Briefings in Bioinformatics, 22(6), 2021, 1–10

https://doi.org/10.1093/bib/bbab191
Problem Solving Protocol

COVID-19 biomarkers and their overlap with comorbidities in a disease biomarker data model

Nikhita Gogate, Daniel Lyman, Amanda Bell, Edmund Cauley, Keith A. Crandall, Ashia Joseph, Robel Kahsay, Darren A. Natale, Lynn M. Schriml, Sabyasach Sen and Raja Mazumder

Corresponding authors. Nikhita Gogate, Department of Biochemistry and Molecular Medicine, The George Washington University Medical Center, Washington, DC 20037, USA. E-mail: nikigogate@gwu.edu; Raja Mazumder, Department of Biochemistry and Molecular Medicine, The George Washington University Medical Center, Washington, DC 20037, USA, and The McCormick Genomic and Proteomic Center, The George Washington University, Washington, DC 20037, USA. Tel: 02-994-5004; Fax: 202-994-8974; E-mail: mazumder@gwu.edu

Abstract

In response to the COVID-19 outbreak, scientists and medical researchers are capturing a wide range of host responses, symptoms and lingering postrecovery problems within the human population. These variable clinical manifestations suggest differences in influential factors, such as innate and adaptive host immunity, existing or underlying health conditions, comorbidities, genetics and other factors—compounding the complexity of COVID-19 pathobiology and potential biomarkers associated with the disease, as they become available. The heterogeneous data pose challenges for efficient extrapolation of information into clinical applications. We have curated 145 COVID-19 biomarkers by developing a

Nikhita Gogate is a Research Associate at the HIVE lab at the George Washington University School of Medicine and Health Sciences. She has more than 5 years of experience in clinical molecular diagnostics and high-throughput sequencing. She has been working on cancer biomarkers and curating FDA approved cancer biomarkers through the OncoMX project. She is now leading the efforts to capture biomarkers for COVID-19 in a robust disease biomarker data model.

Daniel Lyman has graduate training and research in developmental neurogenetics. He is a Senior Research Associate at the George Washington University School of Medicine and Health Sciences, Department of Biochemistry and Molecular Medicine. Currently, his research interests are bio-ontology development and application.

Amanda Bell is a Research Associate at the George Washington University School of Medicine and Health Sciences. She received a bachelor's degree in integrated sciences and biotechnology from George Washington University and a master's degree in molecular biochemistry and bioinformatics. Her current research is in analyzing cancer gene expression and mutation signatures for the discovery of biomarkers.

Edmund (Ned) Cauley is a Research Associate in the Hive Lab at the George Washington University School of Medicine and Health Sciences and has a background studying molecular medicine, neuroscience, and next-generation sequencing data analysis.

Dr Keith Crandall is the Founding Director of the Computational Biology Institute at The George Washington University and a Professor in the Department of Biostatistics and Bioinformatics, Milken Institute School of Public Health. His research focuses on method development and software implementation for DNA sequence analysis with applications in health.

Ashia Joseph is a first-year bioinformatics student at The George Washington University. She received a bachelor's degree in biology from Salisbury University. Her current research interests include cancer biomarkers and translational immuno-oncology.

Dr Robel Kahsay is an Associate Professor at the George Washington University School of Medicine and Health Sciences, Department of Biochemistry and Molecular Medicine. His current research focus is on the use of innovative technologies for building analysis engines that feed on high-throughput heterogeneous biological datasets to extract actionable knowledge and making this knowledge available to the research community through advanced interfaces and automated APIs.

Dr Natale, based in Washington, DC, USA at the Georgetown University Medical Center, is a member of the Protein Information Resource team and leads the Protein Ontology project.

Dr Schriml is an Associate Professor at the University of Maryland, School of Medicine in Baltimore, MD, USA, leading biomedical ontology and metadata standards development as PI of the Human Disease Ontology.

Dr Sen is a Professor of Medicine and Professor of Biochemistry and Molecular Medicine at the George Washington University School of Medicine and Health Sciences. He is a practicing physician in the Division of Endocrinology at the GW Medical Faculty Associates. His current research focus is survival and differentiation of adult stem cells in chronic diseases such as pre-diabetes, diabetes, and HIV with a focus on endothelial dysfunction, adipocyte inflammation, and insulin resistance.

Dr Mazumder has worked closely with his colleagues in developing international biomedical resources and using these resources to identify therapeutics, diagnostics, and vaccines targets. His current research focus includes developing novel methods for data-to-knowledge initiatives in biomedical sciences and community driven bioinformatics projects.

Submitted: 10 February 2021; **Received (in revised form):** 29 March 2021

novel cross-cutting disease biomarker data model that allows integration and evaluation of biomarkers in patients with comorbidities. Most biomarkers are related to the immune (SAA, TNF-$\propto$ and IP-10) or coagulation (D-dimer, antithrombin and VWF) cascades, suggesting complex vascular pathobiology of the disease. Furthermore, we observe commonality with established cancer biomarkers (ACE2, IL-6, IL-4 and IL-2) as well as biomarkers for metabolic syndrome and diabetes (CRP, NLR and LDL). We explore these trends as we put forth a COVID-19 biomarker resource (https://data.oncomx.org/covid19) that will help researchers and diagnosticians alike.

