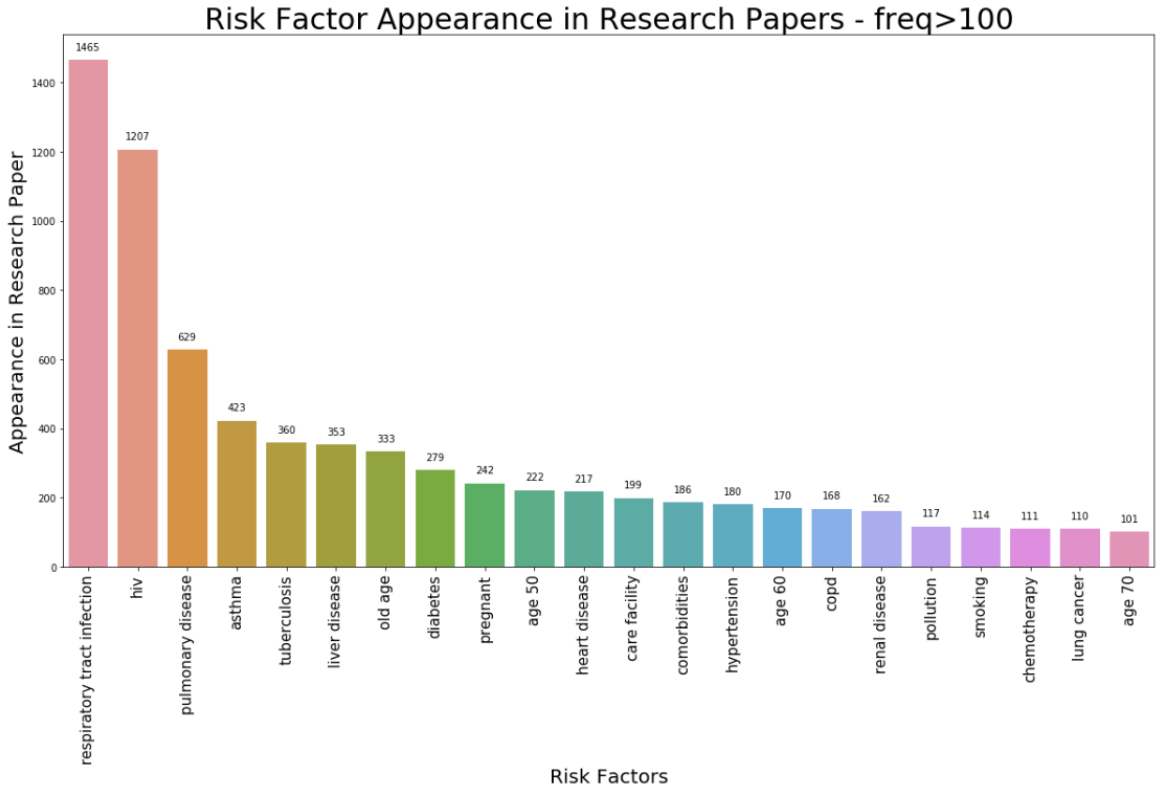
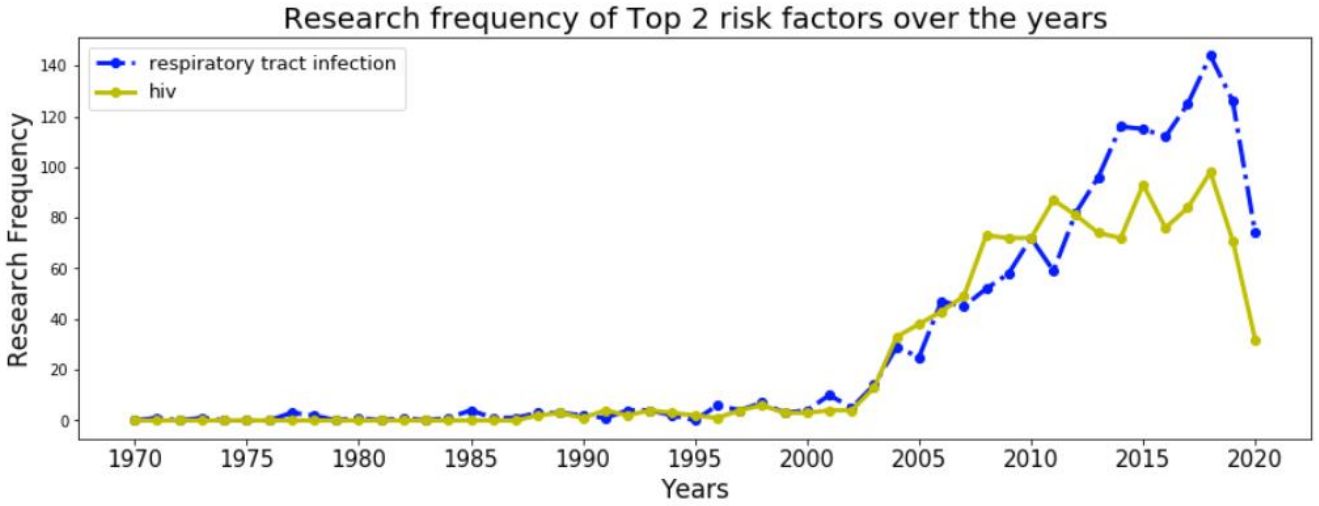
Coronavirus Insights on Risk Factors, Transmission & Affected Organs

