International Journal on Semantic Web and Information Systems
Volume 18 • Issue 1

Learning Disease Causality Knowledge From the Web of Health Data

Hong Qing Yu, University of Derby, UK*
Stephan Reiff-Marganiec, University of Derby, UK

ABSTRACT

Health information becomes importantly valuable for protecting public health in the current coronavirus situation. Knowledge-based information systems can play a crucial role in helping individuals to practice risk assessment and remote diagnosis. The authors introduce a novel approach that will develop causality-focused knowledge learning in a robust and transparent manner. Then, the machine gains the causality and probability knowledge for inference (thinking) and accurate prediction later. In addition, the hidden knowledge can be discovered beyond the existing understanding of the diseases. The whole approach is built on a causal probability description logic framework that combines natural language processing (NLP), causality analysis, and extended knowledge graph (KG) technologies. The experimental work has processed 801 diseases in total (from the UK NHS website linking with DBpedia datasets). As a result, the machine learnt comprehensive health causal knowledge and relations among the diseases, symptoms, and other facts efficiently.

KEYWORDS

Artificial Intelligent, Causality Analysis, Disease Diagnosis, Healthcare, Knowledge Graph, Natural Language Processing, Semantic Web

INTRODUCTION

The development of Artificial Intelligent (AI) technologies makes our daily life much easier than before. For instance, location-based mobile applications help us to find the nearest parking space to your favourite restaurant. In the business domain, BI (Business Intelligent) services assist us to make correct business decisions. In the healthcare domain, AI technologies start to show the strength of detecting diseases in the early stages to minimize the risks of further development. We will see AI technologies becoming one of the most critical future development areas to enhance human healthcare. As a result, symptoms and lifestyle-based disease research and pre-diagnose applications started to show great potential to facilitate self-health care intelligent systems. However, many research problems remain at the current fast AI implementation trends that apply cutting edge technologies such as Deep Learning algorithms and Nature Language Processing (NLP). Some key issues are (but are not limited to): data trust/quality (European Union Agency for Fundamental Right, 2019), security (Pardeep, Masud, Gaba & et al., 2021) transparent prediction (Knight, 2017), and importantly the

DOI: 10.4018/IJSWIS.297145 *Corresponding Author

Causal Analysis (Vorhies, 2019). In contrast to other general domain applications, these open issues are crucial in the healthcare domain. For instance, 'children eating breakfast will avoid teen obesity' (Warner, 2008) and 'eating yoghurts would reduce 19% chances of growing precancerous but only in adenomas for the man (Zheng, Wu, Song, Ogino, Fuchs & Chan, 2019). Both studies only explained associations/correlations discovered from the data observations. However, there is no evidence to tell the possible reasons for 'why'. Recently, causality research in the ML community evidenced that the causal machine learning approach can improve the accuracy of medical diagnosis (Richens, Lee & Johri, 2020). Therefore, knowledge extraction and modelling should be considered as an important step to enhance ML outcomes and expandability, not just focusing on raw data engineering. With the Semantic Web/Knowledge Graph research community growing, knowledge data becomes available and their semantic representations in the semantic cloud. We believe there are enough semantic resources to deal with causality inference and transparent probability calculations collaborating with ML algorithms. For example, we can build Semantic Knowledge Base (SKB) representing relations among symptoms, affecting anatomical structures, most affected groups (age, gender, location), lifestyle effects and drug side effects to a particular group of diseases. Our research work presented in the paper is motivated by such ideas and case studies.

This paper has its distinct contribution to developing a novel semantic modelling framework to generate causality and probability graphs from healthcare information on the Web. Then, the causality knowledge graph data will support more advanced knowledge-based data analysis to address trust, transparency and causality analysis issues. In addition, this paper is a further extension and detailed explanation of the early research outcomes published in (Yu, 2020 & Yu, 2021). The major extension includes merging two separated research methodologies to provide a more inclusive view of the proposed framework and more data evaluations.

The current ML methods and applications for disease recommendation will be reviewed and discuss their critical limitations in section 2. Section 3 will illustrate the proposed framework and its components. Section 4 will demonstrate the benefits of applying the proposed framework in the healthcare AI research domain with our experimental and evaluation results. The conclusion and future work will be drawn at the last section.

RELATED WORK

AI-Supported Diagnoses

AI-supported diagnoses have adapted in many medical domains. Currently, the processes strongly rely on clinical data to support AI learning and prediction algorithms. The research work (Jiang, Jiang, Zhi & et al., 2017) categories seven major data types as raw data sources used in the AI-oriented healthcare data processing such as Imagines, Genetic, Electronic Patient Records (EPR), Physical Examination Notes and Clinical Laboratory Results. Many research outcomes showed that ML combined with raw data would provide huge benefits to healthcare professionals. In a domain-specific or single health condition diagnosis field, most ML algorithms can achieve more than 90% accurate predictions. For example, applying the Multilayer Perceptron algorithm on the Wisconsin diagnostic dataset for breast cancer detection achieves 99.4% accuracy (Agarap, 2018). Meanwhile, The Deep Learning-based automatic detection of chest radiographs algorithm (Kim, Park, & Goo, 2020) can reach 95% accuracy. Deep learning classification algorithm for skin cancer also can provide a highly accurate detection rate developed by the Stanford research team (Esteva, Kuprel, Novoa & et al., 2017). Different types of diseases were investigated to apply ML algorithms for predictions (Lampropoulos, & Tsihrintzis, 2015; Verma & Ghosh, 2018; Yekkala & Dixit 2018; Papachristou & et-al, 2016; Kaur, Sharma & Sohal 2019; Harini & Natesh, 2018; Laskar & et-al, 2016). In the meantime, Jiang & et-al. (2017) suggests that ML and NLP are the two key devices for future AI-based disease prediction technology development.

However, there are two major limitations of existing work:

- The data model focuses on a specific dataset that normally only presents a single category of the patients with fixed features. As result, the learning process purely works on fitting a model that can produce the desired results for the dataset. This is the reason that a well-trained model based on one dataset, may produce worse results for the other dataset. The fundamental problem is the relations between data and datasets are not modelled. Having the insight of healthcare data, many crucial relations need to be modeled such as associated symptoms, history of family health, living style, and the causalities among them. However, these relations cannot be expressed by an isolated single dataset without a comprehensive knowledge modelling framework.
- Hard to explain and have tracing evidence. In some research areas, the explanation ability can be less important, but not for healthcare. The explanations need to support assurance, assist define treatments and provide efficient information to the patients for self-education and self-awareness.

To address the above issues, Richens, Lee & Johri, 2020 suggested causal reasoning based diagnostic principles that include posterior likelihood causal analysis, counter-factual inference and simplicity explanation. However, these principles focused on assisting doctors in using their pre-existing knowledge for disease diagnosis. The knowledge extracting and understanding processes are still missing and challenging.

Knowledge Modelling

Health knowledge modelling and linking is another approach to try to build a domain knowledge infrastructure to support the healthcare system. There are four major reasons that demand knowledge modelling discussed by Mate & et al. 2015 and the approaches can be metadata-driven or ontology-driven processes. For example, Patient Clinical Data (PCD) ontology is developed for EHR clinical data representation for healthcare researchers Boshnak, Abdelgaber, Yehia & Abdo 2019 is a metadata-driven process. Other high-level semantic clinical data models such as OGMS-based MCI introduced by Oberkampf & et al. 2013, SEHR developed by Sheth & et-al. 2006 and SNOMED CT presented by El-Sappagh & et al. 2018 are ontology-driven processes (there are many more related models). However, all these models are used to provide metadata-level interpretations of the clinical data but not at the level of knowledge understanding.

