Big Data Analytics + Virtual Clinical Semantic Network (vCSN): An Approach to Addressing the Increasing Clinical Nuances and Organ Involvement of COVID-19

Fuad Rahman*, Rick Meyer†, John Kriak†, Sidney Goldblatt†, and Marvin J. Slepian*‡§¶

The coronavirus disease 2019 (COVID-19) pandemic has revealed deep gaps in our understanding of the clinical nuances of this extremely infectious viral pathogen. In order for public health, care delivery systems, clinicians, and other stakeholders to be better prepared for the next wave of SARS-CoV-2 infections, which, at this point, seems inevitable, we need to better understand this disease—not only from a clinical diagnosis and treatment perspective—but also from a forecasting, planning, and advanced preparedness point of view. To predict the onset and outcomes of a next wave, we first need to understand the pathologic mechanisms and features of COVID-19 from the point of view of the intricacies of clinical presentation, to the nuances of response to therapy. Here, we present a novel approach to model COVID-19, utilizing patient data from related diseases, combining clinical understanding with artificial intelligence modeling. Our process will serve as a methodology for analysis of the data being collected in the ASAIO database and other data sources worldwide. ASAIO Journal 2021; 67:18–24.

Key Words: big data, SARS-CoV-2, COVID-19, clinical semantic network, electronic health record, neural network, machine learning, multisystem disease, disease modeling

The novel coronavirus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that initially caused a cluster of pneumonia cases in Wuhan, China, in late 2019 rapidly spread across the globe resulting in a worldwide pandemic in only a few short months. The resultant disease syndrome caused by SARS-CoV-2 was termed coronavirus disease 2019 (COVID-19).1 Due to its rapid spread and severe morbidity and mortality, the government and the healthcare community have been desperately searching for the most effective prevention and treatment strategies and recommendations.2 Despite best efforts of the world community, many clinical diagnostic and therapeutic questions remain unanswered. To provide answers, data are needed. Thus, a well-designed data-driven approach may assist in identifying answers to critical COVID-19 disease questions.

What has been particularly challenging about SARS-CoV-2 is the myriad of clinical manifestations that continue to emerge. Initially, COVID-19 was recognized as a primary pulmonary disease, with typical presentations of fever, cough, shortness of breath, followed by progressive dyspnea and evident patchy basal pneumonia.3 Patients either recovered after a severe flu-like illness or rapidly declined requiring intubation and ventilatory support. As our experience with this disease evolved, a subset of clearly worse patients became evident. In this group, decline was rapid, with patients escaping beyond the capacity of ventilator oxygenation, requiring extracorporeal membrane oxygenation.4 With time, an additional subset emerged with evident cardiac compromise, with significant heart failure requiring mechanical circulatory support.2,3 Next, as far as clinical reporting emerged, a burgeoning of progressive renal failure with evidence of direct viral involvement in the kidney with acute kidney injury frequently requiring dialysis.2,3 Neurologic manifestations then emerged with evident stroke.2,3 Pulmonary emboli also emerged with evidence of an overall prothrombotic milieu.4,5 Most recently, a pediatric subset referred to as pediatric multisystem inflammatory syndrome has manifested.5

The chameleon-like and extensive nature of this systemic viral disease, the variable inflammatory and immune response to its infection, and the rapidity of its onset begs for new methodologies to gain insight and paths to best diagnostic and therapeutic practices. In particular, insight is needed for defined patient subset groups at increased risk. As we are dealing with an emergent rapidly progressive crisis, conventional controlled trials are not possible. As such, powerful effective computational means to rapidly sift through heterogeneous data is a methodology that, if brought to bear, can provide insight.7 This is the domain of big data, semantic networks, and machine learning.8

Here, we present a computational approach to answering COVID questions based on detailed analysis of existing healthcare data. We feature this in our development of an analytic backend for use with the ASAIO COVID database.9 The essence of the methodology is to combine big data analytics with a virtual clinical semantic network (vCSN) and machine learning.10 We have chosen this approach as it leverages the strengths of each component technology. Big data technologies are best when used to manage high volumes of data with significantly less up-front rigor on the curation of data, particularly important given the silos of data that exist in our healthcare system.11
Artificial intelligence (AI) domains (including machine learning and natural language processing [NLP]) provide for the interrogation of big data sources for extraction, harmonization, and pattern recognition of data.12 Semantic networks then derive understanding and, in this case, project clinical understanding from the data.11 The resulting operational data stores can then leverage analytics for myriad potential COVID and similar crisis support, including population clinical cohorts, risk, and response projections, and public health surveillance.