Key words: COVID-19 biomarkers; COVID-19; cancer; metabolic syndrome; vascular disease

Introduction

The devastating outbreak of the novel, highly contagious Coronavirus Disease (COVID-19), originating in Wuhan, China, has rapidly spread worldwide since first reported in early January 2020. COVID-19 has created major challenges for worldwide health systems, caused global disruption and has had far-reaching consequences to the global economy [1]. In response, the World Health Organization declared a global pandemic; as of 9 February 2021, there are more than 106 million confirmed cases globally and more than 2.3 million reported fatalities [2]. The causative agent, SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2; SCoV2) is the seventh coronavirus known to infect humans [3]. Coronaviruses SCoV, MERS-CoV and SCoV2 can cause severe disease; while Coronaviruses HKU1, NL63, OC43 and 229E are associated with mild disease states [4, 5]. High recombination rates and the genetic diversity of coronaviruses in the wild suggest that further outbreaks and unpredictable virulence will likely arise in future recombinants [6], which could lead to different and diverse outcomes in patients. Serious clinical manifestations of COVID-19 (in some individuals) include: severe acute respiratory syndrome, inflammatory pneumonitis, hypoxia, blood clots, embolisms, gastrointestinal illness, cardiac and vascular damage and organ damage (lung, heart, kidney, liver, brain) [7]. Severity and mortality of COVID-19 appear to be more prevalent in men than women and, overall, more so in the elderly with underlying health conditions such as hypertension, cardiovascular disease, immunosenescence, immunocompromised systems and diabetes [6–10]. Clinical observations of hospitalized COVID-19 patients report lymphopenia, monocytopenia and hypoalbuminemia, as well as elevated proinflammatory cytokines ('cytokine storm'). In severe cases, pneumonia with a 'ground glass' opacity in chest CT scans, lung injury and pneumonitis are typically observed [11, 12]. Lymphopenia and cytokine storm may initiate severe COVID-19 pathogenesis, viral sepsis, inflammation-induced lung injury and pneumonitis, acute respiratory distress syndrome, respiratory failure, shock, organ failure and death [7, 13, 14]. The probability of severe damage from the direct or indirect effects of SARS-CoV-2 replication can be exacerbated by underlying injuries caused by chronic conditions such as hypertension, diabetes and cancer [9, 15, 16]. Identification of physiological or pathological differences associated with poor outcomes of the COVID-19 in patients with underlying conditions—and discovery of prospective biomarkers predictive of these outcomes—is of paramount importance.

The FDA-NIH Biomarker Working Group (FNBWG) defines a biomarker as a 'characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions' [17]. Molecular biomarkers (also known as molecular markers or signature molecules) may be (for example) genes, proteins, glycans or metabolites that can be used for disease assessment and treatment evaluation and have distinct functions in biomedical research, clinical practice and medical product development. Within the context of the Biomarkers, EndpointS and other Tools (BEST) Resource, the FNBWG further distinguishes important subtypes by role [17], with the Diagnostic, Prognostic and Susceptibility/Risk categories relating biomarkers to disease, the Pharmacodynamic/Response, Predictive and Safety categories relating biomarkers to interventions (drug treatment, for example) and the Monitoring category suitable for both. Measured as objective, reproducible numeric or categorical values, biomarkers play a significant role in highlighting the relationships among environmental exposures, human biology and disease [18]. The BEST categorization provides a constructive framework by which to organize, standardize and integrate data elements.

While researchers race to find drugs or vaccines for the virus, a critical need to identify biomarkers for COVID-19 disease has become evident. A recent article provides an excellent overview of the myriad biomarkers currently being used in the fight against COVID-19 [19]. The article briefly touches upon the biology of the virus and discusses biomarkers from the perspective of therapeutics and drug discovery. However, it does not cover as broad a spectrum of FNBWG-approved segregated biomarkers as is necessary to consider the extensive impact of COVID-19 on patients with comorbidities. It also does not do justice to the amount of COVID-19 biomarker data available in the literature. Indeed, a simple Google Scholar search for COVID-19 biomarkers retrieves more than 17 000 records, indicating that significant data for nucleic acid, protein and other biomarker material await analysis by cross-disciplinary investigation of COVID-19 publications and repositories.

Though preliminary discoveries demonstrate clinical applicability of potential biomarkers, additional research must establish specificity and sensitivity during risk assessment, diagnostic measurements or therapeutic applications to a particular disease state [20, 21]. For both research and clinical applications, improved methods for aggregation of biomarker knowledge must be implemented, a process that comparatively lags behind due to the heterogeneous nature of biomarker data. Indeed, preliminary evaluation reveals that almost none of the biomarker data described in these references are standardized and harmonized to existing ontologies and terms. Here, we describe a publicly available compilation of COVID-19 biomarkers that enables researchers to explore up to date COVID-19 biomarkers in different stages of development and application. Furthermore, we describe features of COVID-19 discovered through the lens of biomarkers that offer insight into COVID-19 pathobiology.

Materials and methods

The overall workflow involves data collection, organization, standardization and integration of the data elements (Figure 1). Further details are provided below.