Early research on applying process-knowledge on health literacy (human-based knowledge extracting) to help self-care is presented in the work of Chin, & et al. 2011. With NLP and Machine Learning development, machine-based knowledge-level modelling has started to explore other domains for ML applications. For example, an ontology-based knowledge modelling framework is proposed in the sustainability assessment domain (Konys, 2018), which has five steps of literature identification, concept extracting, taxonomy construction and ontology construction and finally the ontology validation. The whole process tries to understand how to model the knowledge with a reflected ontology that is not the metadata of the literature document. A similar approach has been identified as one of the important challenges of AI used in healthcare research by Wang & Preininger, 2019. The challenge is how we can extract knowledge from biomedical literature data to power AI algorithms-based health diagnosis. The opportunity from the challenge is to combine advanced NLP and machine learning models to allow machines to understand the learnt knowledge and represent it as a Knowledge Base (KB) for machine learning.

In this paper, we are going to provide a solution to address this challenge by defining a causality knowledge extract framework from a trust health website. The lifted causality knowledge can support further hidden knowledge reasoning and health prediction combining with probability programming. The important novelty comparing to existing approaches is the combinations of NLP, Semantic Web, and robust Causality Analysis.

PROPOSED LOGIC FRAMEWORK AND DEFINITIONS

The framework development methodology has four stones of problem understanding, problem modelling and solution design, experimental analysis, and evaluation. The evaluation applies data analytics methods and simulation testing. To enable achieving our research aim, many programming tools have been used in the project development such as Protégé for ontology modelling, DBPedia APIs for semantic linking and extraction, NLTK Python library for NLP, and RDFLib for implementing the Causality Knowledge Graph and storing sematic data. The research problem (first stone) has been discussed in previous section, this section will start focusing on the problem modelling and framework design.

Causal knowledge extraction and modelling has been applied widely in the areas such as social behavior and healthcare research since Neyman-Rubin published Causal Inference Theory was in 1986. The formal graph-based mathematics models are represented by Pearl, 2010, which clearly provide a method to separate correlation and causality analysis. The model applies probability joint distribution computation. In a directional graph, the causality graph has properties of the back-door criterion and a testing function of do(X=x) instead of doing a random selection of x that has the probability estimation on Y. Thus, the causality relation can be observed/measured when a property is updated, then the remaining properties' probability distributions will be changed accordingly. Now, we can clearly distinct correlations/associations to causality. DeepMind team (Dasgupta, Wang & et al, 2019) applied the Pearl model to the Meta-Reinforcement learning process.

Gutierrez-Basulto, Jung & Lutz, 2017 introduced probability analysis into Semantic Web logic framework with belief rating thresholds called Probabilistic Description Logic (PDL) to deal with subjective uncertainty. PDL introduced an extra probabilistic threshold syntax to the classic Description Logic (DL) notated as $P_{\sim n}$ over Tbox and Abox. The notation $\sim$ can be any one of these operators $\leq, <, =, >, \geq$ and n is the value of the threshold. The problem of PDL is that it needs pre-defined default probability thresholds, which would not support to a highly dynamic system:

$$T ::= C, D \mid C \sqcap D \mid \in r.(C) \mid P_{\sim n} C \tag{1}$$

$$A ::= A \mid A \wedge A' \mid r(a, b) \mid P_{\sim n} A \mid \tag{2}$$

In our proposed logic framework, we introduce the causality probability concept into KB level and we name it as Causal Probability Description Logic (CPDL) framework. We are currently working on DL-$\varepsilon\iota$ only. The definition of Causal Probability Knowledge Base (CPKB) is a triple:

$$CPKB = \{T, A, \phi(T)\} \tag{3}$$

In the definition, T is the terminological structure Tbox (knowledge schema or called class level of ontology), and A is the atomic instances – Abox. Both T and A are the same meaning from the classic DL KB terms. The $\phi(T)$ is our novel element added to the KB that defines the causality relations between any two concepts in Tbox. The $\phi(T)$ is an extra roof (universal) concept comparing to the classic DL. For example, a prediction (relation) definition is unique in DL KB to assign a specific subject (domain) and object (range). In our extension, the $\phi(T)$ can assign to any subject and object that have a causal relationship. Based on the newly defined CPKB definition, The Tbox is:

$$T ::= C, D \mid C \sqcap D \mid \in r.(C) \mid \in r.(C) \Longrightarrow \in \phi.(C) \mid P\mu\ (\phi\ (C)) \mid \tag{4}$$

Here C, D and r are the same terms used in classic DL as terminology elements defined in Tbox. However, CPDL claims that if there is a relation r, then there may have a causal relation ϕ exist mapping to the r according to knowledge certainty. The ϕ may have (depending on if the causal relations have existed in the descriptions of data source, e.g. the text document) explicit direction (a causes b) or bi-direction (a causes b and b also causes a) that need to be defined in Tbox. The bi-directional causal relation is a new term that we define in this paper comparing to classic causality assumption. In a general context, bi-direction causalities will produce confusion but there are truly occurred in the health domain. If the framework is applied to wider research domains, then this bi-directional causality relation can be removed according to the context. The Pμ (ϕ (C)) represents a dynamic probability distribution over all possible ϕ(T) by given observation inputs. Based on Tbox definition, we can define Abox as:

$$A ::= A \mid A \wedge A' \mid r(a, b) \mid \phi(a \mid b) \mid \phi(b \mid a) \mid \phi(b \mid a) \vee \phi(a \mid b) \mid P\mu(\phi) \mid \quad (5)$$

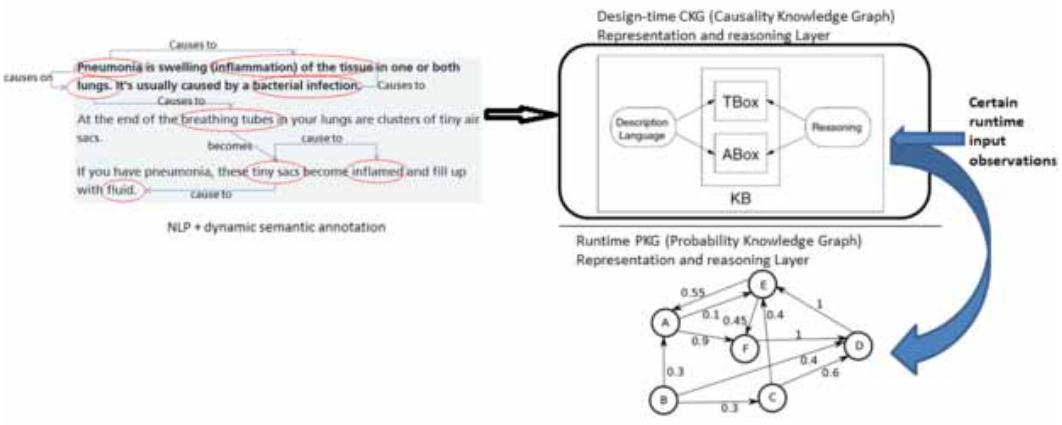
where A, A' and r are the same terms used in classic DL Abox. ϕ (b | a) implies b is one of the reasons to cause an observation condition 'a' according to the Tbox specification and vice versa is for ϕ (a | b). In contrast, ϕ (b | a) ∨ ϕ (a | b) indicates that there are bi-directional causal relations between 'a' and 'b'.

The Pμ(ϕ) distribution can be computed using probability learning algorithms dynamically in the application. We will discuss the Pμ(ϕ) computation in detail later.

The lower bound of the complexity of the reasoning on the CPDL knowledge graph is as same as theorem 4 listed in paper [16] that proved adding probability features in DL-ει is an ExpTime-hard problem. However, our experimental exercises show that the dynamic probability generation can significantly reduce the graph size and the NP-hard problem can be ignored in later sections. In addition, new parallel computing or computing distribution technologies can significantly reduce the computation time e.g. MapReduce algorithm. However, computation optimization and complexity measurements are out of the scope of this paper.

The Causal Probability Description Logic framework can be treated as a two-layer model that is represented in figure 1. The top layer is the knowledge layer which organises and represents the causality knowledge graph to be reasoning. The lower layer is an instance probability knowledge graph generated based on the runtime input of the observations. For example, the runtime inputs can be symptoms or history of health conditions in the healthcare domain.

