Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210187

A Knowledge Graph of Mechanistic Associations Between COVID-19, Diabetes Mellitus and Kidney Diseases

Michael BARRETT[a,1], Ali DAOWD[a], Syed Sibte Raza ABIDI[a] and Samina ABIDI[a,b]
[a] *NICHE Research Group, Faculty of Computer Science, Dalhousie University*
[b] *Dept. of Community Health and Epidemiology, Faculty of Medicine, Dalhousie University, Halifax, Nova Scotia, Canada*

Abstract. This paper proposes an automated knowledge synthesis and discovery framework to analyze published literature to identify and represent underlying mechanistic associations that aggravate chronic conditions due to COVID-19. We present a literature-based discovery approach that integrates text mining, knowledge graphs and ontologies to discover semantic associations between COVID-19 and chronic disease concepts that were represented as a complex disease knowledge network that can be queried to extract plausible mechanisms by which COVID-19 may be exacerbated by underlying chronic conditions.

Keywords. Knowledge graph, Literature-based discovery, Text mining, COVID-19, Diabetes Mellitus, Chronic Kidney Disease

1. Introduction

The SARS-CoV-2 virus (causative agent of COVID-19) has the ability to target multiple organs while activating, in some cases, an intense and systemic immune response. The virus uses the ACE2 receptor to gain entry into cells, which is expressed in the lungs, liver, endocrine pancreas, kidney, endothelium and heart [1], leading to acute multiorgan injuries and death [2], caused by some combination of direct viral involvement, systemic proinflammatory immune response, or systemic hypoxia and coagulopathy [3]. SARS-CoV-2 infections are more likely to progress to severe illness in patients with Diabetes Mellitus (DM) and Chronic Kidney Disease (CKD) [4]. At present, the basic mechanisms, pathophysiological pathways and processes by which COVID-19 *exacerbates* or is *exacerbated by* DM and CKD are not well known [5], despite the rapid accumulation of published evidence on COVID-19. Researchers are actively seeking tools to discover from published evidence how mechanisms involving virus-host interactions, gene-environment networks, and metabolic and immune signaling—referred to as *mechanistic associations*—contribute to increased risk for DM and CKD patients [6].

Discovering new knowledge to understand the mechanistic associations between COVID-19 and DM, CKD or both is challenging as the rapidly emerging evidence is distributed across a large number of publications that are often not overtly related and contain many explicit and implicit causal relationships between diseases, comorbidities

[1] Corresponding Author: Michael Barrett; E-mail: MBarrett@dal.ca

and molecular data (e.g. genes, proteins and metabolites). Recent attempts to integrate COVID-19 literature have resulted in web-based tools like the COVID-19 Knowledge Graph [7] that stores causal mechanistic associations that were manually extracted from the literature and presented as a visual semantic network.

This paper presents a novel COVID-19 knowledge synthesis and discovery approach that involves the integration of (a) Literature-Based Discovery (LBD) [8] to automatically extract both known and new mechanistic associations from published medical literature, such as PubMed; (b) Medical ontologies to further augment and extend the mechanistic associations; and (c) Knowledge Graph (KG) to represent the mechanistic associations for knowledge synthesis and discovery. We applied our approach to answer two queries—(1) what are the drivers of COVID-19 progression in the context of molecular and physiological perturbations that are associated with DM and CKD, and (2) how might these mechanistic associations be implicated in previously hypothesized mechanisms?

2. Methods

We developed an LBD framework (Figure 1) that combines text mining for knowledge extraction from published literature, integration of semantic associations between targeted concepts, representation of associated concepts as a knowledge graph, and pattern analysis to discover mechanistic associations between COVID-19, DM and CKD.

