A Machine Learning Based Modeling of the Cytokine Storm as it Relates to COVID-19 Using a Virtual Clinical Semantic Network (vCSN)

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9378284

Abrar Rahman[1], John Kriak[2], Rick Meyer[3], Sidney Goldblatt[3], Fuad Rahman[3]

[1] The University of California, Berkeley, California, USA
[2] MolecularDx, Windber, Pennsylvania, USA
[3] Goldblatt Systems, Tucson, Arizona, USA

***Abstract*—This paper presents a targeted, machine learning based solution to model the phenomenon known as the 'cytokine storm,' which is suspected to play a major role in explaining the highly variable severity of COVID-19 among patients. It describes how a Natural Language Processing (NLP) approach, augmented by biomedical knowledge databases, can extract pre-existing conditions and relevant clinical markers from Electronic Health Records (EHRs). These extracted variables can be modeled to demonstrate correlation with the severity of infection outcomes, the building blocks of a comprehensive risk assessment and stratification strategy to predict which patients have higher or lower risks in terms of the disease severity and likelihood of hospitalization, exclusively from insights taken from the natural language data. The model has been applied to a cohort of patients from a large database of real, anonymized patients and has displayed demonstrable results.**

Keywords—Big data, Natural Language Processing, Modeling, Clustering, Healthcare, COVID-19, Coronavirus

I. Introduction

A. The Cytokine Storm Phenomenon

An unresolved question of the novel coronavirus is why so many patients are asymptomatic or have mild symptoms, while for others the disease intensifies drastically. COVID-19 seems to do much of its damage by triggering an overzealous immune response, as opposed to direct damage from the virus itself [1]. For example, severe pneumonia is often associated with rapid virus replication, massive inflammatory cell damage, and elevated proinflammatory responses culminating in acute respiratory distress syndrome [2]. This condition is called the "cytokine storm," named after the elevated levels of immune system proteins, called cytokines, in the blood of COVID-19's sickest patients. It is also called Cytokine Release Syndrome (CRS). In CRS, an overstimulated immune system results in the body starting to attack its own cells and tissues, rather than just fighting off the virus [3]. Research into CRS from COVID-19, including who it affects and why, is absolutely critical in order for clinicians, administrators, and policymakers to make informed decisions to handle the COVID-19 crisis.

B. The Role of Big Data in Medical Research

The World Economic Forum anticipates that 463 exabytes of data will be created each day by the year 2025 [4]. The current volume of all electronic data doubles every two years [5]. This skyrocketing volume of data is disproportionately true in healthcare, at least for the next five years. According to an International Data Corporation report, healthcare is expected to be the highest data growth business sector, with a compound annual growth rate of 36% through 2025 [6]. Working with big data is increasingly critical for organizations involved in healthcare, medicine, and pharmaceuticals.

Especially for a public health event of this scale, there is a wealth of data available on patients. However, much of the data is in free format unstructured text and are therefore mostly opaque and uninterpretable [6]. In clinical practice, a plethora of this data exists within Electronic Health Records, or EHRs. This paper attempts to demystify this data by developing a methodology using Natural Language Processing techniques to examine COVID-19 and the cytokine storm.

II. Related Work

A previously published effort presents an approach to model COVID-19, utilizing patient data from related diseases, combining clinical understanding with artificial intelligence modeling [7]. This paper takes a similar approach to the data, but ultimately only lays out a methodology in lieu of actual completed experiments and results. In addition, it focuses on the disease as a whole, as opposed to taking an in-depth look at high-severity cases, including CRS [7]. This prior work establishes a valuable data science methodology, but is nonetheless lacking implementation and analysis, having an entirely different research focus from this current paper.

Another work seeks to forecast the onset and outcomes of a "second wave" of COVID-19, utilizing patient data from related diseases, marrying clinical data with artificial intelligence [8]. This paper used influenza as a proxy for modeling COVID-19, instead of directly engaging with clinical research regarding COVID-19 [8]. Thus, their modeling and conclusions cannot be considered specific to this disease, and this current paper goes far beyond their system's capabilities as a result.

III. Methodology

A. A Big Data Analytics Approach

This section presents a computational approach to analyzing COVID-19 related CRS based on a meta-analysis of existing medical research. The vision here lies in synthesizing big data analytics and machine learning with a virtual Clinical Semantic Network (vCSN). Big data and machine learning do best with analyzing massive volumes of data. On the other hand, having a knowledge network can grant real insight into how various medical topics are related.

The general outline for this approach is straightforward. First, generate a list of biomarkers and other clinical terms that help describe CRS and related phenomena. Afterwards, acquire a database of real patients. Then, build a cohort around the variables to isolate patients of interest via NLP algorithms. From the cohort, as a pre-processing phase, use vCSN to extract clinically-significant information from the raw text. Finally, apply statistical methods and machine learning to model the phenomenon.

B. Biomarkers for COVID-19 and the Cytokine Storm

In order to proceed with an NLP and data analytics centric approach to describing COVID-19 CRS, the first step is to come up with a list of terms related to COVID-19 and CRS.