Figure 1. Two layers architecture of causality and probability knowledge graph generation framework



EXPERIMENTAL: GENERATING CPDL KNOWLEDGE GRAPH FROM UK NHS WEBSITE FOR COMMON DISEASES

For the healthcare application domain, the dedicated top layer CKG ontology is defined in Figure 2.a. The meanings of the key terminologies are:

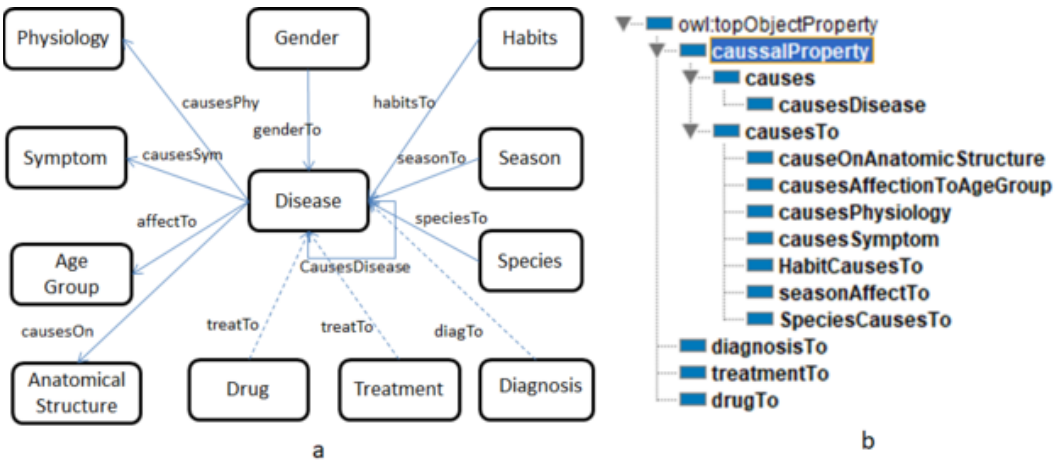
1. **Disease:** It is a central term in the ontology to organize the causality relations among other terms. The atomic disease triples are crawled on UK NHS website of common disease A-Z index firstly and then build link to DBpedia disease terms automatically while crawling.
2. **Symptom and Physiology:** These two terms define symptoms or physiology instances that should be discovered from the dynamic semantic mining approach. The causal relations to the related diseases will be generated at the same time. The direction of the causal relation is that disease causes symptoms and physiology.
3. **AnatomicalStructure:** It presents a class type of instance term that is an organ name of human. The instance of the AnatomicalStructure will be generated from the same semantic mining approach. The causal relation is that some diseases can affect the human anatomical structure.
4. **AgeGroup:** It separates humans into groups by age. Through the semantic mining process, the AgeGroup instances are identified as ageing people/elder, adult, youth, boy, girl, man, woman and baby. The causal relation is bi-direction that means that age can cause an increased risk of getting a disease and the disease can affect a certain age group of people.
5. **Habits, Season, Species:** The other three classes defined for extracting the habit, season and specie related terms and contributing relations to cause the diseases.
6. **Drug, Treatment and Diagnosis (Diagnosis required tests or equipment):** The commonsense classes that define to get more knowledge about diseases but only have the natural relation to disease no causality.

All the predictions that have causal effect relation are the instance relations of $\phi(T)$ in our logic model. The hierarchical structure of the causal relations is represented in Figure 2.b.

Web Information Processing

The core information source for our research is the Web content from UK NHS website. Figure 1 (left side) is an example of a pneumonia webpage that describes all the information related to the

Figure 2. Health domain CKG ontology and causality relation structure



disease, which includes an overview, symptoms, causes, diagnosing, treating, complications and preventing sessions. The information processing process contains 4 steps of information crawling on HTML features of the NHS website, Natural Language Processing pipeline to get keywords and their processing tokens for each sentence, semantic mining to do word tagging and merging and finally lift the semantic relations as triples to the CPDL-based knowledge graph.

Figure 3 shows an example of the NLP result for one paragraph text describing pneumonia disease.

Figure 4 represents the semantic lifting process and major predictions, ontology terms and keywords. The middle part is the predictions that can link to the semantics of the noun tokens. In addition, there are five classes are checked to do the matching and some keywords (bold words) are detected for matching specific terms due to incomplete or missing semantics in the DBpedia RDF graph. After this step, a semantic dictionary is created for containing the matched noun tokens and their matched semantics in the defined T-box (see Figure 1). For example, 'Chest_pain' is one of the noun tokens, the lifting process matched 'Symptom' term through dc:subject prediction in the knowledge graph file of 'Chest_pain' with one of the objects of dbp:Categorty:Symptoms_and_signs.

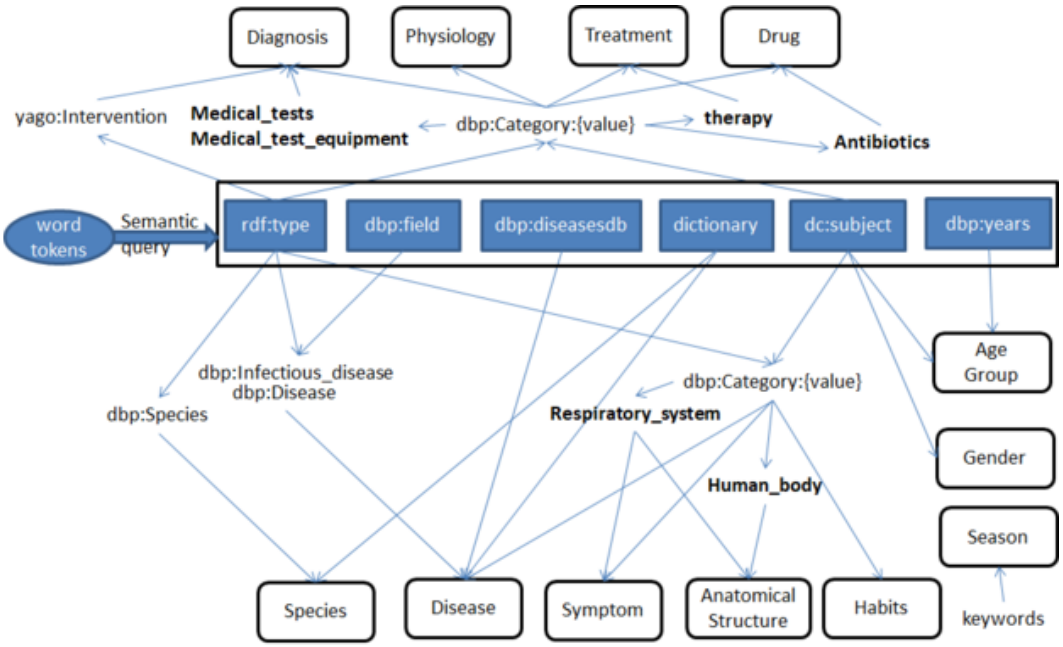
CPKL-Based CKG Generation

The Semantic lifting process provides disease dictionaries that have Noun tokens as keys and semantic annotations as values. A CPKL-based knowledge graph is generated incrementally throughout the dictionaries of all crawled diseases step by step. Currently, we created a graph that contains 383 NHS

Figure 3. NLP-base tokenization result for pneumonia disease

```
***** symptoms *****
 ['Symptom', 'Hour', 'Day', 'Common', 'Cough', 'Yellow', 'Blood', 'Phlegm', 'Difficulty', 'Breathing', 'Shallow', 'He
artbeat', 'Temperature', 'Loss', 'Appetite', 'Chest', 'Pain', 'Less', 'Haemoptysi', 'Headach', 'Muscle', 'Feeling', '
People', 'Xeroderma', 'Chest_pain', 'Muscle_weakness', 'Unwell', 'Common_name', 'Pneumonia', 'Tachycardia', 'Perspira
tion', 'Joint', 'Confusion', 'Emotion', 'Thermoregulation', '48_Hrs.', 'Green', 'The_Guardian', 'Myalgia', 'Brown', '
Aestivation', 'Headache', 'Hematuria']
```

Figure 4. Semantic lifting and merging



diseases that have causal semantic links to extra 418 diseases from the DBpedia knowledge graph. As result, the generated CPKL-based knowledge graph contains 803 diseases. Figure 5 shows a CPKL-based graph representation of pneumonia disease (part of the original graph, all the original data can be seen in our Github repository). The graph identifies the current practice knowledge about pneumonia. For example, blood_test and CT_scan (DBpedia terms) are the possible diagnosis methods. The causality relations include specieCaussesTo, causesDisease (pneumonia has the possibility to cause other diseases), causesSymptom, causeOnAnatomicStrucure and many more. Reflect on the CPDL framework, the lifted causality relations are the subtype of $\phi(u)$ defined in the CPKB definition. The causesDiseas is a bi-directional prediction. The pneumonia causal graph is only one of the integrated 383 diseases.

