An Overview of Healthcare Data Analytics With Applications to the COVID-19 Pandemic

Zhe Fei, Yevgen Ryeznik, Oleksandr Sverdlov, Chee Wei Tan, and Weng Kee Wong

Abstract—In the era of big data, standard analysis tools may be inadequate for making inference and there is a growing need for more efficient and innovative ways to collect, process, analyze and interpret the massive and complex data. We provide an overview of challenges in big data problems and describe how innovative analytical methods, machine learning tools and metaheuristics can tackle general healthcare problems with a focus on the current pandemic. In particular, we give applications of modern digital technology, statistical methods,data platforms and data integration systems to improve diagnosis and treatment of diseases in clinical research and novel epidemiologic tools to tackle infection source problems, such as finding Patient Zero in the spread of epidemics. We make the case that analyzing and interpreting big data is a very challenging task that requires a multi-disciplinary effort to continuously create more effective methodologies and powerful tools to transfer data information into knowledge that enables informed decision making.

Index Terms—COVID-19, digital technologies, epidemiology, high dimensional inference, infection source detection, metaheuristics

1 Introduction

The ongoing global COVID-19 pandemic presents to us daily, if not hourly, updated sets of massive and messy data from all over the world and a continuous series of challenging research questions in multiple areas. They include issues in data management, data analysis and interpretation and public health policies development that range from disease prevention and management to social concerns about mental health of the general public due to prolonged quarantine periods and restrictions in personal freedom. Massive and complex data, structured or unstructured, are now becoming available in practically all disciplines, particularly in health data [1]. The volume and speed at which massive data become available can make conventional methods for analyzing them less efficient or inappropriate.

Identifying quantities of interest and making meaningful summary statistics for trends, patterns and relationships/associations among the the different types of variables become an overwhelming task because of the huge number of variables in the data sets. Even visualizing such data sets correctly can be extremely challenging and easily subject to manipulation and mis-interpretation [2]. Data science is a recognized discipline that develops analytic tools to effectively manage, analyze and interpret big data of various types. The field is rapidly evolving and fuels constant discussion in various disciplines; for instance in statistics and machine learning, recent perspectives on data science can be found in [3], [4] . Its expanding important role to uncover vital insights in big data is now instrumental in many large-scale applications such as healthcare data analytics—the topic of focus in this paper.

Frequently, research questions are formulated into various types of inferential problems, that likely include studying associations among the massive number of different types of the heterogeneous variables, identifying risk factors for selected outcomes and predicting future trends. A distinguishing feature of the current pandemic is that it requires urgent, innovative and effective analytic tools to obtain timely information that enables public health leaders to make data-based strategic decisions with confidence. This paper discusses an overview of innovative analytic approaches for tackling COVID-19 related problems using modern digital technologies, innovative statistical methodology for accurate inference from big and complex data sets, analytic epidemiological tools to track and control disease progression, and state-of-the-art algorithms to compute and search for optimal strategies. The collective tools described herein are not limited to tackling COVID-19 problems and can be applied to solve other types of medical problems, and beyond. For example, the epidemiological tools in Section 4 can be directly modified to detect fraud and news leakage or monitor and identify the key sources of fake news.

In the next few sections, we give an overview of the latest advances in data science with a focus on digital technologies for clinical research, statistical inference for big data and epidemiology. Neural networks, machine learning and metaheuristics are important tools in artificial intelligence and their relevance to solving COVID-19 problems is also mentioned. Table 1 identifies and summarizes specific applications to COVID-19 problems in this paper.

- *Zhe Fei and Weng Kee Wong are with the Department of Biostatistics, University of California, Los Angeles, Los Angeles, CA 90095 USA. E-mail: {feiz, wkwong}@ucla.edu.*
- *Yevgen Ryeznik is with the AstraZeneca, 43150 Gothenburg, Sweden. E-mail: yevgen.ryeznik@gmail.com.*
- *Oleksandr Sverdlov is with the Novartis Pharmaceuticals Corporation, East Hanover, NJ 07936 USA. E-mail: alsverdlov@gmail.com.*
- *Chee Wei Tan is with the Department of Computer Science, City University of Hong Kong, Hong Kong. E-mail: cheewtan@cityu.edu.hk.*

Manuscript received 29 Oct. 2020; revised 29 June 2021; accepted 20 July 2021. Date of publication 0 . 0000; date of current version 0 . 0000.
(Corresponding author: Weng Kee Wong.)
Recommended for acceptance by Y. Xia.
Digital Object Identifier no. 10.1109/TBDATA.2021.3103458

TABLE 1
COVID-19 Related Applications in the Paper

Topic	Methods	Impact	Reference
Digital Health	Wearable devices/technologies, digital therapeutics (DTx)	Remote health monitoring and treatment delivery	Sections 2.1, 2.2
Deep Learning	CNN, RNN	Diagnosis and classification of COVID-19 cases	Section 2.3
Scientific Machine Learning	ODE-based SIR model; Safe Blues	Modeling of limited data at the beginning of the pandemic; contact tracing app	Section 2.4
High Dimensional Inference	LDPE; Debiased LASSO; SSHDI	Statistical inference for future COVID-19 related genomics data	Section 3.4
Computational Epidemiology	Graph-theoretic statistical inference; GNN	Patient Zero search; Infodemic risk management, fake news	Section 4
Metaheuristics	Nature-inspired algorithms: ICA; DE; PSO; CSO	Disease trend prediction, spread monitoring, real-time prediction	Section 5.1

Abbreviations: CNN = Convolutional Neural Networks; RNN = Recurrent Neural Networks; ODE = Ordinary Differential Equations; SIR= Susceptible, Infectious, Recovered; LDPE = Low Dimensional Projection Estimator; LASSO = Least Absolute Shrinkage and Selection Operator; SSHDI = Split and Smoothing for High Dimensional Inference; GNN = Graph Neural Networks; ICA = Imperialist Competitive Algorithm; DE = Differential Evolution; PSO = Particle Swarm Optimization; CSO = Competitive Swarm Optimizer.

Section 2 discusses data science-based approaches to address important clinical research questions. We consider two examples. The first concerns designing a clinical study with exploratory tools, digital technologies and biomarkers to characterize depression, which has seen a spike in the current pandemic [5]. We discuss both conventional and innovative ways to analyze large volume, high frequency data in this setting, and emphasize the importance of careful formulation of research questions to address the scientific goals of the study. Our second example showcases a new branch of artificial intelligence research called Scientific Machine Learning (SciML). SciML enriches mathematical models and facilitates the use of data-driven machine learning techniques to improve the quality of model-based prediction. As an application, we describe some virtues of the SciML approach to improve efficiency of global COVID-19 quarantine policies.

Section 3 describes some state-of-the-art methodologies and algorithms for making inference for big data that can provide some new insights into statistical inference with regression models and many more predictors than samples, referred to as "high dimensional inference." High dimensional inference has broad applications and particular relevance to the joint modeling of features in large data sets, as illustrated by our application using a cancer genomic data set. We demonstrate the utilities of a newly proposed method, where we identify important gene pathways for early diagnosis of a disease via finding significant predictors among hundreds or more of them. The described techniques can be more broadly applied to other areas and data sets, such as the massive COVID-19 data sets continuously generated from the Johns Hopkins University depository, and the inference sought can be estimating or updating estimates of the various risk factors of COVID-19 and accurately identifying significant predictors from a large pool.

Section 4 reviews applications of large-scale computational epidemiology optimization problems such as infection source detection (e.g., searching for Patient Zero) and its related Infodemic management due to the COVID-19 Pandemic. Solutions to these computational epidemiology optimization problems can provide health authorities with *digital contact tracing* to trace the social contacts of an infected person and searching for the outbreak origin [6], [7] or to formulate appropriate healthcare policies in the face of misinformation [8].

Section 5 provides an overview of increasing use of metaheuristics in various disciplines, including recent applications of metaheuristics to tackle multiobjective optimization problems related to COVID-19. The paper concludes in Section 6 by emphasizing on the importance of multidisciplinary research, and the continuing central role of statistical thinking in the era of big data.

2 Data Science in Modern Clinical Research

2.1 Opportunities and Challenges

The 21st century biomedical research arena has and continues to benefit from the increasing computational power, innovative technologies and availability of big data. The term "big" refers to several characteristics of the data that are also referred to as the "V's of big data"[9], [10], [11]. There are at least six V's: volume, variety, velocity, veracity, variability (and complexity), and value. *Volume* refers to the magnitude (amount) of data, which depends on the technology development and is continuously increasing. *Variety* refers to the structural heterogeneity of data sets, e.g., structured, semi-structured, and unstructured data. *Velocity* refers to the rate of data generation and processing, which keeps growing with the advances in technology. *Veracity* corresponds to uncertainty and unreliability in the data sources due to subjectivity of human opinions or in social media. *Variability* (and complexity) refers to the variation in the data flow rates (which can have peaks and troughs) and complexity in data processing due to heterogeneity of the data sources. *Value* corresponds to the benefit that data adds to the enterprise, e.g., increased revenue, decreased operational costs, higher customer satisfaction, etc.

Big data brings tremendous opportunities and new multidisciplinary challenges for clinical research. Here are just a few examples of big data sources in this context.

- *Real World Evidence (RWE)*: In the 20th century, the randomized controlled trials (RCTs) were established as the gold standard of evidence-based

research for evaluating safety and efficacy of new treatment interventions [12]. However, RCTs may be long, expensive, and difficult to conduct. Alternative sources of clinically important data include electronic health records (EHRs), population based registries, and some other real-world data that can supplement and generalize the evidence from RCTs. Development of big electronic databases has enabled collection and integration of such RWE; yet, it is still complex, multi-dimensional, and lacking clear structure. For instance, many drug prescriptions were hand-written and later scanned and saved electronically. How can one extract and then classify this important information? An increasingly promising approach is the *natural language processing* (NLP) [13]. Organization of the NLP to ensure proper data collection, cleaning, restructuring, and getting it to a point when it can undergo a meaningful analysis is both a challenge and an opportunity.

- *Medical Imaging Technology*: Many clinical trials utilize objectively measured biomarkers that capture disease progression over time and provide measurement of treatment effects. The *magnetic resonance imaging* (MRI) has now been widely used in clinical research in Alzheimer's disease [14], multiple sclerosis [15], cancer [16], etc. Analysis and interpretation of MRI data requires high medical expertise and judgement, and it is also very time consuming and expensive. Automating this process could provide a more objective and less costly way of extracting important medical information. *Convolutional neural networks* (CNN) is a class of deep learning methods that can be useful for analyzing MRI data, to produce objective, high quality outcome measures [17]. This can potentially improve signal-to-noise ratio and increase the efficiency of clinical trials.
- *Digital Endpoints*: Novel sensors and wearable technologies (e.g., smart watches) have enabled collection of terabytes of individual health information, such as physical activity, vital signs, quality of daily living, etc. These data can be collected with high frequency over extended time periods, and provide means to identify serious medical problems (e.g., heart abnormalities that could lead to a heart attack). Big data generated by wearable technologies can potentially reduce the need for clinical site visits and streamline the clinical trial research. This can be especially valuable during a global pandemic such as COVID-19, when hospitals and clinical trial research units are overwhelmed and patients are often unable to keep their scheduled in-clinic visits due to quarantine restrictions. However, this promise comes with the need for careful data collection, processing, and development of valid digital endpoints [18].
- *Digital Therapeutics* (*DTx*): A true hallmark of the 21st century medicine is the development of DTx – evidence based therapeutic interventions driven by high quality software programs to prevent, manage, or treat a medical disorder or disease [19]. The unmet medical need addressed by DTx is very diverse. As an example, consider the *precision dosing* paradigm in management of different chronic diseases [20]. *Closed-loop systems* that automatically determine optimal time of blood sampling and perform calculation of the optimal dose and timing of dose delivery provide means for individualizing treatment to patient. For instance, there is emerging clinical trial evidence that closed-loop insulin delivery systems can improve glucose control in patients with type 1 diabetes [21], and these systems are expected to become standard soon. A potential virtue of DTx is magnified during the COVID-19 crisis, as DTx products can deliver safe and effective care remotely (https://www.fda.gov/media/136939/download).

There is clear benefit of analyzing large clinical data sets but there are constant debates on the analytic approaches [22]. In the next subsection, we present an example of a clinical study evaluating different digital technologies, where both big data and traditional clinical data are collected. We argue that most scientific and statistical principles still apply in such experiments, while some novel data science and machine learning techniques can nicely supplement the more traditional and established approaches.

2.2 An Example of a Clinical Study Evaluating Digital Technologies

Depression is a burdensome mental health disorder that often goes undetected and untreated [23]. Symptoms of depression are multi-dimensional and affect emotions, thoughts, behavior, and physical domains. There is a strong need to develop effective methods to diagnose depression and perform efficient monitoring of patients with this condition.

Conventional measures of depression-related symptoms are paper-and-pen outcome assessments, such as the Hamilton Depression Rating Scale (HDRS) [24] and the Montgomery-Åsberg Depression Rating Scale (MADRS) [25]. While these measures are well-established, they are subject to rater bias, may lack clinical relevance and exhibit high variability, which translates into the need for large clinical trials to detect clinically relevant treatment differences. On the other hand, digital technologies have the potential to provide more objective and precise tools to detect depression-related symptoms; yet, these technologies require careful assessment and validation in clinical studies before they can be broadly implemented.

