*Integrated Citizen Centered Digital Health and Social Care*
*A. Värri et al. (Eds.)*

*doi:10.3233/SHTI200684*

# Exploring the Social Drivers of Health During a Pandemic: Leveraging Knowledge Graphs and Population Trends in COVID-19

Joao H. BETTENCOURT-SILVA[a,1], Natasha MULLIGAN[a], Charles JOCHIM[a], Nagesh YADAV[b], Walter SEDLAZEK[b], Vanessa LOPEZ[a] and Martin GLEIZE[a]
[a]*IBM Research Europe, Dublin, Ireland*
[b]*IBM Watson Health*

**Abstract.** Social determinants of health (SDoH) are the factors which lie outside of the traditional health system, such as employment or access to nutritious foods, that influence health outcomes. Some efforts have focused on identifying vulnerable populations during the COVID-19 pandemic, however, both the short- and long-term social impacts of the pandemic on individuals and populations are not well understood. This paper presents a pipeline to discover health outcomes and related social factors based on trending SDoH at population-level using Google Trends. A knowledge graph was built from a corpus of research literature (PubMed) and the social determinants that trended high at the start of the pandemic were examined. This paper reports on related social and health concepts which may be impacted by the COVID-19 outbreak and may be important to monitor as the pandemic evolves. The proposed pipeline should have wider applicability in surfacing related social or clinical characteristics of interest, outbreak surveillance, or to mine relations between social and health concepts that can, in turn, help inform and support citizen-centred services.

**Keywords.** Social determinants of health, Knowledge Graphs, Natural Language Processing, Relation Extraction, Population Trends, COVID-19 risk factors

## 1. Introduction

The World Health Organisation (WHO) defines the Social Determinants of Health (SDoH) as the circumstances in which people grow, live and work that affect their health [1]. Examples of SDoH include socioeconomic status, education or unemployment and addressing them is important to improve health and to reduce longstanding disparities [2]. In recent years, there has been a growing number of government initiatives that tackle SDoH, including nutritional programs addressing food insecurity (i.e. availability and access to healthy foods) or transportation programs boosting access to employment [2]. However, further work is needed to measure the impact of SDoH dimensions and to identify gaps and inconsistencies from data. For example, electronic health record systems have not traditionally been designed to capture SDoH related data and healthcare terminologies such as ICD-10 or SNOMED-CT may not extensively cover social concepts [3]. The COVID-19 pandemic is magnifying disparities across the SDoH and can disproportionately affect low-income, food-insecure households that struggle to meet basic needs [4]. Furthermore, certain social environments or vulnerabilities may increase

[1] Corresponding Author, JH Bettencourt, IBM Research, Dublin, Ireland; E-mail: jbettencourt@ie.ibm.com.

the susceptibility of contracting COVID-19 as well as the risk of developing complications or poorer outcomes. For example, overcrowding and housing insecurity has been shown to lead to increased COVID-19 transmission rates [5]. Therefore, identifying SDoH dimensions that characterise vulnerable populations is of utmost importance, especially during a pandemic.

Social media has been used to track trends and disseminate health information during viral epidemics. Examples include understanding sentiment during COVID-19 [6], and visualising health-related spatial social media data [7]. The latter used Twitter to study the link between healthy/unhealthy food tweets and locations with limited access to affordable and nutritious food. Google Trends has also been used to investigate symptom searches during the COVID-19 outbreak and results showed a strong correlation between the frequency of searches for smell-related symptoms information and the onset of COVID-19 infection in several countries [8].

This paper focuses on the use of natural language processing (NLP) techniques to investigate COVID-19 and its collateral impacts through the SDoH. We propose a pipeline to (1) monitor population-level trends for arising social determinants and (2) utilize a knowledge graph (KG) built from research literature (PubMed) to surface concepts related to those SDoH which may also be valuable to monitor. Previous work in relation extraction has used semi-automatic methods to discover lexico-syntactic patterns of causal relations [9] and KGs have been built from PubMed for COVID-19 [10], yet, to our knowledge, no works have been published describing how such graphs may be used to help monitoring population trends of social-related aspects during public health crises such as COVID-19.