Figure 5. Pneumonia disease top layer causality knowledge graph (partially)

```
<http://dbpedia.org/page/Blood_test> a <http://dbpedia.org/page/Medical_diagnosis> ;
    ns1:diagnosisTo ns1:Pneumonia .

<http://dbpedia.org/page/CT_scan> a <http://dbpedia.org/page/Medical_diagnosis> ;
    ns1:diagnosisTo ns1:Pneumonia .

<http://dbpedia.org/page/Chemotherapy> a <http://dbpedia.org/page/Therapy> ;
    ns1:treatmentTo ns1:Pneumonia .

<http://dbpedia.org/page/HIV> a <http://dbpedia.org/ontology/Species> ;
    ns1:SpeciesCausesTo ns1:Pneumonia .

<http://dbpedia.org/page/Haemophilus> a <http://dbpedia.org/ontology/Species> ;
    ns1:SpeciesCausesTo ns1:Pneumonia .

<http://dbpedia.org/page/Haemophilus_influenzae> a <http://dbpedia.org/ontology/Species> ;
    ns1:SpeciesCausesTo ns1:Pneumonia .

<http://dbpedia.org/page/Human_respiratory_syncytial_virus> a <http://dbpedia.org/ontology/Species>
    ns1:SpeciesCausesTo ns1:Pneumonia .

<http://dbpedia.org/page/Ventilator-associated_pneumonia> a <http://dbpedia.org/ontology/Disease> ;
    ns1:causesDisease ns1:Pneumonia .

<http://dbpedia.org/page/Viral_disease> a <http://dbpedia.org/ontology/Disease> ;
    ns1:causesDisease ns1:Pneumonia .

<http://dbpedia.org/page/Viral_pneumonia> a <http://dbpedia.org/ontology/Disease> ;
    ns1:causesDisease ns1:Pneumonia .
ns1:Pneumonia a <http://dbpedia.org/ontology/Disease> ;
    ns1:causeOnAnatomicStructure <http://dbpedia.org/page/Blood>,
        <http://dbpedia.org/page/Breathing>,
        <http://dbpedia.org/page/Epithelium>,
        <http://dbpedia.org/page/Heart>,
        <http://dbpedia.org/page/Joint>,
        <http://dbpedia.org/page/Kidney>,
        <http://dbpedia.org/page/Leaf>,
        <http://dbpedia.org/page/Lining_(sewing)>,
        <http://dbpedia.org/page/Lung>,
        <http://dbpedia.org/page/Mouth>,
        <http://dbpedia.org/page/Muscle>,
        <http://dbpedia.org/page/Nose>,
        <http://dbpedia.org/page/Pleural_cavity>,
        <http://dbpedia.org/page/Respiratory_system>,
    ns1:causesAffectionToAgeGroup
        <http://dbpedia.org/page/Ageing>,
        <http://dbpedia.org/page/Infant>,
        <http://dbpedia.org/page/Old_age>,
        <http://dbpedia.org/page/Unwell> ;
    ns1:causesPhysiology <http://dbpedia.org/page/Inflammation>,
        <http://dbpedia.org/page/Perspiration>,
        <http://dbpedia.org/page/Phlegm>,
        <http://dbpedia.org/page/Vomiting> ;
    ns1:causesSymptom <http://dbpedia.org/page/Abscess>,
        <http://dbpedia.org/page/Acute_respiratory_distress_syndrome>,
```

CKG Reasoning to Generate Runtime PKG

In this subsection, a Discrete Uniform Distribution (DUD) function is applied to define the causal probability distributions based on the knowledge organized as the CKG. With the probability distribution, the runtime PKG is to be generated for prediction purposes. In classic causal (C) effect (E) calculus definition, $P\left(E \mid C\right)$ means given the C, there is a probability of getting E and $P\left(E \mid C\right) > P\left(E \mid \neg C\right)$. For example, Pneumonia (C) can cause a probability of showing a 'Cough' symptom (E), the probability is very high normally. The difficulty is to provide a probability value to the P in practice. This is the important reason we argue that the current Probability Description Logic is not applicable to many practical tasks. Thus, we bring a CPDL framework into our research. The fundamental difference is that we define probability in a runtime calculation environment using the CKG knowledge in hand.

Definition 1: The probability distribution P' that represented as $P'\left(C \mid E\right)$, where E is the effect and C is the causal fact. The given observation is the effect and P' indicate the probability distribution of individual causal fact.

For instance, if the observed effect is 'Cough', then the probability distribution can be simply calculated at runtime to have equal probabilities to all conditions that can cause the E (Cough). Therefore, the probability of one possible causal fact c_i according to CKG considering one of the observations e_i inside E can be defined as equation 6:

$$P'\left(c_i \ni \mathrm{C} \mid e_i \ni E\right) = \frac{1}{\mathrm{n}} \tag{6}$$

where n is the number of elements in the C.

The example generates a runtime PKG (partially screenshot) based on equation 6 with input conditions of:

Symptoms=['Cough','Breathing','Fever','Heartbeat','Chest_pain','Fatigue','Shivering','Infection','Unwell']
Unwell Body Position=['Lung']
Group = ['Child']
Gender = ['Male']

The runtime PKG (Probability Knowledge Graph) is generated based on the above inputs. The PKG adds extra key probability information to the knowledge graph. For instance, pneumonia is one of the possible causal diseases to the conditioned context (see figure 6) and around 0.0048 and 0.018 causal probabilities for problems of Heartbeat and Cough respect.

Based on the runtime generated PKG, machine learning algorithms can be applied to do disease classification or prediction. According to the PKG features, we investigated on the Naïve Bayesian classification algorithm that has been approved to be efficient for many other diseases' prediction studies (Langarizadeh & Moghbeli, 2016; Wei, Visweswaran & Cooper, 2011; Ehsani-Moghaddam, Queenan, MacKenzie & Birtwhistle, 2018).

Formally we define the probability aggregating function to predication for an individual c_i (equation 7):

$$\mathrm{P'}\left(\mathrm{c_i} \ni \mathrm{C} \mid \mathrm{e_1}, \ldots, \mathrm{e_n} \ni \mathrm{E}\right) = \log\left(\prod_1^n p_i(\mathrm{c_i} \mid \mathrm{e_i})\right) = \sum\nolimits_1^n log(p_i(\mathrm{c_i} \mid \mathrm{e_i})) \tag{7}$$