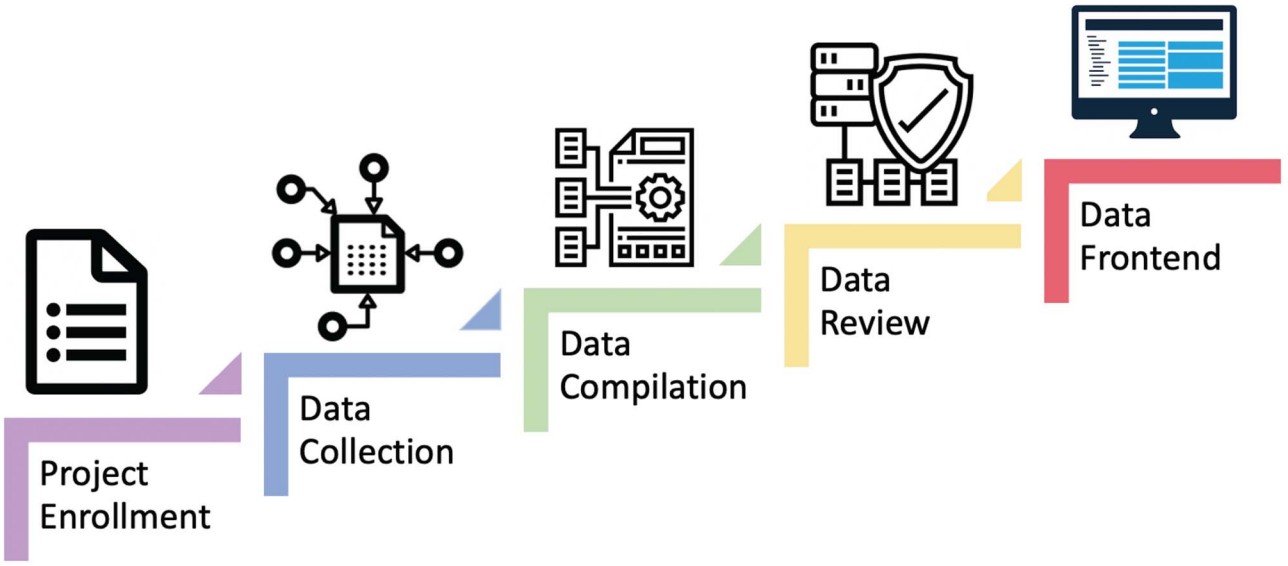


Figure 1. Steps for collection, organization, standardization and integration of the data elements in the COVID-19 biomarker resource data model.

Crowdsourcing

The project was advertised to collaborating faculty members and members of the community, which led to the recruitment of ~30 volunteers. The volunteers ranged from high-school students, undergraduate and graduate students with a biology and basic bioinformatic skills. All volunteers went through a brief training to orient them to the standardized workflow, which involved reading publications and filling in tables. Key staff members of the OncoMX team led the volunteers and served as reviewers for all annotations. Reviewers then re-read the publications to ensure that the cell entries were correct prior to moving them to the reviewed biomarker table. Based on their skills and quality contributions to the project, a few volunteers were given additional training to become project trainers.

Data collection and compilation

Using Google Scholar, curators searched for articles publicly available after January 2020 that mentioned 'COVID-19' and 'biomarker'. Information about a biomarker and its role in COVID-19 was retrieved from selected articles and filled into a structured format (described under 'Disease biomarker data model' in Results and Discussion). The curator included notes in a free text column that documented any comments regarding the data curated. Each curator uploaded their data file into a shared drive once every week. Reviewers would then compile the data from all the curators into a single cohesive dataset file marked Unreviewed.

Biocuration, review processing and quality check

The unreviewed data were scrutinized by at least two reviewers with experience in biomarker curation and ontology mapping. All annotations were checked for content. These checks included: confirmation of the biomarker name mentioned in the article, appropriate mapping of the biomarker accession, suitable representation of the BEST biomarker type based on the article and definitions provided by the FNBWG, and documentation of the specimen type with mapping to Uberon anatomical IDs. Table 1 provides details on the rubric of approved data types in each of the columns. Curator notes were carefully assessed during reviews of the biomarker entries to answer any queries that arose. The data were checked for completeness and adherence to the rubric of data collection (Table 1). A second layer of curation was applied as a sanity check to evaluate reviewers' findings. This is in line with the efforts of meeting the Core criteria set by ELIXIR [22] to build quality and reduced error resources for the community. Furthermore, multiple publications and pre-print articles served as verification to support the biomarker entity. Resulting entries were compiled and quality checked to ensure integrity and format stability.

Mapping COVID-19 biomarkers with comorbidities (cancer and diabetes)

Initial COVID-19 biomarker trends showed an overlap with cancer and metabolic syndrome biomarkers. To further study the commonality, COVID-19 biomarkers were cross-checked against established biomarkers for cancer and diabetes using search parameters comparable to the ones mentioned above. The identified biomarkers were further annotated and curated as described above. Similarly, the biomarker entities were verified with the support of numerous publications and established biomarkers data such as EDRN [23].

Results and discussion

With the rapid increase in the number of publications and preprints mentioning COVID-19 biomarkers, it was clear that a crowdsourcing effort was needed to rapidly collect and cross-validate the data. Biomarker data from publications and bioinformatics databases furthermore require harmonization against a robust standardized data model. We describe our crowdsourced efforts backed by our disease biomarker data model to organize COVID-19 biomarker data that will assist researchers working on the development of diagnostics or drugs.

Table 1. Biomarker table header descriptions and column content

Header	Description	Type
Biomarker ID	Biomarker identifier generated in this resource	Alphanumeric code (A001)
Main x-ref	Accession/identifier with other standardized databases	UPKB (UniProtKB/Swiss-Prot acc) or PCCID (PubChem Compound ID) or CO (Cell Ontology ID) or PRO (Protein Ontology ID) or DO (Disease Ontology ID) or PDB (Protein Database ID) or NCIt (NCI Thesaurus concept code) or CHEBI or LOINC
Assessed biomarker entity	Common name and gene symbol or short name in parenthesis	Free text.
Biomarker	The change measured in disease versus healthy	increased/decreased level or increased/decreased expression or increased/decreased cell count or ratio
BEST biomarker type	Category of BEST Biomarker	monitoring; diagnostic; prognostic; predictive; risk/susceptibility; safety; pharmacodynamic/response
Specimen type	The type of specimen used to access the biomarker	Uberon name (Uberon ID)
LOINC code	LOINC is clinical terminology that is important for laboratory test orders	LOINC numeric code
Disease name	Disease 'Ontology term or name for the disease with DOID in parenthesis'	Disease name (DOID)
Literature evidence	Literature reporting the biomarker	Text from article (PMID or DOI)
Notes	Free text to add meta data to the entry	

Disease biomarker data model

The data model for the COVID-19 biomarkers is based on the OncoMX cancer biomarker model [24], with some revision (Figure 2). Briefly, the model captures information on biomarker name, assessed biomarker entity, specimen type, biomarker description and drug mentioned. The model, furthermore, captures ancillary data regarding the biomarker such as biomarker accession and BEST biomarker type. Importantly, the platform implements strict adherence to accepted standards used in major resources, such as National Center for Biotechnology Information [25], European Bioinformatics Institute (EBI) [26], Alliance of Genome Resources [27] and others. Along with OncoMX, these resources rely heavily on existing biomedical standards and ontologies for semantic unification of datasets, which can enable efficient knowledge modeling, information retrieval and data sharing across otherwise diverse data [22, 28–34]. Accordingly, our biomarkers model requires mapping to accessions that included the canonical UniProtKB/Swiss-Prot accession [35], PubChem Compound ID [36], Cell Ontology ID [37], Protein Ontology ID [38], Disease Ontology ID [39], Uberon Anatomy Ontology [31], Protein Database ID (PDB) [40], NCI Thesaurus concept code (NCIt) [41], Chemical Entities of Biological Interest (CHEBI) [42] and Logical Observation Identifiers Names and Codes (LOINC) [43]. The emphasis on leveraging existing standards and ontologies promotes extensibility and sustainability, allowing the platform to focus on data quality, integration, standardization and knowledgebase maintenance and extension.

