CODCA – COVID-19 ONTOLOGY FOR DATA COLLECTION AND ANALYSIS IN E-HEALTH

Abdullahi Ismail1 and Melike Sah2

1 2 Computer Engineering Department, Near East University, North Cyprus, via Mersin 10 Turkey - (ismail.abdullahi, melike.sah)@neu.edu.tr


ABSTRACT:

Coronavirus (Covid-19) pandemic is one of the most deadly diseases that cause the death of millions around the world. Automatic collection and analysis of Covid-19 patient data will help medical practitioners in containing the virus. For this purpose, Semantic Web technologies can be utilized, which allows machine-processable data and enables data sharing, and reuse across machines. In this paper, we propose a Covid-19 ontology (named CODCA) that helps in collecting, analysing, and sharing medical information about people in the e-health domain. In particular, the proposed ontology uses information about medical history, drug history, vaccination history, and symptoms in order to analyse Covid-19 risk factors of people and their treatment plans. In this way, information about Covid-19 patients can be automatically processed and can be re-usable by other applications. We also demonstrate extensive semantic queries (i.e. SPARQL queries) to search the created metadata. Furthermore, we illustrate the usage of semantic rules (i.e. SWRL) so that treatment plans for individual patients can be inferred from the available knowledge.

1. INTRODUCTION

Covid-19 is a pandemic that jeopardizes the health and safety of everyone worldwide (Sönmez et al., 2020). Early detection of the virus will help in containing the virus. As we are in the age of technology, data is very vital in making decisions, forecasting, and analysis. For this purpose, recent works applied data analysis and machine learning (Sönmez et al., 2020; Direkoglu and Sah, 2020; ArunKumar et al., 2021; Ahmad et al., 2021; Arpaci et al., 2021; Marcos et al., 2021; Shirli et al., 2021). In (Direkoglu and Sah, 2020), a deep learning model is introduced to forecast the spread of Covid-19 with a cumulative number of daily death and a cumulative number of daily cases regionally and worldwide. Time-series statistical model was used in (ArunKumar et al., 2021) to forecast the spread of Covid-19 in the top 16 countries where most cases of Covid-19 are. A detailed review on predicting the confirmed cases of Covid-19 using machine learning and its challenges was conducted (Ahmad et al., 2021). Covid-19 is predicted in (Arpaci et al., 2021) using 14 clinical features as the best alternative to chest X-ray scans to predict Covid-19. In (Marcos et al., 2021) machine learning model was presented to identify the survival of the patients using clinical and laboratory information obtained during hospitalization. (Shirli et al., 2021) proposed a prognostic learning model that used clinical data and quantitative radiomic features of chest CT images of COVID-19 patients to predict the positive cases. These methods utilize information about many Covid-19 patients for predicting trends. On the other hand, there is a need to represent information about each Covid-19 patient so that individual patients' recommendations or analyses of patients in a certain location can be performed automatically. This can be achieved by using Semantic Web technologies, such as Extensive Markup Language (XML), Resource Description Framework (RDF), Ontology, SPARQL queries, and Semantic Web Rule Language (SWRL) (Akerkar, 2009). Semantic Web allows machine-processable data and interoperable data that can easily be used by machines like computers, mobiles, and IoT devices. Semantic interoperability is one of the challenges in the healthcare industry (Itikhar et al., 2012). Our research is aiming to promote the use of e-health systems, and especially collection and analysis of Covid-19 patient data using Semantic Web technologies. We propose a Covid-19 Ontology for Data Collection and Analysis, named CODCA that collects people’s medical history, drug history, vaccination history, and symptoms. In addition, we use SPARQL queries to analyse people based on the risk factors by considering their medical history and symptoms. Furthermore, using semantic rules, we provide recommendations and treatment plans. Although different ontologies were developed for Covid-19, the proposed ontology is unique in design and scope. Specifically, we aim to:

1. Help the medical practitioners to analyse the people's COVID-19 risk factors.

2. Help in recommending treatment and risk factors to patients based on medical history and vaccination history using semantic rules (i.e. SWRL).

3. Help in gathering the statistics regarding patients.

The proposed ontology aims to specifically answer research questions like who is Covid-19 positive, negative, and suspect? What are the statistics of vaccinated and non-vaccinated people? What are the risk factors of Covid-19 patients? What is the medical history of patients (i.e. chronic diseases)? What symptoms exist for severe Covid-19? What treatment plans to follow for different patients types? We use the proposed ontology, SPARQL, and SWRL to answer these questions automatically.