Our example is a single-site, cross-sectional, non-interventional study of novel exploratory tools, digital technologies and biomarkers to characterize depression. The study results have been published in [26]. Here we discuss some important aspects of the study design and illustrate the thinking process and the logic behind selection of appropriate data analysis tools.

The study evaluated 40 subjects (20 patients with major depressive disorder (MDD) and 20 healthy volunteers). There were three in-clinic visits at days 1, 7, and 14. At each visit, study subjects underwent a series of assessments, both conventional (e.g., MADRS) and novel digital technologies. In addition, between visits there was at-home collection of data through mobile apps. The study objectives were threefold:

- To assess feasibility of use of digital technologies.
- To assess utility of these technologies as diagnostic tools (classification of subjects, MDD versus healthy).
- To explore a potential of using digital biomarkers as predictors of the conventional measures (MADRS).

Due to the small sample size, this study was exploratory in nature. It provided only preliminary evidence on virtues of digital technologies, which has to be further confirmed in larger studies. Overall, seven digital technologies were evaluated. These can be broadly categorized as mobile apps that provided real-world data and tests that were performed in-clinic.

One mobile app provided an interactive tool for high-frequency assessments of cognition and mood over the course of the study. Another mobile app was a passive behavioral monitor that integrated smartphone data related to the user's social acts and patterns, e.g., phone calls, social media use, travel, etc. A third app was a platform to perform voice recordings to obtain vocal biomarkers that contain important information on depression-related symptoms.

The in-clinic digital technologies included a neuropsychological test battery; an eye motor tracking system that captures information across multiple domains of mood, cognition, and behavior; an electroencephalogram (EEG)-based technology to analyze the brain network activity; and a task quantifying bias in emotion perception.

The study data was very diverse and varied in structure and complexity. For instance, for a behavioral monitor app, a list of 85 features was derived and later scrutinized to 10 most important features per subject. These features represented various summary measures of social functioning, e.g., total duration of all communication events, entropy of the usage time of social media apps, number of places visited, etc. Therefore, despite high volume and high frequency of the raw data, for each subject we obtained a vector of numeric summaries. By contrast, for the interactive app that provided measurements of cognition and mood, we acquired longitudinal data per subject: a cognitive score and a mood score were calculated each time the subject engaged with the app. For a neuropsychological test battery, the data were acquired at each in-clinic visit, 60, 160, and 230 minutes after admission, and then averages across three time points were taken, and a total of 42 features were derived per subject per visit. The vocal biomarkers were derived by applying signal processing algorithms on various time windows within a sample of speech, for a total of 72 features per subject.

The data from digital technologies was combined with demographic and clinical questionnaire data for analysis, which included exploratory data visualizations and different supervised learning techniques. For instance, we performed classification analysis to predict the class of each subject, MDD or healthy, using logistic regression. Multiple linear regression was used to model and predict MADRS total score as a function of digital biomarkers. Our analysis was organized by technology – to understand utility of each technology and identify digital biomarkers that add most value. In many cases we were able to develop simple, parsimonious models with reasonably high diagnostic accuracy and potential to predict standard clinical outcome in depression [26].

One major lesson learned from this study is that while many novel digital technologies generated large-volume, high-frequency data, the majority of good clinical and statistical research principles were still applicable in this setting. The common techniques of data analysis such as classification and regression could handle many types of the data described. However, analysis of some data types, such as speech samples, require more advanced machine learning techniques. In the next section, we describe approaches based on neural networks that can be potentially useful for analysis of such complex data, and the computational tools for implementing neural networks.

2.3 Neural Networks and Distributed Computing for Clinical Research

Recent developments in computational technologies and increasing computational powers allow exploration and analysis of data with very complex structure. The sound data and the video data might have a potential application in disease diagnostics [27], [28]. One may consider this type of data as a time series with a non-numeric outcome. A video object can be viewed as a time series with picture outcomes, and a sound can be viewed as a time series of multidimensional vectors, or even graphs. Analysis of this type of data requires both new statistical and numerical approaches and highly powerful computational software. A combination of algorithms based on neural networks and distributed computing seems to be a viable approach to perform such an analysis.

A neural network model is based on simplified assumptions of the human brain architecture. The model is highly parameterized and requires computationally intensive optimization for the parameters tune-up. The process of the parameters' optimization is called *learning*. The simplest example of a neural network is a single neuron (perceptron) model [29]. In this model, a neuron accepts n inputs (usually, numerical), referred to as *covariates* in statistics terminology or *features* in machine learning. Each input point x_i has some positive weight w_i. Then, a weighed sum of input points is substituted as an argument of an activation function f, and the value obtained is an output of the model. A common choice is $f(x) = 1/(1 + e^{-x})$, a sigmoid function with a *bias term* b and an output of a single neuron is

$$y = f\left(b + \sum_{i=1}^{n} w_i x_i\right).$$

The perception model was proposed for a binary classification problem in 1958 [29]. The recent dramatic improvements in computational power has resulted in more complex and effective neural network architectures, and they include *deep learning* (DL), where the neural network contains many connected neurons and may have several hidden layers and several outputs. Every single neuron in the network can learn a simple input-output relationship. Then, all the neurons exchange the information learned to make the entire learning feasible. At a glance, it may seem that a neural network is a nonlinear approximator that is not much different from other approximation models. This may be true. For example, [30] proved that a network with a single hidden layer containing a finite number of neurons and a sigmoid activation function can approximate continuous functions on compact subsets of $\mathbf{R}^n$. The theorem thus

TABLE 2
Computational Tools for Neural Networks and Distributed Computing

Tool	Description
Tensorflow	An open software library for ML developed by Google. Accessible via https://tensorflow.org. The main API for working with the library is implemented for Python; there are also implementations for C Sharp, C++, Haskell, Java, Go, and Swift.
Caffe	Caffe stands for *Convolutional Architecture for Fast Feature Embedding*. It is an open-source DL framework written in C++ with a Python interface. It was developed at the University of California, Berkeley and is accessible via http://caffe.berkeleyvision.org/. It supports various DL architectures for image classification and segmentation, as well as GPU- and CPU-based acceleration computational kernel libraries such as NVIDIA cuDNN and Intel MKL.
PyTorch	An open-source ML library used for applications such as computer vision and NLP. The supported programming languages include Python, C++, and CUDA. PyTorch has been primarily developed by Facebook's AI research group. At the end of March 2018, Caffe and PyTorch were merged.
Keras	An open-source NN library written in Python on top of TensorFlow and some other DL libraries. It was designed with a focus on being user-friendly, modular, and extensible, and it allows fast experimentation with DNN, as well as CNN and RNN. Accessible via https://keras.io/.
MapReduce [41]	Provides a framework for computing some sets of distributed tasks using a large number of computers (called "nodes") that make up a cluster. The scope of MapReduce consists of two steps: Map and Reduce. At the Map step, one of the computers (called the master node) receives the input data of the task, splits it into parts, and transfers it to other computers (work nodes) for preliminary processing. At the Reduce step, the pre-processed data is collapsed. The main node receives responses from the working nodes and on their basis forms the result, i.e., the solution to the originally formulated problem.
Apache Spark	An open-source framework for implementing distributed processing of unstructured and weakly structured data. It provides computations built around resilient distributed data sets (RDDs). Unlike MapReduce that operates with disk storage, Apache Spark uses RDDs for recursive processing in RAM, thereby enabling it to perform more efficiently for some classes of tasks. It supports high-level tools for SQL queries and structured data processing (Spark SQL), ML problems (MLlib), graph processing (GraphX), and stream processing of live data streaming (Spark Streaming). Apache Spark is a key platform for distributed DL; it allows embedding of TensorFlow, and other DL frameworks in Spark workflows, to build distributed DL applications. Accessible via https://spark.apache.org/.
Databricks	A platform from the Apache Spark creators that provides functionality for reproducible research. It supports several programming languages (e.g., Spark, SQL, Java, Scala, Python, R) and allows easy switching between different languages within a project. Accessible via https://databricks.com.

Abbreviations: CNN = Convolutional Neural Networks; DL = Deep Learning; DNN = Deep Neural Networks; ML = Machine Learning; NLP = Natural Language Processing; NN = Neural Networks; RNN = Recurrent Neural Networks.

states that simple neural networks can represent a wide variety of nonlinear functions when given appropriate parameters. However, neural networks can be scaled, extended and generalized in a variety of ways: more hidden units in a layer, multiple hidden layers, weight sharing, innovative learning algorithms for massive data sets [31].

There are two special classes of deep neural networks: *convolutional neural networks* (CNNs) and *recurrent neural networks* (RNNs). The CNNs' architecture is inspired by the structure of the animal visual cortex, which makes CNNs extremely useful in applications dealing with imaging data, particularly, in medical image analysis. Thus, the algorithms using CNNs for processing and analysis of computer tomography, CT Scans, and chest X-ray images have been found efficient for diagnostic and classification of COVID-19 cases [32], [33], [34]. The RNNs are designed in such a way that connections between their nodes represent a directed graph along a temporal sequence, which allows to model dynamic behavior over time. The RNNs' architecture makes them capable of using their internal state as a memory to process the inputs of variable length and to model time series with more complex observations. This makes RNNs applicable for solving such problems as handwriting or speech recognition. Some of the recent examples of using RNNs to monitor the COVID-19 situation can be found in [35], [36]. A computational issue with classic RNNs is that when training it using back-propagation, the gradients which are back-propagated can "vanish" (i.e., go to zero) or "explode" (i.e., go to infinity). As a special case of RNNs, long short-term memory (LSTM) avoids the vanishing gradient problem by using recurrent gates called "forget gates" which allow gradients to flow backwards and unchanged. However, LSTM networks can still suffer from the exploding gradient problem [37], [38]. Extensions of RNNs continue to develop; a recent example is the Transformer, which is a new model used primarily in the field of natural language processing (NLP), which, unlike RNNs, does not require that the sequential data be processed in order.

The procedure of training a neural network includes optimization of weights by minimizing some loss function, given a training data set. Usually, the loss measures how different are the outputs produced by a neural network and the true responses taken from a training data set. In most of the scenarios, optimization is performed using gradient-based methods. The backpropagation algorithm [39] is used for efficient calculation of gradients. While the gradient-based methods are good optimization techniques which work excellently for convex functions and low dimensional space, one may expect much better results with the particle swarm optimization (PSO) technique [13], [40]. However, there appears to be very few publications using PSO as a tool to train neural networks, and it certainly merits investigation.

Table 2 provides a short list of current computational tools for implementing neural networks in practice.

2.4 Scientific Machine Learning

Recently, a new branch of AI research has evolved at the edge of scientific computing and machine learning (ML)—Scientific Machine Learning (SciML, https://sciml.ai). Scientific computing deals with mathematical models of real-life systems based on physical laws, the so-called *mechanistic* models, e.g., models utilizing ordinary and partial differential equations (ODEs and PDEs) or integral equations (IEs). ML models are usually data-driven models, and the more training data are available, the better are the model-derived outcomes. Scientific computing models typically involve a small number of parameters to describe the system to predict the system's outcome, whereas ML models may depend on a large number of parameters that have to be tuned by the data available.

It is difficult to say which approach (mechanistic models or ML models) is better. Both have pros and cons. Mechanistic models do not depend on data availability and are easy to interpret. However, they require that a modeler knows a mechanism underlying the model, which may be elusive. In this case, data-driven non-mechanistic models can give a very accurate prediction directly form data.

But what if no data are available and the mechanism is known only partially? A good example of such a situation is the COVID-19 pandemic. Very limited data are available at the beginning of the pandemic. On the other hand, we have an ODE-based SIR (Susceptible, Infectious, Recovered) model [42] that is described using only a few parameters and can be fitted by a small dataset. At the same time, this ODE-based model has components that are subject to high uncertainty and require specific strategies to understand (learn) this uncertainty.

SciML provides a scientifically sound approach to handle uncertainty of a physical or biological model using ML algorithms. Reference [43] provides an example of a SciML approach to gain insights of the COVID-19 pandemic using the SIR model, which in its simplest form has three ODEs

$$\begin{cases} \dfrac{dS}{dt} = -\beta\dfrac{SI}{N} + \gamma I \\ \dfrac{dI}{dt} = \beta\dfrac{SI}{N} - \gamma I \\ \dfrac{dR}{dt} = \gamma I. \end{cases}$$

Here, $S(t)$, $I(t)$ and $R(t)$ are the number of susceptible, infected, and recovered subjects at time t, respectively. The constants β and γ are the infection and the recovery rates. The total number of subjects in the population $N = S(t) + I(t) + R(t)$ is regarded as a constant, that is, births and deaths are not taken into account. The goal is to study the effect of quarantine. For this purpose, the SIR model is augmented by adding a time-dependent quarantine strength rate term $Q(t)$ and a quarantined population $T(t)$, which is prevented from having any further contact with the susceptible population. Therefore, the system of ODEs takes the form.

$$\begin{cases} \dfrac{dS}{dt} = -\beta\dfrac{SI}{N} + \gamma I \\ \dfrac{dI}{dt} = \beta\dfrac{SI}{N} - (\gamma + Q)I \\ \dfrac{dR}{dt} = \gamma I + \delta T \\ \dfrac{dT}{dt} = QI - \delta T. \end{cases}$$

Thus, the term $I(t)$ denotes the infected population still having contact with the susceptibles (as done in the standard SIR model), whereas the term $T(t)$ denotes the infected population of subjects who are effectively quarantined and isolated. The constant δ is an additional recovery rate that quantifies the rate of recovery of the quarantined subjects. Thus, we can write an expression for the quarantined infected population $T(t)$ as $T(t) = Q(t) \times I(t)$. Due to the universal approximation theorem, the quarantine term is replaced by a neural network (NN) and the deterministic system of ODEs is approximated by neural ODEs.