COVID-19 biomarker data curation and integration

Crowdsourcing allowed us to annotate and cross-validate 145 biomarker type combinations. These curated biomarkers are classified into the appropriate diagnostic, monitoring, prognostic disease biomarkers as shown in Figure 3. The majority of the biomarkers are prognostic in nature. Such biomarkers help to predict the likelihood of disease progression or severity, providing justification for patients needing special treatment such as ICU admission. The second category is monitoring biomarkers, which are serially measured to determine the disease status in an individual exposed to SARS-CoV-2 virus. Diagnostic biomarkers enable the detection of COVID-19 in patients with different clinical manifestations ranging from pulmonary distress to gastrointestinal illness. Comorbidities such as cardiovascular disease and diabetes pose an increased chance of contracting the disease and are evaluated as risk factors. Our results indicate the emergence of a pattern that points to specific pathways and cell types targeted by SARS-CoV-2. Our manually curated resource shows that most biomarkers belong to biological processes within specific tissue systems and supports the need for further investigations using multidrug combination therapies targeting these biological processes.

Immune system biomarkers are elevated in COVID-19

Various manuscripts asserted that certain biomarkers strongly correlated with COVID-19 [e.g. C-reactive protein (CRP), interleukin-6 (IL-6), D-dimer], while other biomarkers [e.g. Von Willebrand factor (VWF), citrullinated histone H3 (Cit-H3), macrophage colony stimulating factor, RNA-binding protein EWS (EWS)] showed latent connections requiring further investigation. Figure 4 shows the leading assessed biomarker entities as indicated by the number of articles supporting the biomarker, along with the direction of the trend (increased or decreased values) detected in COVID-19 patients. A high level of CRP, identified as a top biomarker appearing in 41 independent studies, indicates disease progression and has been positively correlated with lung lesions [44]. CRP level has been used as a monitoring biomarker in early stages of the disease to determine progression from mild to severe [44, 45]. IL-6 is a known cancer

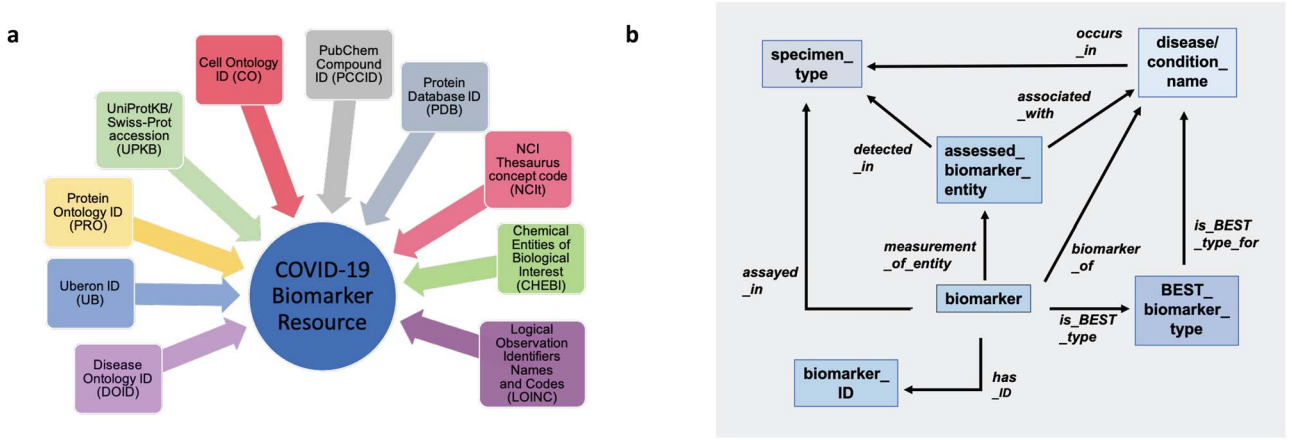


Figure 2. Disease Biomarker Framework: Harmonization of biomarker-centric knowledge (including terms, definitions, synonyms), enriched with objects imported from related reference ontologies under a unified framework. (a) Biomarker are mapped to accessions that included the canonical UniProtKB/Swiss-Prot accession [33], PubChem Compound ID [34], Cell Ontology ID [35], Protein Ontology ID [36], Disease Ontology ID [37], Uberon Anatomy Ontology [28], Protein Database ID (PDB) [38], NCI Thesaurus concept code (NCIt) [39], CHEBI [40] and LOINC [41]. (b) The schema depicts some named relations (black arrows) existing among some named types of biomarker-related data (colored rectangles/squares). BEST (Biomarkers, EndpointS and other Tools) categories of biomarkers are described by FDA-NIH Biomarker Working Group (FNBWG) [18]; assessed_biomarker_entity is the name of the entity assayed as a biomarker (e.g. IL-6); biomarker is the name of the measured biomarker (e.g. increased IL-6 level); specimen_type is the name of the tissue (from Uberon ontology) in which the named biomarker was measured.