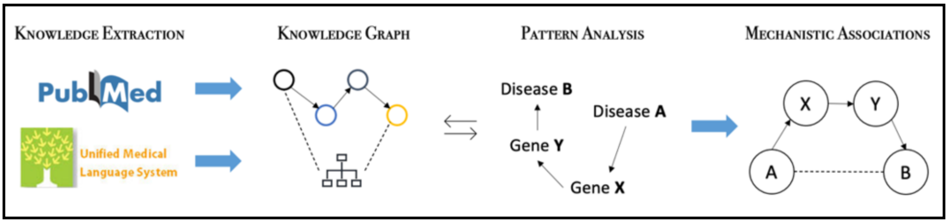

Figure 1. LBD framework.

2.1. Knowledge Extraction

We analyzed research articles related to COVID-19 and chronic diseases in PubMed (2020 – onwards). Candidate articles were selected using an iterative-feedback-based search process based on a set of pre-defined MeSH index terms. Article identifiers (i.e. PubMed IDs) were automatically retrieved and used to index the Semantic MEDLINE Database (SemMedDB), a repository of semantic associations extracted from PubMed articles [9] using an NLP tool called SemRep. Given textual abstracts as input, SemRep reads each sentence and infers logical relationships—called *semantic associations*—between concepts of interest. For instance, given the sentence *"The capacity for autophagy in both podocytes and renal tubular cells is markedly impaired in type 2 diabetes"*, it provided: (1) Tubular Cells (subject)–LOCATION_OF (relation)→ Autophagy (object) and (2) Podocytes (subject)–LOCATION_OF (relation)→ Autophagy (object). Each association consists of a standardized subject and object mapped to Unified Medical Language System (UMLS) semantic types (e.g. *cytokine production* is mapped to *Organism Function*) and a high-level relation type. SemRep extracts assertions based on relations in comorbidity (e.g. COEXISTS_WITH), substance interactions (e.g. INTERACTS_WITH) and physiological disturbances (e.g. DISRUPTS). A novel aspect of our approach is the integration of external knowledge sources (ontologies) and public databases to further extend the semantic associations to cover concepts related to gene function. This step involved semantic integration using the UMLS MRREL dataset that comprises associations between biomedical vocabularies (i.e. ontologies). Gene function and pathway associations were extracted from MRREL and linked to semantic associations where applicable. As a result, the KG was enriched with biomedical knowledge specific to DM, CKD, and COVID-19.

2.2. Knowledge Graph

The full set of semantic associations were represented as an interactive visual knowledge graph using a Neo4j graph database—the knowledge graph represents each subject and object concept as a node and the predicate (i.e. semantic association between the concepts) as a directed link between nodes. Each concept node was annotated with its name, UMLS identifier, semantic type, and frequency of occurrence. Predicates were assigned a type, frequency of occurrence, and unique identifier (article PMID or database reference) to establish provenance. Finally, concepts were assigned to one of five high-level UMLS semantic groups (*anatomy*, *chemicals*, *disorders*, *living beings*, or *physiology*) based on their semantic type. For example, genes/gene products and metabolites were assigned to chemicals and diseases and comorbidities to disorders.

We developed a two-step strategy to extract mechanistic associations using Neo4j's query language Cypher by specifying direct (concept A → B) or indirect (A → … → B) associations. An initial search was used to find *chemicals* indirectly (2 concepts away) associated with DM or CKD, which refers to molecular mechanisms (e.g. substance interactions) that may be implicated in disease. Conditions were set to ensure that A (chemicals) and B (DM or CKD) were not already directly related. Intermediate concepts were limited to other chemicals, physiology, or disorders (not COVID-19 or related). Then, a second search was developed to find indirect links (2 or more concepts away) between concepts returned by the previous search and COVID-19. This step generated multi-node patterns—called *discovery patterns*—that comprised chemical interactions, physiological disturbances, and disease-disease relations (e.g. comorbidity).

