, 54, 56	50.43	Ruptured
IA25	Un-Ruptured	3	55, 56, 57	56	Ruptured
IA26	Un-Ruptured	5	39, 41, 42, 49, 50	44.2	Un-Ruptured
IA27	Un-Ruptured	2	49, 52	50.5	Ruptured
IA28	Un-Ruptured	0		0	Un-Ruptured
IA29	Un-Ruptured	5	34, 39, 46, 48, 51	43.6	Un-Ruptured
IA30	Un-Ruptured	5	41, 43, 44, 48, 49	45	Un-Ruptured
IA31	Un-Ruptured	5	51, 34, 46, 38, 39	41.6	Un-Ruptured
IA32	Un-Ruptured	2	40,43	41.5	Un-Ruptured
IA33	Un-Ruptured	0		0	Un-Ruptured
IA34	Un-Ruptured	6	42,54,35,36,38,39	40.67	Un-Ruptured
IA35	Un-Ruptured	10	40,42,41,43,56,61,46,37,38,39	44.3	Un-Ruptured
IA36	Un-Ruptured	0		0	Un-Ruptured
IA37	Un-Ruptured	3	44,46,48	46	Un-Ruptured
IA38	Un-Ruptured	3	51,43,49	47.67	Un-Ruptured

IA39	Un-Ruptured	3	41,43,49	44.33	Un-Ruptured
IA40	Un-Ruptured	4	40,41,43,44	42	Un-Ruptured
IA41	Un-Ruptured	2	49,50	49.5	Un-Ruptured
IA42	Un-Ruptured	1	38	38	Un-Ruptured
IA43	Un-Ruptured	0		0	Un-Ruptured
IA44	Un-Ruptured	1	45	45	Un-Ruptured
IA45	Un-Ruptured	1	45	45	Un-Ruptured
IA46	Un-Ruptured	0		0	Un-Ruptured
IA47	Un-Ruptured	6	41,43,34,46,38,39	40.16	Un-Ruptured
IA48	Un-Ruptured	7	52,50,56,54,45,47,49	50.43	Ruptured
IA49	Un-Ruptured	0		0	Un-Ruptured
IA50	Un-Ruptured	1	45	45	Un-Ruptured

The use of OBIE is particularly important to develop relationships in KG. For example, annotating each concept with semantic type makes it easier to establish a meaningful relationship. Considering the sentence “Patient X has blurred vision and COPD”, blurred vision and COPD seem to be conditions, however, blurred vision will be represented accurately as symptom and COPD as disease in the final knowledge graph due to semantic enrichment process. Existing studies have considered statistical analysis in identifying the significance of relationships, however, they have not included statistical values as a fact itself. In our proposed work, we tag statistical facts for each of the features that can provide a versatile insight into disease studies.

Yet another novelty of the proposed ASKG is to offer knowledge in form of production rules for applications to consume the required knowledge. To develop production rules, we have employed ensemble learning and an Apriori association mining algorithm.

6. Conclusion

Predicting a subarachnoid hemorrhage (SAH) is a complex, unsolved clinical challenge. There doesn’t exist any formal knowledge for SAH that could be used by consumer applications such as clinical decision support system tool. Moreover, no domain specific automated curation framework does exist that can extract knowledge from disparate sources and build formal knowledge to have following: a) representation of structured knowledge of each patient record in form of concepts, relationships; b) statistical knowledge based on complete dataset; c) predictive knowledge. Accordingly, this paper has presented an automated domain specific knowledge curation framework to develop a unified knowledge graph having different types of knowledge.

The framework supports different knowledge representations in a curated knowledge graph such as structural knowledge having concepts and relationships between them, statistical knowledge derived from semantic and predictive knowledge layers, and predictive knowledge represented in form of production rules.

The structured knowledge represented in form the concepts and relationship at semantic layer uses bioportal ontologies and rule-based approach along with word embedding approach. To develop predictive knowledge, association rules were developed using Apriori algorithm, and production rules were calculated by concatenating each association rule with its probability of rupture. The probability of rupture was calculated using ensemble learning of following machine learning classifiers: Random Forest, Decision Tree, Support Vector Machine, Adaptive Boosting, and Gradient Boosting.

The primary advantage of the proposed framework is the extensible nature of knowledge graph that can incorporate and dynamically account for newly discovered factors that may contribute to the progression and rupture of intracranial aneurysm. Also, the flexible nature of the framework allows incremental evaluation, accommodating newer multi-variable assessments (in form of rules) for aneurysm rupture.

The quantitative evaluation of the proposed framework shows that concepts and relationship extraction from the unstructured clinical text, statistical knowledge, and predictive knowledge using unseen patient data, are reliable. Future work will focus extending current KG to develop knowledge to understand natural history of cerebral aneurysms (which is unknown to clinicians) and for predictions of other types of stroke.

Acknowledgement

This research was funded partially by the Brain Aneurysm Foundation, Oakland University, and graciously supported by the invaluable data and domain expert assistance rendered by Neuroscience Institute, Henry Ford Health System, Michigan.