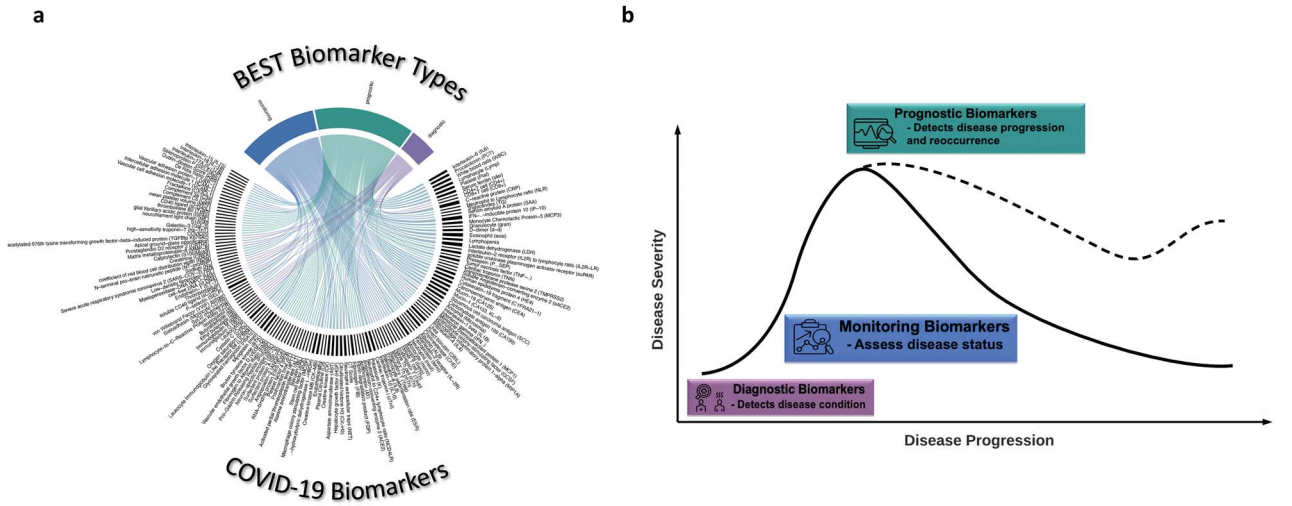
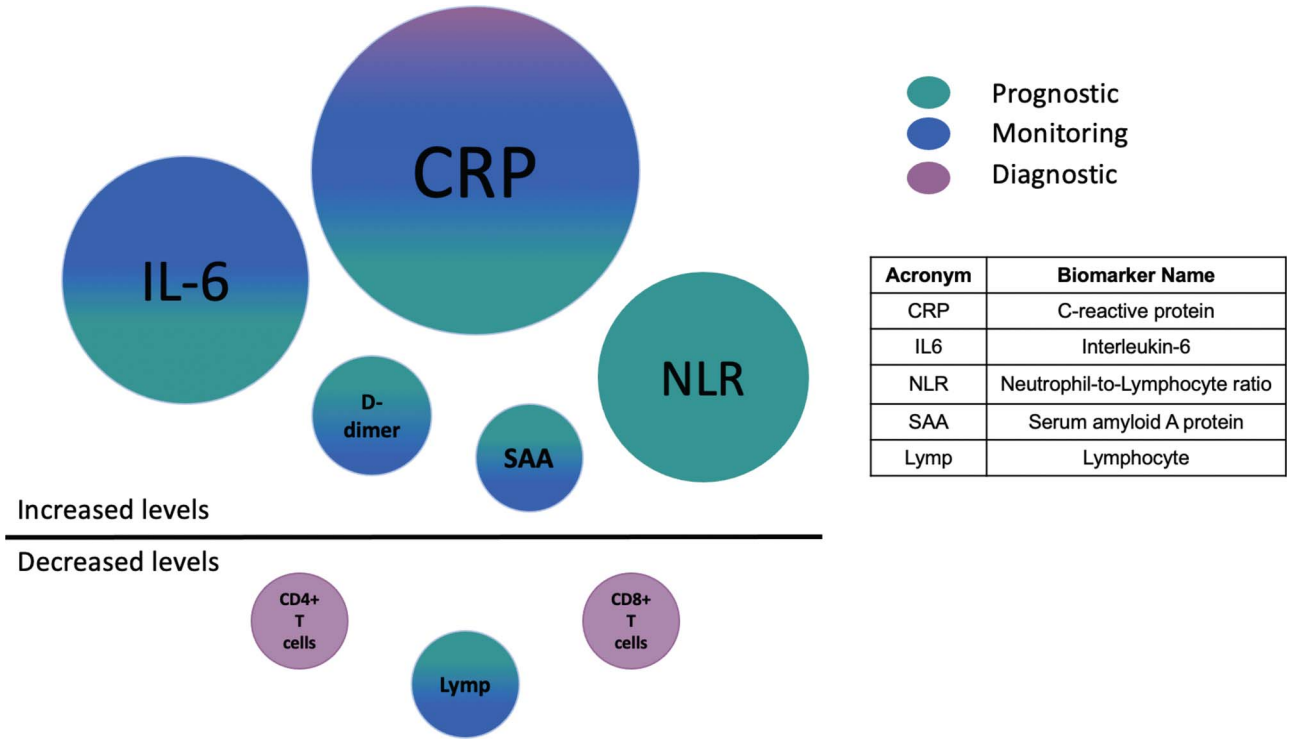


Figure 3. Classification of COVID-19 biomarkers into 7 BEST categories. (a) The collated 145 COVID-19 biomarkers are mapped to the Monitoring, Prognostic and Diagnostic BEST biomarker categories. The instances of potential COVID-19 biomarkers provide measurable evidence data of existing or potential health status. (b) Stylized graph of BEST biomarker types modeled on disease progression. Each category of BEST biomarker fulfils a distinct role 'as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention.'