Thereafter, a neural network was trained by using a small portion of data available, and the developed model was able to predict the infected and recovered counts for highly affected countries in Europe, North America, Asia and South America with a good accuracy. Also, the proposed approach allowed to extract valuable information regarding the efficiency of different quarantine policies [43].

Another interesting and important example of SciML application in the context of COVID-19 pandemic is given in [44]. It shows how to utilize SciML approach with the data obtained via the contact-tracing apps. The method presented in the paper is called Safe Blues (https://safeblues.org), and it uses Bluetooth signals similarly to the existing technologies but the method does not require to record information about individuals and their interactions. Instead, it helps to understand population wide dynamics in a privacy-preserving manner.

There are software packages for implementing SciML, such as the *DiffEqFlux.jl* package [45] implemented in the Julia programming language (https://julialang.org). It combines the differential equations-based modeling approach with machine learning and neural networks algorithms, and they collectively provide ready solutions, such as, neural differential equations [46] and universal differential equations [47] to support research in SciML.

3 High Dimensional Inference (HDI)

In the current pandemic, large data sets are increasingly available for data mining, analysis and interpretation. The same is true in many other health science areas, especially in cancer research, where high throughput genomic measurements are available [48], [49]. To properly make inference from complex data sets, novel statistical models and methods are needed to account for the high dimensionality, including cases when the number of predictors may be much larger than the sample size. This section first briefly reviews modern methods for analyzing big data before we introduce the SSHDI method in the context of high dimensional inference. Although, to the best of our knowledge, there are currently no COVID-19 data sets with a large

number of predictors, i.e., genetic information, we expect that once an appropriate COVID-19 data becomes available, the proposed methodology will be helpful for gaining insights into the many puzzling questions on the pathogenesis of COVID-19.

For moderate sized data sets, there are well-established statistical methods when the number of predictors diverges with the sample size. Some examples are marginal screening in Genome-Wide Association Studies (GWAS) [50], [51] and penalization/regularization methods for joint modeling and variable selection [52]. There are also tree-based methods for feature selection that optimize the information gain or gain ratio when generating the decision trees [53], [54]. To enhance weak learners, such as trees, boosting and resampling methods have been proposed, along with XGBoost, random forests, among others [55], [56], [57]. However, efficient methods for drawing inferences from large and complex data sets have somewhat lagged but there is now intense research in high dimensional inference (HDI), where the focus is on assessing the uncertainty measures of model parameters, finding asymptotic distributions of estimated parameters and deriving significance tests or confidence bands.

In the traditional low-dimensional setting when $n > p$ and p fixed, it is well known that the least squares estimator $\hat{\beta}_{LS} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ converges to a normal distribution and exact inference through p-values and confidence intervals is possible. However, when $n < p$, the least squares estimation becomes problematic because the sample covariance matrix $\hat{\boldsymbol{\Sigma}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/n$ is singular. Such problems have become increasingly relevant in the past two decades when high-throughput data becomes common. The goal is often to find a parsimonious model to study the response variable when there are many covariates.

3.1 Current HDI Methods

Consider the homoscedastic linear model

$$\boldsymbol{Y} = \boldsymbol{X}\beta^* + \varepsilon, \tag{1}$$

where $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$ is the n-vector of responses; $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is the $n \times p$ design matrix where the columns contain p covariate vectors X_j's; $\beta^* = (\beta_1^*, \ldots, \beta_p^*)^{\mathrm{T}}$ is the true parameter vector of interest and ε is the random noise vector with $\mathbf{E}(\varepsilon) = \mathbf{0}_n$.

The high dimensionality referred to herein includes, but is not limited to the usual case when "$p > n$," such as when $n = 500$ samples with $p = 1000$ covariates. Even in a "$p < n$" setting with $n = 1000$ samples and $p = 500$ predictors, direct applications of classic regression models can lead to ambiguous and meaningless estimations and inferences if the number of predictors p is allowed to increase with the sample size n. In other words, the classic inferential results for the *fixed* p case would not directly apply to the *diverging* p case [58], [59]. To solve the joint estimation problem, penalized regressions have been widely used, including LASSO [52] and some of its many adaptive variants [60], [61], [62], [63], [64]. The estimators from penalized regressions are shrunk and thus "irregular" as their asymptotics become difficult to track. There are three directions of current research in high dimensional inference:

- *De-biased methods* derive the p-vector of the coeffcients by correcting the known bias of a sparse estimator, for examples, Low Dimensional Projection Estimator (LDPE) in [65] and de-sparsified LASSO estimator [66], [67], [68]. The de-biased estimators of β^* are for the joint effects of all p predictors and are shown to be asymptotically unbiased and normally distributed under some regularity conditions and when p is much larger than n. Therefore, the individual p-values and confidence intervals of the effects can be derived based on the asymptotics. However, such approaches have limitations [69], [70]. As the de-biasing procedure relies on accurate estimation of the $p \times p$ precision matrix of the predictors, which itself is a challenging problem [71], [72], finite sample estimation errors are expected. The optimization procedure also involves excessive number of tuning parameters to achieve the desired theoretical properties.
- *Post-selection inference* focuses on valid inference given a *selected* model and can be considered as a twin of HDI. [73] proposed a post double selection procedure for estimation and inference with continued work in [74], [75], and [76] characterized the distribution of a post-LASSO-selection estimator conditioned on the *selected variables*, but only for linear regressions. The apparent limitation is that the post-selection inference cannot detect or correct any errors already made by the selection. For example, if an important feature/predictor was not selected in the first place, the post-selection inference would not retrive it either.
- *High dimensional testing* solves the HDI problem without estimating the coefficients, but derives test statistics for the hypotheses such as $H_0 : \beta_j^* = 0$ and $H_0 : a^{\mathrm{T}}\beta^* = c$, where a can be a $p \times q$ matrix but $\mathrm{rank}(a)$ needs to be fixed and not increasing with n or p. [77] proposed the decorrelated score tests for penalized M-estimators; [78] introduced a similar procedure based on proportional hazards model; [79] also proposed a method for testing linear hypothesis in high-dimensional linear models. While simplifying the estimation and inference problem to testing could gain robustness and computational advantages, it also loses important information regarding effect sizes and directions.

3.2 SSHDI Method

A recent and novel framework solves the HDI problem from a different angle and potentially avoids the related limitations. We have shown such an approach has both theoretical and empirical advantages over existing methods [69], [70]. By using multi-sample splitting and smoothing techniques, the novel method converts the challenging high dimensional estimation problem to a sequence of low dimensional estimations. In each of the lower dimensional estimation, the sample size is sufficiently large for the number of predictors [58]. Algorithm 1 describes the base procedure, so-called one-split estimator, where we first split the original data into equal halves, then apply a general

variable selection procedure to choose a subset of covariates using one half of the data. Next we use the other half of the data to fit partial regressions iteratively using each covariate and the selected covariates. In other words, each coefficient is estimated jointly with the selected subset, which achieves dimension reduction. We show that when the selected covariates contain the sparse active set, the resulting coefficient estimator is unbiased, whether the actual effect $\beta_j^*,\ j = 1, 2, .., p$ is significant or not.

Algorithm 1. One-Split Estimator

Require: A selection procedure $\mathcal{S}_\lambda$ with tuning parameter λ
Input: Data $(\mathbf{Y}, \mathbf{X})$
Output: Coefficient estimator $\widetilde{\beta}$
1: Split the data into equal halves $\mathbf{D}_1$ and $\mathbf{D}_2$, with sample sizes $|\mathbf{D}_1| = \lfloor n/2 \rfloor$, $|\mathbf{D}_2| = \lceil n/2 \rceil$
2: Apply $\mathcal{S}_\lambda$ on $\mathbf{D}_2$ to select a subset of important covariates $S \subset [p]$
3: **for** $j = 1, .., p$ **do**
4: On $\mathbf{D}_1 = (\mathbf{Y}^1, \mathbf{X}^1)$, let $S_{+j} = \{j\} \cup S$ and let $\widetilde{\beta}^1$ be the coefficient estimator obtained from the partial regression of $\mathbf{Y}^1$ on $\mathbf{X}^1_{S_{+j}}$
5: Define $\widetilde{\beta}_j = \left(\widetilde{\beta}^1\right)_j$, which is the coefficient for covariate X_j
6: **end for**
7: Define $\widetilde{\beta} = (\widetilde{\beta}_1, \widetilde{\beta}_2, .., \widetilde{\beta}_p)$

Algorithm 2. SSHDI Estimator

Require: A selection procedure $\mathcal{S}_\lambda$ with tuning parameter λ
Input: Data $(\mathbf{Y}, \mathbf{X})$, number of re-samples B
Output: Coefficient estimator $\widehat{\beta}$
1: **for** $b = 1, 2, .., B$ **do**
2: Run Algorithm 1 with random data split $(\mathbf{D}_1^b, \mathbf{D}_2^b)$
3: Denote the output estimator as $\widetilde{\beta}^b = (\widetilde{\beta}_1^b, \widetilde{\beta}_2^b, .., \widetilde{\beta}_p^b)$
4: **end for**
5: Define $\widehat{\beta} = (\widehat{\beta}_1, \widehat{\beta}_2, .., \widehat{\beta}_p)$, where $\widehat{\beta}_j = \frac{1}{B}\sum_{b=1}^B \widetilde{\beta}_j^b$ is the average

The estimator based on a single split is highly variable due to the random data split and the variation in the selection. To this end, we use the idea of bagging [80] and multi-sample splitting to reduce the variability and increase the power of detecting signals. As shown in Algorithm 2, by repeating the split and estimation a number of times B, and aggregating the one-split estimators, SSHDI, shortened for *Split and Smoothing for High Dimensional Inference*, solves the estimation problem of the whole coefficient vector β^* in the assumed model (1) with increased power. We show that each coefficient estimator is asymptotically unbiased and normal. More importantly, the procedure accounts for the variation in model selection, which is largely neglected in most existing works. We also highlight that the final estimator $\widehat{\beta}$ is robust to various selection methods, such as, sure independent screening (SIS) [81], or regularized regressions with different penalties, like LASSO and SCAD. We denote such a selection method with tuning parameter λ by $\mathcal{S}_\lambda$. We further derive a model-free variance estimator based on non-parametric delta method and sub-sampling properties (Algorithm 3) [82], [83]. The variance estimators $\widehat{V}_j^B$'s are asymptotically consistent, and possess satisfying empirical performance when the number of re-samples B is of the same order as the sample size, i.e., $B = O(n)$.

Algorithm 3. Model-Free Variance Estimator

Input: $n, B, \widetilde{\beta}^b, b = 1, 2, .., B$ and $\widehat{\beta}$
Output: Variance estimator $\widehat{V}_j^B$ for $\widehat{\beta}_j, j = 0, 1, .., p$
1: For $i = 1, 2, \ldots, N$ and $b = 1, 2, .., B$, let $J_{bi} \in \{0, 1\}$ be the indicator of the i^{th} observation from the b^{th} sub-sample $\mathbf{D}_1^b$ in Algorithm 2 and let $J_{\cdot i} = \left(\sum_{b=1}^B J_{bi}\right)/B$.
2: **for** $j = 0, 1, .., p$ **do**
3: Define

$$\widehat{V}_j = \frac{4(n-1)}{n}\sum_{i=1}^n \widehat{\mathrm{cov}}_{ij}^2$$

where

$$\widehat{\mathrm{cov}}_{ij} = \frac{1}{B}\sum_{b=1}^B (J_{bi} - J_{\cdot i})\left(\widetilde{\beta}_j^b - \widehat{\beta}_j\right)$$

4: Define

$$\widehat{V}_j^B = \widehat{V}_j - \frac{n}{B^2}\sum_{b=1}^B (\widetilde{\beta}_j^b - \widehat{\beta}_j)^2$$

5: **end for**
6: Set $\widehat{V}^B = \left(\widehat{V}_1^B, \widehat{V}_2^B, \ldots, \widehat{V}_p^B\right)$

The theoretical properties of the one-split estimator and the SSHDI estimator are available [69], [70]. To show the asymptotic consistency and normality of the SSHDI estimators, the selection method $\mathcal{S}_\lambda$ has to satisfy the "sure screening" property, which requires the selected subsets to pick out the true active set S^* with probability approaching 1 as the sample size n goes to infinity. As an example, LASSO with a proper order of the tuning parameter λ [84] and sure independence screening (SIS) with a "beta-min" condition [85], among others, satisfy the sure screening property. Both the one-split estimator and the SSHDI estimator are asymptotically unbiased and normal as the sample size n goes to infinity. The SSHDI estimator has a smaller variance because of the bagging effect [86].