This approach can be utilized to help the medical and scientific community, as well as governments and health agencies, answer questions such as what unique attributes predispose susceptibility and immunity?; What traits drive specific end-organ manifestations? What characteristics contribute to COVID-19 recovery with or without complications? What factors contribute to the success or failure of a COVID-19 treatment? Who is likely to develop a recurrence?; and What measures can be taken to promote positive outcomes overall?

**Methodology—Disease Modeling**

To answer the questions outlined above, we first need to model COVID-19 and its progression. Therefore, we need to answer two questions upfront: (1) how do we model diseases such as COVID-19 and (2) what does a model really mean and look like in this context?

A model is a mathematical construct that is built based on “observed” samples and calculation of the relationships between measurable characteristics of a phenomenon or “variables” in the form of mathematical equations.14 In general, modeling physical systems requires numerous “labelled” data points, data that have been tagged and verified manually with well-recognized and established tagsets.

Disease modeling consists of two distinct but interrelated components: scientific (mathematical) disease modeling and clinical disease modeling. A scientific disease model is a mathematical construct that is built based on “observed” samples and calculating the relationships between measurable characteristics of a phenomenon or “variables” in the form of mathematical equations.14 One of the primary benefits of a mathematical disease model is the establishment of associations between a given disease and other factors (epidemiology, symptomatology, risk factors, etc.). However, it does not necessarily consider the severity and importance of the aforementioned factors. For example, a CDC report indicated that persons with select underlying health conditions (diabetes mellitus, chronic lung disease, and cardiovascular disease) or other recognized risk factors for severe outcomes from respiratory infections appear to be at a higher risk for severe disease from COVID-19 than persons without these conditions.15 This report was likely based on a mathematical model that made associations between these conditions and COVID-19 outcomes. However, the CDC report did not rank the importance of these conditions (or combination thereof) or identify factors related to these conditions that are more likely to result in severe outcomes. Fortunately, this deficiency can be corrected when clinical disease modeling is combined with scientific disease modeling.

Clinical disease modeling can help determine the underlying disease factors and degree and severity of each associated factor. For example, once the mathematical model establishes an association, clinicians will analyze the data and help determine exactly what additional data/information is needed from a clinical perspective to help properly characterize a disease. Thus, clinical disease modeling involves mathematical disease modeling with input from clinicians.

Both influenza and COVID-19 disease are highly contagious respiratory illnesses caused by different viruses. Influenza is caused by influenza viruses (e.g., Influenza A & B), and COVID-19 disease is caused by a novel coronavirus (SARS-CoV-2). However, both viral illnesses are contracted in a similar manner (droplet infection by close contact [cough, sneeze, talking]) and have numerous overlapping symptoms. For example, in humans, both infections cause fever, chills, cough, shortness of breath, dyspnea, fatigue, headache, myalgia, and others. Thus, influenza was chosen as the “proxy” condition by which COVID-19 disease would be modeled.

Our platform uses a combination of both scientific (mathematical) disease modeling and clinical disease modeling to properly characterize new and emerging disease states as well as recharacterizing existing disease states.

**Methodology—Machine Learning and Semantic Interoperability**

AI is the generic study of how human intelligence can be incorporated into computers.16 Machine learning, which is a subset of AI, concentrates on the theoretical foundations used by computational aspects of these algorithms.17 Machine learning belongs to the field of computational intelligence and soft computing, some examples of which are neural networks, fuzzy systems, and evolutionary algorithms.18 More simplistically, a machine learning algorithm (either supervised or unsupervised) is used to determine the relationship between a system’s inputs and outputs using a learning data set that is representative of all the behavior found in the system using various data modeling techniques—mostly statistical modeling.19

Machine learning has been widely applied in clinical research and problem solving in recent years.20 The reasons for this are manifold including improved understanding of how these algorithms work, access to very powerful but affordable computing resources, availability of electronic and tagged data, and an emergence of a genre of cross-disciplinary researchers, academics, and industry professionals collaborating their efforts to build practical solutions with real data and validation. Finally, adding machine learning–based semantic interpretation of data extracted from multivariate sources dramatically increases the ability to both aggregate and transfer data with a baseline clinical comprehension.