Figure 6. Runtime generated PKG (partially) based on the certain inputs

```
<https://www.nhs.uk/conditions/Pleurisy> a <http://dbpedia.org/ontology/Disease> ;
    ns1:hasCausalProbablity [ ns1:causalityTo <http://dbpedia.org/page/Chest_pain> ;
            ns1:pvalue "0.022222222222222223" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Lung> ;
            ns1:pvalue "0.022222222222222223" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Cough> ;
            ns1:pvalue "0.018181818181818818" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Breathing> ;
            ns1:pvalue "0.017857142857142856" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Infection> ;
            ns1:pvalue "0.005376344086021506" ] .

<https://www.nhs.uk/conditions/Pneumonia> a <http://dbpedia.org/ontology/Disease> ;
    ns1:hasCausalProbablity [ ns1:causalityTo <http://dbpedia.org/page/Infection> ;
            ns1:pvalue "0.005376344086021506" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Cough> ;
            ns1:pvalue "0.018181818181818818" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Heartbeat> ;
            ns1:pvalue "0.047619047619047616" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Breathing> ;
            ns1:pvalue "0.017857142857142856" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Chest_pain> ;
            ns1:pvalue "0.022222222222222223" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Fever> ;
            ns1:pvalue "0.008333333333333333" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Lung> ;
            ns1:pvalue "0.022222222222222223" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Unwell> ;
            ns1:pvalue "0.037037037037037035" ] .

<https://www.nhs.uk/conditions/Poisoning> a <http://dbpedia.org/ontology/Disease> ;
    ns1:hasCausalProbablity [ ns1:causalityTo <http://dbpedia.org/page/Child> ;
            ns1:pvalue "0.007518796992481203" ] .

<https://www.nhs.uk/conditions/Polio> a <http://dbpedia.org/ontology/Disease> ;
    ns1:hasCausalProbablity [ ns1:causalityTo <http://dbpedia.org/page/Infection> ;
            ns1:pvalue "0.005376344086021506" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Fever> ;
            ns1:pvalue "0.008333333333333333" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Breathing> ;
            ns1:pvalue "0.017857142857142856" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Childhood> ;
            ns1:pvalue "0.007518796992481203" ] .

<https://www.nhs.uk/conditions/Polymorphic-light-eruption> a <http://dbpedia.org/ont
    ns1:hasCausalProbablity [ ns1:causalityTo <http://dbpedia.org/page/Fever> ;
            ns1:pvalue "0.008333333333333333" ],
        [ ns1:causalityTo <http://dbpedia.org/page/Male> ;
```

Applying to the example inputs presented earlier, the top 10 ranking results of the prediction in percentage is displayed in the figure 7. The result indicates the most possible health condition according to the inputs is 'Pneumonia' (80% probability). However, the other 9 conditions are also possible with more than 60% probabilities. In fact, the input example is created based on a pneumonia case.

EVALUATION OF KNOWLEDGE DATA AND A PROTOTYPE APPLICATION

Causality Knowledge Graph Evaluation

The current version of CPKG presents 801 diseases that are semantically lifted from the NHS and the DBpedia mining process. These 801 diseases link to 1078 377 treatments and drugs, physiologies/ symptoms, 8 categorized habits. In addition, 113 species and 66 different human groups are connected.

Figure 7. The prediction result

	d	pred_v
241	Pneumonia	80.000000
294	Cystic-fibrosis	78.838983
119	Chest-infection	75.965158
4	Bronchiolitis	75.603460
111	Epiglottitis	75.603460
146	Asthma	74.211864
228	Granulomatosis-with-polyangiitis	71.004332
227	Schistosomiasis	70.279575
67	Tuberculosis-tb	70.254011
239	Sickle-cell-disease	69.917877

Details of the implementation, CPKG datasets and evaluation process are available on the Github repository[1] for you to review.

The first interest evaluation we did is to examine which diseases or health conditions have the most causal relations to other diseases. From figure 8, we can see the top 25 conditions that have the most causal relations with other diseases. For example, 'Infection' and 'Fever' are the most common conditions that can cause or caused by other diseases.

The second evaluation we did is to explore symptoms. We present the top 25 symptoms or physiological in figure 9 according to the numbers of connected diseases and causal relations. From figure 9, we can see that Schizophrenia is the top symptom condition (a kind of mental health condition) that has the most connections to the diseases (it can be developed from 264 diseases).

Figure 10 shows that surgery and antibiotics are the most effective treatments to health conditions. The knowledge also indicates that traditional Chinese medicines can have positive impacts on 26 health conditions. Interestingly, bread as a kind of food is recommended in 11 health conditions, e.g., kidney disease. Except to check back the original UK NHS articles, we did some extra research regarding to this discovery. For example, Sheridan (2012) suggests suitable bread can provide important sodium and phosphorus to help easing chronic kidney disease.

Figure 8. Top 25 diseases that are most connective to other diseases

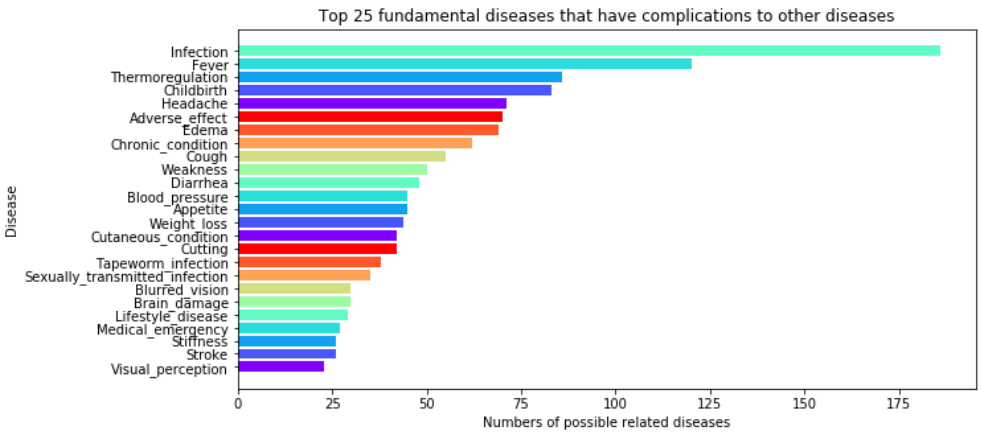


Figure 9. Top 25 symptoms that are most connective to diseases

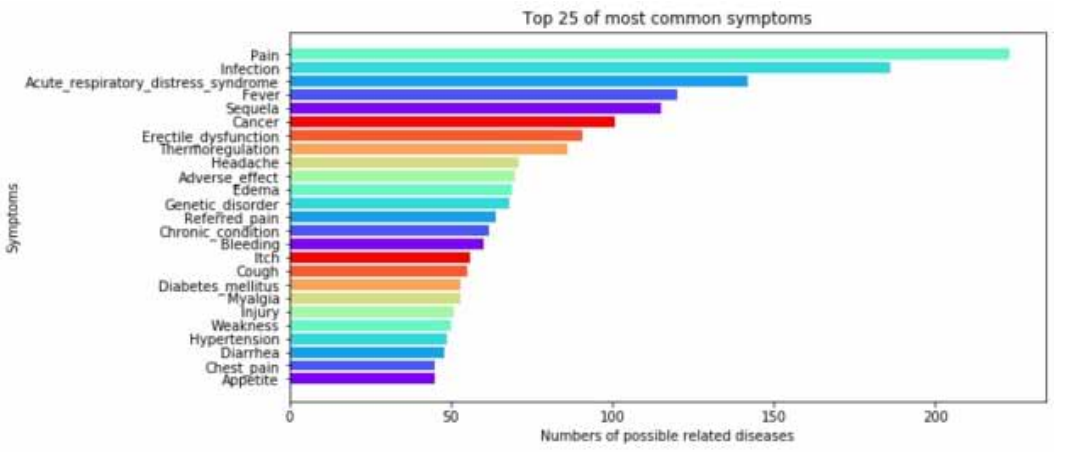
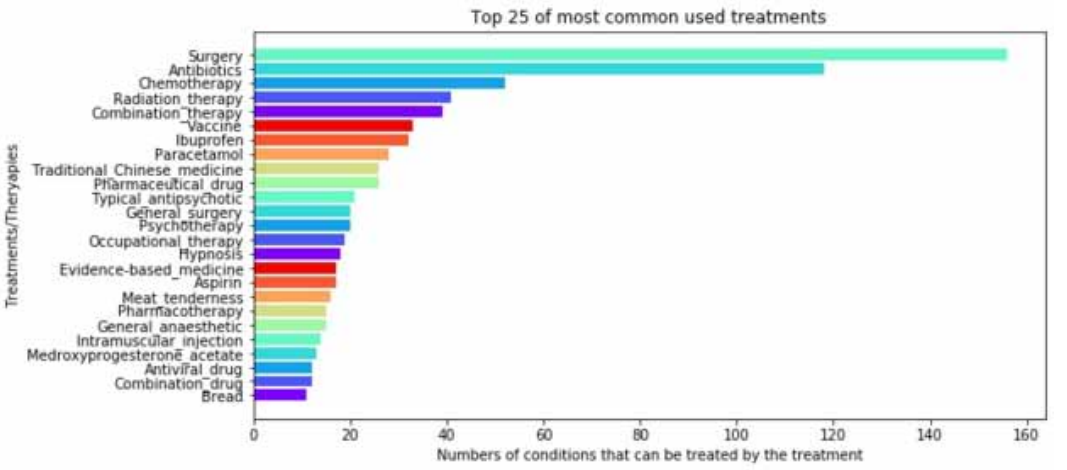