3.3 Numerical Studies

There have been extensive numerical experiments comparing the SSHDI procedure with current methods [69], and examples with non-linear models in [70]. Here we compare SHDI with two de-biased LASSO estimators, LASSO-Pro [87] and SSLASSO [66]. Under the high dimensional linear model (1), we set $n = 200$, $p = 500$, the active set $S^* \subset \{1, 2, .., p\}$ was a fixed random realization with size $|S^*| = 5$, and $\beta_{S^*}^*$ was a fixed realization of 5 i.i.d. random variables from $U[0.5, 2]$. We consider three correlation structures of the covariate vector $\mathbf{x}_i$:

- Identity: $\mathbf{\Sigma}_{p\times p} = \mathbf{I}_{p\times p}$;
- Autoregressive (AR(1)): $\mathbf{\Sigma}_{p\times p} : (\mathbf{\Sigma})_{jk} = (0.8)^{|j-k|}$;
- Compound symmetry (CS) $\mathbf{\Sigma}_{p\times p}$: $= 0.5\mathbf{I}_{p\times p} + 0.5\mathbf{1_p}\mathbf{1_p}^T$ and $\mathbf{1_p}$ is the $p \times 1$ vector of 1's.

Table 3 displays the estimated biases and coverage probabilities, where the coverage probability is defined as the

TABLE 3
Comparisons of SSHDI With LASSO-Pro and SSLASSO

		Identity		AR(1)		CS	
Index	β_j^*	Bias ($\times 10^{-3}$)	Cov Prob (%)	Bias ($\times 10^{-3}$)	Cov Prob (%)	Bias ($\times 10^{-3}$)	Cov Prob (%)
		-1.77	90.5	10.43	92.5	-0.35	96.5
		-81.78	70.5	-44.09	86	-38.43	92.5
78	1.07	-79.33	90.5	-101.95	84.5	-113.72	92.5
		-1.04	96.5	9.70	92	2.44	95
		-80.28	76	-44.54	87	-32.42	89
102	1.04	-77.72	93.5	-99.66	82	-105.60	92
		-1.62	94	15.58	93.5	-4.67	96.5
		-89.43	71.5	-47.57	88.5	-40.39	91.5
242	1.19	-88.69	87.5	-104.25	84	-115.51	92
		-0.14	94	2.98	96.5	2.01	95
		-75.87	81	-41.40	88	-30.61	91
359	1.43	-80.91	94	-98.14	85	-107.5	89
		-3.57	95.5	0.54	93	5.88	91.5
		-84.86	75	-60.80	88	-24.20	86.5
380	0.62	-85.73	89.5	-111.11	81.5	-99.26	90.5
		-0.46	95	0.65	94.82	3.26	95.16
		-0.40	97	3.16	96.46	5.24	96.34
-	0	-0.27	99.5	4.15	99.69	26.88	99.94

Rows consist of 5 true signals and the average of noise variables. In each cell, top number is for SSHDI; middle number is for LASSO-Pro; lower number is for SSLASSO.

proportion of simulations when the derived 95 percent confidence intervals cover the true parameter β^*. We observe that across the board, SSHDI gives less biased point estimates for the true signals, and provides reliable confidence intervals around the nominal level for both true signals and noise variables. In contrast, both LASSO-Pro and SSLASSO have visible discrepancies in terms of point estimation and inference for the true signals and noise variables.

3.3.1 Multiple Myeloma Genomics Data

We analyzed a cancer genomic data with $n = 163$ multiple myeloma patients [69]. Our interest is to detect association between the β-2 microglobulin (B2M) and gene expressions. B2M is a continuous variable measuring a small membrane protein produced by malignant myeloma cells, indicating the severity of disease. Identifying genes that are related to B2M is clinically important as it helps construct molecular prognostic tools for early diagnosis of disease.

We used the target gene approach KEGG [88] to identify gene pathways that were shown to be related to cancer development and progression. There were $p = 789$ unique probes from the identified pathways and we took the logarithm of both the B2M test value and the gene expressions, respectively, as the response and predictors for model (1). We applied SSHDI with LASSO as the selection method and $B = 500$ re-samples were drawn for smoothing.

Table 4 shows the proposed method offered new biological insights with two significant probes at the 5 percent family-wise error rate level after adjusting for the Bonferroni correction: 204171_at (RPS6KB1) and 202076_at (BIRC2). The two de-biased LASSO estimators found no significant probes. Both detected genes are highly associated with malignant tumor cells: RPS6KB1, member of the ribosomal protein S6 kinase (RPS6K) family, altercation/mutation has been related to numerous types of cancer including breast cancer, colon cancer, non-small-cell lung cancer, and prostate cancer [89], [90], [91], [92]; BIRC2, whose encoded protein is a member of inhibitors of apoptotic proteins (IAPs) that inhibits apoptosis by binding to tumor necrosis factor receptor-associated factors TRAF1 and TRAF2 [93], has been related to lung cancer and lymphoma [94], [95].

3.4 Extensions

The SSHDI method takes advantage of the multi sample-splitting in [96] and the bagging idea in [86]. It is thus

TABLE 4
Top 6 Most Significant and Bottom 6 Least Significant Genes From SSHDI on Multiple Myeloma Genomic Data

Gene	$\widehat{\beta}$	SE	p
204171_at (RPS6KB1)	-0.20	0.042	0.002
202076_at (BIRC2)	-0.17	0.041	0.037
220414_at	-0.20	0.05	0.14
220394_at	-0.18	0.05	0.59
206493_at	-0.19	0.06	0.63
209878_s_at	-0.17	0.05	0.69
...			
207924_x_at	5×10^{-4}	0.07	1
205289_at	-4.4×10^{-4}	0.06	1
203591_s_at	4.7×10^{-4}	0.07	1
224229_s_at	2.4×10^{-4}	0.06	1
217576_x_at	2.5×10^{-4}	0.07	1
201656_at	2.5×10^{-4}	0.08	1

p-values are after Bonferroni adjustment.

fundamentally different from methods based on penalized regressions for high dimensional predictors. As the data split separates selection and estimation, the SSHDI estimator and inferences are not sensitive to the tuning parameters used for variable/model selection, which is a major drawback of the current methods [67], [77]. Furthermore, the variance estimator is free of the parametric model and achieves variance reduction from the effect of bagging.

There are clear computational advantages of the proposed procedure. First, the concise algorithms are justified by theory and straightforward to implement in real data applications. Second, as SSHDI uses multiple data-splitting, it is naturally suitable for parallel computing to greatly speed up the computing time. In particular, the procedure can be paralleled for both the number of splits B and the partial regressions iterating among p covariates, thus taking full advantage of the multi-core CPU or GPU computing. On the other hand, SSHDI performance might scale up with the sample size, since the number of re-samples required is $B = O(n)$. The estimation and inference accuracy also depends on the quality of the model selection results, where the sure screening property is crucial. Further technical details are available in [69], [70] and in a new application that further demonstrates the flexibility of the procedure to make accurate inference from complex data in a different type of medical trial.

Specifically, [70] applied the SSHDI procedure to a lung cancer study with high dimensional genetics data. The goal was to find significant predictors among 13,663 SNPs and SNP-smoking interactions that are associated with lung cancer patients in a case-control study with sample size $N = 1,459$. SSHDI was extended from linear regression to generalized linear models (GLM), and comparison between SSHDI and LASSO-Pro showed that the former was able to identify more significant predictors than the latter (9 versus 2) and with faster computation time.

Because the core of the SSHDI procedure is an aggregation method that involves re-sampling, averaging of base learners, and model-free inferences, it has many applications and extensions. For example, it is straightforward to extend the framework to generalized linear models [70], survival models, mixed effects models, among others. Similar ideas for inferences have also been studied in the context of random forests and predictions [83]. The idea of split and smoothing can also be applied to problems beyond regression models. It can be used for dimension reduction and tackle estimation problems without resorting to penalization but still have desired properties. In addition, the model-free variance estimation and inference approaches add extra flexibility to the framework.

There are recent research studies on understanding the disease and genome sequencing of the SARS-CoV-2 virus [97], [98], including a comprehensive data collection effort on the population level for COVID-19 cases and deaths [99]. There is now a data gathering and repository [100] that helps to model trends of the spread and make predictions of COVID-19 [101]. Recently, [102] proposed a hierarchical agglomerative algorithm for pooled testing with a social graph that could lead to roughly 20-35 percent cost reduction compared to random pooling by using the Dorfman two-stage method when samples within a group are positively correlated. As large scale phenotype and genotype data at the individual level become available in this repository and other sources, we expect HDI analysis to play an important role in understanding the COVID-19 pathology. For example, HDI models will be able to estimate and test the significance of risk factors related to health outcomes in the presence of many confounders. We also see potential extensions of the SSHDI method to prediction and learning problems with a large number of features.

4 Computational Epidemiology Advances

Recent years have seen a proliferation of healthcare-related data inventory and cloud-driven software that are used to solve computationally challenging data science problems, especially those related to computational epidemiology. One example is infection source detection (i.e., searching for superspreaders during the COVID-19 pandemic) or to enforce quarantine measures efficiently, which finds applications in *digital contact tracing* that employs human contact tracers or mobile technologies (e.g., wireless Bluetooth to measure social connectivity) to trace the social contacts of infected person as well as searching for the outbreak origin [6], [7]. These computational epidemiology problems typically involve networks that arise from social relationship and mobility in the network and require algorithmic advances for computational speedup.

Epidemic spreading patterns can be discovered by knowing who infects whom in an outbreak by modeling users and social contacts between users as vertices and edges respectively. However, digital contact tracing in a viral outbreak may lead to huge graphs whose veracity, volume and size may impede a direct use of standard graph algorithms. For example, running standard graph algorithms such as the breadth-first-search algorithm for every single vertex of a massive graph can become computationally impractical. Scalable cloud computing or machine learning can possibly alleviate this data challenge to some extent.

In addition, computational epidemiology problems often have statistical features that cannot be easily modeled mathematically. For example, the social network topology in digital contact tracing may have missing or noisy information and runs the risk of being out-dated. The data may also have local and global statistical dependencies that affect the problem-solving approach and its solution quality. *In general, the inherent statistics of data influences algorithmic tuning and consequentially computational performance.* Designing computational techniques that exploit statistical features is thus key to algorithmic speedup without incurring significant information loss or degraded solution quality. We give an overview of *Network Centrality as Statistical Inference* [103], [104], [105], [106] and Graph Neural Networks as useful frameworks to design scalable algorithms for digital contact tracing and other computational epidemiology optimization problems.

Consider the problem of tracing infection source in [103], [105], [106]: Given a snapshot observation of the social graph with the infected users, who is the Patient Zero that causes the outbreak? Let us model the cascading over a graph G by the susceptible-infectious (SI) model in the epidemiology literature. The SI model assumes that a user once

infected possesses the disease and in turn spreads the disease to one of his or her susceptible neighbors. A snapshot observation of the cascade is G_n, where n is the number of infected users (modelled by the vertices in the graph G_n, which is a subgraph of G). For a given social graph G_n over the underlying graph G, v^* is a maximum likelihood estimator for the source in G_n, i.e., $P(v^*|G_n) = \max_{v_i \in G_n} P(v_i|G_n)$. By Bayes theorem, $P(G_n|v)$ is the probability that v is the actual Patient Zero whose initial infection leads to the social graph data G_n. Now, let σ be a possible spreading order (defined as a sequence of distinct vertices sequence starting from v and containing all the infected subgraph vertices), and let $M(v, G_n)$ be the collection of all the spreading orders starting from v as the source in G_n. The likelihood function is

$$P(G_n|v) = \Sigma_{\sigma \in M(v,G_n)} P(\sigma|v).$$

Given the observation G_N, the node that is most likely to be the epidemic source can be obtained by solving the maximum-likelihood estimation problem: [103], [105], [106]

$$\hat{v} \in \arg\max_{v \in G_N} P(G_N|v). \quad (2)$$

Both the size and the combinatorial nature of the problem makes solving (2) computationally challenging. For instance, when G_N is a general tree graph, (2) is still an open problem even when we consider the simplest Susceptible-Infectious (SI) spreading model [103], [105]. However, if G is an infinite degree-regular tree graph, one can show that $P(\sigma|v)$ for any vertex v is equal, and thus solving (2) reduces to simply counting $M(v, G_n)$ (also known as the rumor centrality [103]). The node with the largest rumour centrality is the rumour center, which is equivalent to the tree centroid in graph theory [106]. However, when G is finite or G_n is a general graph with cycles, then each $P(\sigma|v)$ in (2) is different, and so evaluating the likelihood function requires computing $M(v, G_n)$ and also tracking $P(\sigma|v)$ for all spreading orders, making (2) harder to solve. Apart from identifying special cases in which (2) can be solved optimally (e.g., degree-regular trees with an underlying infinite graph), it is interesting to solve (2) by a network centrality-based approach that allows a graph-theoretic interpretation and retains some of its intuition.

4.1 Reverse Engineering Approach

In *reverse engineering*, we ask: Given a well-known network centrality, what are relevant inference problems to computational epidemiology that they implicitly solve?