## 2. Methods

This section describes the steps taken to develop the pipeline and their respective components. Figure 1 shows an overview of the proposed pipeline. A typical use-case begins by monitoring population trends for a predefined set of keywords. In this paper, a well-established set of SDoH keywords was monitored using Google Trends and this is described in detail in section 2.1. Specific SDoH keywords are then identified by performing a statistical analysis of population data (e.g. keywords trending higher in a particular time period compared to historical data). Such keywords become *seeded terms* to be found as nodes in a knowledge graph (KG) of related concepts. Finding the nodes connected to the *seeded terms* by traversing the KG yields additional nodes with insights of potentially relevant concepts to be investigated further. The knowledge graph in this paper was built by first mining co-occurring concepts from the literature (section 2.2) and then extracting relations between those concepts (section 2.3) using a trained classifier. A graph database was subsequently used to store, query and visualise the mined concepts (section 2.4).

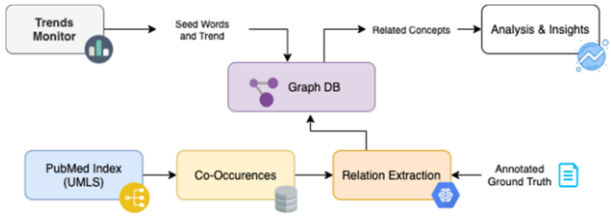


**Figure 1.** Overview of the pipeline.

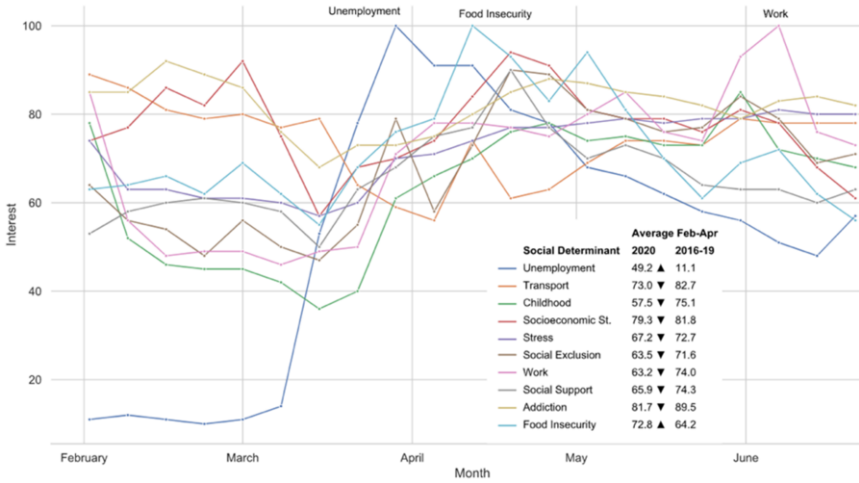


**Figure 2.** Chart showing Google Trends' interest across 10 SDoH dimensions between February-June 2020 and a summary of the average interest at the start of the pandemic (February-April 2020) compared with previous years.

### *2.1. Monitoring Trends with Google Analytics*

In order to identify relevant SDoH trends, a list of ten seeded terms (e.g. *Unemployment*, *Social exclusion*) based on the WHO's definition of social determinants of health [1] was selected, each corresponding to a SDoH dimension. This list was then mapped to the exact (n=6) or nearest Google Trends' Topic (n=4, e.g. *early life* mapped to *Childhood*) so that trends data could be collected for the past 5-year period (the longest period available for collection) and in English language. The collected trends data was then used to monitor the interest of SDoH dimensions from Google with particular attention to the months when the COVID-19 pandemic flared beyond China (defined in this paper as the time period between February to April 2020). Figure 2 shows the trends during this time period and also reveals the averages for the same months over the past four years for all ten SDoH dimensions. From the list of ten Google Topics, *Unemployment* and *Food Insecurity* were the two that peaked the most during the start of the pandemic and also saw their highest 5-year peaks in the same period. These two concepts were selected for the case study presented in this paper to illustrate the developed pipeline. Other methods, techniques and data sources may be used in this step for trend surveillance.