Figure 10. Top 25 of most common used treatments



The reasoning queries find that 8 bad lifestyles will contribute to diseases (see Figure 11). unsurprisingly, the smoking related habits are most dangerous and connect to more than 100 diseases. The wider concept of bad habit and civility are connecting with many different types of diseases.

The data analysis also shows that some diseases are related to seasons. Autumn and winter are associated to more diseases than other two seasons.

Figure 12 shows the Top 10 vulnerable groups of people. We can see that Child and Ageing groups are the most vulnerable. We can also see that Men are two times more vulnerable than Women. The other interesting find is that the 'higher education' keyword appeared, which suggests the higher education group may have relevant connections to more than 20 diseases.

A causal knowledge graph provides a huge benefit for providing traceable evidence to explain the outcomes of machine learning. The traceable evidence is the causal knowledge chain. In the current state of the art of the project, the longest traceable chain of a certain disease or symptom is 5. Here are concrete examples:

Figure 11. 8 bad habits

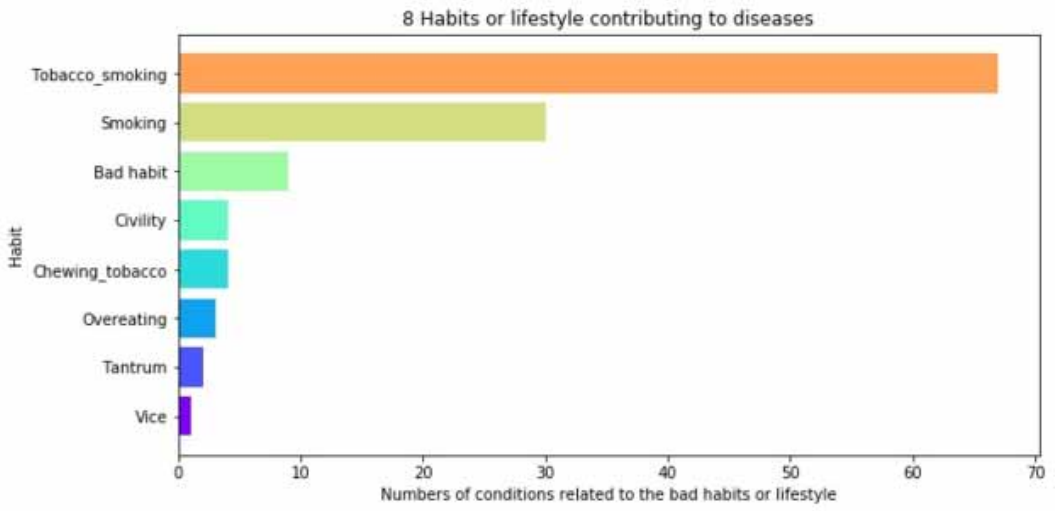
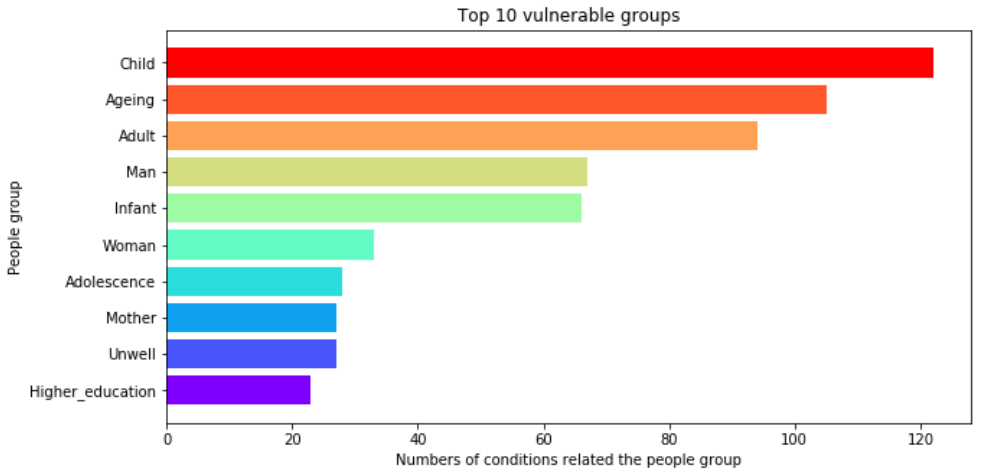


Figure 12. Top 10 vulnerable group terms

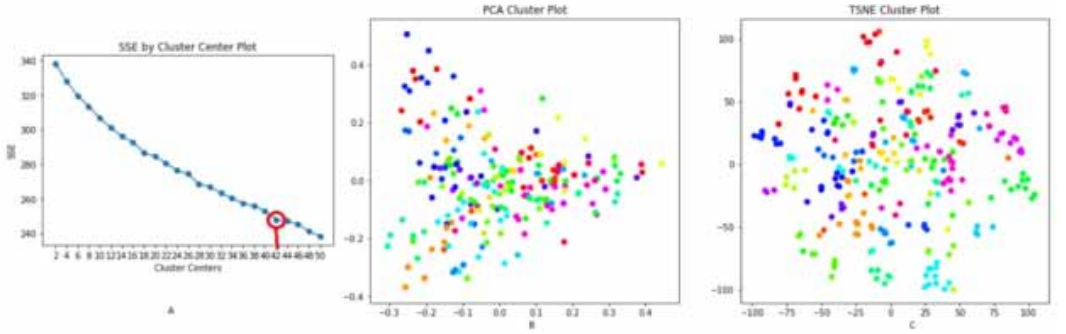


Rheumatoid_arthritis->Psoriasis->Psoriatic_arthritis->Myositis->Inclusion_body_myositis
Rheumatoid_arthritis->Psoriasis->Pagets_disease_nipple->Breast_cancer->Weight_loss

In total, we detected 111186 3-length chains, 3847 4-length-chains and 3683 5-length-chains in our CKB space.

The other valuable benefit of applying a causal knowledge graph is to have a new type of disease cluster based on multidimensional causal relationships. The unsupervised K-mean clustering algorithm is applied to build disease clusters based on related diseases, symptoms, human groups and affected anatomical structures. To gain a better clustering result, the cluster number optimization is performed first that shows in Figure 13(A) and the elbow point of the optimization is on 42. Figure 13(B, C) shows two different graphic visualizations of the 42 K-mean clusters using Principal Component Analysis and t-Distributed Stochastic Neighbor Embedding plots. The visualisations clearly show that

Figure 13. Disease clustering based on the knowledge data



the clusters are created reasonably well. The trained clustering K-mean model can start to predict a given list of observations such as symptoms and history of the disease to a cluster that contains the most potential health conditions. For example, a list of observations ['headache, influenza, fever, throat, children'] is mostly related to the health condition in Cluster 0 that contains 12 diseases of ['Bornholm-disease', 'Common-cold', 'Diphtheria', 'Chickenpox', 'Flu', 'Hand-foot-mouth-disease', 'Polio', 'Q-fever', 'Roseola', 'Rubella', 'Slapped-cheek-syndrome', 'Tonsillitis'].