The appropriate network centrality can succinctly capture the effect of stochastic processes on the graph, and its algorithms can be useful for computing exact or approximate solution to statistical inference optimization problems. For example, the rumour centrality in [103] is statistically optimal only when the graph is degree-regular, and can serve as a good heuristic to find approximately good solution. Hence, a network centrality perspective can provide guiding principles on algorithm design even when the original problem is hard to solve. The value of reverse engineering thus lies in giving theoretical insights to the solvability of the problem and whether a solution is near-optimal or not.

4.2 Forward Engineering Approach

In *forward engineering*, we ask: Given a stochastic optimization formulation over a network, how to transform it or to decompose it to one whose subproblems are graph-theoretic and can utilize network centrality, then solve or approximate the overall problem? Answering these questions thus entails an algorithmic approach that seeks to simplify the original problem, making the problem-solving methodology scalable to accommodate practical situations and low-complexity data algorithms.

For instance, even though the rumor centrality approach in [103] is optimal only for graphs that are degree-regular, the fact that the rumor center is equivalent to either the distance center [103] or the graph centroid in [106] opens doors to new algorithmic methodology associated with distance centrality or branch weight centrality respectively. This can lead to fast algorithms for processing graphs that are not degree-regular, serving as a good heuristic to solve (2) for the general case. In other words, the forward engineering approach enables the reuse of existing algorithms or a performance comparison between different graph algorithms. It also provides a message-passing (i.e., belief propagation) algorithmic perspective to improve existing network centrality-based algorithms [103], [104], [105], [106].

Another instance of forward engineering is the problem of minimizing the disease spread, where the vaccine centrality is proposed in [106] as an approximation algorithm to solve a statistical estimation problem. The approach of using network centrality as a statistical tool for inference can be generalized from a static network to time-dependent networks when real-time data or more accurate spreading models for COVID-19 are available [107], [108], [109]. Finding the appropriate network centrality to explain flow patterns or temporal scales of changes in the network is of practical importance. There are also connections between network centrality as statistical inference and graph signal processing, which include methods for sampling, filtering or machine learning over graphs. The confluence of these research directions can lead to mathematically rigorous graph analytics for analyzing contact tracing and other computational epidemiology problems in large networks.

4.3 Graph Neural Network Approach

In this section, we describe a graph neural network (GNN) learning methodology to solve (2). The idea of GNN learning is to encode structured features of the graph data into a neural network by applying recurrent layers to each node of the graph along with some form of approximation as in recent applications of GNN to combinatorial optimization over graphs [110]. As the neural network weights are trained using semi-supervised examples with labels to capture structural properties of all the nodes in the graph input data, GNNs can be leveraged to address problems related to network centrality. An advantage of using GNN lies in complexity reduction. For example, graph algorithms like the Breadth-First Search algorithm have complexity $O(N + |E|)$ where E is the edge set of the graph. To compute a solution to (2) with low complexity, one approach is to approximate the spreading order probabilities in the graph instead of keeping track of all possible probabilities. Another

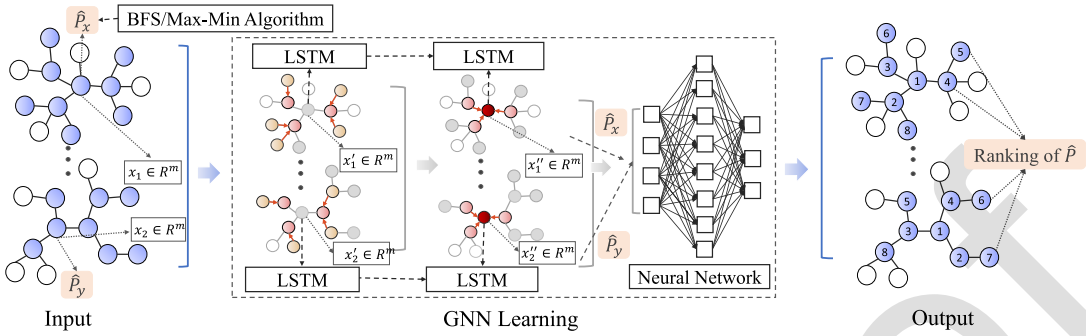

Fig. 1. The overall architecture of an epidemic source inference by node regression using a graph neural network. The input data is a number of smaller contact tracing networks, where each node has a few structural features, labeled with an approximation of the permutation probability. We then use the GraphSage algorithm in [111] with LSTM aggregators as our training model to output a prediction of the spreading order probability for each node as the solution of (2) for the input of bigger networks.

approach is dimension reduction by ignoring nodes near to the graph boundary when the graph is sufficiently large so that accurate GNN models can be trained by only nodes of interests in order to reduce computational time.

The training stage of GNN is important to capture simultaneously the inherent statistical and topological features of a graph data. One possible way is to first generate a training set using a number of graphs that are typically small in size (e.g., hundreds of nodes) and to augment descriptors with the structural features for each node of the graphs as the input data. For these smaller graphs, we solve (2) to find approximately the permutation probabilities of these nodes that are then used as the training set labels for the GNN regression as shown by the GNN architecture in Fig. 1.

4.3.1 Node Feature Selection and Labeling

There are several possibilities to identify useful features and construct labels for each node of the graph input data. Let us consider some basic graph-theoretic features such as the degree and distance. For example, given a snapshot of epidemic network (with the vertex set $\mathcal{V}$) as shown in Fig. 2, some node features can be obtained as follows:

Degree Ratio. This is the ratio of the degree of a node v_i, say $d(v_i)$, to the sum of the degrees of all the other nodes

$$r(v_i) = \frac{d(v_i)}{\sum_{v\in\mathcal{V}} d(v)}.$$

For example, the degree rate of nodes 1,2,3,4, and 5 in Fig. 2 are $\frac{1}{16}, \frac{3}{32}, \frac{1}{8}, \frac{5}{32}$ and $\frac{3}{16}$, respectively.

Infected Proportion Ratio. This is the ratio of the number of infected neighbors of a node to the sum of the uninfected nodes in the graph

$$\widehat{r}(v_i) = \frac{\widetilde{d}(v_i)}{\sum_{v\in\mathcal{V}} \widetilde{d}(v_i)}.$$

For example, the infected proportion ratio of nodes 1,2,3,4, and 5 in Fig. 2 is $\frac{1}{12}, \frac{1}{12}, \frac{1}{4}, \frac{1}{12}$, and $\frac{1}{12}$, respectively.

Labels. The training label of a node is an approximation or averaging of its permutation probabilities obtained by solving (2) using any network centrality algorithm or standard graph algorithms (e.g., Breadth-First Search).

4.3.2 Node Regression Using Graph Convolutional Network

Once the feature selection stage is completed, we train the GNN to learn a function by generating the node embedding based on the selected features and the topological structures of each node in the graph data by some form of iterative updates [112]. At each layer of the neural networks, the vertex v in GNN can be updated as follows:

$$\begin{aligned}\alpha_v^{(k)} =& \text{Aggregate}^{(k)}(\{\beta_u^{(k-1)} : u \in \mathcal{N}(v)\}, \{\gamma_x : x \in \varepsilon(v)\}) \\ \beta_v^{(k)} =& \text{Combine}^{(k)}(\beta_u^{(k-1)}, \alpha_v^{(k)}),\end{aligned} \tag{3}$$

where $\mathcal{N}(v)$ denotes the set of the neighbors of v, $\varepsilon(v)$ denotes the set of edges with v as one end node, $\beta_v^{(k)}$ denotes the kth layer's output feature of vertex v, and $\alpha_v^{(k)}$ is the aggregate iterate. The learning process can be accomplished by inductive graph neural network training, e.g., GraphSage in [111], and the following LSTM aggregator:

$$\alpha_v^{(k)} = \text{LSTM}(\{\beta_u^{(k-1)} : u \in \mathcal{N}(v)\}),$$

and the Rectified Linear Unit (ReLU) combination function

$$\beta_v^{(k)} = \max\left\{0, W^{(k)} \cdot \frac{1}{|\mathcal{N}(v)|+1} \sum_u \beta_v^{(k-1)}\right\},$$

where $u \in \mathcal{N}(v) \cup \{v\}$ and $\{W^{(k)}\}$ are the weight matrices to be updated. This GNN framework can be extended with more advanced deep learning techniques or integrated with other network centrality-based algorithms.

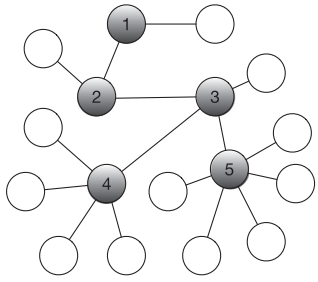

Fig. 2. An infection network with a degree-irregular tree as the underlying graph, where the degrees of the shaded (infected) nodes 1,2,3,4,5 are 2,3,4,5,6, respectively.

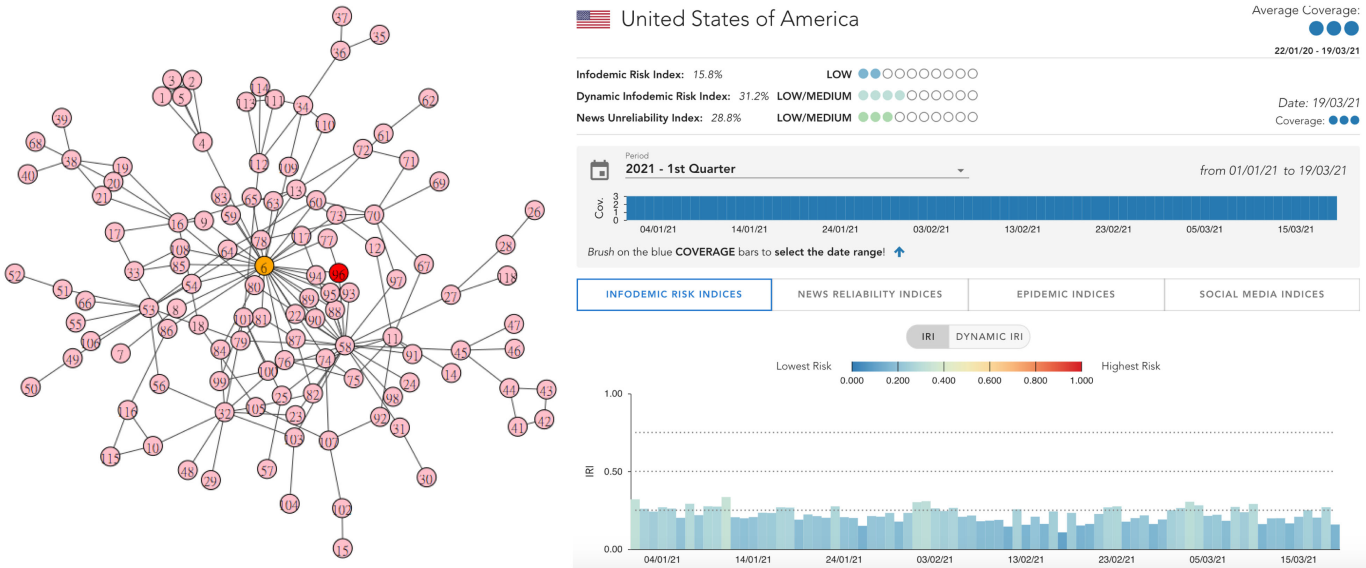

Fig. 3. (a) SARS-CoV2003 Contact Tracing Network in Taiwan. Each vertex represents either a confirmed case or a hospital. The orange and red vertices represent the estimated source determined by contact tracing algorithms in [104] and in [103] respectively. (b) Photo courtesy of the World Health Organization Infodemic risk management system in [8] to assess the spread of COVID-19 related misinformation.

We describe briefly a Contact Tracing Algorithm 4 in [104] that uses a weighted distance centrality measures where the weights are computed for node regression using the aforementioned GNN approach to solve (2). As an illustration, by reconstructing the contact tracing network data of SARS-CoV2003 (a virus very similar to the COVID-19 coronavirus) in Taiwan [104], this algorithm can correctly identify the first place, Taipei Municipal Heping Hospital (now Taipei City Hospital Heping Branch), of an infection cluster in April 2003 in Taiwan as compared to a breadth-first search heuristic in [103], which chooses the red vertex modeling a confirmed index case (not the first case) who had been to the Taipei Municipal Heping Hospital as shown in Fig. 3a. In addition, the network centrality approach can enable visualization tools (e.g., dot distribution map) to visualize the likelihood of infection source that may be of value to public healthcare policymakers.

Algorithm 4. Contact Tracing Algorithm by Ranking Infection Source by Network Centrality and GNN

Input: Infection networks $\{G_1, G_2, \ldots, G_N\}$ harvested by forward contact tracing

Output: A ranking of the outbreak source probability of each node for backward contact tracing

1: Calculate the structure features (e.g., degree ratio and infected proportion ratio) for each node in the network.

2: Calculate the labels (i.e., approximate solution to (2)) for each node according to a network centrality algorithm (e.g., rumor centrality in [103], [105] or statistical distance centrality algorithm in [104]).

3: Train the regression model by GraphSage [111] to output a ranking of the probability for each node to be the outbreak source in each network $\{G_1, G_2, \ldots, G_N\}$.

