Integrative clinical, genomics and metabolomics data analysis for mainstream precision medicine to investigate COVID-19

Zeeshan Ahmed,1,2 Saman Zeeshan,3 David J Foran,3 Lawrence C Kleinman,4 Fredric E Wondisford,2 XinQi Dong1,2

ABSTRACT

Despite significant scientific and medical discoveries, the genetics of novel infectious diseases like COVID-19 remains far from understanding. SARS-CoV-2 is a single-stranded RNA respiratory virus that causes COVID-19 by binding to the ACE2 receptor in the lung and other organs. Understanding its clinical presentation and metabolomic and genetic profile will lead to the discovery of diagnostic, prognostic and predictive biomarkers, which may lead to more effective medical therapy. It is important to investigate correlations and overlap between reported diagnoses of a patient with COVID-19 in clinical data with identified germline and somatic mutations, and highly expressed genes from genomics data analysis. Timely model clinical, genomics and metabolomics data to find statistical patterns across millions of features to identify underlying biological pathways, modifiable risk factors and actionable information that supports early detection and prevention of COVID-19, and development of new therapies for better patient care. Next, ensuring security reconcile noise, need to build and train machine learning prognostic models to find actionable information that supports early detection and prevention of COVID-19. Based on the myriad data, applying appropriate machine learning algorithms to stratify patients, understand scenarios, optimise decision-making, identify high-risk rare variants (including ACE2, TMPRSS2) and making medically relevant predictions. Innovative and intelligent solutions are required to improve the traditional symptom-driven practice, and allow earlier interventions using predictive diagnostics and tailor better personalised treatments, when confronted with the challenges of pandemic situations.

BACKGROUND

The time has never been more critical for drug discovery data and innovative solutions development based on artificial intelligence (AI) to win the battle against COVID-19.1 As of this day (19 August 2020), over 5 498 384 cases have been confirmed and over 1 718 000 deaths have been reported in USA as a result of COVID-19, when over 21 991 954 people have been affected and over 777 018 have died worldwide (data sources are available in online supplementary material). Scientists are focusing on discovering treatment and vaccines while physicians are managing the care of their patients. However, one of the most effective treatments for overcoming this infection is a strong immune system. In case of its absence or weakness in seriously affected patients with COVID-19 lead to the provision of unlimited medical resources (eg, ventilators, intensive care units, and so on).2 Due to the rapid increase in the number of people testing positive for COVID-19, one of the biggest challenges is to classify and prioritise the patients who require urgent care, and identify which can self-recover in quarantine.3 All the findings made so far in the quest to discover a viable treatment for COVID-19 may be underappreciated as one can argue that they require further investigation,4 5 but it would not be irrational to use them to guide clinical practice in efficiently managing the available resources and saving lives.3 Despite current progress, there is still no stand-alone solution available to help current and future pandemic situations than efficiently integrating clinical, environmental, genomics, metabolomics, epidemiological data acquisition,
and enabling more efficient management of data analytics with a user-friendly physician and nurse-oriented clinical interface.\textsuperscript{6–8}

COVID-19 is a respiratory illness caused by the novel SARS-CoV-2. COVID-19\textsuperscript{9} symptomatology includes but is not limited to fever, headache, fatigue, dry cough, pneumonia, dyspnoea,\textsuperscript{9} stroke, and smell impairment.\textsuperscript{10} However, due to severe cytopathic effects and induced acute inflammatory response, the lung is one of the most affected organs by SARS-CoV-2 infection.\textsuperscript{11} Furthermore, presence of SARS-CoV-2 in cerebrospinal fluid and neurotropic effects are one of the causes of high morbidity and mortalities among patients with COVID-19.\textsuperscript{12} SARS-CoV-2 is a single-stranded RNA virus that enters the human cells by binding to the ACE2 receptor.\textsuperscript{13} ACE2 supports the regulation of metabolism, blood pressure and vascular function, and provides defence against the lung, heart, brain and injuries.\textsuperscript{12} Due to various roles and expression levels in different tissues, it is important to investigate whether SARS-CoV-2 interferes with ACE2 expression and analyse the structural and sequence variants of ACE2. Its RNA can be detected for research purposes using assays involving nasopharyngeal swabs and sputum test; genotyping of ACE polymorphisms (insertion/deletion) in white cells; evaluation of ACE2 receptor expression in whole blood;\textsuperscript{13} and whole genomics sequencing of SARS-CoV-2 to strain typing and surveillance. Clinical-grade services in genomics, transcriptomics, molecular pathology and clinical accredited laboratories need support in timely patient testing, designing novel clinical assays, customising research projects, and analysing and reporting data. Different types of clinical and scientific data have been collected and generated for various studies worldwide, which include genome sequencing of COVID-19 (SARS-CoV-2),\textsuperscript{14} sequence and annotation data made available by Ensembl, UCSC Genome Browser, and many other data sources (online supplementary material provides information on further available COVID-19 data sources). All of these observations are complicated by the capacity of the virus to stimulate immune and inflammatory responses, acutely with a ‘cytokine storm’ or macrophage activation syndrome-like illness and later, for example, the postinfectious multisystem inflammatory syndrome in children (MIS-C).\textsuperscript{15}