The rest of the paper is structured as follows: Section 2 reviews and discusses related Covid-19 ontologies and semantic approaches. Section 3 introduces the proposed ontology. Section 4 discusses extensive SPARQL queries to analyse data. Section 5 introduces SWRL for the recommended treatment plan and risk factors to the patients, and in Section 6 we provide discussions. Finally, we conclude in Section 7.
2. LITERATURE REVIEW

Different methodologies for Covid-19 data collection were reported in the literature. We review existing Covid-19 ontologies in terms of the application domain, software, aim, and purpose (Table 1). (Kouamé and McHeick, 2021) proposed a Covid-19 ontology called suspectedCOPDcovidology which uses important Covid-19 sign parameters and symptoms like fever temperature and so forth to diagnose the suspected Covid-19 cases from COPD patients. Protégé was used for developing the model. Data is dynamically collected from COPD patients to detect suspected cases. They used python, SQL Server, and the ONTYP platform for system/application development. Authors (Dutta and DeBellis, 2020) introduced ontology for collecting and analysis of Covid-19 data called CODO base on the FAIR data principle. The main aim of their research is: (1) to provide the explicit ontology to data and services providers, (2) an application that is semantic and heterogeneous, and (3) to help in providing a standard base vocabulary that is reusable for the organizations like the hospital. Their model is developed using the protégé software. CODO has 50 classes, 62 objects, and 45 data properties that are used in the ontology model design. SPARQL is used for querying. (Visweswaran et al., 2021) proposed an ontology that encompasses over 50,000 concepts from the sub-medical domain like diagnosis, procedures, medication, and laboratory. The research is aiming at the causes and outcomes of an illness (Covid-19), 14.5m patients’ data is used in this research. Their ontology model is developed using protégé software to enable the querying and harmonization of Electronic Health Records (EHRs). (Sherimo et al., 2020) presented an ontology model to help the medical practitioners in diagnosing the Covid-19 patients and to be used in an ontology-based clinical decision support system (CDSS) that assists in the treatment and diagnosis of pathologies. The main aim of their research is to design a Covid-19 ontology to enhance support system decisions. Symptoms, diagnosis, clinical diagnosis, lab test, epidemiological test recommendation, and treatment are parameters that are used in the decision system which comes from the context of the sultanate of Oman. SWRL rules and semantic query enhance web rules language (SQWRL) is used in data retrieval.

In (Sargsyan et al., 2020) researchers introduced an ontology that encompasses the chemical entities that are best for repurposing drugs. Their ontology includes 2270 classes and 38,987 axioms, which were developed and tested on Medline and Covid-19 corpus given by the Allen Institute. Protégé ontology editor is used to construct the ontology. They reported that their Covid-19 ontology can be modified and used in text mining. Authors (Santipantakis et al., 2020) presented a system that collects Covid-19 data from different public services. Their work shows that the integrated data can be used to answer complex questions regarding COVID-19. They have used SPARQL to query their ontology in Protégé editor, and also developed a web-based GUI for search results.

In (Qundus et al., 2021), an ontology for country Covid-19 responses called ROC is proposed. Authors provided an ontology that collects, link, and share Covid-19 data. The ontology focuses on country data and evaluating the side effect of government responses in different countries. Protégé is used for development and SPARQL is used for querying.

In (John Beverley, Shane Babcock,2, Lindsay Cowell,4, n.d.), two different ontologies were presented, one for infectious diseases, and the second one for Covid-19 infectious diseases. The latter is more specific to Covid-19, which is an extension to the ontology model called Coronavirus Infectious Diseases Ontology (CIDO) for investigation. (Dong et al., 2021) proposed an emergency ontology model called EOM-POSEs. They gathered emergency data from different sources to form an emergency dataset. They construct their ontology using the emergency dataset and public opinion elements of emergencies, then they later considered Covid-19 emergency in the construction of their system. Protégé tool was used in development.

Authors in (Celik Ertağrul and Çelik Ulusoy, 2021) presented the design of a rule-based expert system for the self-pre-diagnosis of Covid-19. Their system collects and shares Covid-19 data with health experts as well as assist in predicting the possible Covid-19 risk by analyzing the user symptoms. They developed a mobile application for smartphone users, health experts, and the government health authority. The system captured the needed information from the users according to TRNC law. Protégé tool was utilized in constructing the ontology, SWRL is used for semantic rules and SQWRL is used for querying.

In (Sayeb et al., 2021), the authors presented a web-based ontology for Covid-19 Crisis Health Care Information System. They aim to prove and manage the life and death in the Covid-19 period on health workers using their profile. Protégé ontology editor is used in building the ontology. Web application interface was developed to provide services so that health workers can manage the Covid-19 crisis and provide better decision making. In the reported literature review, none considered vaccine and vaccinated or non-vaccinated peoples history, and its effect on treatment, decision making. The proposed ontology also takes into account vaccination, and other patient information for recommending treatment plans and deciding risk factors of the patient, which is novel.