A number of these terms are derived from a University of Chicago paper [9]. They found severe COVID-19 was associated with impaired T cell responses, showing that key immune cells were being underused. This was backed up by low levels of certain expected interferons, which are specific signaling proteins in the immune system [9]. Additionally, some patients in their sample suffered acute respiratory distress syndrome (ARDS), reflecting overzealous cytokine production and excessive inflammation, the hallmarks of CRS [9].

Specific medical observations for the cytokine storm in COVID-19 were published by a University of Paris team [1], helping the software specifically distinguish high-severity cases from more normal immune responses. They note striking downregulation of interferon-stimulated genes in critical patients compared to mild-to-moderate patients, showing that patients who fared worse may have had weaker immune responses in the earliest stages of the disease [1].

C. Biological Aging & Lifestyle Factors

The most vulnerable groups for severe COVID-19 outcomes are often the elderly, so some variables specifically focusing on the disease in this patient cohort were sought out. COVID-19 infection shows increased levels of plasma proinflammatory mediators, including IL1-β, IL1RA, and IL8 [10]. MCP1, MIP1α, and TNFα were isolated as being among mediators marking disease severity [10].

In order to get a more nuanced look into the role of aging in COVID-19, additional variables are still needed. After all, chronological aging and "biological aging" are not quite the same thing: as an extreme example, a 50-year-old with sustained healthy diet and exercise habits may well be better protected from infection than a 30-year-old morbidly obese smoker [11]. Smoking is a particular area of focus, because, of course, COVID-19 is primarily a respiratory disease, and as so, smoking-related terms were collected [12]. Markers of biological aging were found in yet another paper. These terms included obesity, diabetes, cholesterol, triglycerides, creatine, etc. [11], as well as the formula in Fig 1.

$$\begin{aligned} \mathrm{BAS} = 0.02\mathrm{SBP} - 2.189\mathrm{FEV}_1/\mathrm{Ht}^2 - 0.104\mathrm{HCT} \\ - 1.541\mathrm{ALBU} + 0.077\mathrm{BUN} + 9.19 \end{aligned}$$

Fig. 1. Nakamura's biological aging index formula. SBP = systolic blood pressure, FEV1 = forced expiratory volume, Ht = height, HCT = hematocrit, ALBU = albumin, BUN = blood urea nitrogen [11]

IV. Implementation

A. Training Data

The data was acquired from a health system organization serving the eastern United States. The database includes 97,500 individual patients and their full medical records. For reasons of privacy and regulation, all patient records were fully de-identified and anonymized. The dataset is all-encompassing, including extensive doctors' notes (a vital source of data for our NLP apparatus), immunization records, etc. It is important to stress that this is not a COVID-19 database, presenting additional challenges for the team to work through.

B. The Virtual Clinical Semantic Network (vCSN)

In the context of healthcare research, NLP has significant limitations worth addressing. Sample text is often ungrammatical, filled with bullet points and sentence fragments. Moreover, these corpora make heavy use of acronyms and abbreviations, which are a massive challenge for any computational approach to address. In addition, clinical notes often contain terms or phrases that have more than one meaning. For example, the abbreviation MD can be interpreted as the credential for "Doctor of Medicine" or as an abbreviation for "major depression" [13].

One such solution is the standardization of medical language such as the Unified Medical Language System (UMLS), a platform that brings together many standards and thesauruses to enable interoperability between computer systems [14]. The UMLS metathesaurus is a repository of over 100 biomedical vocabularies, including CPT®, ICD-10-CM, LOINC®, MeSH®, RxNorm, and SNOMED CT®. Terms across vocabularies are grouped together based on concepts, allowing users to account for the huge variations in language and expressions. Each concept has a specific tag and numerical code.

The Clinical Semantic Network (CSN) has established relationships among various clinical concepts, encapsulated using a tree structure [7]. The CSN tree structure allows users not only to know exactly how a concept is related to another concept, but also how closely two concepts are related.

The vCSN converts this rich and complex data structure into a virtual model. This requires us to build a machine-learning based probabilistic model of all these concepts [7]. We have applied NLP techniques to model these concepts in terms of how closely they are mapped with respect to the description text. We then used machine learning techniques. More specifically, it comprises of a convolutional neural network (CNN) with transfer learning, to learn how these concepts are mapped ("CSN

Number") and how these are related ("pathway"). The overall process is shown in Fig 2.

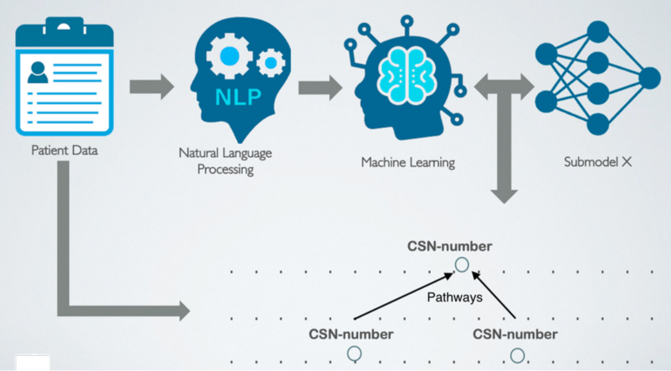

Fig. 2. The relationship between CSN and vCSN

V. Results

A. Exploratory Queries

Provided the variables collected from our extensive research on COVID-19, the cytokine storm, and biological aging factors, the next step was seeing how frequently the terms appeared in the database. For each term unearthed, the program combed through the 'Encounter' and 'Encounter Notes' sections from the patient data, returning independent hits, all via Stanford's CoreNLP tool. Thus, the output counts do not represent unique individual patients. In total, 97 terms were used.