By – Gautam Dawar, ID:

Section 1. Data Exploration



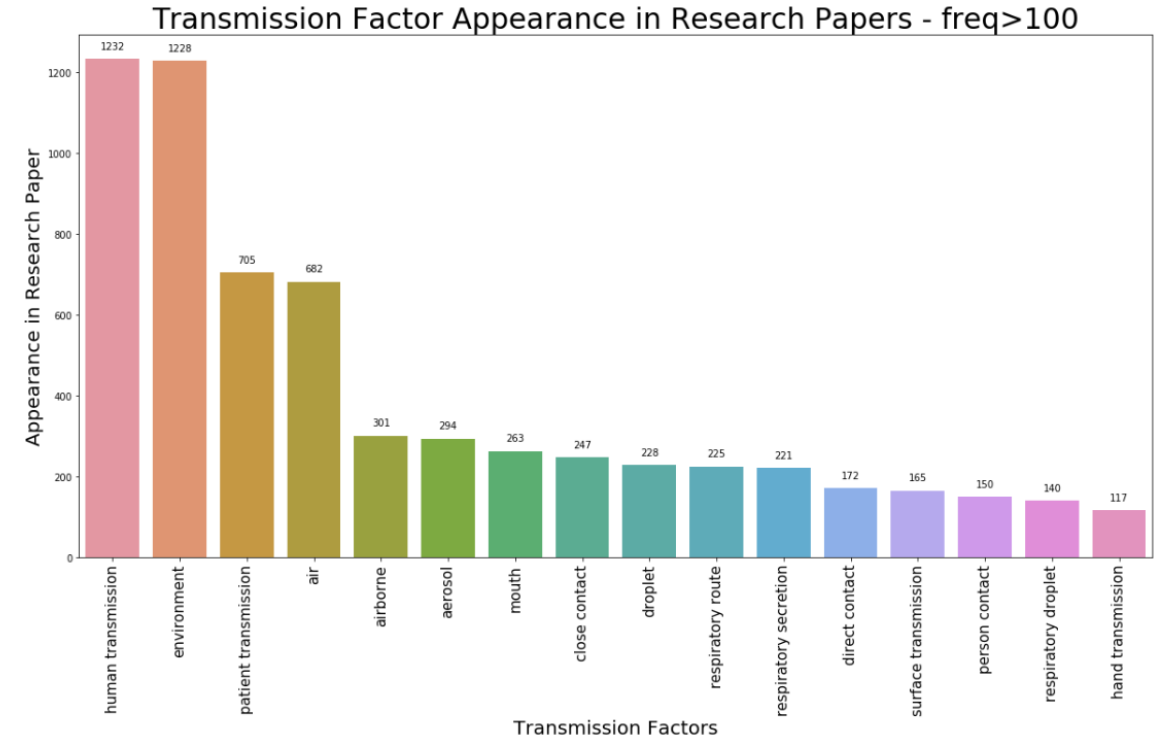


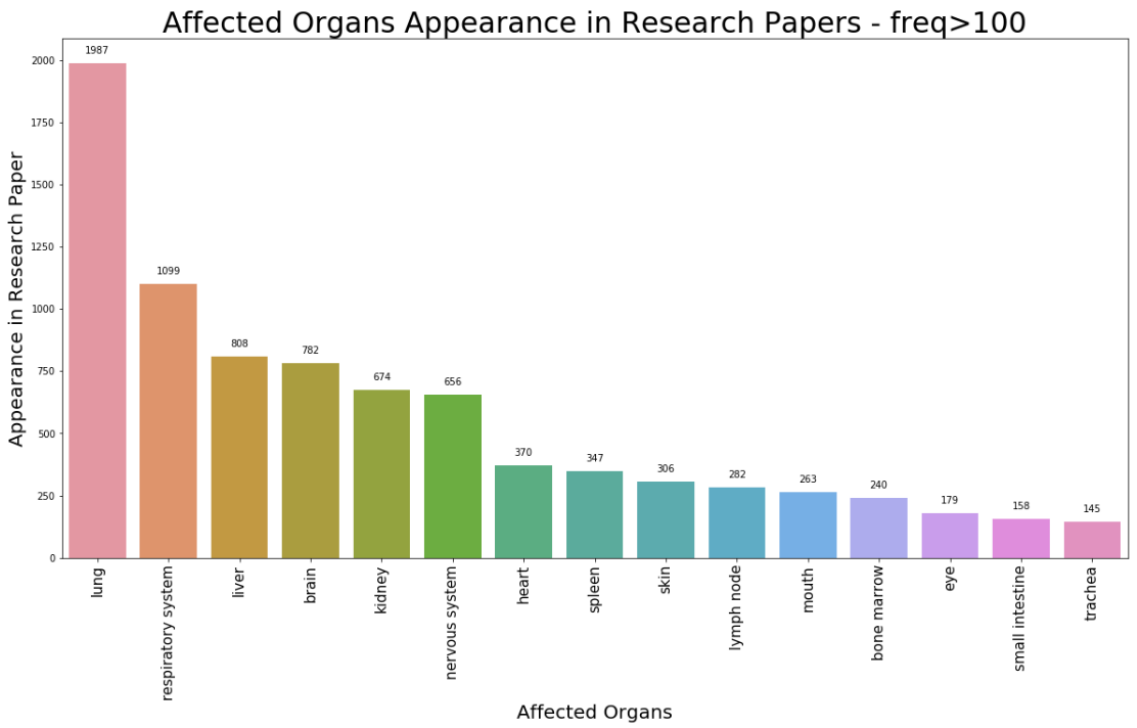
↑ Research frequency of Top 2 **Risk Factors** –

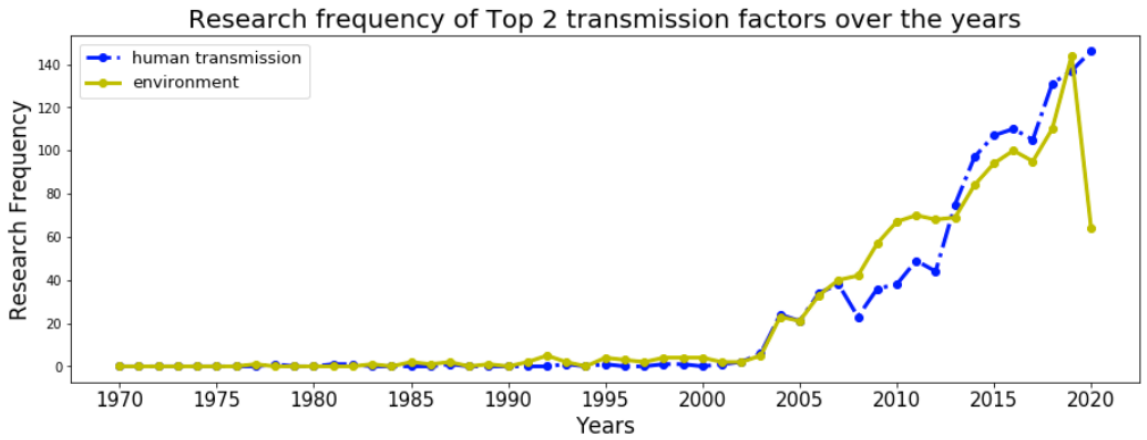
Also, in Risk Factor Time Series Graph, it is evident that the research of the risk factors of **HIV** and **Respiratory Track Infection** started growing after 2003. The reason being, **2003 was hit by one of a kind SARS pandemic** which is similar to the current coronavirus pandemic in terms of symptoms. Hence the research post the 2003 pandemic increased in **weak immunity and breathing based risk factors** such as HIV and Respiratory Track Infection.