<table>
<thead>
<tr>
<th>Research</th>
<th>Software</th>
<th>User Interface</th>
<th>Querying Language</th>
<th>Aim</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Kouamé and McHeick, 2021)</td>
<td>Protégé</td>
<td>Yes</td>
<td>_</td>
<td>Data collection and analysis</td>
</tr>
<tr>
<td>(Dutta and DeBellis, 2020)</td>
<td>Protégé</td>
<td>No</td>
<td>SPARQL</td>
<td>Data collection and analysis</td>
</tr>
<tr>
<td>(Visweswaran et al., 2021)</td>
<td>Protégé</td>
<td>Yes</td>
<td>_</td>
<td>Cause and outcomes of an illness.</td>
</tr>
<tr>
<td>(Sherimont et al., 2020)</td>
<td>Protégé</td>
<td>Yes</td>
<td>SQWRL</td>
<td>Diagnosis, treatment, and knowledge base representation.</td>
</tr>
<tr>
<td>(Sargsyan et al., 2020)</td>
<td>Protégé</td>
<td>Yes</td>
<td>_</td>
<td>Investigating and use of the best suitable entities for drug repurposing.</td>
</tr>
<tr>
<td>(Santipantakis et al., 2020)</td>
<td>_</td>
<td>Yes</td>
<td>SPARQL</td>
<td>Collecting and integrating COVID-19 data from different sources to help to answer complex COVID-19 questions.</td>
</tr>
<tr>
<td>(Qundus et al., 2021)</td>
<td>Protégé</td>
<td>No</td>
<td>SPARQL</td>
<td>Collection, linking, and sharing of country COVID-19</td>
</tr>
<tr>
<td>Year</td>
<td>Method</td>
<td>Protégé</td>
<td>SWRL</td>
<td>Data Description</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------</td>
<td>---------</td>
<td>------</td>
<td>----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>(Dong et al., 2021)</td>
<td>Protégé</td>
<td>No</td>
<td>–</td>
<td>Construction of emergency COVID-19 dataset.</td>
</tr>
<tr>
<td>(Sayeb et al., 2021)</td>
<td>Protégé</td>
<td>Yes</td>
<td>–</td>
<td>Developing an ontology aiming to prove and manage the life and death in the COVID-19 period on health workers.</td>
</tr>
</tbody>
</table>

Table 1. Literature review summary and comparison

3. CODCA – PROPOSED COVID-19 ONTOLOGY

We provide the description of our ontology design, where we explain details of classes, object properties, and data properties of the ontology model. The conceptual schema of the developed ontology model is shown in Figure 1.

![Figure 1. Schema of the CODCA ontology.](image)

3.1 Classes of CODCA

The main classes are shown in Figure 2. The patient class is the superclass of all.

![Figure 2. Main class hierarchy of the CODCA ontology.](image)

Medical history has a subclass of chronic and non-chronic diseases (See Figure 3). Vaccination history has a subclass of fully vaccinated for people that finished taking their vaccine shots, start vaccination for people that start taking their shots, and not vaccinated subclass for those that don’t take Covid-19 vaccine. Drug history has a subclass of 0-1, 1-3, and over 3 weeks on drugs. Risk has a subclass of high, medium, and low risk (Bhopal et al., 2021) as shown in Table 2. Risk factors are automatically determined according to SWRL rules as we discuss in Section 5. Symptoms have a subclass of common, serious, and no symptoms, those with serious symptoms are usually hospitalized, those with common symptoms observe self-quarantine, and those that show no symptoms are advised to stay home for some days to confirm the situation. The diagnosis class has a subclass of Covid-19 positive, negative, and suspect. Covid-19 suspect is for those that are suspected to contain the virus either by direct contact to Covid-19 positive people or by traveling to a Covid-19 area. The recommendation class has a subclass of home quarantine and no quarantine. Table 2 shows the recommendation for COVID-19 positive, negative, and suspect respectively. Treatment class provides the treatment based on the diagnosis and medical history, it has self-quarantine with no medication, self-quarantine with medication, hospitalize with medication, and intensive care with medication subclasses. Treatment plans are automatically inferred by SWRL rules, which will be discussed in Section 5. The vaccine class contains the nine most well-known and available vaccines that can be given to people like Pfizer, AstraZeneca, and the rest.
4. QUERYING CODCA ONTOLOGY WITH SPARQL QUERIES

The SPARQL query engine is used to query ontology. SPARQL constructs a graph to search patterns (Tally, 2010). In this section, because of space restrictions, the snippets of some SPARQL queries are shown. SPARQL query in Figure 5 shows Covid-19 positive patients who are above 65, as well as their personal and previous medical history data is retrieved. SPARQL query in Figure 6 finds out the most popular vaccine preferred by people. In our work, patient data is manually entered using Protégé but in the future, a semantic annotation tool like (Sah and Direkoglu, 2017) can be used to automatically annotate patient data according to CODCA ontology.