### *2.2. Indexing Evidence from PubMed*

A natural way to mine relationships between socio-medical concepts is to look for their co-occurrence in published literature [11]. We indexed the full 2019 MEDLINE PubMedBaseline[1], which notably includes the abstracts of articles. We used MetaMap [12] to tokenise and identify UMLS concepts in the sentences of the abstracts and indexed each single sentence so that it could be retrieved using multiple annotation layers, like words and phrases, UMLS semantic types, or UMLS concepts. We then queried the index for any pair of a SDoH identified in Section 2.1, and another of the same SDoH or

[1] https://www.nlm.nih.gov/databases/download/pubmed_medline.html

a medical concept (listed in UMLS). UMLS is a very extensive ontology [13] and the concepts identified by MetaMap vary widely in nature, so we restricted the medical concepts to only those of the following UMLS semantic types[2]: *Disease or Syndrome*, *Individual Behavior, Mental or Behavioral Dysfunction*. These concepts seemed to be the most relevant to our use case, which is to identify potential socio-medical issues in the context of COVID-19. We additionally filtered out of the results the sentences containing three concepts or more, which we believed would prove too difficult to use to extract accurate pairwise relations. In total, these queries yielded 20,244 sentences.

### *2.3. Relation Extraction*

Given co-occurring concepts of interest and their sentence context, we then predicted the relation between them within sentences. Previous approaches either used co-occurrence counting [11] or syntactic rules [14] to predict a Bayesian probabilistic relation between biomedical concepts. In this work, we captured more fine-grained relations, using a supervised sentence classification model.

#### *2.3.1. Ground Truth Annotation*

We sampled 550 of the context sentences and manually annotated them with 5 labels according to the statement made on the relation between the concepts in the article: *positive* if the concepts were found to be in positive correlation, *negative* for a negative correlation, *complex* for a more complex relation not easily classified as the first two (e.g. a relation conditioned on a specific characteristic of the population), *nocor* if the authors did not find a correlation, and *n/a* for sentences not expressing any sort of statement on the relation at all. In terms of balance, each label made up respectively 39, 3, 19, 1, 37 percents of the dataset. Four independent annotators labeled 50 sentences as the pilot, with a Fleiss' kappa of 0.732, then pairs of annotators labelled the remaining 500 sentences.

#### *2.3.2. Sentence Classification Experiments*

We fine-tuned a transformer BERT-base-uncased model [15] with a dense last layer on the train portion of our dataset and evaluated it on the test, with an 80/20 split. Most learning parameters were kept as default, batch size was set to 8 and we ran 2 epochs. Given the class imbalance found in our annotations (*positive, complex* and *n/a* make up 96% of the instances), we also performed the same experiment on a 2-class restriction of the problem, turning *positive/negative/complex* into *positive*, and *nocor* and *n/a* into *n/a*, to model a binary "relation"-"no relation" classification. For each setup we saw a performance accuracy of 63% (5-class) and 83% (2-class) compared with baselines for predicting the most frequent label in the train set of 41% (5-class) and 57% (2-class). We report in each case a higher accuracy for the trained classifier than the basic baseline, and also logically see a higher accuracy for the 2-class classification compared to the 5-class. We then used the 2-class classifier to validate an edge between co-occurring concepts in order to ultimately build a graph. When the same pair of concepts occurs in multiple sentence contexts, we validate their relation using a majority vote of all the predictions.

[2] https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html

### *2.4. Graph Database*

A graph database (Apache Tinkerpop stack) was used to store and query the co-occurring concepts and relations. A property graph was modeled in GraphML language and visualised using Graphexp connected to a Tinkerpop Gremlin server. A first version of the pipeline was built on a smaller subset of nodes and their related concepts (top-5 relative co-occurrences).

