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

doi:10.3233/SHTI210271

Social Determinant Trends of COVID-19: An Analysis Using Knowledge Graphs from Published Evidence and Online Trends

Martin GLEIZE[a], Natasha MULLIGAN[a], Alessandro DI BARI[b] and Joao H. BETTENCOURT-SILVA[a,1]
[a] *IBM Research Europe,*
[b] *IBM Watson Health*

Abstract. This paper presents the results of a new approach to discover related health and social factors during the COVID-19 pandemic. The approach leverages a knowledge graph of related concepts mined from a corpus of published evidence (PubMed) prior to the pandemic. Population trends from online searches were used to identify social determinants of health (SDoH) concepts that trended high at the outset of the pandemic from a list of SDoH topics from the World Health Organization (WHO). The trending concepts were then mapped to the knowledge graph and a subsequent analysis of the derived insights, spanning two years, was conducted. This paper suggests an approach to derive new related health and social factors that may have either played a role in, or been affected by, the onset of the global COVID-19 pandemic. In particular, our results show how, from a list of SDoH topics, Food Security, Unemployment trended the highest at the start of the pandemic. Further work is needed to continue to ascertain the validity of the derived relations in a population health context and to improve mining insights from published evidence.

Keywords. Social Determinants of Health, Knowledge Graphs, Infodemiology, Population Trends, Natural Language Processing, COVID-19 risk factors

1. Introduction

Global efforts are ongoing to identify vulnerable and at-risk populations during the COVID-19 pandemic, however, both the short- and long-term health and social impacts of the pandemic on individuals and populations are still not well understood. Studying how social determinants impact disadvantaged and other sub-populations during times of crisis can help governments better plan and manage health emergencies while ensuring every individual has an equal opportunity to staying healthy [1]. Examples of the impact of social determinants of health (SDoH) on the coronavirus pandemic are emerging in the literature [1-3]: Individuals living in food deserts, i.e. locations with poor access to healthy foods, were found to have a poor diet and this increases their risk of obesity [2]. The latter has also shown to have a strong association with an increased risk of hospital admissions as well as a risk of critical illness among patients with COVID-19 [3]. The growing wealth of information available in published literature, both before and after the

[1] Corresponding Author, Joao H Bettencourt-Silva; E-mail: jbettencourt@ie.ibm.com.

onset of the coronavirus pandemic, could be used to identify other important associations between health conditions and even social factors. Understanding these complex associations and how they may relate to population trends is important to better analyse and inform public health trends. Studies of online search engine trends, such as Google Trends2, have shown statistically significant correlations with COVID-19 data such as geographical case distributions [4]. Estimated models using the same trends' data have also shown strong COVID-19 predictability [4]. Indeed, internet sources may be employed to inform public health and policy and can be valuable to monitor and forecast outbreaks and epidemics [4]. Our previous work [5] focused on monitoring concepts of interest in Google Trends and exploring, in PubMed, their connection to other concepts in the health and social domains. A knowledge graph was built by drawing an edge between two concepts found in correlation using a state-of-the-art neural NLP model [5] and other approaches are possible [6,7]. This paper reports the subsequent analysis of a later span of Google Trends data and gives a first glimpse of the predictive ability of such a graph. This paper uses 2019 PubMed3 data to study whether the relations found from previously published evidence could help to provide other relevant population insights by observing online search trends. This should help inform further research avenues in health informatics, public health and infodemiology.

2. Methods

The approach consists of two steps, described in detail below: (1) a knowledge graph of relations between concepts is mined from PubMed using NLP techniques [5] and (2) population trends are mapped to the knowledge graph in order to study online search trends and how they might relate to the insights found from published evidence.

2.1 Knowledge Graph from Published Evidence

The knowledge graph was created with PubMed information up to 2019, therefore guaranteeing it was not built on associations discovered during or after the pandemic. MetaMap [8] was used to tokenize and identify UMLS concepts in the sentences of the abstracts. Sentences containing pairs of UMLS concepts were retrieved and the relationship between concepts was assessed using the same fine-tuned BERT transformer model as in [5]. UMLS[4] is a very extensive and diverse ontology, so we restricted the concepts to SDoH and health concepts of the semantic types most relevant to our use case: *Disease or Syndrome*, *Individual Behavior*, *Mental or Behavioral Dysfunction, Findings*. The classes used in the annotations and as output of the BERT model were the binary restriction of the 5-class system described in [5], and simply put, they model if the concepts in the sentence are stated to be in a correlation relation or not.

By aggregating sentences containing the same two concepts, each judged for the correlation relation, we could decide if an edge was drawn between the two concepts and the knowledge graph was created. In this paper, we chose to create an edge between concepts when more than half the sentences featuring them were indeed judged to contain a correlation relation.

