CORD-19 visualization using dynamic evidence gap maps

Aravind Mohanoor, Founder of Mining Business Data

Background

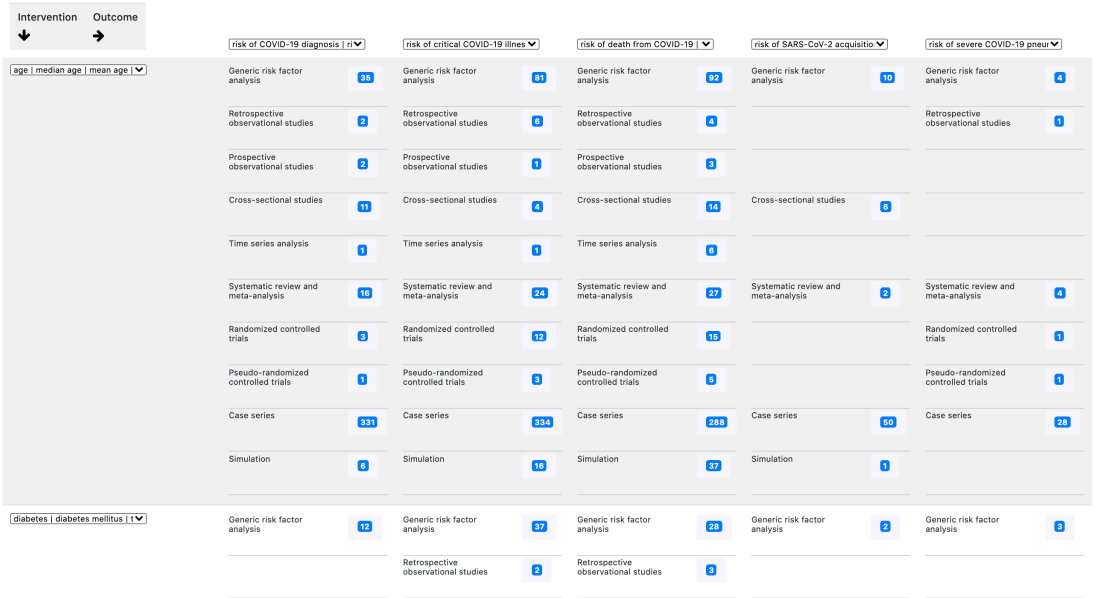
The CORD-19 dataset [1] was created to find ways to use Natural Language Processing (NLP) techniques "to develop text and data mining tools that can help the medical community develop answers to high priority scientific questions" about COVID-19. One of these questions is to find how certain factors such as age and diabetic condition (also called "risk factors") [2] affect events such as COVID-19 acquisition and COVID-19 diagnosis (also called "outcomes").

Research papers discussing risk factors can be categorized into different study types based on the "level of evidence" they provide to support their conclusions. For example, some papers report the results of stringent randomized controlled clinical trials with a large study population (high level of evidence), some report results of a single patient (low level of evidence) etc. Various checklists and guidelines have been published to establish standards for the different study types [3, 4, 5, 6, 7].

Evidence gap maps

Evidence gap maps [8, 9] provide a quick way to summarize and visualize existing research on a specific topic by plotting a matrix where the rows are the risk factors (also called "interventions") being studied (e.g. age, diabetes) and the columns are the outcomes being measured (e.g. risk of COVID-19 acquisition or diagnosis). In addition, each cell in the 2D matrix is further subdivided into the different study types to indicate the level of evidence supporting various risk factors which have been analyzed. The gaps in this matrix point to gaps in evidence, which in turn indicates that a particular risk factor would require further analysis (in the form of more clinical trials etc). While the original evidence gap maps used static values for risk factors and outcomes, the pandemic nature of COVID-19 is forcing epidemiologists to study a large number of risk factors and outcomes to better understand the different factors at play. So I proposed creating a *dynamic* evidence gap map which allows the researcher to dynamically select from a large list of risk factors and outcomes, and automatically update the evidence gap map based on the selected values. In fact, a medical dictionary containing a list of risk factors, outcomes and study types was created specifically for the CORD-19 challenge [10].

I have built a dynamic evidence gap map [11] using the faceted search service Algolia. The NLP library spaCy [12] was used for identifying risk factors, outcomes and study types using rules-based approaches. Then the risk factors, outcomes and study types were added as facets during search index creation. The figure below shows a screenshot of the dynamic evidence gap map. As future work, I plan to incorporate ML techniques to improve the identification of risk factors, outcomes and study types.



Acknowledgments

I would like to thank Savanna Reid (epidemiologist and Kaggle challenge co-participant) for the idea for creating evidence gap maps and Algolia for providing free access to their paid search service for COVID-19 related apps [13].

[1] Competition homepage: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/
[2] Risk factors task: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks?taskId=558
[3] Combined STROBE checklist for observational studies: https://www.equator-network.org/wp-content/uploads/2015/10/STROBE_checklist_v4_combined.pdf, STROBE statement and checklist for cohort studies: https://www.strobe-statement.org/fileadmin/Strobe/uploads/checklists/STROBE_checklist_v4_cohort.pdf and good practices in retrospective chart reviews: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3853868/
[4] PRISMA checklist: http://www.prisma-statement.org/documents/PRISMA%202009%20checklist.pdf
[5] CONSORT checklist: http://www.consort-statement.org/consort-2010
[6] CARE checklist: https://static1.squarespace.com/static/5db7b349364ff063a6c58ab8/t/5db7bf175f869e5812fd4293/1572323098501/CARE-checklist-English-2013.pdf
[7] STRESS guidelines: https://www.equator-network.org/reporting-guidelines/strengthening-the-reporting-of-empirical-simulation-studies-introducing-the-stress-guidelines/
[8] Snilstveit, B., Vojtkova, M., Bhavsar, A., & Gaarder, M. (2013). Evidence gap maps–a tool for promoting evidence-informed policy and prioritizing future research.
[9] See an epidemiologist's view on why evidence gap maps are useful for analyzing the CORD-19 dataset: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/discussion/137645
[10] Link to medical dictionary: https://docs.google.com/spreadsheets/d/1t2e3CHGxHJBiFgHeW0dfwtvCG4x0CDCzcTFX7yz9Z2E/
[11] https://aravindmohanoor.github.io/cord19search/dynamic-egm.html
[12] https://spacy.io/
[13] https://blog.algolia.com/supporting-our-communities-during-this-time-of-need/