The problem in (2) can potentially be useful to other kinds of computational epidemiology problems, such as COVID-19 Infodemic management as introduced by the World Health Organization. In [8], an Infodemic risk management system, as shown in Fig. 3b, is developed to visualize and assess the spread of misinformation concerning vaccine and erroneous treatment. Given the volume of COVID-19 related misinformation, finding a rumor source can help limit the damage and spread of false information [8]. In summary, the network centrality as statistical inference and GNN machine learning frameworks are examples of first step towards a theoretically sound and computationally efficient approach to digital contact tracing and computational epidemiology.

5 Metaheuristics

This section briefly reviews the increasing role of metaheuristics in big data research. There are many metaheuristic algorithms and for space consideration, we consider a popular subclass of them called nature-inspired metaheuristic algorithms. These algorithms are widely used to tackle high-dimensional and complex optimization problems in engineering and computer science, and are increasingly used in other disciplines. They are appealing for several reasons. They are general-purpose optimization algorithms, assumptions-free, fast, powerful and easy to implement for solving all kinds of complex optimization problems with hundreds or thousands of variables. Codes for many of them are available in MatLab, and freely on many websites. Their recent meteoric rise in popularity in industry and even in academia is nicely documented in [113], [114]. The promise and excitement of investigating how well metaheuristics performs for tackling problems with millions of variables was the focus of a special issue in Information Sciences [115].

Some examples of nature-inspired metaheuristic algorithms that seem more popular are Particle Swarm optimization (PSO), Differential Evolution (DE), Imperialist Competitive Algorithm (ICA), Competitive Swarm Optimizer (CSO) and Cuckoo search, just to name a few. Each algorithm has a different motivation from nature and works

differently. A commonality is that each algorithm has a few tuning parameters and a few stochastic elements and each has its own way of updating its trajectory via a couple of equations that model a natural phenomenon or an animal's behavior. For example, PSO mimics a flock of birds fling in the sky and looking for food on the ground. Each bird has its own idea where the food is (local optimum) but they communicate with one another and collectively make a decision where the food is on the ground (global optimum) and each bird flies toward it without completely relinquishing its take where the food is (local optimum). [116] gave a concise description of many such algorithms with illustrative applications and sample codes. However, theoretical properties and rigorous proofs of convergence for metaheuristic algorithms are generally elusive but they remain popular because of their widely reported ability to find an optimum or a nearly optimum for all kinds of optimization problems [113], [114].

In the last decade, medical researchers have increasingly resorted to and continue to use metaheuristic for tackling all kinds of medical problems that involve optimization. A very common problem is to use clinical data and cluster patients into various categories of disease progression given baseline data. For example, we want to know in six months, whether a patient will likely have a stable disease, or whether the patient will improve or deteriorate. Such problems are challenging because the data set is large, and there is a large number of different types of explanatory variables potentially useful for predicting outcomes accurately. The problem then translates to selecting a small number of features in the whole data set that best predict the outcome of interest. For instance, [117] predicted disease progression in Idiopathic Pulmonary Fibrosis patients by combining random forest and Quantum Particle Swarm Optimization. This is an increasingly common optimizing technique where one algorithm is hybridized with another to enhance the search capability by exploiting the particular strengths of the two algorithms so that the hybridized version performs better than each of the individual algorithms. The algorithms involved can be two or more and they can be metaheuristic or deterministic. A recent application is [118], who showed the Grey Wolf algorithm can be hybridized with PSO for accelerated convergence. A monograph on hybridized algorithms for enhanced performance with applications is [119].

5.1 Metaheuristics for the COVID-19 Pandemic

The impact of metaheuristic algorithms, hybridized or not, can be seen in its increasing use in many sub-specialties in medicine and beyond. For instance, in cardiology research, [120] applied ICA to optimally select a minimum number of features best for diagnosing heart problems. Similarly, [121] used a modified DE algorithm and [122] used PSO and a Bayesian paradigm to predict heart diseases. In systems biology, [123] applied PSO to select biological model and estimate parameters in the model, and [124] gave a review of metaheuristics for estimating parameters. Likewise, they are also increasingly used in disciplines that traditionally rely on analytical approaches. For example, there is notable and recent use of metaheuristics to find optimal experimental designs for nonlinear regression models in the biostatistical literature. Some examples are [125], [126], [127], [128], who respectively applied quantum PSO, DE and a modified CSO algorithm, to search for various types of optimal designs for generalized linear models with several interacting factors and some have random effects.

Not surprisingly, nature-inspired algorithms have been promptly applied to better understand the various aspects of COVID-19. Various such algorithms were used and they include PSO, DE and ICA to tackle different aspects of the pandemic. For example, [129] implemented ICA to predict trends in the COVID-19 pandemic in Hungary, [130] used DE to monitor spread of the COVID-19 virus in Italy, [131], [132] applied PSO to estimate model parameters in SEIR models or used PSO to use real time data to estimate and predict death rates caused by COVID-19, and [133] used DE to classify COVID-19 patients from chest CT images. Most recently, [134] proposed a COVID-19 optimizer algorithm specifically for modeling and controlling the coronavirus distribution process and one its objectives is to minimize the number of infected countries to slow down the spread. The authors also showed their algorithm outperformed PSO and GA by 53 and 59 percent, respectively, and newer created metaheuristic algorithms, like the Volcano Eruption Algorithm and the Grey Wolf optimizer, by 15 and 37 percent, respectively.

Pareto Optimization (PO) is a common approach to solve optimization problems with multiple objectives. [135] applied PO to tackle problems posed by COVID-19, which can infect many people quickly resulting in huge and sudden requests of medical care at various levels. Coping with how, when and where to admit COVID-19 patients efficiently is a complex multiobjective optimization problem. For instance, to decrease the in-bed time, save lives and resources, the choice of the most suitable hospital for the patient has to be balanced by expected admission time, hospital readiness and severity of the COVID-19 patient. These are multiobjective optimization problems and the author showed their strategy using data from 254 patients in Saudi Arabia outperformed the lexicographic multiobjective optimization method. Recently, evolutionary algorithms have made remarkable progress in tackling many types of multiobjective optimization problems [136], [137], [138] and we expect metaheuristic algorithms will make important contributions to solve COVID-19 multiobjective optimization problems, especially when combined with the latest machine learning advances for tackling COVID-19 problems [139], [140].

Metaheuristics is not a panacea for optimization problems. A perennial problem is how to tune the parameters for accelerated convergence and ensure that the algorithm converges to the theoretical global optimum. Both issues have been active research areas for a long time; recent advances include [141], [142]. Other open healthcare problems in metaheuristics are described in [143].

6 Conclusion

We provide an overview of innovative analytic approaches for gaining insights into big data problems. We focus on handling healthcare issues relevant to the current pandemic

indicated in Table 1 and reiterate that the methodologies are applicable to other types of big data problems.

There are open challenges in data science for healthcare diagnosis, inference, and pandemic response. For example, consider the statistical and computational issues for digital contact tracing and its applications in epidemiology. It is important to have accurate infection spreading models and parameters before robust predictive analytics can be developed to solve the large-scale problems. This means that we have to address challenges for high-fidelity computational algorithms and statistical exploration of the data, where new principles are needed to combine these two aspects. Examples of open issues are: can existing network centrality be reverse-engineered to find optimal estimates for the parameters of most interests in large-scale infection spreading like COVID-19 pandemic? A forward-engineering approach may create new forms of network centrality that possess desirable statistical or computational traits for solving the problem. Can machine learning techniques and massive graph neural networks provide an impetus for breakthrough technologies in analyzing past pandemic behaviors to fight against newly-emerging pandemics? Future research in big data for the health sciences concern three important areas of applications of machine- and deep-learning approaches in modern drug development, namely, adverse event detection, trial recruitment optimization, and clinical drug repurposing, including big data analytics for various stages in drug discovery and development [144].

We close with two remarks. First, page limits and the breadth of the field did not allow us to discuss all relevant topics. Some omitted topics include social media text data [145], virus lineage [146], public health data monitoring [147] and analysis of administrative and translation data [148]. Second, analysis for big data may alternatively begin with a selected subset of the big data with some optimality properties; see [149]. Modern statistical methods are then applied to infer the key messages from the subset data to the massive data. Invariably, the task to properly analyze big data is challenging and requires collaboration among statisticians, engineers and computer scientists to jointly create powerful computing environments, new software, data platforms, data integration systems and state-of-the-art statistical and machine learning techniques.

Acknowledgments

The authors would like to thank three anonymous referees and the associate editor for their thorough review of their manuscript and constructive feedback. This work was supported in part by a grant from the Hong Kong ITF Project No. ITS/188/20, UGC Teaching Award Project No. 6989041 and an Institute for Pure and Applied Mathematics (IPAM) Senior Fellowship.

References

[1] J. Andreu-Perez, C. C. Poon, R. D. Merrifield, S. T. Wong, and G.-Z. Yang, "Big Data for health," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 4, pp. 1193–1208, Jul. 2015.

[2] C. Engledowl and T. Weiland, "Data (mis)representation and COVID-19: Leveraging misleading data visualizations for developing statistical literacy across grades 6–16," *J. Statist. Data Sci. Educ.*, vol. 1-6, pp. 1–6, 2021.

[3] E. Smirnova, A. Ivanescu, J. Bai, and C. M. Crainiceanu, "A practical guide to Big Data," *Stat. Probability Lett.*, vol. 136, pp. 25–29, 2018.

[4] P. Diggle, "Statistics: A data science for the 21st century," *J. Roy Stat. Soc. A*, vol. 178, no. 4, pp. 793–813, 2015.

[5] C. K. Ettman, S. M. Abdalla, G. H. Cohen, L. Sampson, P. M. Vivier, and S. Galea, "Prevalence of depression symptoms in US adults before and during the COVID-19 pandemic," *JAMA Netw. Open*, vol. 3, no. 9, 2020, Art. no. e2019686. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2020.19686

[6] S. Kojaku, L. Hébert-Dufresne, E. Mones, S. Lehmann, and Y.-Y. Ahn, "The effectiveness of backward contact tracing in networks," *Nat. Hum. Behav.*, vol. 17, pp. 652–658, 2021.

[7] G. Cencetti *et al.*, "Digital proximity tracing on empirical contact networks for pandemic control," *Nat. Commun.*, vol. 12, no. 1655, 2021, Art. no. e2019686.

[8] R. Gallotti, F. Valle, N. Castaldo, P. Sacco, and M. D. Domenico, "Assessing the risks of 'infodemics' in response to COVID-19 epidemics," *Nat. Hum. Behav.*, vol. 4, pp. 1285–1293, 2020.

[9] J. L. Torrecilla and J. Romo, "Data learning from Big Data," *Stat. Probability Lett.*, vol. 136, pp. 15–19, 2018.

[10] A. Gandomi and M. Haider, "Beyond the hype: Big Data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, pp. 137–144, 2015.

[11] I. Lee, "Big data: Dimensions, evolution, impacts, and challenges," *Bus. Horiz.*, vol. 60, pp. 293–303, 2017.

[12] D. P. Harrington, "The randomized clinical trial," *J. Amer. Stat. Assoc.*, vol. 95, no. 449, pp. 312–315, 2000.

[13] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

[14] D. M. Cash, J. D. Rohrer, N. S. Ryan, S. Ourselin, and N. C. Fox, "Imaging endpoints for clinical trials in Alzheimer's disease," *Alzheimer's Res. Ther.*, vol. 6, no. 9, 2014, Art. no. 87.

[15] D. Bar-Zohar, F. Agosta, D. Goldstaub, and M. Filippi, "Magnetic resonance imaging metrics and their correlation with clinical outcomes in multiple sclerosis: A review of the literature and future perspectives," *Mul. Scler. J.*, vol. 42, no. 6, pp. 719–727, 2008.

[16] M. Haris *et al.*, "Molecular magnetic resonance imaging in cancer," *J. Transl. Med.*, vol. 13, pp. 313–313, 2015.

[17] S. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. Khan, "Medical image analysis using convolutional neural networks: A review," *J. Med. Syst.*, vol. 42, pp. 226–226, 2018.

[18] L. M. Barbak *et al.*, "Traditional and digital biomarkers: Two worlds apart?," *Digit. Biomarkers*, vol. 3, pp. 92–102, 2019.

[19] O. Sverdlov, J. van Dam, K. Hannesdottir, and T. Thornton-Wells, "Digital therapeutics: An integral component of digital innovation in drug development," *Clin. Pharmacol. Therapeutics*, vol. 104, no. 1, pp. 72–80, 2018.

[20] T. M. Polasek, S. Shakib, and A. Rostami-Hodjegan, "Precision dosing in clinical medicine: Present and future," *Expert Rev. Clin. Pharmacol.*, vol. 11, no. 8, pp. 743–746, 2018.

[21] P.-Y. Benhamou *et al.*, "Closed-loop insulin delivery in adults with type 1 diabetes in real-life conditions: A 12-week multicentre, open-label randomised controlled crossover trial," *Lancet Digit. Health*, vol. 1, pp. e17–e25, 2019.

[22] A. Caliebe, F. Leverkus, G. Antes, and M. Krawczak, "Does Big Data require a methodological change in medical research?," *BMC Med. Res. Methodol.*, vol. 19, 2019, Art. no. 125.

[23] A. J. Ferrari *et al.*, "Burden of depressive disorders by country, sex, age, and year: Findings from the global burden of disease study," *PloS Med.*, vol. 10, 2010, Art. no. e1001547.