Figure 5. SPARQL Query for COVID-19 positive cases who are above age 65

```
PREFIX rdfs:...
SELECT DISTINCT ?Name ?gender ?age ?Diagnosis
    ?patients cvd:Name ?Name.
    FILTER(?age>65)
    ?patients cvd:is_diagnose_with ?Diagnosis.
ORDER BY (?Name)
```

Figure 6. SPARQL query for vaccine jab the most.

```
PREFIX rdfs:...
SELECT ?vaccine (COUNT(?vaccine) AS ?NumberOfVaccineUsed)
WHERE { ?patients cvd:has_vaccinated_with ?vaccine}
GROUP BY ?vaccine
ORDER BY DESC(?NumberOfVaccineUsed)
LIMIT 1
```

5. SEMANTIC RULES TO INFERENCE TREATMENT PLANS AND RISK FACTORS OF COVID-19 PATIENTS

We have constructed a SWRL rules to automatically recommend a treatment plan and risk factors of different patients cases. Four different SWRL rules were created for self-quarantine-no-medication (Table 4, rule 1), self-quarantine with medication (Table 4, rule 2), hospitalization with medication (Table 4, rule 3), and intensive care with medication (Table 4,
rule 4). Table 4 shows the SWRL for the treatment plan recommendation. If a patient shows no symptoms but has contact with the COVID-19 patient then the treatment is self-quarantined with no medication. If a patient shows common symptoms like fever, cough, and tiredness with age less than 60 years old, low or medium risk factor, and has a medical history of non-chronic diseases like cancer, diabetes, etc. then, the treatment plan is self-quarantined with medication. If a patient shows serious symptoms like difficulty breathing with age less than 60 years old, medium or high risk factor, and has a medical history of chronic diseases then the treatment plan is to hospitalize with medication. If a patient shows serious symptoms like difficulty breathing with age more than 60 years old, high-risk factors, and has a medical history of chronic diseases then the treatment plan is intensive care with medication.

Three SWRL rules are constructed for automatically recommending high, medium, and low-risk factors for Covid-19 patients as shown in Table 5. As mention in Table 3, the patient is at high risk if has a chronic diseases medical history and is more than 60 years old, the patient is at medium risk of Severe Covid-19 if has non-chronic medical history and age ranging between 19 and 60, and the patient is at low risk if has a non-chronic diseases medical history and age less than 19 years old.

<table>
<thead>
<tr>
<th>No</th>
<th>Treatment Plan</th>
<th>SWRL rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Self Quarantine No Medication</td>
<td>Patients(?p) ^ has_symptoms_of(?p, no_symptoms) ^ ContactWithCOVID-19Patients(?p, false) -&gt; SelfQuarantineNoMedication(?p)</td>
</tr>
<tr>
<td>2</td>
<td>Self Quarantine With Medication</td>
<td>Patients(?p) ^ has_symptoms_of(?p, chronic_diseases) ^ has_medical_history_of(?p, non_chronic_diseases) -&gt; SelfQuarantineWithMedication(?p)</td>
</tr>
<tr>
<td>3</td>
<td>Hospitalize with Medication</td>
<td>Patients(?p) ^ has_symptoms_of(?p, serious_symptoms) ^ has_medical_history_of(?p, chronic_diseases) -&gt; HospitalizeWithMedication(?p)</td>
</tr>
<tr>
<td>4</td>
<td>Intensive Care With Medication</td>
<td>Patients(?p) ^ has_symptoms_of(?p, serious_symptoms) ^ has_medical_history_of(?p, chronic_diseases) -&gt; IntensiveCareWithMedication(?p)</td>
</tr>
</tbody>
</table>

Table 4. SWRL rules for recommending Covid-19 treatment plan

<table>
<thead>
<tr>
<th>No</th>
<th>Risk Factor</th>
<th>SWRL rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>High Risk</td>
<td>Patients(?p) ^ Age(?p, ?age) ^ has_medical_history_of(?p, chronic_diseases) -&gt; HighRisk(?p)</td>
</tr>
<tr>
<td>2</td>
<td>Medium Risk</td>
<td>Patients(?p) ^ Age(?p, ?age) ^ has_medical_history_of(?p, non_chronic_diseases) -&gt; MediumRisk(?p)</td>
</tr>
<tr>
<td>3</td>
<td>Low Risk</td>
<td>Patients(?p) ^ Age(?p, ?age) ^ has_medical_history_of(?p, non_chronic_diseases) -&gt; LowRisk(?p)</td>
</tr>
</tbody>
</table>