A semantic network is an elegant way to capture relationships of related units of information.21 For example, “shortness of breath” may be a very specific symptom in some diseases. This is usually called a “concept,” a unique entity defining a phenomenon. This shortness of breath may then be associated with other broader concepts, such as “respiratory distress,” which is a “type-of” distress related to respiratory illnesses, of which “shortness of breath” may be one of its many manifestations. How closely two concepts are connected are determined by how closely related they are to each other, denoted by a “link.” Even when two concepts are related, they may be strongly or weakly related, which in turn is denoted by “weights.” So, in general, concepts are represented as nodes
with labeled links (e.g., Type-of, Is-a, or Part-of), within a complex network of nodes, called a semantic network. This way, a semantic network is a process of capturing knowledge, using concepts and how those concepts relate to each other. A clinical semantic network (CSN) is simply a network that is built to encapsulate clinical knowledge. As seen in Figure 1, the left-hand side represents a network based solely on relating terms and concepts between industry-standard ontologies, like the unified medical language system (UMLS) discussed in more detail later. The right side of the figure shows how the application of the structure of a semantic network allows “proximity” of these concepts denoting how closely these concepts are related with one another—derived from real patient data—in terms of humanly annotated weights representing the relationships. In other words, a basic network relates points of data that can be shared and a CSN allows the data to be shared and compiled as clinically comprehensive information.

A working model of COVID-19 starts with methods to understand clinical context within a computable framework. There has been a long-standing effort to make clinical data computable, such as the UMLS. This is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems. The UMLS Metathesaurus is a repository of over 100 biomedical vocabularies, including CPT®, ICD-10-CM, LOINC®, MeSH®, RxNorm, and SNOMED-CT®, all designed for tagging clinical data. Within the Metathesaurus, terms across vocabularies are grouped together based on meaning—forming concepts—allowing us to capture and account for the huge variations in language and expressions.

Integrating industry standards with even deeper, proprietary ontologies can be accomplished with a semantic network that pulls together the attributes of various clinical “concepts” with respect to interdependence, comorbidity, diagnosis, and outcomes. One such solution is termed the CSN. The clinical framework for this innovative system was derived from the expertise of a team of clinicians and client users. At its core, the CSN is a semantically built taxonomy and ontology—allowing monumental strides towards large scale clinical analytics—potentially a very useful resource for processing healthcare data. It comprises a network that readily captures, classifies, labels, and cross-references usable medication information, data, and clinical concepts—medical information that typically must be abstracted by a clinician directly from each individual medical note.

The CSN is a network of clinical knowledge that maps concepts across multiple medical data sources, thus allowing for discovery of new (and validation of existing) clinical relationships, from initial symptom/s to diagnosis. For example, dyspnea, or breathing discomfort, is a common symptom that

Figure 1. Moving beyond related concepts to a clinical semantic knowledge base.
afflicts millions of patients with pulmonary disease and may be the primary manifestation of lung disease, myocardial ischemia or dysfunction, anemia, neuromuscular disorders, obesity, or deconditioning. Dyspnea alone is not suggestive of COVID-19 as it is not universally present, being detected in 30–50% of patients. When a second symptom is added to dyspnea, one or more of the aforementioned conditions can be ruled out. However, the diagnostic weight or importance of a particular symptom is variable for any given condition. For example, myalgia, another symptom often associated with COVID-19, is only found in about a third of SARS-CoV-2 positive patients and may be present in many of the previously mentioned conditions. Thus, the combination of dyspnea and myalgia would not have the same diagnostic weight as the combination of dyspnea and fever, as fever is present in most cases of COVID-19, and less frequently present, broadly in conditions where dyspnea is found. Thus, the nature of a (clinical) semantic network is to provide weights to the interrelated aspects (e.g., nodes and arcs) of the knowledge base making it directly impactful on the ability of the CSN to weight the importance of symptoms and findings discovered in the AI/ML modeling process. Furthermore, the CSN ultimately supports clinical decisions via a precision medicine approach, built on data, that is, capable of aggregating and disseminating the data not simply as data but repre- sented with and “loaded” with clinical knowledge. Established relationships among various clinical concepts within the CSN complemented by standard medical terminologies such as SNOMED-CT and RXNorm can be used as “targeting” mechanisms to derive useful and impactful relationships within text for extraction and use. But in order for us to use this framework as a “knowledge-backbone,” we need to convert the encapsulated relationships, such as “what-is,” “how-is,” “who-is,” “who-else-is,” into a set of probabilistic pathways—a representation that an AI/ML engine can consume and build upon. The vCSN is a virtual model of this very rich and complex data structure, which is a way to computationally estimate the relationship of various interconnected clinical concepts based on knowledge captured within CSN. The “targeting” provided by the vCSN combined with standard terminologies can then power NLP and machine learning techniques to literally extract structured clinical data from the “dark matter” of text. This platform can then be configured to seek more specific clinical data and relationships, for example, signals in the data relative to COVID-19.