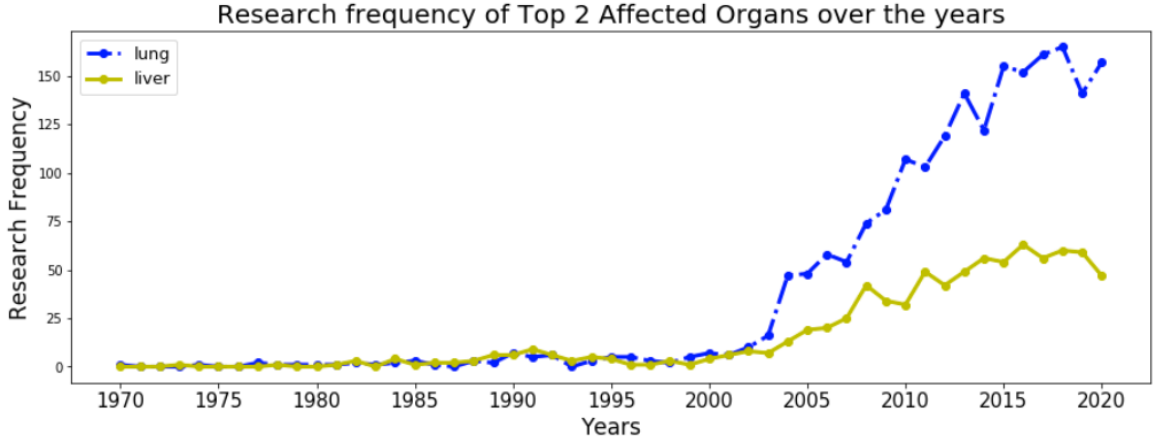
↑ **Risk Factor** Appearance in Research Paper –

Preexisting condition of **respiratory track infection, HIV, Pulmonary Disease and Asthma** are the **major risk factors** discussed in the research papers related to coronaviruses. Hence people with these pre-existing conditions should be extra careful and under care during the corona virus pandemic. Also, people with age 50 appear to be at most risk (age wise) during the pandemic. This figure corresponds to the data available on CDC and WHO website









↑ **Transmission** Appearance in Research Paper–

Human Transmission, **Patient Transmission and air** appear to be most potent to spread coronaviruses. Hence social distancing during the pandemic would be an effective option to curb the spread of coronaviruses through such means.

↑ **Affected Organs** Appearance in Research Paper–

Lungs, liver and kidney are affected the most. Hence care must be taken during organ transplant, smoking and consumption of fried food and alcohol should be avoided during this pandemic to avoid hampering the health of these organs

Section 2. Model Feature Importance

Risk Factors Model Feature was created via 3 methods and quality of hierarchical clustering was evaluated for each

	respiratory tract infection	hiv	pulmonary disease	asthma
0	0.000000	0.0	0.000000	0.0
1	0.000000	0.0	0.042553	0.0
2	0.000000	0.0	0.000000	0.0
3	0.000000	0.0	0.000000	0.0
4	0.006329	0.0	0.018987	0.0

	respiratory tract infection	hiv	pulmonary disease	asthma
0	0.000000	0.0	0.000000	0.0
1	0.000000	0.0	1.113088	0.0
2	0.000000	0.0	0.000000	0.0
3	0.000000	0.0	0.000000	0.0
4	0.292451	0.0	0.927577	0.0

	respiratory tract infection	hiv	pulmonary disease	asthma
0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0

Modified Bag of Ngrams using REGEX (self developed) – **(best performer with Hierarchical Clustering)**