## 3. Results and Discussion

In this paper, a pipeline was built to identify and extract related social and health concepts of relevance to COVID-19. The analysis of trending SDoH dimensions at the start of the pandemic identified *Unemployment* and *Food Insecurity*. Relative frequencies were computed for all concepts that co-occurred with *Unemployment* (n=16,314) and *Food Insecurity* (n=7,876). A sub-graph (Figure 3) showing the two SDoH dimension concepts and their most relevant neighbours based on relative frequency was produced. Figure 3 illustrates disease concepts associated with *Unemployment* such as *Tuberculosis* and mental health disorders. Similarly, health conditions related to *Food Insecurity* include *Malnutrition*, *Diabetes* and *Anemia*. It is also reassuring that only a small number of noisy nodes are seen in this sub-graph (e.g. *Likely*). Noise can be controlled by selecting semantic types and it is likely to increase as thresholds for selecting neighbours are relaxed. Most interesting are the nodes connected to both SDoH dimensions (e.g. *Obesity* or *Depression*). It can be argued that any of these concepts should be closely monitored and analysed in the time period following the start of the pandemic. For example, a simple analysis of Google Trends (Worldwide) from May to June 2020 revealed peaks for *Obesity* (Google Trend class: medical condition) and *Coping* (topic) in May 2020 and for *Anxiety* (emotional disorder) in June. These examples show the largest interest recorded in the past 5-years. Further work is needed to analyse this data, inspect other geographical levels, and understand the causes for the sudden rise in these concepts. However, these first results indicate that a pipeline such as the one presented in this paper may be a useful first step to extract structured knowledge that can be used, for example, to help identify upcoming trends that may affect services and populations.


**Figure 3.** Visualisation of a sub-graph showing two SDoH dimensions (Unemployment and Food Insecurity) and their top-5 related concepts (nodes) based on selected UMLS Semantic Types.

## 4. Conclusions

We present a pipeline for mining relations between health and social concepts from published literature based on trending SDoH dimensions at the start of the COVID-19 pandemic. Future work will explore ways to extend our Knowledge Graph with additional social concepts, to learn better relation type labels and weights for edges, link social concepts to other ontologies and. Further work is also needed to continue analysing population data.

## References

[1] Wilkinson RG, Marmot M, for Europe WHORO, Project WHC, for Health WIC, Society. Social Determinants of Health: The Solid Facts. Academic Search Complete. World Health Organization; 2003.
[2] Artiga S, Hinton E. Beyond health care: the role of social determinants in promoting health and health equity. Health. 2018;20(10):1–13.
[3] Bettencourt-Silva J, Mulligan N, et al. Discovering New Social Determinants of Health Concepts from Unstructured Data: Framework and Evaluation. Stud Health Technol Inform. 2020 Jun;270:173–177.
[4] Wolfson JA, Leung CW. Food Insecurity and COVID-19: Disparities in Early Effects for US Adults. Nutrients. 2020;12(6):1648.
[5] Deziel NC, Allen JG, et al. The COVID-19 pandemic: a moment for exposure science. Journal of Exposure Science & Environmental Epidemiology. 2020; p. 1–3.
[6] Medford R, Saleh S, et al. An "Info-demic": Leveraging High-Volume Twitter Data to Understand Public Sentiment for the COVID-19 Outbreak. medRxiv; 2020.
[7] Widener MJ, Li W. Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. Applied Geography. 2014;54:189-197.
[8] Walker A, Hopkins C, Surda P. Use of Google Trends to investigate loss-of-smell-related searches during the COVID-19 outbreak. Int Forum Allergy Rhinol. 2020.
[9] Girju R, Moldovan DI, et al. Text mining for causal relations. In: Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference, May 14-16, 2002, Florida, USA; 2002.
[10] Oniani D, Jiang G, Liu H, Shen F. Constructing Co-occurrence Network Embed-dings to Assist Association Extraction for COVID-19 and Other Coronavirus In-fectious Diseases. J of the Am Med Informatics Ass. 2020.
[11] Theobald M, Shah N, Shrager J. Extraction of Conditional Probabilities of the Relationships Between Drugs, Diseases, and Genes from PubMed Guided by Relationships in PharmGKB. In: 2009 AMIA Summit on Translational Bioinformatics. American Medical Informatics Association. AMIA; 2009. p. 124–128.
[12] Aronson A, Lang FM. An Overview of MetaMap: Historical Perspective and Recent Advances. Journal of the American Medical Informatics Association : JAMIA. 2010. 05;17:229–36.
[13] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004;32(suppl1):D267–D270.
[14] Trovati M, Hayes J, Palmieri F, Bessis N. Automated extraction of fragments of Bayesian networks from textual sources. Applied Soft Computing. 2017;60:508–519.
[15] Devlin J, Chang M, et al. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the ACL, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.