For the sake of organization, querying was conducted in multiple batches. The first set of variables were regarding immune cells. The second set centered around inflammatory pathways. The third set of variables was associated with interferons, specific signaling proteins key to immune response. A fourth set consisted of terms related to the cytokine storm that did not fit in the first three categories. A fifth set of variables were markers for biological aging, including terms associated with smoking. Excluding the aging terms, the query emerged with the counts for results in Table I.

TABLE I. Initial Query Output and Encounters

	Encounter Metadata	Encounter Notes
Total Hits	16,415	164,801
Unique Patients	206	2,637

B. Cohort Generation & Tagging

The next step was to construct a cohort of patients to focus on. In order to be selected for the cohort, an encounter had to either include a certain number of terms, or include rarer but cytokine-specific variables in its place. The overall size of the combined cohort was 421 unique patients, comprising 1,570 unique encounters. From the "Visit_Info" subheading alone, this meant 11,666 columns, 14,135,301 characters, 2,011,758 individual words, and 14.1 MB of raw textual data.

Afterwards, this text data was processed through vCSN, generating a wealth of CSN tags. Known clinical terms were pulled out, identified, and placed into a tree of disease pathways, showing relationships between various conditions. In addition to the main CSN data silo, there were two databases available within vCSN to tag against: 'Findings' and 'Diseases', each with their own internal relationships. All data was tagged three times, marked with respect to "CSN," "Findings," and "Diseases," resulting in three slightly-different interpretations. See Table II for details.

TABLE II. Findings and Disease Tagging Results

	Findings	Diseases
Named Entities Recognized (NERs)	21,875	19,795
CSN Phrases	113,594	53,483
CSN Medical Words	112,254	72,358
CSN NERs in Medical Words	1,731	509
Phrases	177,109	177,119
Medical Words	128,159	128,165
NERs in Medical Words	8,066	8,067

C. Cohort Data Processing, Phase I

This is the post-processing phase to clean the tagging outputs. One issue was the inclusion of multiple redundant tags for a singular disease. In order to solve this, we developed a simple anti-redundancy algorithm. The vCSN representation can be reduced down to a tree structure. From there, the system does a depth-first search of the entire graph. Each node is hashed by title and CSN number. When encountering a new node, if the CSN number or titles are repeated from a prior node, the system merges the two together. The merge function preserves multiple pathways to/from the combined node, maintaining all variant CSN numbers for the same phenomena.

As an example to illustrate what these tags actually mean, two visualizations of the CSN tagging for rheumatoid arthritis are presented below, before the redundancy reduction (Fig 3) and after (Fig 4). Note that Fig 3 has five identical copies of rheumatoid arthritis, in lavender, whereas they all converge to one in Fig 4. These figures are just screenshots from the full html output, which are fully interactive in 3D.