There is an urgent need to facilitate clinicians in decision-making, especially in extremely demanding low-resource settings by predicting burdensome physiological measurements to find the phenotypes before it reaches end stages.\textsuperscript{8} Clinicians need to be advantageous by integrating them to integrate healthcare data with targeted assays and tests to identify and assess disease risks, determine genetic variants (eg, ACE2, TMPRSS2\textsuperscript{13}) in patients, obtain a view of the metabolome and map metabolites to disease pathways. We need to support practice transformation with a precision medicine approach by the development of an intelligent digital solution for predictive diagnostics and effective therapeutics in healthcare. AI might not be the adequate solution for quick vaccine development, but it may be useful to elevate the patient’s recovery process with efficient diagnosis and appropriate allocation of available medical resources to critical patients with COVID-19.\textsuperscript{16}

**PERSPECTIVE: MAINSTREAM PRECISION MEDICINE**

Spanning clinically actionable health discoveries in precision medicine, many diverse and targeted studies have emerged to identify novel risk factors and disease markers.\textsuperscript{7} To enable the adoption of precision medicine in critical pandemic situations, various integrative clinical and multiomics data challenges must be addressed.\textsuperscript{7} A high-throughput precision medicine platform can facilitate an efficient and effective information sharing process maximising the utility of the collective data. It will foster a cross-disciplinary collaborative research environment that can be expected to lead to new fundamental insights into COVID-19 by analysing original, annotated and aggregated data (figure 1).\textsuperscript{8} Actual models may be more complex as climatological, environmental, population health, economic and healthcare resources data are among the community health data that may define the context in which these other variables impact the population. Such contextual data may be equivalent to more biological data or may define distinct environments that require distinct computational models.

Developing these intelligent systems will empower bench scientists with user-friendly graphical interface and reduce time for computational scientists to perform extensive analysis.\textsuperscript{7,18} Its impactful implementation requires coordinated efforts between disparate groups with non-aligned data formats and massive amounts of computing time—time that is essential in many cases to positively impact treatment outcomes. This will allow clinicians to systematically predict the most appropriate course of action for a patient.\textsuperscript{19} Economically, it will be cost-effective by saving time and reducing human and computational resources, especially when dealing with heterogeneous big data management, analysis and sharing.

Viral pathogenesis is altered in the obese host with obesity impacting the antiviral response, viral shed and viral evolution.\textsuperscript{20} As obesity is one of the most common medical problems in the USA, it should come as no surprise that an increasing prevalence of obesity has emerged as a significant independent predictor of disease severity.\textsuperscript{21} Individuals having a propensity for visceral fat storage will have more lipotoxic profile than those having subcutaneous fat depots. Epidemiological studies indicate that obesity increases the risk of severe complications and death from influenza virus infections.\textsuperscript{21} Plasma
metabolomics profile in chronic pulmonary diseases has been associated with metabolic dysfunctions leading to local and systemic inflammation, with more divergence as the disease gets more severe like in chronic obstructive pulmonary disease and chronic obstructive lung disease. Evidence also suggests that obesity might impact more than just disease severity as it not only alters innate and adaptive immune response, characterised by a state of chronic and low-grade inflammation, it can also prolong virus shedding and increase the risk of viral transmission making the individual potentially more contagious. These factors increase the risk for severe complications of COVID-19. Beyond obesity, the observations of more serious illness in smokers and in adolescents and young adults who vape offer further opportunity for insight. The use of AI may be particularly helpful in the COVID-19 context where no test represents a full-on gold standard. Clinicians are puzzled by illnesses that clinically are considered COVID-19 related but for which neither acute nor postacute antibody tests demonstrate infection. The largest series of MIS-C cases found that not all children had biomarkers that confirmed infection. By definition, each of the cases that were both PCR and antibody negative had documented exposure to COVID-19 in the appropriate time period for it to be causal. Contemporaneous computational evidence of a likely COVID-19 diagnosis can improve clinical management, while evidence of unlikely infection can steer clinicians to look for other causes.