[24] M. Hamilton, "A rating scale for depression," *J. Neurol., Neurosurgery, Psychiatry*, vol. 23, pp. 56–62, 1960.

[25] S. A. Montgomery and M. Åsberg, "A new depression scale designed to be sensitive to change," *British J. Psychiatry*, vol. 134, pp. 382–389, 1979.

[26] O. Sverdlov *et al.*, "A study of novel exploratory tools, digital technologies, and central nervous system biomarkers to characterize unipolar depression," *Front. Psychiatry*, vol. 12, 2021, Art. no. 640741. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyt.2021.640741

[27] C. Bourke, K. Douglas, and R. Porter, "Processing of facial emotion expression in major depression: A review," *Australian New Zealand J. Psychiatry*, vol. 44, pp. 681–696, 2010.

[28] Y. Ozkanca, M. G. Öztürk, M. Nur Ekmekci, D. C. Atkins, C. Demiroglu, and R. Hosseini Ghomi, "Depression screening from voice samples of patients affected by Parkinson's disease," *Digit. Biomarkers*, vol. 3, pp. 72–82, 2019.

[29] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *British J. Psychiatry*, vol. 65, no. 6, pp. 386–408, 1958.

[30] G. Cybenko, "Approximations by superpositions of sigmoidal functions," *Math. Controls, Signals, Syst.*, vol. 2, no. 4, pp. 303–314, 1989.

[31] B. Efron and T. Hastie, *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. New York, NY, USA: Cambridge Univ. Press, 2016.

[32] T. D. Pham, "A comprehensive study on classification of COVID-19 on computed tomography with pretrained convolutional neural networks," *Sci. Rep.*, vol. 10, 2020, Art. no. 16942.

[33] L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," *Sci. Rep.*, vol. 10, 2020, Art. no. 19549.

[34] H. Mukherjee, S. Ghosh, A. Dhar, S. M. Obaidullah, K. C. Santosh, and K. Roy, "Deep neural network to detect COVID-19: One architecture for both CT scans and chest X-rays," *Appl. Intell.*, vol. 51, pp. 2777–2789, 2021.

[35] H. T. Rauf *et al.*, "Time series forecasting of COVID-19 transmission in Asia Pacific countries using deep neural networks," *Personal Ubiquitous Comput.*, vol. 10, pp. 1–18, 2021.

[36] A. Hassan, I. Shahin, and M. B. Alsabek, "COVID-19 detection system using recurrent neural networks," in *Proc. Int. Conf. Commun., Comput., Cybersecur., Informat.*, 2020, pp. 1–5.

[37] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *J. Mach. Learn. Res.*, vol. 3, pp. 115–143, 2002.

[38] J. Bayer, D. Wierstra, J. Togelius, and J. Schmidhuber, "Evolving memory cell structures for sequence learning," in *Proc. Int. Conf. Artif. Neural Netw.*, 2009, pp. 755–764.

[39] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[40] J. Kennedy and R. Eberhart, "Particle swarm optimization," *Proc. IEEE Int. Conf. Neural Netw.*, 1995, pp. 1942–1948.

[41] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Mag Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[42] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," *Proc. Roy. Soc. London.*, vol. 115, no. 772, pp. 700–721, 1927.

[43] R. Dandekar, C. Rackauckas, and G. Barbastathis, "A machine learning aided global diagnostic and comparative tool to assess effect of quarantine control in COVID-19 spread," *Patterns*, vol. 1, no. 9, 2020, Art. no. 100145.

[44] R. A. Dandekar *et al.*, "Safe blues: A method for estimation and control in the fight against COVID-19," *medRxiv*, 2020. [Online]. Available: https://doi.org/10.1101/2020.05.04.20090258

[45] C. Rackauckas, M. Innes, Y. Ma, J. Bettencourt, L. White, and V. Dixit, "DiffEqFlux.jl – A Julia library for neural differential equations," 2019, *arXiv:1902.02376*.

[46] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," in *Adv. Neural Inf. Process. Syst.*, 2018, pp. 6571–6583.

[47] C. Rackauckas *et al.*, "Universal differential equations for scientific machine learning," 2020, *arXiv:2001.04385*.

[48] H. Nakagawa and M. Fujita, "Whole genome sequencing analysis for cancer genomics and precision medicine," *Cancer Sci.*, vol. 109, no. 3, pp. 513–522, 2018.

[49] J. N. Rosenbaum *et al.*, "Genomic heterogeneity of ALK fusion breakpoints in non-small-cell lung cancer," *Modern Pathol.*, vol. 31, no. 5, 2018, Art. no. 791.

[50] W. Y. Wang, B. J. Barratt, D. G. Clayton, and J. A. Todd, "Genome-wide association studies: Theoretical and practical concerns," *Nat. Rev. Genet.*, vol. 6, no. 2, pp. 109–118, 2005.

[51] J. Wu, B. Devlin, S. Ringquist, M. Trucco, and K. Roeder, "Screen and clean: A tool for identifying interactions in genome-wide association studies," *Genet. Epidemiol. Official Pub. Int. Genet. Epidemiol. Soci.*, vol. 34, no. 3, pp. 275–285, 2010.

[52] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Stat. Soc. Ser. B Stat. Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[53] S. B. Kotsiantis, "Decision trees: A recent overview," *Artif. Intell. Rev.*, vol. 39, no. 4, pp. 261–283, 2013.

[54] N. Prasad and M. M. Naidu, "Gain ratio as attribute selection measure in elegant decision tree to predict precipitation," in *Proc. 8th EUROSIM Congr. Model. Simul.*, 2013, pp. 141–150.

[55] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[56] R. E. Schapire, "The boosting approach to machine learning: An overview," *Nonlinear Estimation Classification*, vol. 171, pp. 149–171, 2003.

[57] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 785–794.

[58] S. Portnoy, "Asymptotic behavior of M estimators of p regression parameters when p^2/n is large; ii. normal approximation," *Annal. Stat.*, vol. 13, no. 4, pp. 1403–1417, 1985.

[59] W. Niemiro, "Asymptotics for m-estimators defined by convex minimization," *Annal. Stat.*, vol. 20, no. 3, pp. 1514–1533, 1992.

[60] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Stat. Assoc.*, vol. 96, pp. 1348–1360, 2001.

[61] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Stat. Soc Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.

[62] E. Candès and T. Tao, "The Dantzig selector: Statistical estimation when P is much larger than n," *Annal. Stat.*, vol. 35, no. 6, pp. 2313–2351, 2007.

[63] J. Huang, S. Ma, and C.-H. Zhang, "Adaptive LASSO for sparse high-dimensional regression models," *Stat. Sinica*, vol. 18, pp. 1603–1618, 2008.

[64] J. Lv and Y. Fan, "A unified approach to model selection and sparse recovery using regularized least squares," *Annal. Stat.*, vol. 37, no. 6A, pp. 3498–3528, 2009.

[65] C.-H. Zhang and S. S. Zhang, "Confidence intervals for low dimensional parameters in high dimensional linear models," *J. Roy. Stat. Soc. Ser.B Stat. Methodol.*, vol. 76, no. 1, pp. 217–242, 2014.

[66] A. Javanmard and A. Montanari, "Confidence intervals and hypothesis testing for high-dimensional regression," *J. Mach. Learn. Res.*, vol. 15, pp. 2869–2909, 2014.

[67] P. Bühlmann, M. Kalisch, and L. Meier, "High-dimensional statistics with a view toward applications in biology," *Ann. Rev. Stat. Appl.*, vol. 1, pp. 255–278, 2014.

[68] R. Dezeure *et al.*, "High-dimensional inference: Confidence intervals, p-values and R-software HDI," *Stat. Sci.*, vol. 30, no. 4, pp. 533–558, 2015.

[69] Z. Fei, J. Zhu, M. Banerjee, and Y. Li, "Drawing inferences for high-dimensional linear models: A selection-assisted partial regression and smoothing approach," *Biometrics*, vol. 75, no. 2, pp. 551–561, 2019.

[70] Z. Fei and Y. Li, "Estimation and inference for high dimensional generalized linear models: A splitting and smoothing approach," *J. Mach. Learn. Res.*, vol. 22, no. 58, pp. 1–32, 2021.

[71] L. Wang, X. Ren, and Q. Gu, "Precision matrix estimation in high dimensional Gaussian graphical models with faster rates," in *Proc. Artif. Intell. Stat.*, 2016, pp. 177–185.

[72] P.-L. Loh *et al.*, "High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination," *Electronic J. Stat.*, vol. 12, no. 1, pp. 1429–1467, 2018.

[73] A. Belloni, V. Chernozhukov, and C. Hansen, "Inference on treatment effects after selection among high-dimensional controls," *Rev. Econ. Studies*, vol. 81, no. 2, pp. 608–650, 2014.

[74] A. Belloni, V. Chernozhukov, and Y. Wei, "Post-selection inference for generalized linear models with many controls," *J. Bus. Econ. Stat.*, vol. 34, no. 4, pp. 606–619, 2016.

[75] A. Belloni, V. Chernozhukov, and K. Kato, "Valid post-selection inference in high-dimensional approximately sparse quantile regression models," *J. Amer. Stat. Assoc.*, vol. 114, no. 526, pp. 749–758, 2019.

[76] J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor, "Exact post-selection inference, with application to the Lasso," *Annal. Stat.*, vol. 44, no. 3, pp. 907–927, 2016.

[77] Y. Ning and H. Liu, "A general theory of hypothesis tests and confidence regions for sparse high dimensional models," *Annals Stat.*, vol. 45, no. 1, pp. 158–195, 2017.

[78] E. X. Fang, Y. Ning, and H. Liu, "Testing and confidence intervals for high dimensional proportional hazards models," *J. Roy. Stat. Soc. Ser. B Stat. Methodol.*, vol. 79, no. 5, pp. 1415–1437, 2017.

[79] Y. Zhu and J. Bradic, "Linear hypothesis testing in dense high-dimensional linear models," *J. Amer. Stat. Assoc.*, vol. 113, no. 524, pp. 1583–1600, 2018.

[80] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[81] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *J. Roy. Stat. Soc. Ser. B Stat. Methodol.*, vol. 70, no. 5, pp. 849–911, 2008.

[82] B. Efron, "Estimation and accuracy after model selection," *J. Amer, Stat. Assoc.*, vol. 109, no. 507, pp. 991–1007, 2014.

[83] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *J. Amer. Stat. Assoc.*, vol. 113, no. 523, pp. 1228–1242, 2018.

[84] P. Zhao and B. Yu, "On model selection consistency of Lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, 2006.

[85] J. Fan and R. Song, "Sure independence screening in generalized linear models with np-dimensionality," *Annal. Stat.*, vol. 38, no. 6, pp. 3567–3604, 2010.

[86] P. Bühlmann and B. Yu, "Analyzing bagging," *Annal. Stat.*, vol. 30, no. 4, pp. 927–961, 2002.

[87] S. Van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure, "On asymptotically optimal confidence regions and tests for high-dimensional models," *Annal. Stat.*, vol. 42, no. 3, pp. 1166–1202, 2014.

[88] M. Carlson, *hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2)*, 2015, R package version 3.2.2.

[89] C. S. Sinclair, M. Rowley, A. Naderi, and F. J. Couch, "The 17q23 amplicon and breast cancer," *Breast Cancer Res. Treat.*, vol. 78, no. 3, pp. 313–322, 2003.

[90] M. L. Slattery, A. Lundgreen, J. S. Herrick, and R. K. Wolff, "Genetic variation in RPS6KA1, RPS6KA2, RPS6KB1, RPS6KB2, and PDK1 and risk of colon or rectal cancer," *Mutat. Res./Fundam. Mol. Mechanisms Mutagenesis*, vol. 706, no. 1, pp. 13–20, 2011.

[91] Y. Zhang, H.-J. Ni, and D.-Y. Cheng, "Prognostic value of phosphorylated mTOR/RPS6KB1 in non-small cell lung cancer," *Asian Pacific J. Cancer Prevention*, vol. 14, no. 6, pp. 3725–3728, 2013.

[92] C. Cai *et al.*, "miR-195 inhibits tumor progression by targeting RPS6KB1 in human prostate cancer," *Clin. Cancer Res.*, vol. 21, no. 21, pp. 4922–4934, 2015.

[93] M. Saleem, M. I. Qadir, N. Perveen, B. Ahmad, U. Saleem, and T. Irshad, "Inhibitors of apoptotic proteins: New targets for anticancer therapy," *Chem. Biol. Drug Des.*, vol. 82, no. 3, pp. 243–251, 2013.

[94] Y. Wang, Q. Dong, Q. Zhang, Z. Li, E. Wang, and X. Qiu, "Overexpression of yes-associated protein contributes to progression and poor prognosis of non-small-cell lung cancer," *Cancer Sci.*, vol. 101, no. 5, pp. 1279–1285, 2010.

[95] R. Rahal *et al.*, "Pharmacological and genomic profiling identifies NF-κB-targeted treatment strategies for mantle cell lymphoma," *Nat. Med.*, vol. 20, no. 1, pp. 87–92, 2014.