Application Example and Accuracy Evaluation

Based on the same logic framework and knowledge extracting process, we develop a health Chatbot application for answering health condition questions and providing advice on symptom-based predictions. The implementation detail can be accessed in Github repository[2]. Figure 14 shows that if the question has not been asked before, then the processed knowledge is provided at runtime. Otherwise, the organised knowledge from the knowledge base will be provided based on an answering template.

To evaluate the accuracy of the Chatbot prediction based on the symptoms, we created two evaluation data pools. The first pool contains 10 testing datasets and each of the datasets contains more than 100 disease cases. Individual-disease-prediction test case is first generated by a selection of random amount (3 to 10) of symptoms and related position together with the randomly picked age and gender from a certain disease described by the NHS website. The second pool contains 10 testing datasets with noisy information. The test case generating method is the same as the first pool but applies one extra step of random injection of two noisy symptoms from the top 15 most common symptoms discovered by our CKB. The testing environment setting is based on the defined evaluation settings. The evaluation results show the framework can achieve an average of 93.63% and 89.60% accuracy rates for the two testing data pools (see Figure 15.a). The accuracy results are surely promising even with noisy datasets.

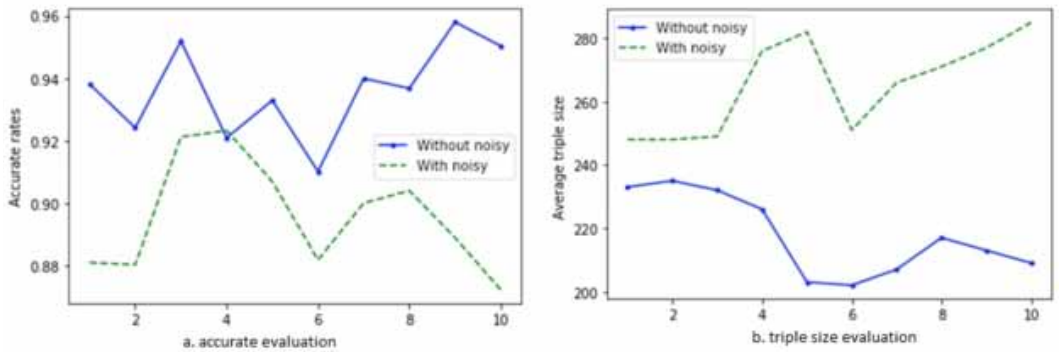
Figure 15.b presents the average triple size of runtime second layer PKG in the 20 testing datasets of the two different pools. The result shows that the largest graph contains about 280 triples, which means the probability-based graph ExpTime-hard complexity issue is not essential at PKG layer.

CONCLUSION AND FUTURE WORK

In summary, knowledge-based health information extracting and learning framework is introduced, explained, implemented and evaluated in the paper. The framework has three key differences comparing to the state of art ML approaches:

Figure 14. Chatbot prototype interface.

Figure 15. Evaluation results 1



- The proposed framework built a humankind machine learning approach by learning knowledge, creating causality network and runtime probability calculations according to different input scenarios.
- The proposed framework emphasized causality relations as the core knowledge to support probability inference.
- The proposed framework has strong reusability to enable dealing with general AI tasks in contrast to current fixed problem solution attached with a fixed data. In other words, the framework will continue learning new knowledge to solve runtime problems.

The current framework has three extending directions in future work:

1. Extracting negating causal relations from the text and adding the negations to the logic framework to express 'not causes' semantics. There are two major challenges of mining and understanding negating causal relation through NLP including the probability calculation and negations with positive facts.
2. At this stage, probability runtime computation uses the Naïve Bayes algorithm. To apply Naïve Bayes, it requires a strong assumption that believes all observations are independent events. However, the independence assumption does not always hold in healthcare or other application domains. For example, symptoms mostly have association relations with each other, e.g., a long period of high temperature and infection always appears together. In our early initial exploration, we applied the Apriori association rule-mining algorithm (Agrawal & Srikant, 1994) to investigate the association relations among symptoms. We managed to identify 26 association connections. Therefore, we need to find a way to add such an association into probability distributions, which is one of the important future research topics for us. We will investigate other probability functions, such as Multinomial Naive Bayes proposed by Kibriya, Frank, Pfahringer & Holme 2004 and Partially Observed Markov Decision Processes (Krishnamurthy 2016).
3. Clearly understand the weight distributions among different semantic dimensions (which type of semantics has more influence on the accuracy of diagnosis).

FUNDING INFORMATION

The publisher has waived the Open Access Publication fee for this article.

REFERENCES

Agarap, A. F. M. (2018) On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. In *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing (ICMLSC '18)*. Association for Computing Machinery. doi:10.1145/3184066.3184080

Agrawal, R., & Srikant, R. (1994) Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*. Morgan Kaufmann Publishers Inc.

<bok> Pardeep, K., Masud, M., Gaba, G. S., Alqahtani, S., Muhammad, G., Gupta, B. B., & Ghoneim, A. (2021). A Lightweight and Robust Secure Key Establishment Protocol for Internet of Medical Things in COVID-19 Patients Care. IEEE Internet of Things Journal. doi:10.1109/JIOT.2020.3047662</bok>

Boshnak, H., Abdelgaber, S., Yehia, E., & Abdo, A. (2019, April). Ontology-Based Knowledge Modelling for Clinical Data Representation in Electronic Health Records. *International Journal of Computer Science and Information Security*, *16*(10), 68–86.

Chin, J., Morrow, D.G., Stine-Morrow, E.A., Conner-Garcia, T., Galich, J.F., & Murray, M.D. (2011). The process-knowledge model of health literacy: evidence from a componential analysis of two commonly used measures. *J Health Communed, 16*(3), 222-41. doi: 10.1080/10810730.2011.604702

Dasgupta, I., & Wang, J. (2019). *Causal Reasoning from Meta-reinforcement Learning*. arXiv preprint arXiv:1901.08162.

Ehsani-Moghaddam, B., Queenan, J. A., MacKenzie, J., & Birtwhistle, R. V. (2018). Mucopolysaccharidosis type II detection by Naïve Bayes Classifier: An example of patient classification for a rare disease using electronic medical records from the Canadian Primary Care Sentinel Surveillance Network. *PLoS One, 13*(12). doi:10.1371/journal.pone.0209018

El-Sappagh, S., Franda, F., & Ali, F. (2018). SNOMED CT standard ontology based on the ontology for general medical science. *BMC Medical Informatics and Decision Making*, *18*, 76. https://doi.org/10.1186/s12911-018-0651-5

Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. doi:10.1038/nature21056 PMID:28117445

European Union Agency for Fundamental Right. (2019). *Data quality and artificial intelligence-mitigating bias and error to protect fundamental right*. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-data-quality-and-ai_en.pdf

Gutierrez-Basulto, V., Jung, J. C., & Lutz, C. (2017). Probabilistic Description Logics for Subjective Uncertainty. *Journal of Artificial Intelligence Research*, *58*, 1–66.

Harini, D. K., & Natesh, M. (2018). Prediction of Probability of Disease based on Symptoms Using Machine Learning Algorithm. *International Research Journal of Engineering and Technology, 5*(5).

Holland, P. (1986) Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230–243. Advance online publication. doi:10.1136/svn-2017-000101 PMID:29507784

Kaur, C., Sharma, K., & Sohal, A. (2019). Disease Prediction System using Improved K-means Clustering Algorithm and Machine Learning. *International Journal on Computer Science and Engineering*, *7*(5), 1148–1153.

Kibriya, A. M., Frank, E., Pfahringer, B., & Holme, G. (2004). Multinomial Naive Bayes for Text Categorization Revisited. In G. I. Webb & X. Yu (Eds.), Lecture Notes in Computer Science: Vol. 3339. *AI 2004: Advances in Artificial Intelligence. AI 2004*. Springer.