A novel method was developed to search for the labels by splitting the label into tokens and taking the intersection of the outputs of each token search. This would ensure that all the tokens of the label appeared in a abstract selected (irrespective of the order of tokens)

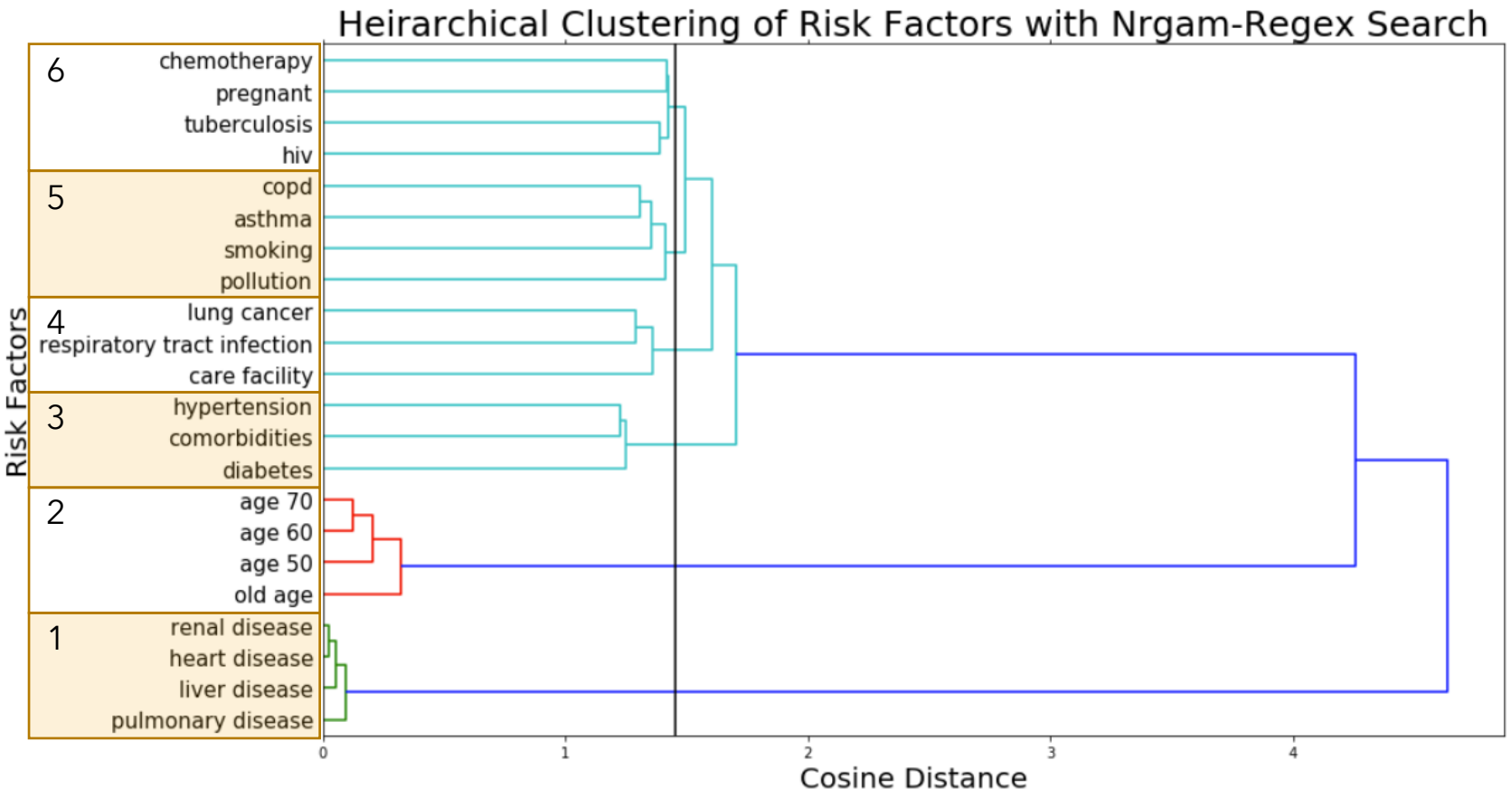
Okapi BM25–

In information retrieval, Okapi BM25 (BM stands for Best Matching) is a ranking function used by search engines to rank matching documents according to their relevance to a given search query. It is based on the probabilistic retrieval framework developed in the 1970s and 1980s by Stephen E. Robertson, Karen Spärck Jones, and others.