We need to accelerate the discovery with gene-based collapsing analysis to identify monogenic and polygenic variants that cause emergence of metabolic multimodality and affect body mass index (BMI). A high BMI is a major risk factor for several chronic and infectious diseases, but the biology underlying these associations is not well understood. Generating a patient-specific metabolic profile associated with BMI can identify potential metabolic markers that may be associated to COVID-19. However, we also need to pay attention to racial and gender disparities in obesity and its role in the incidence, prevalence and severity of lung viral pathogenesis. Furthermore, mapping genes to the patient profiles, and at the same time, associating pre-existing diseases to their respective clinical codes to efficiently link medical records with identified causative genes and variants will provide a cross cut analysis for more efficient identification of drugs and therapeutics for specific patients with COVID-19. Intelligently linking curated clinical data obtained from healthcare platforms with computationally processed genomics and metabolomics data is the key to identify common and rare functional variants, and examine relations between genomic variations and metabolite levels across multiple health disparities.
Epidemiologists observe time courses and leading and trailing indicators of COVID-19. One of our ongoing research suggests the appearance of symptoms and precedes test positivity, which precedes hospitalisations, which precedes ventilator use and deaths as well as postinfectious complications. Others have found Google searches can anticipate the arrival of confirmed COVID-19 infections. Layering machine learning (ML) techniques to quantify the meaning of leading indicators in population data can help enhance the public health response to bring resources and prepare regions for what is to come. Appropriate use of AI and ML has the potential to predict the occurrence of COVID-19, and even other life-threatening chronic and acute diseases’ risk susceptibility, starting from the most common to rare. We need to use potential of AI with the implementation of ML algorithms for analysing metabolic and genomic read-outs offering assessments of disease states, causes of the emergent metabolic patterns of multimodal distributions for candidate genes to illuminate biological pathways underlying standard measures, penetration using listed features and abnormalities and enable exploration of disease progression. This will support predictive diagnostics and mapping molecular pathways with associated sequelae and diseases. Along with analytics, it is important to address the issues related to data privacy and security. In most of the cases, research environments do not have access to the electronic medical records, mainly due to lack of trust by healthcare institutions. We suggest developing Health Insurance Portability and Accountability Act-compliant research environments to overcome such problems.

**CONCLUSION**

We need to improve processes for clinical genomics testing, integrate genomics and metabolomics into the clinical workflow, develop and evaluate prevention and therapeutic strategies, and build knowledge bases for predictive genomic and precision medicine. Robust scientific solutions are required for everyday clinical and public health practices to perform combined analysis of clinical, metabolomics and genomics data with community and population data to support the application of AI (figure 1). It will help detect new predictive models, understand disease mechanisms, identify major causes of morbidity and mortality, develop personalised therapies and reduce medical costs.

**Correction notice** This article has been corrected since it was published. Funding statement has been updated.

**Acknowledgements** We appreciate the great support by the Institute for Health, Health Care Policy and Aging Research (IH), and Robert Wood Johnson Medical School, Rutgers Cancer Institute of New Jersey, Rutgers Biomedical and Health Sciences, at Rutgers, The State University of New Jersey.

**Contributors** ZA proposed the study and drafted the manuscript. ZA, SZ, DJF, LCK, FEW and XQD participated in writing of the manuscript, and approved the final version for publication. DJF, LCK, FEW and XQD guided the study.

**Funding** This work was supported by the Institute for Health, Health Care Policy and Aging Research, and Robert Wood Johnson Medical School, at Rutgers, The State University of New Jersey. LCK was supported by the grant U3DMC32755-02 from the Health Resources & Services Administration.

**Competing interests** None declared.

**Patient consent for publication** Not required.

**Provenance and peer review** Not commissioned; externally peer reviewed.

This article is made freely available for use in accordance with BMJ’s website terms and conditions for the duration of the covid-19 pandemic or until otherwise determined by BMJ. You may use, download and print the article for any lawful, non-commercial purpose (including text and data mining) provided that all copyright notices and trade marks are retained.

**ORCID iD** Zeeshan Ahmed http://orcid.org/0000-0002-7065-1699

**REFERENCES**