2.3. Pattern Analysis

We analyzed the discovery patterns to focus on highly relevant findings. Semantic filtering to exclude concepts belonging to unrelated semantic types (e.g. *Laboratory Procedure*) as these were too generic or uninformative. Ranking methods were applied to narrow down the list of potential hypotheses generated by the LBD approach. We compared (i) indirect association and (ii) graph-theoretic measures, namely linking term count (LTC) and PageRank, respectively. LTC considers the number of intermediate concepts when A and B are 2 concepts away while PageRank is a measure of each node's connectivity in the network. LTC was chosen because of its empirically proven performance and PageRank is considered an improvement on other metrics (e.g. Degree centrality) that are biased toward well-known concepts. We used the average score of each pattern to rank the associations as a criterion for discovery of mechanistic associations. Ontology terms were included to express concepts in finer detail, which was especially useful for ambiguous concepts such as "Immunity". As a result, we were able to identify mechanistic associations beyond the published literature.

3. Results

We found 605 papers on COVID-19 related to DM or CKD. The three most frequently occurring index terms were therapy, complications (i.e. comorbidities) and virology. We retrieved 1,911 unique associations from these articles in SemMedDB and represented them in a knowledge graph. The result was a complex network of comorbid conditions with molecular, physiologic, and pathologic levels. Rankings were similar for both LTC and PageRank, and the three top-ranking discovery patterns that are regarded as candidate mechanistic associations are shown in Table 1.

Table 1. Pathways of semantic predications linking chemicals and disorders.

Node 1	Relation	Node 2	Relation	Node 3	Relation	Node 4
S protein	*Interacts with*	ACE2	*Disrupts*	Diabetes Mellitus	*Coexists with*	COVID-19
S protein	*Interacts with*	ACE2	*Disrupts*	Hypertension	*Predisposes*	COVID-19
S protein	*Interacts with*	ACE2	*Associated with*	Heart failure	*Predisposes*	COVID-19

In Table 1, S protein refers to the protein used by SARS-CoV-2 to enter human cells via angiotensin-converting enzyme 2 (ACE2). The relation ACE2 *disrupts* Diabetes Mellitus/Hypertension is misleading since ACE2 may actually play a protective role in these disorders [10]. This relation should convey that the change in ACE2 due to COVID-19 somehow disturbs physiologic processes important in DM or cardiovascular diseases. Figure 2 shows the role of ACE2 in the disease network including its interactions with other chemicals, effect on physiologic disturbances, and association with disease concepts.

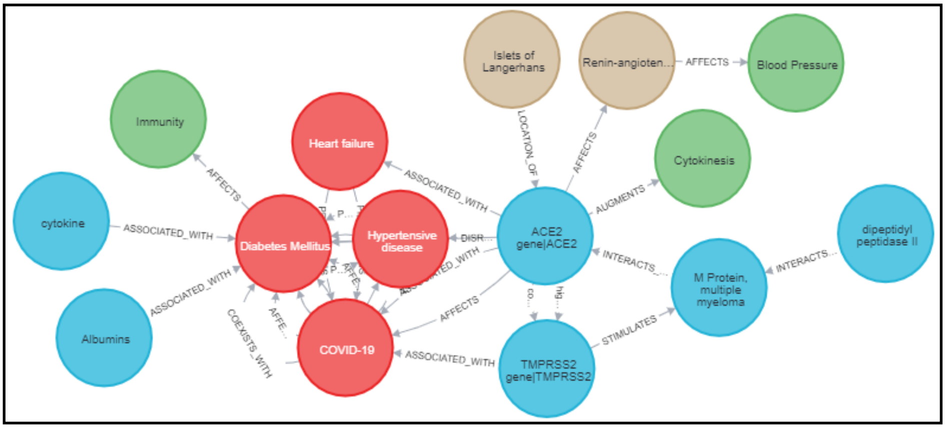

Figure 2. Segment of the KG. Blue nodes(chemicals): Cytokine, Albumins, ACE2, S protein, DPP2, TMPRSS2. Red nodes(disorders): COVID-19, DM, Hypertension, Heart failure. Brown nodes(anatomy): Islets of Langerhans, Renin-angiotensin. Green nodes(physiology): Immunity, Cytokinesis, Blood Pressure.