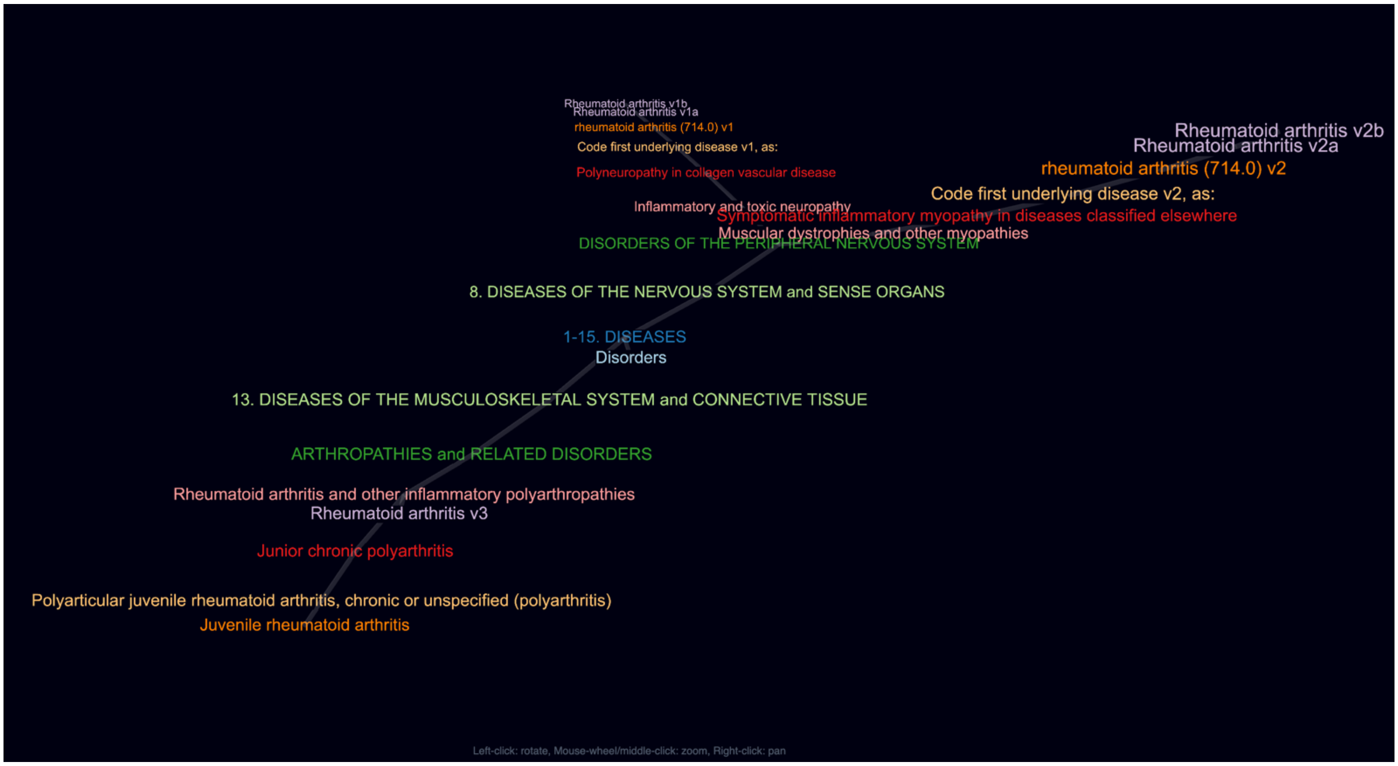
Rheumatoid arthritis v1b
Rheumatoid arthritis v1a
rheumatoid arthritis (714.0) v1
Code first underlying disease v1, as:
Polyneuropathy in collagen vascular disease
Inflammatory and toxic neuropathy
DISORDERS OF THE PERIPHERAL NERVOUS SYSTEM
Rheumatoid arthritis v2b
Rheumatoid arthritis v2a
rheumatoid arthritis (714.0) v2
Code first underlying disease v2, as:
Symptomatic inflammatory myopathy in diseases classified elsewhere
Muscular dystrophies and other myopathies
8. DISEASES OF THE NERVOUS SYSTEM and SENSE ORGANS
1-15. DISEASES
Disorders
13. DISEASES OF THE MUSCULOSKELETAL SYSTEM and CONNECTIVE TISSUE
ARTHROPATHIES and RELATED DISORDERS
Rheumatoid arthritis and other inflammatory polyarthropathies
Rheumatoid arthritis v3
Junior chronic polyarthritis
Polyarticular juvenile rheumatoid arthritis, chronic or unspecified (polyarthritis)
Juvenile rheumatoid arthritis
Left-click: rotate, Mouse-wheel/middle-click: zoom, Right-click: pan