[96] N. Meinshausen, L. Meier, and P. Bühlmann, "P-values for high-dimensional regression," *J. Amer. Stat. Assoc.*, vol. 104, no. 488, pp. 1671–1681, 2009.

[97] P. Forster, L. Forster, C. Renfrew, and M. Forster, "Phylogenetic network analysis of SARS-CoV-2 genomes," *Proc. Nat. Acad. Sci.*, vol. 117, no. 17, pp. 9241–9243, 2020.

[98] Y.-Z. Zhang and E. C. Holmes, "A genomic perspective on the origin and emergence of SARS-CoV-2," *Cell*, vol. 181, no. 2, pp. 223–227, 2020.

[99] CDC., "COVID-19 projections," Accessed: Aug. 30, 2020. [Online]. Available: https://covid.cdc.gov/covid-data-tracker/#cases

[100] F. B. Hamzah *et al.*, "Coronatracker: Worldwide COVID-19 outbreak data analysis and prediction," *Bull World Health Organ*, vol. 1, 2020, Art. no. 32.

[101] F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus COVID-19," *PloS One*, vol. 15, no. 3, 2020, Art. no. e0231236.

[102] Y.-J. Lin, C.-H. Yu, T.-H. Liu, C.-S. Chang, and W.-T. Chen, "Positively correlated samples save pooled testing costs," *IEEE Trans. Netw. Sci. Eng.*, to be published, doi: 10.1109/TNSE.2021.3081759.

[103] D. Shah and T. Zaman, "Rumors in a network: Who's the culprit?," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5163–5181, Aug. 2011.

[104] P. Yu, C. W. Tan, and H. Fu, "Epidemic source detection in contact tracing networks: Epidemic centrality in graphs and message passing algorithms," 2020, *arXiv:2006.11913v2*.

[105] Z. Wang, W. Dong, W. Zhang, and C. W. Tan, "Rumor source detection with multiple observations: Fundamental limits and algorithms," in *Proc. ACM Int. Conf. Measur. Model. Comput. Syst.*, 2014, pp. 1–13.

[106] P. Yu, C. W. Tan, and H. L. Fu, "Averting cascading failures in networked infrastructures: Poset-constrained graph algorithms," *IEEE J. Sel. Top. Signal Process.*, vol. 12, no. 4, pp. 733–748, Aug. 2018.

[107] Y.-C. Chen, P.-E. Lu, C.-S. Chang, and T.-H. Liu, "A time-dependent sir model for COVID-19 with undetectable infected persons," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 3279–3294, Oct.–Dec. 2020.

[108] R. Eletreby, Y. Zhuang, K. M. Carley, O. Yagan, and H. V. Poor, "The effects of evolutionary adaptations on spreading processes in complex networks," *Proc. Nat. Acad. Sci.*, vol. 117, no. 11, pp. 5664–5670, 2020.

[109] O. Yagan, A. Sridhar, R. Eletreby, S. Levin, J. B. Plotkin, and H. V. Poor, "Modeling and analysis of the spread of COVID-19 under a multiple-strain model with mutations," *Harvard Data Sci. Rev.*, 2021.

[110] E. Khalil, H. Dai, Y. Zhang, B. Dilkina, and L. Song, "Learning combinatorial optimization algorithms over graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6348–6358.

[111] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, 1025–1035.

[112] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

[113] J. M. Whitacre, "Recent trends indicate rapid growth of nature-inspired optimization in academia and industry," *Computing*, vol. 93, pp. 121–133, 2011.

[114] J. M. Whitacre, "Survival of the flexible: Explaining the recent popularity of nature-inspired optimization within a rapidly evolving world," *Computing*, vol. 93, pp. 135–146, 2011.

[115] X. Li, K. Tang, P. N. Suganthan, and Z. Yang, "Editorial for the special issue of information sciences journal on "nature-inspired algorithms for large scale global optimization"," *Inf. Sci.*, vol. 316, pp. 437–439, 2015.

[116] X. S. Yang, *Particle Swarm Optimization*. Hoboken, NJ, USA: Wiley, 2010.

[117] Y. Shi, W. K. Wong, J. Goldin, M. S. Brown, and H. J. Kim, "Prediction of progression in idiopathic pulmonary fibrosis using quantum particle swarm optimization hybridized random forest," *Artif. Intell. Med.*, vol. 100, 2019, Art. no. 101709.

[118] N. Singh and S. B. Singh, "Hybrid algorithm of particle swarm optimization and grey wolf optimizer for improving convergence performance," *J. Appl. Math.*, vol. 2017, 2017, Art. no. 2030489.

[119] C. Blum and G. R. Raidl, *Hybrid Metaheuristics: Powerful Tools for Optimization*, Berlin, Germany: Springer, 2016.

[120] J. Nourmohammadi-Khiarak, M.-R. Feizi-Derakhshi, K. Behrouzi, S. Mazaheri, Y. Zamani-Harghalani, and R. M. Tayebi, "New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection," *Health Technol.*, vol. 10, pp. 667—678, 2020.

[121] T. Vivekanandan and N. C. S. N. Iyengar, "Optimal feature selection using a modified Differential Evolution algorithm and its effectiveness for prediction of heart disease," *Comput. Biol. Med.*, vol. 90, no. 2, pp. 125–136, 2017.

[122] U. N. Dulhare, "Prediction system for heart disease using Naive Bayes and particle swarm optimization," *Biomed. Res.*, vol. 29, no. 212, pp. 2646–2649, 2018.

[123] A. Abdullah, S. A. Deris, M. S. Mohamad, and S. Anwar, "An improved swarm optimization for parameter estimation and biological model selection," *Plos One*, vol. 8, no. 4, 2013, Art. no. e61258.

[124] J. Sun, J. M. Garibaldi, and C. Hodgman, "Parameter estimation using metaheuristics in systems biology: A comprehensive review," *IEEE/ACM Trans. Computati. Biol. Bioinformat.*, vol. 8, no. 1, pp. 185–202, Jan. 2012.

[125] J. Lukemire, A. Mandal, and W. K. Wong, "d-QPSO: A quantum-behaved particle swarm technique for finding D-optimal designs with discrete and continuous factors and a binary response," *Technometrics*, vol. 61, no. 1, pp. 77–87, 2018.

[126] W. Xu, W. K. Wong, K. C. Tan, and J. X. Xu, "Finding high-dimensional D-optimal designs for logistic models via differential evolution," *IEEE Access*, vol. 7, pp. 7133–7146, 2019.
[127] Z. Zhang, W. K. Wong, and K. C. Tan, "Competitive swarm optimizer with mutated agents for finding optimal designs for nonlinear regression models with multiple interacting factors," *Memetic Comput.*, vol. 12, no. 3, pp. 219–233, 2020.
[128] J. Lukemire, A. Mandal, and W. K. Wong, "Optimal experimental designs for ordinal models with mixed factors for industrial and healthcare applications," *J. Quality Technol.*, pp. 1–13, 2020.
[129] G. Pinter, I. Felde, A. Mosavi, P. Ghamisi, and R. Gloaguen, "COVID-19 pandemic prediction for Hungary, a hybrid machine learning approach," *MDPI Math.*, vol. 8, 2020, Art. no. 890.
[130] I. D. Falco, A. D. Cioppa, U. Scafuri, and E. Tarantino, "Coronavirus COVID-19 spreading in Italy: Optimizing an epidemiological model with dynamic social distancing through Differential Evolution," 2020, *arXiv:2004.00553v3*.
[131] R. G. Makade, S. Chakrabarti, and B. Jamil, "Real time estimation and prediction of the mortality caused due to COVID-19 using particle swarm optimization and finding the most influential parameter," *Infect. Dis, Model.*, vol. 5, pp. 772–782, 2020.
[132] S. He, Y. Peng, and K. Sun, "SEIR modeling of the COVID-19 and its dynamics," *Nonlinear Dyn.*, vol. 101, pp. 1667–1680, 2020. [Online]. Available: http://doi.org/10.10007/s11071–020-05743-y
[133] D. Singh, V. Kumar, Vaishali, and M. Kaur, "Classification of COVID-19 patients from chest CT images using multi-objective Differential Evolution-based convolutional neural networks," *Eur. J. Clin. Microbiol. Infect. Dis.*, vol. 39, pp. 1–11, 2020.
[134] E. Hosseini, K. Z. Ghafoor, A. S. Sadiq, M. Guizani, and A. Emrouznejad, "COVID-19 optimizer algorithm, modeling and controlling of coronavirus distribution process," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2765–2775, Oct. 2020.
[135] A. M. AbdelAziz, L. Alarabi, S. Basalamah, and A. Hendawi, "A multi-objective optimization method for hospital admission problem - a case study on COVID-19 patients," *Algorithms*, vol. 14, no. 2, 2021, Art. no. 38. [Online]. Available: https://doi.org/10.3390/a14020038
[136] Y. Tian, X. Zhang, C. Wang, and Y. Jin, "An evolutionary algorithm for large-scale sparse multiobjective optimization problems," *IEEE Trans. Evol. Comput.*, vol. 24, no. 2, pp. 380–393, Apr. 2020.
[137] Y. Tian, S. Yang, and X. Zhang, "An evolutionary multiobjective optimization based fuzzy method for overlapping community detection," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 11, pp. 2841–2855, Nov. 2020.
[138] Y. Tian, X. Su, Y. Su, and X. Zhang, "EMODMI: A multi-objective optimization based method to identify disease modules," *IEEE Trans. Emerg. Top. Computat. Intell.*, vol. 5, no. 4, pp. 570–582, Aug. 2021.
[139] D. C. dos Gomes and G. L. de Oliveira Serra, "Machine learning model for computational tracking forecasting the COVID-19 dynamic propagation," *IEEE J. Biomed. Health Informa.*, vol. 25, no. 3, pp. 515–622, Mar. 2021.
[140] Y. Zoabi, S. Deri-Rozov, and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," *NPJ Digit. Med.*, vol. 4, no. 1, pp. 1–5, 2021.
[141] T. X. Tong, K. P. Choi, T. L. Lai, and W. K. Wong, "Stability bounds and almost sure convergence of improved particle swarm optimization methods," *Res. Math. Sci.*, vol. 8, 2021, Art. no. 30.
[142] T. L. Lai, K. P. Choi, T. X. Tong, and W. K. Wong, "A statistical approach to adaptive parameter tuning in nature-inspired optimization and optimal sequential design of dose-finding trials," *Stat. Sinica*, vol. 31, pp. 1–21, 2021.
[143] C. W. Tsai, M. C. Chiang, A. Ksentini, and M. Chen, "Metaheuristic algorithms for healthcare: Open issues and challenges," *Comput. Electrical Eng.*, vol. 53, pp. 421–434, 2016.
[144] Z. Chen, X. Liu, W. Hogan, E. Shenkman, and J. Bian, "Applications of artificial intelligence in drug development using real-world data," *Drug Discov. Today*, vol. 26, no. 5, pp. 1256–1264, 2020.
[145] N. Chambers, B. Fry, and J. McMasters, "Detecting denial-of-service attacks from social media text: Applying NLP to computer security," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2018, pp. 1626–1635.
[146] M. F. Boni *et al.*, "Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic," *Nat. Microbiol.*, vol. 5, pp. 1408–1417, 2020.
[147] J. M. Soucie, "Public health surveillance and data collection: General principles and impact on hemophilia care," *Hematology*, vol. 17, no. sup1, pp. s144–s146, 2012.
[148] P. Diggle, "Statistical challenges of administrative and transaction data," *J. Roy. Stat. Soc.*, vol. 181, no. 3, pp. 555–605, 2018.
[149] L. Wang, J. Elmstedt, W. K. Wong, and H. Xu, "Orthogonal subsampling for big data linear regression," 2021, *arXiv:2105.14647*.

Zhe Fei received the PhD degree in biostatistics in 2019. He is currently an assistant professor In-Residence with the Department of Biostatistics, UCLA. His research interests include statistical methods and theories for big data, machine learning and statistical computing, survival analysis, genetics, and epigenetics.

Yevgen Ryeznik is currently a senior statistician with Early Biometrics and Statistical Innovation Group, AstraZeneca. His research interests include biostatistics, pharmacometrics, machine learning, integral and differential equations, and their applications. He designed and taught two Master or PhD level courses on optimal designs and innovative clinical trials with Uppsala University.

Oleksandr Sverdlov received the PhD degree in information technology with concentration in statistical science from George Mason University in 2007. He is currently Neuroscience Disease area lead statistician with Early Clinical Development, Novartis. He has been actively involved in methodological research and applications on clinical trials in drug development. His most recent work involves design and analysis of proof-of-endpoint clinical studies evaluating novel digital technologies.

Chee Wei Tan received the PhD degree from Princeton University. He was a senior fellow with the Institute for Pure and Applied Mathematics for the Program on Science, Extreme Scales: where Big Data Meets Large-Scale Computing. His research interests include artificial intelligence, networks, data science, and convex optimization theory. He was an editor of the *IEEE/ACM Transactions on Networking*.

Weng Kee Wong received the PhD degree in statistics from the University of Minnesota. He is currently a professor of biostatistics with UCLA, a fellow of the Institute of Mathematical Statistics, American Statistical Association, and the American Association for the Advancement of Science. His current research focuses on applications of natureinspired metaheuristic algorithms to solve complex design problems in the biomedical arena.