Inverse Document Frequency Data Preparation (TF-IDF) –

Tf-idf stands for term frequency-inverse document frequency, and the tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

Section 3. Results and Visuals



Hierarchical Clustering of **Risk Factors** & Discussion

Cluster 1 is relating to risk factor associated with preexisting conditions in organs like liver, heart and renal (kidney)

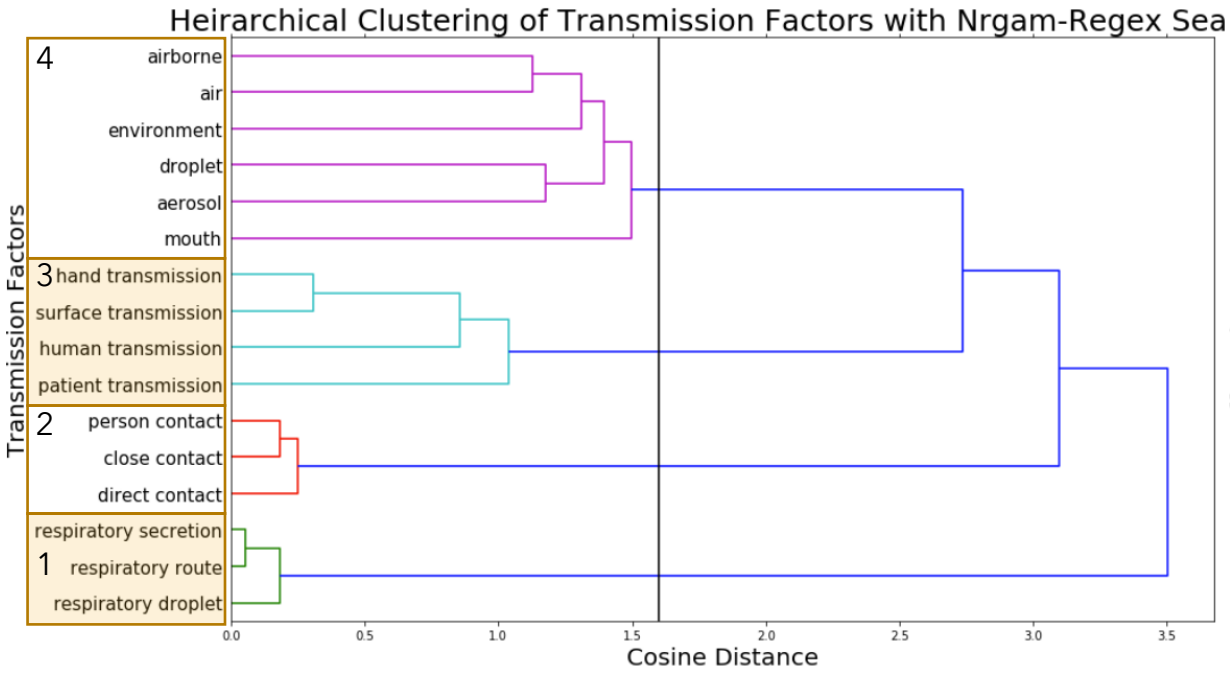
Cluster 2 is related to the risk factors associated with the age-related immunity loss. Also, the graph of Risk Factor Appearance in Research Paper suggests that people in their 50's/old age are at highest risk in the COVID-19 pandemic.