biomarker [24]. It was also used as a disease progression monitoring biomarker in multiple studies and its elevated levels were shown to be strongly associated with respiratory failure in symptomatic COVID-19 patients. Diagnosticians have used this biomarker to determine the need for mechanical ventilation [46]. Other interleukins, such as IL-4 and IL-2, were also proposed for biomarker assessment in some studies. Notably, interleukins can be both pro- and anti-inflammatory and, as such, use of these biomarkers should be coupled with other indicators of disease progression. Interestingly, Herold *et al.* [46] found no correlation between IL-6 levels and age, comorbidities, radiological findings, respiratory rate or qSofa score of patients. On the other hand, IL-4 was shown to inhibit SARS-CoV replication partially by down regulating ACE2 expression *in vitro* [47]. While this study investigated SARS-CoV and not SARS-CoV-2, research advises screening of all patients for hyper inflammation [48] and the use of inflammation biomarkers to assess the severity of disease. Neutrophil-to-lymphocyte ratio (NLR) was also suggested as a prognostic biomarker in over 10 independent studies. It has been commonly used as a marker for subclinical or systemic inflammation, and a high NLR has been linked to poor clinical outcome in many solid tumors. For COVID-19 patients of advanced age, this ratio, when elevated, could serve as a prognostic biomarker to determine access to valuable limited clinical resources such as intensive care units (ICUs) or ventilators [49]. Another biomarker reported by multiple research groups, in agreement with aggravated inflammation observed in COVID-19 patients, is serum amyloid A protein (SAA). SAA was used as both a monitoring and a prognostic biomarker by multiple independent research groups to evaluate severity and prognosis of COVID-19. Dynamic changes in SAA have been proposed as prognostic markers in COVID-19



Acronym	Biomarker Name
CRP	C-reactive protein
IL6	Interleukin-6
NLR	Neutrophil-to-Lymphocyte ratio
SAA	Serum amyloid A protein
Lymp	Lymphocyte

Figure 4. COVID-19 Biomarkers highlights. Top biomarkers are depicted as increased or decreased levels of the indicated assessed entity. The size of the circle is indicative of the number of articles supporting the biomarker. The color of the circle corresponds to the BEST biomarker type and the proportion in the circle with the number of articles supporting the biomarker in that category.

progression [50]. This is because it belongs to the family of apolipoproteins that are constitutively expressed in plasma. SAA is a potential therapeutic target in chronic inflammation [51]. A common theme from our resource points to the immune system of the patients and the inflammatory response to the disease. Other concurrent biomarkers in our resource are tumor necrosis factor-α (TNF-α), interferon-Υ inducible protein-10 (IP-10), CD4+ counts and CD8+ counts, all of which suggest the immune system of the patient as a central target to determine disease progression, therapy and possibly prevention.

Coagulation cascade biomarker levels are important in COVID-19 response

Another biomarker substance that has been extensively studied is D-dimer, a degradation product of cross linked fibrin resulting from plasmin cleavage [52]. Several independent studies from Wuhan, China, showed that elevated levels of D-dimer in COVID-19 patients are associated with higher mortality. Indeed, it has been used as a prognostic biomarker to predict mortality rates in patients with COVID-19 [53]. However, since it is a product of cross-linked fibrin, there are many other common conditions in which it can be elevated and, consequently, use of this biomarker warrants caution. The most common substances or processes resulting in analytical interference with D-dimer levels are paraproteins, bilirubin, lipids and hemolysis [52]. As such, establishing a fold-change cutoff specific for the patient for this prognostic biomarker before drawing conclusions is essential. Interestingly, our resource suggests the use of other biomarkers from similar biological processes, though these have been studied less extensively. These include increased VWF and decreased antithrombin levels, among others. VWF, made within endothelial cells, helps platelets stick together, assists clot formation and transports coagulation factor VIII to areas of clot formation [54]. High levels of VWF have been linked to potential efficacy of COVID-19 treatment and have also been used as a prognostic biomarker for endothelial damage [55]. Escher *et al.* [56] observed an ~500% increase in VWF and coagulation factor VIII expression in COVID-19 patients during later stages of stay in an ICU. This increase was observed in patients that registered an increase in D-dimer expression in the earlier stages of their stay in the ICU. These patients underwent extensive endothelial stimulation and damage, which can be explained by the presence of ACE2, the receptor for SARS-CoV-2, on the surface of endothelial cells [56]. Similarly, antithrombin, a glycoprotein that plays a critical role in controlling coagulation [57], was also proposed as a biomarker by at least two independent studies. Decreased levels of antithrombin, along with increased levels of D-dimer and VWF, in conjunction with other proposed biomarkers such as increased fibrinogen expression and decreased platelet counts, point to recurrent coagulopathies in COVID-19 patients.

Immune and coagulation cascade biomarkers link metabolic syndrome to COVID-19

A biomarker that seems to have given interesting results, in the context of metabolic syndrome, diabetes and hyperlipidemia, is Low Density Lipoprotein (LDL) [58, 59]. Subjects with pre-existing dyslipidemia and metabolic syndrome (defined as high triglyceride and low HDL) appear to have decreased LDL at the onset of the disease, with lower levels predicting worse outcome for mortality [58]. It is possible that virus particles require LDL to proliferate and replicate, in which case the drop in LDL at

the onset would indicate huge virus reproduction capability, rather than a good indicator in the context of metabolic syndrome. Thus, even in these subjects, lowering of endogenous LDL production with medications such as statins may help to reduce or at least impair further virus production capability [58]. It has also been shown previously that NLR in diabetic patients is significantly higher than in healthy individuals and is positively correlated with insulin resistance. This study also suggested that NLR is a reliable predictive biomarker for insulin resistance [60]. High levels of NLR observed in both COVID-19 and diabetic patients point to the reliability of this biomarker in both diseases while informing on how patients with metabolic syndrome are at an increased risk of worse prognosis if they contract COVID-19. Another biomarker from our resource, D-dimer, which is part of the coagulation cascade, has been shown to be significantly elevated in COVID-19 patients with diabetes, as compared with COVID-19 patients without this comorbidity. This results in worse prognosis for patients with this comorbidity, as hypercoagulation might lead to further complications [61]. These data indicate D-dimer as an extremely useful biomarker, especially in patients with these comorbidities. Other biomarkers included in our resource that have previously been suggested for the detection of both metabolic syndrome and COVID-19 are IL-6, IL-10 and TNF-$\propto$; all are part of the immune cascade. Elevated levels of IL-6 and TNF-$\propto$ positively correlate with both metabolic syndrome and COVID-19; however, IL-10 is found to be suppressed in metabolic syndrome but elevated in COVID-19 patients [62]. These studies highlight the importance of carefully curating biomarker information in different disease settings.