[2] Google Trends, http://www.google.com/trends
[3] PubMed, https://www.nlm.nih.gov/databases/download/pubmed_medline.html
[4] UMLS, https://www.nlm.nih.gov/research/umls

2.2 Population Trends

A list of Social Determinant of Health (SDoH) concepts (or terms) from the World Health Organization definition [9] was used as seed terms, i.e. the terms to be observed at the onset of the pandemic in March 2020. From this list, the terms that peaked the most at the start of the pandemic, when compared with the homologous period in the previous 4 years, were selected (*Unemployment, Food Security*) and supporting information is given in Section 3.

Using the method described in section 2.1, the top-5 terms connected with *Unemployment* and *Food Security* were obtained for the four selected UMLS semantic types. The selected semantic types should help capture relevant information such as diseases and syndromes and filter out less useful information such as qualitative concepts. Despite this, some concepts were still filtered out either because they were too generic (e.g. *Skills*) or because they were not easily mapped to Google Trends' topics. Mapping each concept to topics was needed to ensure an accurate coverage of the concept from the search engine results. The resulting list of connected terms from the knowledge graph represents those concepts mined from PubMed articles that are related to the seeded terms (Figure 1).

Subsequently, data from Google Trends (Worldwide) was observed between the week of the 6th January 2019 and the week of the 10th January 2021, ensuring enough time before the pandemic was used to reduce bias from seasonal trends. This information was analysed to identify related concepts trending high (i.e. those that reached the peak Google Trends' interest value of 100) in the months following the onset of the pandemic.

3. Results and Discussion

Figure 1 shows the knowledge graph and the concepts related to the two seeded terms (N=29). The two SDoH seeded terms showed a substantial increase in interest also when only considering the two-year time period. An intersection of the top-5 resulting concepts for both seed terms is observed in Figure 1 and covers 38% (N=11) of the concepts. 9 concepts were removed and represent noise which could be filtered by applying additional NLP techniques. Online search trend data was overlayed with the knowledge graph and concepts that reached their highest value after the seeded terms peaked are represented in triangle shapes (N=10). The online trend data was normalised for the time period from January 2019 to January 2021 to account for seasonal bias. We also observe that, after removing noise, the majority of concepts that overlap with both *Food Security* and *Unemployment* had rising trends after the onset of the pandemic (N=6/8).

Figure 1. Knowledge graph mined from PubMed abstracts showing the concepts connected to Food Security and Unemployment. Triangle shapes indicate concepts that rose to their top interest value after the start of the pandemic. Concepts that were removed are shown in brackets.

A more thorough analysis of the trends is given in Figure 2 for those 8 concepts that overlap the two seeded terms. Figure 2 shows a chart for each of the 8 concepts (in blue) overlayed on top of the two seeded terms. A vertical line marks March 2020 which was used here to represent the start of the coronavirus pandemic as it expanded beyond China.

Concepts *Smoking* and *Disease* were the only two that did not reach an overall interest of 100 after the seed terms peaked, however, *Disease* had just peaked earlier in February 2020. Overall, the remaining 6 terms all peaked to the maximum interest value in the observed two-year time window.

In addition to the chart in Figure 2, correlation coefficients and a correlation matrix was computed (not shown). Based on this analysis, and as confirmed by the trends in Figure 2, the most positive correlation observed was between *Unemployment* and *Anxiety* (r=.72) suggesting that anxiety might be an increasing concern as the pandemic continued in time. Other observations include: *Anxiety* and *Distress* (r =.66); *Coping* and *Distress* (r=.65); *Coping* and *Unemployment* (r =.61); *Distress* and *Unemployment* (r =.64); *Distress* and *Food Security* (r =.62).