Kim, H., Park, C. M., & Goo, J. M. (2020). Test-retest reproducibility of a deep learning–based automatic detection algorithm for the chest radiograph. *European Radiology*, *30*(4), 2346–2355. doi:10.1007/s00330-019-06589-8 PMID:31900698

Knight, W. (2017). The Dark Secret at the Heart of AI. *MIT Technology Review*. https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

Konys, A. (2018). An Ontology-Based Knowledge Modelling for a Sustainability Assessment Domain. *Sustainability*, *2018*(10), 300.

Krishnamurthy, V. (2016). *Partially Observed Markov Decision Processing – from filtering to controlled sensing*. Cambridge University Press.

Lampropoulos, A. S., & Tsihrintzis, G. A. (2015). *Machine Learning Paradigms: Application in Reconmmender System*. Springer Nature. doi:10.1007/978-3-319-19135-5

Langarizadeh, M., & Moghbeli, F. (2016). Applying Naive Bayesian Networks to Disease Prediction: A Systematic Review. *Acta Informatica Medica*, *24*(5), 364–369. doi:10.5455/aim.2016.24.364-369

Laskar, , Rahman, Raihan, & Rashid. (2016). Automated Disease Prediction System (ADPS): A User Input-based Reliable Architecture for Disease Prediction. *International Journal of Computers and Applications*, *133*, 24–29.

Mate, S. (2015) Ontology-based data integration between clinical and research systems. *PloS One, 10*(1). doi:10.1371/journal.pone.0116656

Oberkampf, H. (2013). *An OGMS-based Model for Clinical Information (MCI).* ICBO.

Papachristou, N., Miaskowski, C., & Barnaghi, P. (2016). Comparing machine learning clustering with latent class analysis on cancer symptoms' data. *Proceedings of the IEEE Healthcare Innovation Point-of-Care Technologies Conference 2016*.

Pearl, J. (2010). An Introduction to Causal Inference. *The International Journal of Biostatistics*, *6*, 2.

Richens, J. G., Lee, C. M., & Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications*, *11*(1), 3923. doi:10.1038/s41467-020-17419-7 PMID:32782264

Sheridan, K. (2012). Choosing a Bread That Will Fit the Chronic Kidney Disease Diet: An Emphasis on Sodium and Phosphorus. *Journal of Renal Nutrition*, *22*(3), e27–e35. https://doi.org/10.1053/j.jrn.2012.02.002

Sheth, A., Agrawal, S., & Lathem, A. J. (2006) Active Semantic Electronic Medical Record. In Proceedings of the Semantic Web 2006 (pp. 913-926). Springer Berlin Heidelberg.

Verma, C., & Ghosh, D. (2018). Prediction of Heart Disease in Diabetic patients using Naive Bayes Classification Technique. *International Journal of Computer Applications Technology and Research.*, *7*(7), 255–258. doi:10.7753/IJCATR0707.1002

Vorhies, W. (2019). *The Next Most Important Thing in AI/ML.* https://www.datasciencecentral.com/profiles/blogs/causality-the-next-most-important-thing-in-ai-ml

Wang, F., & Preininger, A. (2019). AI in Health: State of the Art, Challenges, and Future Directions. *Yearbook of Medical Informatics*, *28*(1), 16–26. https://doi.org/10.1055/s-0039-1677908

Warner, J. (2008). *Eating Breakfast May Beat Teen Obesity*. https://www.webmd.com/diet/news/20080303/eating-breakfast-may-beat-teen-obesity

Wei, W., Visweswaran, S., & Cooper, G. F. (2011). The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. *Journal of the American Medical Informatics Association*, *18*(4), 370–375. doi:10.1136/amiajnl-2011-000101

Yekkala, I., & Dixit, S. (2018). Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection. *International Journal of Big Data and Analytics in Healthcare*, *3*(1), 1–12. doi:10.4018/IJBDAH.2018010101

Yu, H. Q. (2020). Health Causal Probability Knowledge Graph: Another Intelligent Health Knowledge Discovery Approach. In *2020 7th International Conference on Bioinformatics Research and Applications (ICBRA 2020).* Association for Computing Machinery. doi:10.1145/3440067.3440077

Yu, H. Q. (2021). Dynamic Causality Knowledge Graph Generation for Supporting the Chatbot Healthcare System. In *Proceedings of the Future Technologies Conference (FTC) 2020* (vol. 3). Springer International Publishing. doi:10.1007/978-3-030-63092-8_3

Zheng, X., Wu, K., Song, M., Ogino, S., Fuchs, C., Chan, A., Giovannucci, E., Cao, Y., & Zhang, X. (2019). Yogurt consumption and risk of conventional and serrated precursors of colorectal cancer. *Gut*.

ENDNOTES

[1] https://github.com/semanticmachinelearning/nhscausalknolwedgegraph
[2] https://github.com/semanticmachinelearning/HealthChatbot

Hong Qing Yu received the Ph.D. degree from the University of Leicester, in 2010. He is a Senior Lecturer with the School of Computing and Engineering, University of Derby, U.K. He leads the University's M.Sc. Big Data Analytics Program. Before he gained the Senior Lectureship, he has involved in more than ten research projects as a Research Associate and the Lecturer with the University of Leicester, Knowledge Media Institute (The Open University), and the University of Bedfordshire. These projects mostly were supported by European Committee Research Frameworks. His research interests include in the fields of data/text mining, semantic technologies, deep machine learning, natural language processing for healthcare, and e-learning systems. He is currently serving as a reviewer for many highly ranked international journals and program committee members for conferences and workshops.

Stephan Reiff-Marganiec (Member, IEEE) was an Associate Professor of Informatics with the University of Leicester, in 2003, and he has been the Director of the ERDF funded Leicester Innovation Hub, since 2017. He had worked with the computer industry in Germany and Luxembourg. From 1998 to 2001, he was a Research Assistant with the University of Glasgow, while at the same time reading for the Ph.D. degree in computing science. The work performed at Glasgow investigated hybrid approaches to the feature interaction problem. From 2001 to 2003, he was a Research Fellow with the University of Stirling, where he was investigating policies, emerging features, and associated conflict resolution techniques. He was responsible for organizing the British Colloquium for Theoretical Computer Science, in 2001 and 2004, and was a Treasurer of BCTCS, from 2004 to 2018. From July 2009 to 2013, he was appointed as a Guest Professor with the China University of Petroleum. He was a Visiting Professor with Lamsade, University of Dauphine, France, and a Visiting Researcher of ICMC with USP, Brazil, and UNIFEI, Brazil. He conducting the research in the broad area of service computing, the IoT, and cloud computing. He was a Principal Investigator of the project Ad-Hoc Web Applications funded by the Nuffield Foundation and the Leader of work packages and tasks in the EU funded projects Leg2Net, Sensoria, and inContext focusing on automatic service adaption, context aware service selection, and workflows and rule based service composition. He has led a KTP and a shorter KTP as an Academic Supervisor and an Academic Lead on a double KTP, all investigating various issues of data processing, management, and system optimization. He is currently a Professor of computer science and the Head of the School of Computing and Engineering, University of Derby. He has published more than 100 papers in international conferences, workshops, and journals. Prof. Reiff-Marganiec has been a member of a large number of program committees. He has served as a Panel Member and Reviewer for Funding Bodies in Brazil, Austria, and U.K. He was elected the member of the BCS (MBCS) in November 2002, the Fellow of the BCS (FBCS) in May 2009. He was a member of ACM and the Senior Member of the Steering Committee for YR-SOC until 2010. He was the Co-Chair of the 8th and 10th International Conference on Feature Interactions in Telecommunications and Software Systems and the second, third and fourth Young Researchers Workshop in Service Oriented Computing (YR-SOC 2007, 2008, and 2009). Most recently he was the PC Chair of the IEEE International Conference on Web Services (ICWS) 2016, the General Chair of ICWS 2017, and the PC Co-Chair of the Workshop Track at Services in 2019 and 2020 and the Symposium on Service Oriented Software Engineering (SOSE) in 2019 and 2020. He was the Co-Editor of the Handbook of Research on Service-Oriented Systems and Non-Functional Properties: Future Directions (2011).