4. Discussion and Concluding Remarks

This paper presents an LBD framework that offers a novel integration of published literature with biomedical knowledge to synthesize published evidence to discover new knowledge about COVID-19 impact on chronic conditions. Previous LBD studies only consider indirect links as 2 concepts away, such as the work to discover associations between biomarkers of arterial stiffness [11], using a single pattern ranking technique. In our work, we have generated patterns to link distant literature sources and thus increase the likelihood of discovering hidden or unknown associations, whilst also using multiple pattern-ranking methods. We conclude that further investigation is needed to assess the relative performance of LTC and PageRank, as our initial results suggest that both are able to focus the user's attention on highly relevant patterns.

The associations found by our approach are intuitive and reflect suggested hypotheses in the COVID literature. DM is closely intertwined with hypertension and cardiovascular disease, which are known contributors to COVID-19 severity. Untangling interactions between these disorders in the context of COVID-19 was facilitated by our LBD framework. We were able to identify plausible links between molecular and disease concepts that were enriched with contextual information i.e. the nature of those relationships. The relation extraction task was prone to error when assessing the relatedness of chemicals (ACE2) and disorders (DM), but relations between chemicals were accurate and relevant. SARS-CoV-2 binds the ACE2 receptor with high affinity, which may lead to loss of cardiovascular protection and anti-inflammatory regulation in the lungs and adipose tissue [6]. We conclude that more research is needed to address the association between virus-host interactions and physiological perturbations in DM or CKD. There may be downstream effects of these interactions that alter important host pathways or physiologic processes, such as blood pressure and immune response regulation. Future work will integrate additional external knowledge sources to address the noted paucity of biomedical evidence pertaining to COVID-19 and DM or CKD.

References

[1] Iwai M, Horiuchi M. Devil and angel in the renin-angiotensin system: ACE-angiotensin II-AT1 receptor axis vs. ACE2-angiotensin-(1-7)-Mas receptor axis. Hypertens Res. 2009;32(7):533–6.

[2] Potere N, et al. Acute complications and mortality in hospitalized patients with coronavirus disease 2019: A systematic review and meta-analysis. Vol. 24, Critical Care. BioMed Central; 2020.

[3] Wu T, et al. Multi-organ Dysfunction in Patients with COVID-19: A Systematic Review and Meta-analysis. Aging Dis. 2020 Jul 30;11(4):874.

[4] Singh AK, et al. Prevalence of co-morbidities and their association with mortality in patients with COVID-19: A systematic review and meta-analysis. Diabetes, Obes Metab. 2020 Jul 16;22(10).

[5] Gupta R, Hussain A, Misra A. Diabetes and COVID-19: evidence, current status and unanswered research questions. European Journal of Clinical Nutrition. Springer Nature; 2020.

[6] Apicella M, et al. COVID-19 in people with diabetes: understanding the reasons for worse outcomes . Vol. 8, Lancet Diabet and Endocrin. Lancet Publishing Group; 2020 p. 782–92.

[7] Domingo-Fernandez D, et al. COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. bioRxiv. 2020 Apr 15.

[8] Henry S, McInnes BT. Literature Based Discovery: Models, methods, and trends. J Biomed Inform. 2017;74:20–32.

[9] Kilicoglu H, et al. SemMedDB: A PubMed-scale repository of biomedical semantic predications. Bioinformatics. 2012;28(23):3158–60.

[10] Gheblawi M, et al. Angiotensin-Converting Enzyme 2: SARS-CoV-2 Receptor and Regulator of the Renin-Angiotensin System. Circ Res. 2020;1456–74.

[11] Baek SH, et al. Enriching plausible new hypothesis generation in PubMed. Smalheiser NR, editor. PLoS One. 2017 Jul 5;12(7):e0180539.