Table 5. SWRL rules for recommending Covid-19 risk factors

6. DISCUSSIONS/LIMITATIONS

It is a known fact that the information about Covid-19 in the world is still incomplete. It seems quite challenging to reach a final solution with the knowledge obtained from Covid-19 patients that are hospitalized (Çelik Erdoğan and Çelik Ulusoy, 2021). Hence, there is a need for more accurate and machine-processable information/data about patients. In our research, we proposed an ontology, called CODCA that collects Covid-19 patients’ important data like medical history, vaccination history, treatment, vaccine uses, and much more to help data collection in a machine-processable format. Our ontology also helps in recommending risk factors and treatments plans to patients using SWRL rules and depending on their symptoms, medical history, and risk levels. In addition, like in other medical domains such as diabetes (Göğebakan and Sah, 2021), using our ontology, a user interface can be developed for search and analysis of patient data.

Different than other related works shown in Table 6, in our approach, Semantic Web Rule Language (SWRL) is used for automatically recommending the possible treatment plans and risk levels to Covid-19 patients. For this purpose, different parameters like age, medical history, symptoms, and risk level are the most vital parameters used in recommending patient with the treatment plans and the level of COVID-19 risk. Young people are likely to have low risk, but depending on their medical history they may have high-level COVID-19 risk especially if they have chronic diseases like diabetes. For the treatment, depending on the patients’ symptoms they may be recommended to self-quarantine or to be hospitalized.

A limitation to the research is that some new symptoms and vaccines are continuously introduced continuously due to the mutation of the COVID-19 virus that was discovered to a new one called Delta variant; the COVID-19 mutant virus was recently discovered. Another limitation to this research is that our ontology collected and analyses data that is collected, it does not use other COVID-19 data from different sources that were already collected in performing the analyses, but that is targeted in our feature work.
Table 6. Comparison of similar research studies

<table>
<thead>
<tr>
<th>Work</th>
<th>COVID-19 ontology</th>
<th>COVID-19 sign parameters and symptoms to design their ontology model that diagnoses the suspected COPD patients. Their data was dynamically collected from COPD patients.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Dutta and DeBellis, 2020)</td>
<td>Data collection and analysis of COVID-19 called CODO.</td>
<td>Developing an ontology that is explicit to service providers like a hospital. They design their ontology based on the FAIR data principle.</td>
</tr>
<tr>
<td>(Shermon et al., 2020)</td>
<td>Covid-19 Ontology Engineering -Knowledge Modeling of SARS-CoV-2</td>
<td>Developing an ontology model by considering symptoms, lab tests, epidemiological test recommendations, and treatment to collect and make a decision on the patients.</td>
</tr>
<tr>
<td>(Qundus et al., 2021)</td>
<td>ROC: Ontology for country COVID-19 responses.</td>
<td>They design an ontology that collect government responses and used it to evaluate the side effect on the country.</td>
</tr>
<tr>
<td>Proposed Work</td>
<td>CODCA: COVID-19 ontology for data collection and analyses in e-Health.</td>
<td>The ontology model is designed to collect and analyze COVID-19 data; important data is collected like medical history, vaccination history, vaccine, symptoms, diagnoses, risk, treatments, and recommendations. SWRL is used to set recommendation rules for treatment and risk. The research is to help medical practitioners and other researchers in fighting against the pandemic.</td>
</tr>
</tbody>
</table>

To conclude, we have presented the design and development of COVID-19 ontology for data collection and analysis (CODCA), where we presented the ontology’s classes, properties, and sample SPARQL queries. Furthermore, we demonstrate the usage of SWRL rules for an automatic treatment plan and risk factor recommendations. Our research is aiming at collecting as much data as possible to help medical practitioners and other researchers in performing analysis, forecasting, and classifications. In future work, we plan to develop a semantic annotation interface so that Covid-19 patient data can be automatically annotated according to CODCA ontology.

REFERENCES


7. CONCLUSIONS AND FUTURE WORK

To conclude, we have presented the design and development of COVID-19 ontology for data collection and analysis (CODCA),


Tally, R.J., 2010. Sartre, marcus, and the Utopian project today, CLCWeb - Comparative Literature and Culture. https://doi.org/10.7771/1481-4374.1572