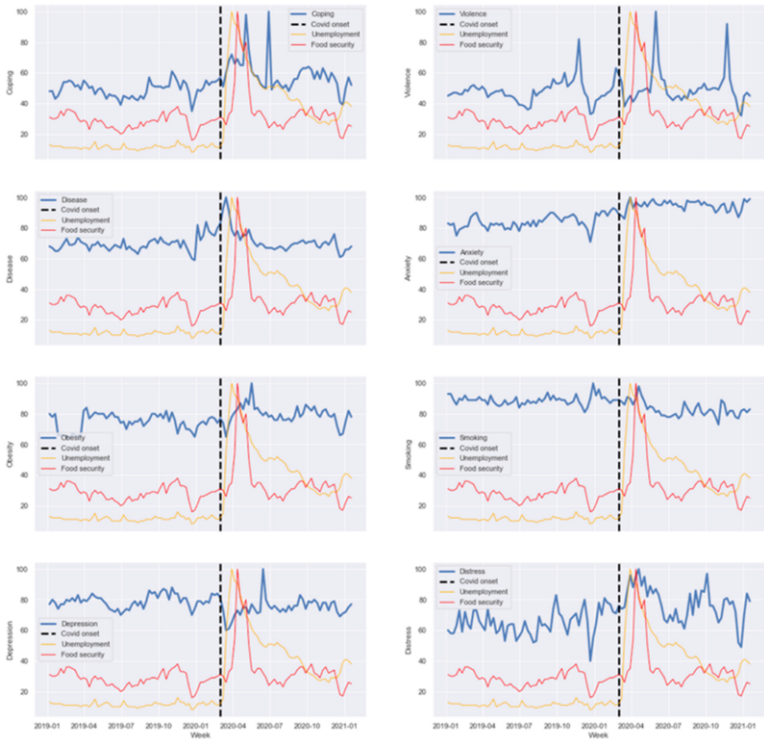

Figure 2. Online search trend results for key concepts (in blue) related to *Food Security* (red) and *Unemployment* (yellow). Each chart shows search interest (y-axis) where 100 is the most interest observed between January 2019 and January 2021 (x-axis). A vertical dashed line denotes March 2020.

The association between unemployment, anxiety and depression has already been established in the literature, however, new research has emerged since the start of the pandemic and one paper concluded that young adults in the U.S. experience significant symptoms of anxiety as a result of job insecurity amidst the COVID-19 pandemic [10]. Another 2020 study has also shown an association between household food insecurity and anxiety and mental health symptoms among Canadians in the early months of the COVID-19 pandemic [11]. These suggest that the information mined from PubMed prior

to the start of the pandemic might be used, for example, to help identify research areas of interest or potential future public health concerns. The proposed approach can also be used to retrieve additional concepts or semantic types, beyond the top-5 for four semantic types shown in this analysis, and these could help surface less obvious or more complex insights to be studied. Limitations of the proposed approach include outliers and noise in the graph and other sources of bias in the population trends. For example, *Violence* (Figure 2) peaks twice after the start of the pandemic and a close inspection reveals that this may be driven by racial and other incidents in the U.S. Further work is needed to account for other sources of bias, and analyses can be more geographically selective using the proposed approach. The proposed approach could also be used with databases other than PubMed as long as an annotator is used to identify key concepts from available abstracts or full texts.

4. Conclusions

There is a wealth of published evidence about the impacts of specific social determinants on health outcomes, however, this evidence may have not been explored to provide new insights for population trend analyses. This paper presents a first analysis of a new approach to combine insights mined from PubMed with population trends. This approach has shown that insights from previous published evidence may be used to identify areas of research or to suggest public health studies or interventions. Further work is needed to interpret other sources of bias that may influence the results, to expand the search to insights beyond top-5 and across different semantic types, or to better mine relationships from the published literature. This approach can also lead to further work on automatic surveillance systems to monitor trends that could impact public health.

References

[1] Singu S, Acharya A, Challagundla K, Byrareddy SN. Impact of Social Determinants of Health on the Emerging COVID-19 Pandemic in the United States. Front Public Health. 2020 Jul 21;8:406.
[2] Cooksey-Stowers K, Schwartz MB, Brownell KD. Food swamps predict obesity rates better than food deserts in the United States. Int J Environ Res Public Health. 2017;14:1366. 10.3390/ijerph14111366.
[3] Petrilli CM, Jones S, Yang J, Rajagopalan H, et al. Factors associated with hospitalization and critical illness among 4,103 patients with COVID-19 disease in New York City. medRxiv. 2020;1–25.
[4] Mavragani A, Gkillas K. COVID-19 predictability in the United States using Google Trends time series. Nature Sci Rep. 2020 Nov 26;10:20693. https://doi.org/10.1038/s41598-020-77275-9.
[5] Bettencourt-Silva JH, Mulligan N, Jochim C, Yadav N, Sedlazek W, Lopez V, Gleize M. Exploring the Social Drivers of Health During a Pandemic: Leveraging Knowledge Graphs and Population Trends in COVID-19. Stud Health Technol Inform. 2020 Nov 23;275:6-11. PMID: 33227730.
[6] Xu J, Kim S, Song M, et al. Building a PubMed knowledge graph. Sci Data. 20207;205.
[7] Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. Comput Struct Biotechnol J. 2020 Jun 2;18:1414-1428.
[8] 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.
[9] Wilkinson RG, Marmot M. Social Determinants of Health: The Solid Facts. World Health Organization; 2003.
[10] Ganson KT, Tsai AC, Weiser SD, Benabou SE, Nagata JM. Job Insecurity and Symptoms of Anxiety and Depression Among U.S. Young Adults During COVID-19. J Adolesc Health. 2021 Jan;68(1):53-56.
[11] Polsky JY, Gilmour H. Food insecurity and mental health during the COVID-19 pandemic. Health Rep. 2020 Dec 16;31(12):3-11. doi: 10.25318/82-003-x202001200001-eng. PMID: 33325672.