COVID-19 is a rapidly unfolding multisystem disease. Currently, efforts are primarily focused on saving as many lives as possible and imposing best efforts to disrupt the transmission of the infection among the world population. As such, the “inputs” required for a data-driven approach, as introduced here, are not yet fully available. Our hopes with the ASAIO database is to provide a channel for heterogeneous aggregation of clinical data on COVID patients and outcomes. What data are available, independent of this database, however, focuses on infection rates, mortality statistics, recovery rates, resource utilization, and other logistical and related to “managing” the outbreak, and almost always lacks the hard clinical data needed to model this infection effectively. We plan, therefore, to utilize data from prior events to dictate and guide and prove the principle of our approach—specifically utilizing influenza, pneumonia, and acute respiratory distress syndrome data/clinical cases—which are closely related to COVID-19 and for which we have adequate clinical data available. The approach, as depicted in Figure 2, is to collect existing data from one or more diseases that are closely related to COVID-19 with respect to symptoms, comorbidity, treatment, and outcomes. In Figure 2, data are collected from various “silos” such as electronic health records, claims such as Dx (diagnosis), Sx (symptom), and Rx (prescription), and other sources such as social networks or case studies. These data will be then ingested into a big data environment and clinically curated (“tagged”) using the vCSN platform. Once the data are tagged, we will utilize a second set of machine learning algorithms to apply clinical “meaning” (semantics) and model the disease, as seen in Figure 3, that can eventually result in our ability to deliver or complement advanced analytics, such as response recasting, risk stratification, contact tracing engagement, and so on.

Figure 3 is a simple representation of how this meaning or semantics is derived, specifically within the vCSN. vCSN “scans” patient data—often written in free-form “natural” or “unstructured” form—NLP identifies clinically significant phrases, which in turn is then modeled using machine learning algorithms resulting in identifying clinical pathways for various relevant clinical concepts embedded within the patient data. As we have access to past data, we will be able to fine-tune the model by validating against known outcomes. This will establish the base models, which then can be retrained to model COVID-19 when patient data are widely available, as shown in Figure 4.

Figure 4 shows additional details of how machine learning may be used to train the vCSN. As is shown, we use long short-term memory recurrent neural network and concurrent neural networks to build these models. This figure specifically shows how this modeling approach can be used to retrain based on a new set of data related to a disease. Each machine learning model is a result of extensive training using the source data related to one specific disease, but then can be easily retrained to model other diseases as long as relevant data are available. The resulting representational model is a set of relational matrices, which captures the essential variable or parameters of the disease and how they influence the progression and output of that disease.

Machine Learning Models—Bias and Accuracy

Clinical researchers understand that machine learning modeling is often susceptible to bias. This is an active area of research, and it is beyond the scope of this article, but since our proposed model is based on machine learning methodologies, a short discussion about the issues related to these emerging technologies is warranted.

Mehrabi et al. have identified 23 different biases in machine learning models broadly categorized as computational bias (such as algorithm bias), data bias (such as sampling bias), cultural bias (such as observer bias), user bias (such as selection and presentation bias), and other bias (such as funding bias). In our case, although we need to be careful about all these sources of bias, we primarily need to focus on computational bias and data sampling bias.