Fig. 3. 3D graph of CSN classification of rheumatoid arthritis, raw output

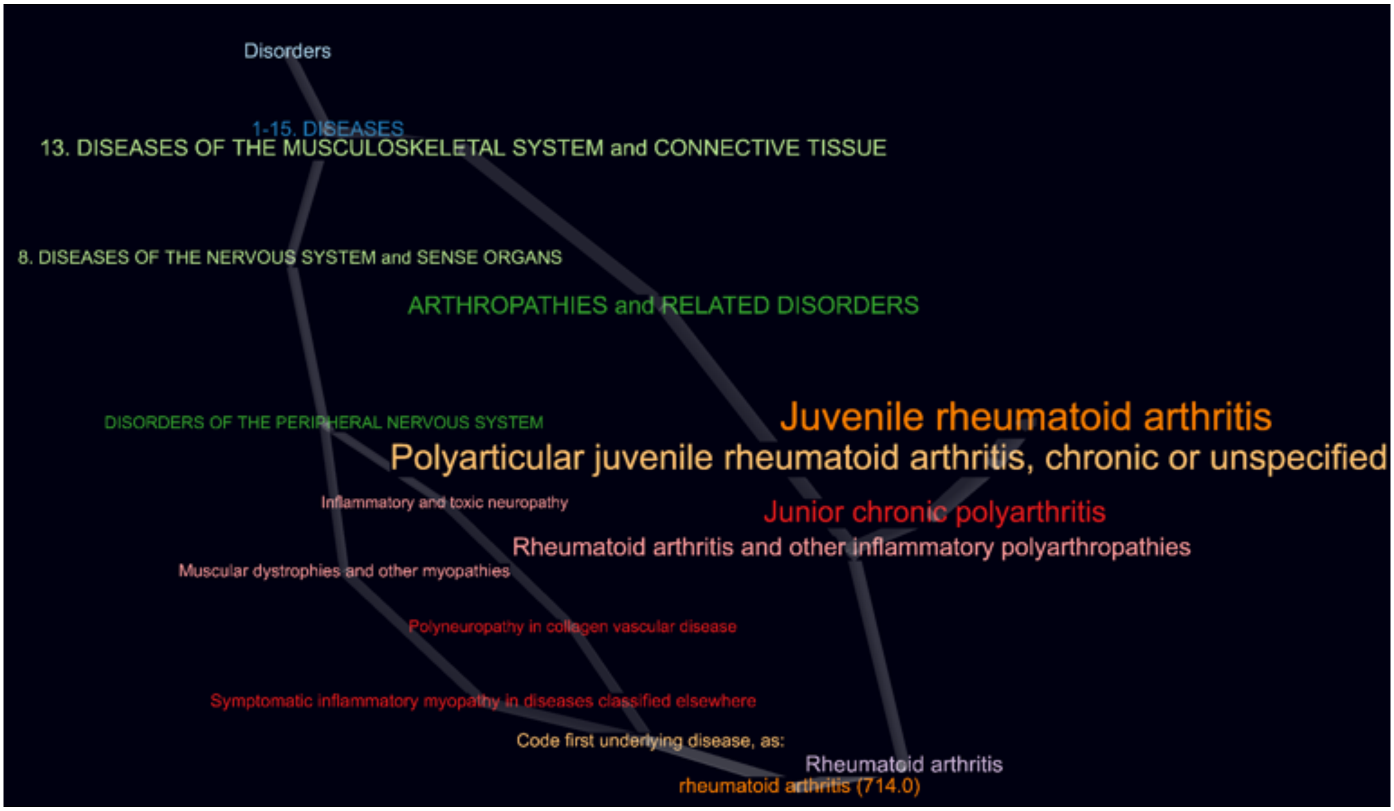
Disorders
1-15. DISEASES
13. DISEASES OF THE MUSCULOSKELETAL SYSTEM and CONNECTIVE TISSUE
8. DISEASES OF THE NERVOUS SYSTEM and SENSE ORGANS
ARTHROPATHIES and RELATED DISORDERS
DISORDERS OF THE PERIPHERAL NERVOUS SYSTEM
Juvenile rheumatoid arthritis
Polyarticular juvenile rheumatoid arthritis, chronic or unspecified
Inflammatory and toxic neuropathy
Junior chronic polyarthritis
Rheumatoid arthritis and other inflammatory polyarthropathies
Muscular dystrophies and other myopathies
Polyneuropathy in collagen vascular disease
Symptomatic inflammatory myopathy in diseases classified elsewhere
Code first underlying disease, as:
Rheumatoid arthritis
rheumatoid arthritis (714.0)

Fig. 4. 3D graph of CSN classification of rheumatoid arthritis, after running anti-redundancy algorithm

D. Cohort Data Extraction, Phases II, III, & IV

Though vCSN breaks down the natural language samples from the cohort into digestible components, the output here is still mostly in the form of English words. This is a fundamental aspect of the technology. In order to make the inputs more comprehensible for a computer, the system performed patient-by-patient iteration through the "visit_info" section, getting a count for each of the 97 terms originally used to construct the cohort per patient, using the tagging data. These tags were cleaned using the above anti-redundancy algorithm. This became the input for a machine learning algorithm to generate insights into the comparative interactions between these terms.

Two other numerical approaches to breaking down the data were also explored. In the first, a more straightforward approach, the team generated a localized count of the 97 terms straight from the "visit_info" section instead of from the CSN tags. This is a more targeted approach than looking through the myriad sections of patient data. As this data had a narrower focus, omitting entire sections from the sample space, the speed of the program improved dramatically. And finally, one can always work directly with all the tags in the dataset instead of fishing for specific terms in the tagging data. These two alternative outputs were then condensed into a single joint distribution for further analysis.

F. Machine Learning

In order to better understand and stratify within the patient cohort, the next step was to find an appropriate clustering algorithm. One of the most commonly used technique is k-means [15], which clusters the data into multiple subcategories based on a variety of factors. However, as k-means relies on vector quantization, it is fundamentally incompatible with categorical data, requiring numerical data instead. Though the numerical previous section represented a step towards full quantization of the dataset, they are ultimately still categorical.

A solution to this was to create a heuristic to convert these categorical outputs into something simpler. This was built directly from the tagging outputs. The indices constructed from the 97 original terms were condensed down based on a simplified version of the query batch structure from section 5.1. That is to say, given tags for 1) immune cells and interferons, 2) inflammation and cytokine storm, and 3) biological age and smoking, within each subcategory all hits were tallied together, and then used to create a tri-variate coordinate system for the data. All patients in the cohort were plotted on the "immune," "inflammation," and "aging" axes. This is finally valid input for the k-means clustering algorithm.

Now that the data was transformed into a format that k-means would be able to interpret, there was nothing left to do but train the algorithm. This was done with Sci-Kit Learn, a free machine learning library developed for the Python language [16]. The goal was to write a program to divide the patient cohort into multiple clusters. Unfortunately, as the size of the cohort was 421 patients and 10 clusters were found in three dimensions, it is difficult to visualize these results in a figure in any meaningful way.