References

Alobaidi, M., Malik, K. M., & Sabra, S. (2018). Linked open data-based framework for automatic biomedical ontology generation. *BMC Bioinformatics*, *19*(1), 319. https://doi.org/10.1186/s12859-018-2339-3

API Reference — scikit-learn 0.21.3 documentation. (n.d.). Retrieved September 1, 2019, from https://scikit-learn.org/stable/modules/classes.html

Bizer, C. (2011). *Linked Data [BOOK]*. https://doi.org/10.4018/978-1-60960-593-3.ch008

Bresnick, J. (2017). Understanding the Many V's of Healthcare Big Data Analytics. Retrieved August 31, 2019, from https://healthitanalytics.com/news/understanding-the-many-vs-of-healthcare-big-data-analytics

Brownlee, J. (2018). A Gentle Introduction to k-fold Cross-Validation. Retrieved September 1, 2019, from https://machinelearningmastery.com/k-fold-cross-validation/

Costa, R. D. D. da. (2015). *Semantic enrichment of knowledge sources supported by domain ontologies*. *FCT: DEE - Teses de Doutoramento*. Retrieved from https://run.unl.pt/handle/10362/14076

Disease Ontology - Institute for Genome Sciences @ University of Maryland. (n.d.). Retrieved September 1, 2019, from http://disease-ontology.org/

Dutta, A., Meilicke, C., & Stuckenschmidt, H. (2015). Enriching structured knowledge with open information. *WWW 2015 - Proceedings of the 24th International Conference on World Wide Web*, 267–277. https://doi.org/10.1145/2736277.2741139

Finlayson, S. G., LePendu, P., & Shah, N. H. (2014). Building the graph of medicine from millions of clinical narratives. *Scientific Data*, *1*(1), 140032. https://doi.org/10.1038/sdata.2014.32

Gensim · PyPI. (2018). Retrieved September 1, 2019, from https://pypi.org/project/gensim/

GitHub - TripleIE/ASKG (Source Code). (n.d.). Retrieved September 1, 2019, from https://github.com/TripleIE/ASKG

GitHub - Word2Vec Implementation. (n.d.). Retrieved November 13, 2019, from https://github.com/fakharealam/Oakland-BrainAneurysm/blob/master/SemnaticAnalysis.py

GitHub - Predicitiveknowledge Implementation. (n.d.) Reterieved Nove 13, 2019, from https://github.com/fakharealam/Oakland-BrainAneurysm/blob/master/Machine%20Learning-%20Predictive%20Knowledge.ipynb

Gyrard, A., Gaur, M., Shekarpour, S., Thirunarayan, K., & Sheth, A. (2018). Personalized Health Knowledge Graph. *ISWC 2018 Contextualized Knowledge Graph Workshop*. Retrieved from http://www.who.int

Hobson, S. P., Dorr, B. J., Monz, C., & Schwartz, R. (2007). Task-based evaluation of text summarization using Relevance Prediction. *Information Processing & Management*, *43*(6), 1482–1499. https://doi.org/10.1016/J.IPM.2007.01.002

Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2013). YAGO3: A spatially and temporally enhanced knowledge base from Wikipedia. *IJCAI International Joint Conference on Artificial Intelligence*, 3161–3165. https://doi.org/10.1016/j.artint.2012.06.001

Home - PubMed - NCBI. (n.d.). Retrieved September 1, 2019, from https://www.ncbi.nlm.nih.gov/pubmed/

Hsu, W., Gonzalez, N. R., Chien, A., Pablo Villablanca, J., Pajukanta, P., Viñuela, F., & Bui, A. A. T. (2015). An integrated, ontology-driven approach to constructing observational databases for research. *Journal of Biomedical Informatics*, *55*, 132–142. https://doi.org/10.1016/J.JBI.2015.03.008

Jurafsky, D., & Martin, J. H. (n.d.). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition Second Edition*. PEARSON. Retrieved from http://www.cs.colorado.edu/~martin/SLP/Updates/1.pdf

Lee, C. H., & Yoon, H.-J. (2017). Medical big data: promise and challenges. *Kidney Research and Clinical Practice*, *36*(1), 3–11. https://doi.org/10.23876/j.krcp.2017.36.1.3

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., … Bizer, C. (2015). DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, *6*(2), 167–195. https://doi.org/10.3233/SW-140134

Martinez-Rodriguez, J. L., Lopez-Arevalo, I., & Rios-Alvarado, A. B. (2018). OpenIE-based approach for Knowledge Graph construction from text. *Expert Systems with Applications*, *113*, 339–355. https://doi.org/10.1016/J.ESWA.2018.07.017

NeuroAssist - BA Foundation. (n.d.). Retrieved August 31, 2019, from https://www.oakland.edu/research/centers/cyber-security

Ng, A. (n.d.). Association Rules and the Apriori Algorithm: A Tutorial. Retrieved from https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html

Ping, P., Watson, K., Han, J., & Bui, A. (2017). Individualized Knowledge Graph: A Viable Informatics Path to Precision Medicine. *Circulation Research*, *120*(7), 1078–1080. https://doi.org/10.1161/CIRCRESAHA.116.310024

Ringer, A. J. (n.d.). *Intracranial aneurysms* (1st ed.). Academic Press - Elsveir.