One of the ways we can counteract the effect of computational bias is by identifying the outliers in the data. Wrongful
inclusion of outlier data in building a predictive model can be an important indicator of problems since including an outlier skews data, diminishing accuracy for machine learning initiatives. This automatically leads to another requirement in designing a dataset for training these models, the importance of understanding how the data are distributed. Machine learning models are underpinned by statistical methods and, therefore, susceptible to the accurate distribution of the sampling data. This departure from a normal bell curve is often measured by skewness, indicating how diffused the data are around a mean, median, and mode.

These observations lead us to adopt two important design principles. These problems are not just a machine learning issue but involve subject matter experts at a very early stage to validate the quality of the data. The second requires a very focused and well-defined problem statement to be validated using our machine learning models.

Another important aspect to discuss here is the limitations of adopting a machine learning solution to our problem. Machine learning is not without its share of concerns. Two broad categories of these limitations are limitations inherent to machine learning solutions in general and limitations related to the specific problem and the associated domain.

In general, it may be stated that these models encode correlation but not necessarily causation or ontological relationships. This may be related to the fact that, underneath all the mathematics and computational algorithms, these are a series of geometric transformations and that they are not designed for creating high-level, symbolic reasoning. Thus, each narrow application needs to be specially trained and requires large amounts of hand-crafted, structured training data. Finally, in most cases, learning is supervised with a large amount of often human-annotated training data.
The modeling case outlined in this article addresses additional important issues. For example, computer processing of natural languages is still very narrow. In addition, the very nature of the healthcare data we are processing has its own share of issues, as described by Jarrett et al. in their article on radiation oncology or by Maria and Seymour, in their article about pain neuroimaging. In our application, a major issue was the sparseness of the data requiring methods to solve this problem for effective training of these machine learning models.

**Pathway to a Solution**

We present here an approach to model complex diseases such as COVID-19 using machine learning and AI. Modeling a disease has particular advantages. Since the model is derived from real patients with detailed healthcare data including pre-existing conditions, symptoms, treatment paths, medications, lab results and device interventions, such as intubations and extracorporeal membrane oxygenation, the model is supposed to be able to answer questions related to how COVID-19 affects organ systems, how the disease progresses, how pre-existing conditions contribute to various manifestations of this disease, and so on. For example, it is entirely possible to build different models that address issues related to renal and respiratory organs. If we can identify the risks of patients based on their past medical and special histories, it will be possible to assess risk in terms of possible manifestation of the disease and devise an appropriate treatment plan. Some patients may be better off at home with adequate home support, some may fare better in the hospital while others, asymptomatic patients, may perhaps be better off in home isolation. Estimating what to expect, although never an exact science, may be a great starting position to successfully manage a resurgence of COVID or next pandemic. Very similar conclusions about how these computational models may be integrated in the ICU or other settings may be drawn based on dynamic risk stratification since appropriate level of care has assured a very high survival in COVID-19 patients.

Although the methodology presented is a combination of clinical considerations and machine learning, the learning or modeling aspect is completely data-driven. Every steps of the analysis and the transition to the subsequent steps are data-driven and not based or a hypothesis. This, however, poses its own set of challenges, some of which were addressed earlier in the section elaborating on the limitations of these models in practice.

**Conclusion**

We present a data-driven approach to model COVID-19. While the pandemic is still ongoing, it is critical that we collect data and document as many case studies as possible. As such, ASAIO has organized a rich database for this specifically. We present here a novel tool to build a COVID-19 disease model using AI and machine learning—especially exploiting the power of NLP and deep learning neural networks—via designing a semantic network-based computational platform. These computational models can now sift through years of healthcare data and automatically build connections and dependencies so that modeling a complex disease such as COVID-19 is now computationally possible. In addition, since the models are underpinned by a significant human-annotated
semantic network, as described earlier in the article, we have confidence that the quality of the model will be relatively high. The methodology outlined above will serve as a platform to provide insight from data collected to answer a myriad of questions. During the data collection and standardization process, we will train working models of diseases that are closely related to COVID-19. This will put us in a stronger position to apply these models to COVID-19 as soon as relevant data accumulate and contribute to better preparation for COVID progression, resurgence, or the next pandemic.

References