A few groups seemed to be clustered together for little discernible real-world reason. However, four of the groups (sub-cohorts) had obvious overlaps when looking over both the extracted quantized data outputs and the patient record texts themselves. 285 patients appear to have exhibited severe flu-like symptoms for nonsmokers with inflammation, postulated to be the cytokine storm. 15 patients showed severe respiratory symptoms, this time with history of smoking. 107 patients were bound together by common metabolic problems, including cardiovascular issues and diabetes. Finally, 23 patients that were highly divergent from the other clusters, mostly outliers on the "aging" axis of the coordinate system, had bodily injury and/or accidents. These extraneous results likely result from the aging terms, as "body ache" was one of a few common threads among those final 23 patients.

G. Statistical Methods & Final Visualizations

Though the machine learning model produced some highly encouraging results, a drawback of the system as described is that the outputs and clusters are difficult to represent visually. In order to gain a more intuitive understanding of the patients in cohort and the biomarkers that are used to define them, alternative methods are required. Thus, statistical methods may be a fruitful area to examine.

The correlations of terms from the hybridized approach, condensing tagging outputs and "visit_info" direct searches together, are illustrated in Fig 5, a tri-surface plot correlation matrix. The most prominent feature on this graph are the two large spikes on the left-hand side, which represent, from left to right, the correlations of diabetes/blood pressure and between triglycerides/cholesterol.

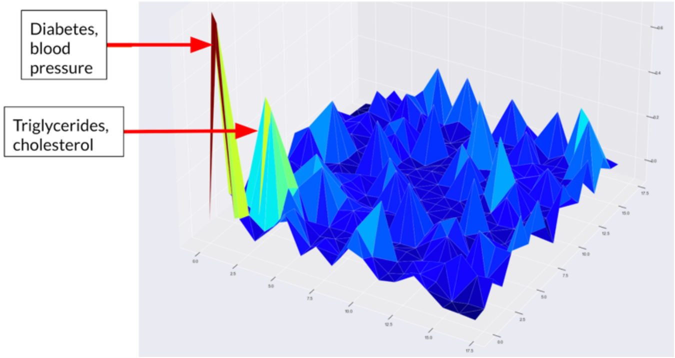

Fig. 5. Plot of correlations from CSN/Diseases/Findings tagging metadata

The combined distribution of both is also plotted in Fig 6, which is a correlation plotting heatmap. A few hotspots to note in Fig 6: fever/viral load, monocytes/viral load, loss of taste/smell, smoking/blood pressure, shortness of breath/ smoking, interferons/viral load, viral load/shortness of breath, and viral load/cholesterol. These derived relationships line up neatly with existing clinical observations. For example, the effects of smoking on blood pressure and respiration have been well-documented in medical science for decades. Similarly, it makes sense that immunology-specific terms like monocytes and interferons both lined up with viral load, as these concepts are highly interconnected. This demonstrates that the concept of data extraction, despite all these layers of abstraction, are still grounded in and correctly reflective of their real-world counterparts.