Cancer patients have an elevated risk and a poorer prognosis of COVID-19

Our resource has also registered biomarkers, such as Cit-H3 and ACE2, that have extensive ramifications in various cancers. Cit-H3 plays an important role in neutrophil release of nuclear chromatin, also called neutrophil extracellular traps (NETs), which have been associated with tumor progression in colon cancer [63, 64]. Notably, NETs have also been proposed as a biomarker for SARS-CoV-2 infection in COVID-19 patients and are known to play a role in thrombosis, thereby strengthening our observation that specific biological processes are activated during SARS-CoV-2 infection. Finally, the angiotensin converting enzyme ACE2, which serves as a receptor for the spike glycoprotein of SARS-CoV-2, shows stabilized protein levels in colorectal and renal cancers and has also been proposed to be used as a biomarker [65]. This suggests that the majority of cancer patients, and not just immunocompromised patients, also have an elevated risk of contracting the disease and might have a poorer prognosis when compared with non-cancer individuals with COVID-19.

Mapping COVID-19 biomarkers with cancer and diabetes has identified similar expression profiles, emphasizing the underlying physiological or pathological association. Preliminary analysis has shown an overlap in established EDRN [23] cancer biomarkers (such as IL-6, CRP, TMPRSS2, HE4, CA125, IL1B, IL-4, IL-10, HGF and VWF) with COVID-19. Taken together, these data suggest that hyper-activation of the immune system, coagulopathies and the targeting of specific types of cells that are indispensable for vasculature (such as endothelial cells) are the primary modus operandi of SARS-CoV-2, and combination therapies targeting these biological processes may be beneficial for COVID-19 patients. Additionally, risk factors such as cardiovascular disease, hypertension, thrombocytopenia and cancer need to be taken into consideration while devising therapeutic regimens.

Phases of biomarker development and use

Specific biomarkers approved for clinical use for certain diseases can potentially be used for emerging diseases such as COVID-19. However, such repurposed use would need to be tested and experimentally validated. We propose six phases of biomarker development and use based on previous discussions on this matter [66, 67]. The phases are preclinical exploratory, clinical assay and validation, retrospective longitudinal, prospective screening, disease control and finally regulatory body guidance or approval for clinical use for the specific disease/condition and BEST category. The COVID-19 biomarkers presented here would thus be considered in the preclinical exploratory stage and would have to go through the other phases to be accepted as a bona fide biomarker.

There is an urgent need within the research community to have an integrated and harmonized COVID-19 biomarker resource. Rapidly accumulating, dispersed and heterogeneous COVID-19 datasets pose challenges to data comparison and extrapolation of meaningful observations, hindering translation of information into clinical applications. In this effort, we have found a number of issues. For example, we note that the same biomarker is often known by multiple different names—which might differ from the name used in standardized databases—thus potentially making the connection between biomarker and underlying biology less amenable to discovery. One such case is Carbohydrate Antigen 15–3, also commonly known as Krebs von den Lungen-6, which is called Mucin-1 in UniProtKB. Our resource has solved these discrepancies by including all such entries identified by respective studies, and then unifying them under common identifier links to standardized databases. Furthermore, we have assigned unique five-character alphanumeric codes to each of the biomarkers in our resource for ease of referencing. Another issue involves discerning exactly what is being measured. For example, what substance is being assayed when using alanine aminotransferase (ALT) as a biomarker? Is it alanine aminotransferase 1, alanine aminotransferase 2, or both? (It is both). Finally, biomarkers may be assigned to different categories for a particular disease in different resources. For example, soluble urokinase plasminogen activator receptor (suPAR) isolated from ovarian cysts appears to distinguish between malignant versus benign cysts [68] and is therefore diagnostic, but that same substance is considered prognostic (for a number of diseases) when obtained from blood [69]. Such common issues, often faced while compiling of a knowledgebase based on an extensive literature, have been addressed in our resource.

Future directions

Our goal is to build a stable and sustainable infrastructure for both common and rare COVID-19 biomarkers using the criteria identified by ELIXIR [22], and to provide regular updates that will help the research and medical community stay abreast of the extensive research being conducted in the face of the ongoing pandemic. To further develop and improve the COVID-19 Biomarker resource, we envision extending the search for new biomarkers in future curation rounds with adjusted query terms and parameters—for example expanding the assessed biomarkers entities to include glycan panels. We have initiated the expansion of the OncoMX biomarker model (manuscript in

preparation) to record a panel of data types associated with MALDI-MS measured N-glycans covalently attached to immunocaptured serum glycoproteins from cirrhosis, hepatocellular carcinoma and liver transplant cohorts. Additionally, we anticipate expanding the data model to capture patient demographic and comorbidity data that can further aid in data analyses with the application of machine learning to search for possible common correlations/associations between patient features (demography, comorbidity, etc.) and observed biomarkers. This will be further valuable to examine interesting biological questions about COVID-19 and may help improve treatment. To these ends, we recognize the need for standardization and formalization of the collected information in an ontology that connects biomarkers with their indications. Construction of such an ontology is underway, as semantic modeling of biomarkers will empower novel comparisons across related medical domains, support programmatic data science methods of discovery and make biomarker information available for easy integration into other resources. Furthermore, we are set up for continued crowdsourced contributions at https://data.oncomx.org/covid19. This will also enable addition of annotation as well as validation of biomarkers overtime.