Risk factors in health and disease - EUPATI. (n.d.). Retrieved September 1, 2019, from https://www.eupati.eu/pharmacoepidemiology/risk-factors-health-disease/

Rokach, L., Schclar, A., & Itach, E. (2014). Ensemble methods for multi-label classification. *Expert Systems with Applications*, *41*(16), 7507–7523. https://doi.org/10.1016/J.ESWA.2014.06.015

Ross, M. K., Wei, W., & Ohno-Machado, L. (2014). “Big Data” and the Electronic Health Record. *Yearbook of Medical Informatics*, *23*(01), 97–104. https://doi.org/10.15265/IY-2014-0003

Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., & Sontag, D. (2017). Learning a Health Knowledge Graph from Electronic Medical Records. *Scientific Reports*, *7*(1), 5994. https://doi.org/10.1038/s41598-017-05778-z

Russert, R. C. (1922). *US1435663A*. US. Retrieved from patentimages.storage.googleapis.com/82/e0/32/7b94720218b2d0/US1435663.pdf

Sabra, S., Mahmood Malik, K., & Alobaidi, M. (2018). Prediction of venous thromboembolism using semantic and sentiment analyses of clinical narratives. *Computers in Biology and Medicine*, *94*, 1–10. https://doi.org/10.1016/J.COMPBIOMED.2017.12.026

Sheth, A., Yip, H. Y., Iyengar, A., & Tepper, P. (2019). Cognitive services and intelligent chatbots: Current perspectives and special issue introduction. *IEEE Internet Computing*, *23*(2), 6–12. https://doi.org/10.1109/MIC.2018.2889231

Shi, L., Li, S., Yang, X., Qi, J., Pan, G., & Zhou, B. (2017). Semantic Health Knowledge Graph: Semantic Integration of Heterogeneous Medical Knowledge and Services. *BioMed Research International*, *2017*, 1–12. https://doi.org/10.1155/2017/2858423

Software | Epic. (n.d.). Retrieved September 1, 2019, from https://www.epic.com/software

Song, M., Kim, W. C., Lee, D., Heo, G. E., & Kang, K. Y. (2015). PKDE4J: Entity and relation extraction for public knowledge discovery. *Journal of Biomedical Informatics*, *57*, 320–332. https://doi.org/10.1016/J.JBI.2015.08.008

Statistics and Facts - Brain Aneurysm Foundation. (n.d.). Retrieved August 31, 2019, from https://bafound.org/about-brain-aneurysms/brain-aneurysm-basics/brain-aneurysm-statistics-and-facts/

Steiner, T., Verborgh, R., Troncy, R., Gabarro, J., & Van De Walle, R. (2012). Adding realtime coverage to the google knowledge graph. In *11th International Semantic Web Conference (ISWC 2012)* (Vol. 914, pp. 65–68).

Stroke Information | cdc.gov. (n.d.). Retrieved August 31, 2019, from https://www.cdc.gov/stroke/index.htm

The Levenshtein Distance Algorithm - DZone Big Data. (n.d.). Retrieved September 1, 2019, from https://dzone.com/articles/the-levenshtein-algorithm-1

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78–85. https://doi.org/10.1145/2629489

Welcome to the NCBO BioPortal | NCBO BioPortal. (n.d.). Retrieved August 31, 2019, from https://bioportal.bioontology.org/

Weng, H., Liu, Z., Yan, S., Fan, M., Ou, A., Chen, D., & Hao, T. (2017). A Framework for Automated Knowledge Graph Construction Towards Traditional Chinese Medicine. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10594 LNCS*, 170–181. https://doi.org/10.1007/978-3-319-69182-4_18

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Yu, T., Li, J., Yu, Q., Tian, Y., Shun, X., Xu, L., … Gao, H. (2017). Knowledge graph for TCM health preservation: Design, construction, and applications. *Artificial Intelligence in Medicine*, *77*, 48–52. https://doi.org/10.1016/J.ARTMED.2017.04.001

Yuan, J., Jin, Z., Guo, H., Jin, H., Zhang, X., Smith, T., & Luo, J. (2019). Constructing biomedical domain-specific knowledge graph with minimum supervision. *Knowledge and Information Systems*, 1–20. https://doi.org/10.1007/s10115-019-01351-4

Zhao, Q., Kang, Y., Li, J., & Wang, D. (2018). Exploiting the semantic graph for the representation and retrieval of medical documents. *Computers in Biology and Medicine*, *101*(May), 39–50. https://doi.org/10.1016/j.compbiomed.2018.08.009