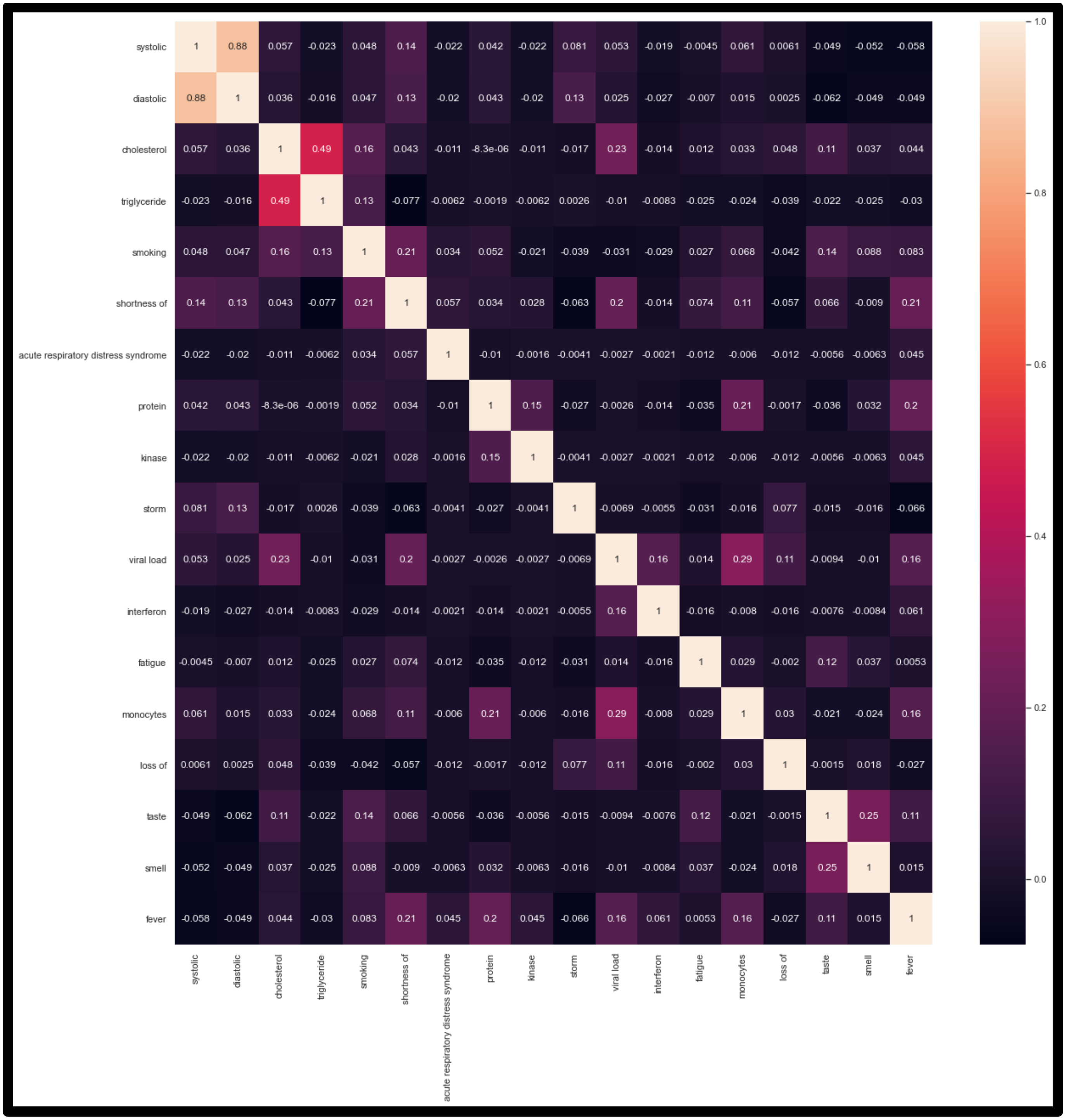

Fig. 6. Heatmap of hybridized dataset correlation

VI. Discussion

By and large, our results are highly encouraging. This paper successfully demonstrates a big data analytics approach to healthcare. More specifically, we demonstrate that by using Natural Language Processing (NLP) techniques in concert with a virtual Clinical Semantic Network, it is possible to gather extensive data about the cytokine storm and COVID-19 in general exclusively from the raw text data extracted from patient health records.

This project began from scouring medical journals, coming up with a list of 97 clinically-significant terms relating to COVID-19, the cytokine storm, and factors of biological aging. Using these terms, computational means were used to successfully isolate a cohort of 421 unique patients from among 97,500 overall in the EHR database. This output was put through the virtual Clinical Semantic Network, the Disease silo, and the Findings silo, resulting in three different sets of clinical knowledge mark-ups. Some of these were converted into numerical outputs, being used to construct a rudimentary clustering schema via k-means. The rest were used to create correlation maps, showing that related terms did in fact appear together with stunning regularity. This indicates that this method of data-gathering from unstructured sources, particularly for the cytokine storm, is a viable avenue for the clinical research community to pursue.

During the course of this project, a common difficulty was in interpreting the meaning of a vCSN tagging. For this reason, the 3D force-directed graphs were developed, to visualize these complex branching relationships.

Both the Findings and Disease silo results have very sparse counts for any tags that line up nicely with those 97 terms. Note that out of 97 query terms, only 18 even overlapped at all (Fig 6). The majority of the cytokine storm-specific terminology were never found to coincide with another clinical term, leaving a subset behind which is more general. That is why in Fig 5, the largest spikes are for non-COVID related terms that will always be highly interconnected, such as triglycerides and cholesterol. This is expected behavior, since none of these patients actually were COVID-19 patients. No clinical semantic network should even consider including specific items like gene expressions, hundreds of different signaling proteins, etc. There was a real mismatch between the tagging approach and the querying approach, one which worked off each other's strengths complementarity in some regards, but for more sparse fields, made it more difficult to glean useful insights from the data.

A related concern is that the most frequently-measured terms are normally tied to more common standard measurements in clinical examinations (for example, systolic blood pressure, temperature, fever are all part of the essential and routinely recorded "patient vitals"). Again, this is a fully expected behavior. In order to counteract this, an additional post-processing step before constructing the machine learning model may be of benefit, forward selection [17]. This will allow the computer to create a hierarchy of the relative significance of each term, simplifying the model generation.

VII. Future Work

The work in this paper outlines a methodology for conducting an NLP- and big data-based analysis of unstructured text from Electronic Medical Records to assess the prognosis of a given patient for COVID-19, specifically analyzing severe cases and the cytokine storm that causes them. The limitation of lacking a database with specific COVID-19 patients does not undermine the usefulness of developing this procedure.

In fact, all that is needed to develop a better model is a more current database. The team is currently in the process of gaining access to a new dataset, populated extensively with COVID-19 patients. Future efforts will build off of the currently established data-processing pipeline. Within the next few months, as the data and human capital becomes available, the team will attempt to reuse this methodology with the new dataset, hopefully resulting in a model with greater predictive power, as opposed to mere descriptive abilities.