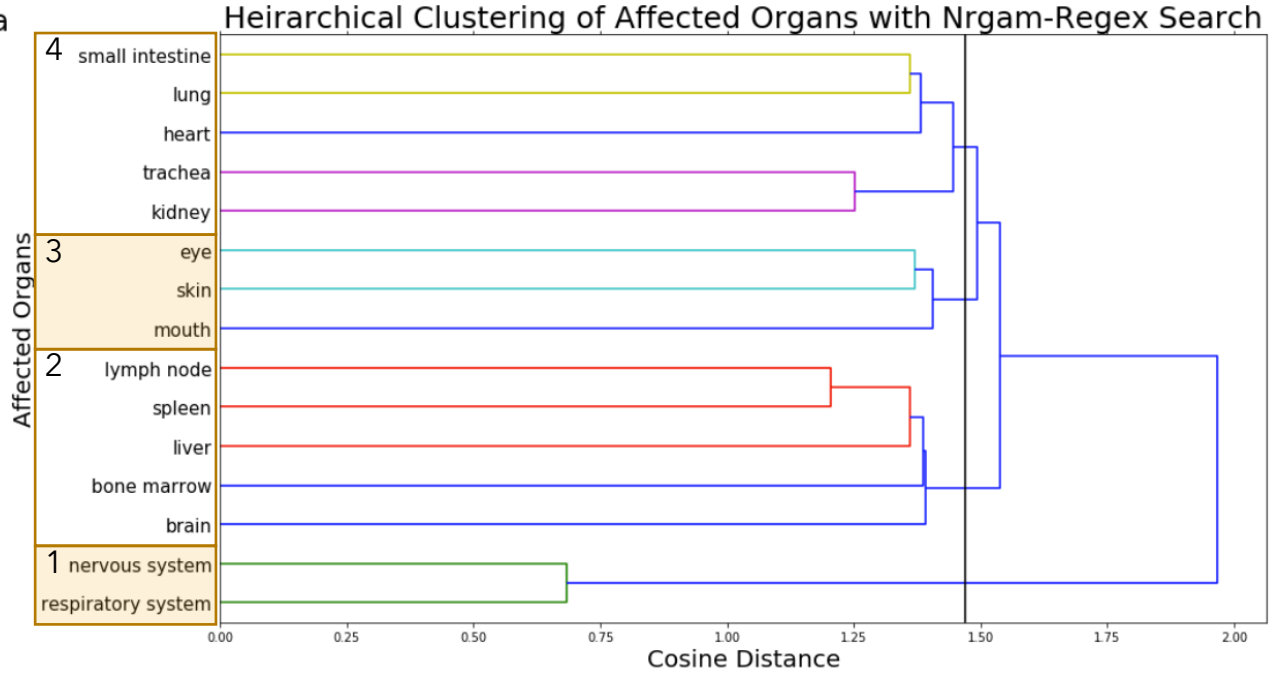
Cluster 3 suggests that risk factors related to obesity such as diabetes, comorbidities and hypertension. The combined impact of diabetes and high blood pressure can increase the risk of cardiovascular disease, kidney disease, and other health problems in the COVID-19 Pandemic.

Cluster 4 suggests that risk factors related to breathing complications arising pre-existing conditions like lung cancer or RTI. Moreover working in a care facility for corona virus also puts one at risk of contracting the virus.

Cluster 5 suggests that risk factors related to pre-existing conditions of air pollution like pollution, smoking, asthma and COPD. Living in a polluted area or smoking will put one at higher risk during COVID pandemic.

Cluster 6 suggests that risk factors related to pre-existing conditions of weak immunity due to HIV Aids, Tuberculosis, Pregnancy and Chemotherapy.





Hierarchical Clustering of **Transmission Factors** & Discussion

Cluster 1 is relating to Transmission factor associated with Air to Human (if in vicinity of an infected person) through inhaling the respiratory droplets of an infected person.

Cluster 2 is related to the Transmission factors associated with Surface to Human. Factors like direct contact with a surface having active virus might trigger the spread.

Cluster 3 suggests that Transmission factors related to Human to Human contact. Handshakes or direct contact with a carrier of coronavirus will trigger the spread.

Cluster 4 suggests that Transmission factors related to Air to Human (airborne) transmission. This cluster suggest the spread of the disease due to the inhaling the infected aerosol in the environment.

Hierarchical Clustering of **Affected Organs** & Discussion

Cluster 1 is relating to damage to the respiratory system.

Cluster 2 is related to the damage associated with liver and linked organs. The bone marrow is the source of mature liver cells and spleen is affected due to the operation of the liver.

Cluster 3 is related to the damage to the human body due to transmission with eye or mouth contact.

Cluster 4 is related to the damage associated with lung and linked organs. Bad lung can hamper the functioning of kidneys and make heart erratic.

NOTE: All the visuals are discussed in detail along with policy suggestions in the python Notebook