VIII. Conclusion

This paper presents a machine learning based modeling solution to analyze the relationships of specific pre-existing conditions and clinical data markers in understanding the phenomenon of cytokine storm as it relates to COVID-19 patients. During the construction of this model, the research team still did not have access to COVID-19 patient data, so we applied the solution to a carefully chosen cohort of patients from an existing dataset of patients. The initial results are extremely promising and demonstrate that it is possible to generate correlation of clinical markers and pre-existing conditions to a risk stratification of severity of COVID-19 patients, especially for the severest patients experiencing the cytokine storm. Currently we are in the process of being given access to real COVID-19 patients through a research study on patients from four nursing homes. As a next step on this research, the team looks forward to applying these models on these patients' data and reporting detailed results soon.

Acknowledgment

We thank Molecular DX for funding and guiding the research every step of the way. Additional thanks goes to Goldblatt Systems for providing additional technical support.

References

[1] Hadjadj, Jerome, et al. "Impaired Type I Interferon Activity and Exacerbated Inflammatory Responses in Severe Covid-19 Patients." MedRxiv, Cold Spring Harbor Laboratory Press, 1 Jan. 2020, www.medrxiv.org/content/10.1101/2020.04.19.20068015v1

[2] Channappanavar, Rudragouda, and Stanley Perlman. "Pathogenic Human Coronavirus Infections: Causes and Consequences of Cytokine Storm and Immunopathology." *Seminars in Immunopathology*, U.S. National Library of Medicine, 10 Apr. 2017, pubmed.ncbi.nlm.nih.gov/28466096/.

[3] Tisoncik, Jennifer R, et al. "Into the Eye of the Cytokine Storm." Microbiology and Molecular Biology Reviews : MMBR, American Society for Microbiology, Mar. 2012, www.ncbi.nlm.nih.gov/pmc/articles/PMC3294426/.

[4] Desjardins, Jeff. "How Much Data Is Generated Each Day?" World Economic Forum, 17 Apr. 2019, www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/.

[5] "The Exponential Growth of Data." Inside Big Data, 17 Apr. 2018, insidebigdata.com/2017/02/16/the-exponential-growth-of-data/.

[6] Farrier, Helen. "Seagate Launches New Data-Readiness Index Revealing Impact Across Four Global Industries as 30 Percent of Data Forecasted to Be Real-Time by 2025." Business Wire, Berkshire Hathaway, 27 Nov. 2018, www.businesswire.com/news/home/20181126005585/en/Seagate-Launches-New-Data-Readiness-Index-Revealing-Impact.

[7] Rahman, Ahmad, et al. "Big Data Analytics + Virtual Clinical Semantic Network: ASAIO Journal." *ASAIO Journal*, 12 Aug. 2020, journals.lww.com/asaiojournal/Abstract/9000/Big_Data_Analytics___Virtual_Clinical_Semantic.98460.aspx.

[8] Rahman, Ahmad, et al. "AI Based Health Signals Discovery Engine." *SNOMED CT Expo*, 2019, confluence.ihtsdotools.org/display/FT/201955+AI+based+health+signals+discovery+engine.

[9] Acharya, Dhiraj, et al. "Dysregulation of Type I Interferon Responses in COVID-19." Nature News, Nature Publishing Group, 26 May 2020, www.nature.com/articles/s41577-020-0346-x.

[10] Nikolich-Zugich, Janko, et al. "SARS-CoV-2 and COVID-19 in Older Adults: What We May Expect Regarding Pathogenesis, Immune Responses, and Outcomes." GeroScience, Springer International Publishing, Apr. 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7145538/.

[11] Nakamura, Eitaro, and Kenji Miyao. "A Method for Identifying Biomarkers of Aging and Constructing an Index of Biological Age in Humans." Journal of Gerontology: Biological Sciences, 2007.

[12] Nicita-Mauro, Vittorio, et al. "Smoking, Health and Ageing." Immunity and Aging, BioMed Central, 16 Sept. 2008, www.ncbi.nlm.nih.gov/pmc/articles/PMC2564903/.

[13] Boat, Thomas F. "Acronyms and Abbreviations." Mental Disorders and Disabilities Among Low-Income Children, U.S. National Library of Medicine, 28 Oct. 2015, www.ncbi.nlm.nih.gov/books/NBK332905/.

[14] "UMLS Metathesaurus Vocabulary Documentation." U.S. National Library of Medicine, National Institutes of Health, 4 May 2020, www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html.

[15] Arthur, David, and Sergei Vassilvitskii. "k-Means++: The Advantages of Careful Seeding." *Stanford*, 2006, ilpubs.stanford.edu:8090/778/1/2006-13.pdf.

[16] "Sci-Kit Learn." *Scikit*, 2020, scikit-learn.org/stable/.

[17] Blanchet, F. Guillaume, et al. "Forward Selection of Explanatory Variables." Ecological Society of America, 2008.