Conclusion

Variations in symptoms have not only exacerbated the diagnosis, prognosis and monitoring of COVID-19 but have also made it difficult to identify and develop vaccines and drugs. Our COVID-19 Biomarker resource has drawn from our extensive experience in integrating large biomarker (OncoMX [24]) and glycoprotein (GlyGen [70]) datasets. As such, our biomarker resource currently includes over 500 biomarker articles encompassing both vastly studied and largely cross-referenced biomarkers, as well as rare and risk biomarkers. An overview of the current repository shows that the COVID-19 impacts the patient's immune system and is involved in various coagulopathies. Risk biomarkers included in our resource also shed light on at-risk patient populations and allow investigation of underlying comorbidities that might affect prognosis and treatment outcome. In the stages of understanding the pathology of this infectious disease, we provide this biomarker resource to support continued research around the world to better understand and manage COVID-19. Collective analyses of these biomarkers using a resource such as ours will help researchers gain a wider perspective of the disease state, with potential positive clinical impact.

Key Points

- Most comprehensive COVID-19 biomarkers with a robust biomarker data model.
- Annotated COVID-19 biomarkers suggest complex vascular pathobiology of the disease.
- COVID-19 biomarkers display a large extent of commonality with established cancer biomarkers (ACE2, IL-6, IL-4 and IL-2) as well as biomarkers for metabolic syndrome and diabetes (CRP, NLR, LDL).

Data availability and License

Authors' Contributions

Conception and design: NG and RM.

Collection and assembly of data: NG, AB, DL and HIVE lab volunteers.

Data analysis and interpretation: NG, AB, DL, KC, RK, DN, LS, SS, RM.

Manuscript writing: All authors.

Final approval of manuscript: All authors.

Accountable for all aspects of the work: All authors.

Acknowledgements

HIVE Lab (https://hive.biochemistry.gwu.edu): Sneh Talwar (content developer); Ashia Joseph (bioinformatics curator).
Collaborators: Dr. Shant Ayanyan (physician);
Volunteers: Anders Gyllenhoff (bioinformatics curator); Andy Cao (bioinformatics curator); Anjali Shankar (bioinformatics curator); Antarjot Kaur (bioinformatics curator); Arya Adake (bioinformatics curator); Ashia Joseph (bioinformatics curator); Avery Ye (bioinformatics curator); Chakshu Gandhi (bioinformatics curator); Dia Jhaveri (bioinformatics curator); Gracelyn Hill (bioinformatics curator); Helen Ibeawuchi (bioinformatics curator); Jonathan Ye (bioinformatics curator); Kathryn Cowie (bioinformatics curator); Kristina Ayers (bioinformatics curator); Mariana Escalante (bioinformatics curator); Meylakh Barshay (bioinformatics curator); Miguel Mazumder (bioinformatics curator); Niharika Chandna (bioinformatics curator); Nikita Wagle (bioinformatics curator); Nora Shepherd (bioinformatics curator); Noyanika Vattathara (bioinformatics curator); Nuerye Ainiwan (bioinformatics curator); Pranav Mishra (bioinformatics curator); Renee Long (bioinformatics curator); Rishab Desai (bioinformatics curator); Rita Mazumder (Asst. coordinator and bioinformatics curator); Ruqaia Al-Kohlany (bioinformatics curator); Sahana Ramesh (bioinformatics curator); Sara Burr (bioinformatics curator); Sejal Singh (bioinformatics curator); Siddharth Krishnan (bioinformatics curator).

Funding

National Cancer Institute (Grant No. U01CA215010 to R.M.); National Science Foundation (DEB-2028280 to K.A.C.).

References

1. Lai CC, Shih TP, Ko WC, *et al.* Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): the epidemic and the challenges. *Int J Antimicrob Agents* 2020;**55**:105924.
2. World_Health_Organization *WHO Coronavirus Disease (COVID-19) Dashboard*. https://covid19.who.int/ (2020).
3. Coronaviridae Study Group of the International Committee on Taxonomy of, V. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* 2020;**5**:536–44.
4. Andersen KG, Rambaut A, Lipkin WI, *et al.* The proximal origin of SARS-CoV-2. *Nat Med* 2020;**26**:450–2.
5. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol* 2019;**17**:181–92.

6. Su S, Wong G, Shi W, *et al*. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends Microbiol* 2016;**24**:490–502.
7. Prompetchara E, Ketloy C, Palaga T. Immune responses in COVID-19 and potential vaccines: lessons learned from SARS and MERS epidemic. *Asian Pac J Allergy Immunol* 2020;**38**:1–9.
8. Tang B, Bragazzi NL, Li Q, *et al*. An updated estimation of the risk of transmission of the novel coronavirus (2019-nCov). *Infect Dis Model* 2020;**5**:248–55.
9. Wang B, Li R, Lu Z, *et al*. Does comorbidity increase the risk of patients with COVID-19: evidence from meta-analysis. *Aging* 2020;**12**:6049–57.
10. Yi Y, Lagniton PNP, Ye S, *et al*. COVID-19: what has been learned and to be learned about the novel coronavirus disease. *Int J Biol Sci* 2020;**16**:1753–66.
11. Huang C, Wang Y, Li X, *et al*. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020;**395**:497–506.
12. Wu F, Zhao S, Yu B, *et al*. A new coronavirus associated with human respiratory disease in China. *Nature* 2020;**579**:265–9.
13. Nicholls JM, Poon LLM, Lee KC, *et al*. Lung pathology of fatal severe acute respiratory syndrome. *Lancet* 2003;**361**:1773–8.
14. WONG CK, Lam CWK, Wu AKL, *et al*. Plasma inflammatory cytokines and chemokines in severe acute respiratory syndrome. *Clin Exp Immunol* 2004;**136**:95–103.
15. Xia Y, Jin R, Zhao J, *et al*. Risk of COVID-19 for cancer patients. *Lancet Oncol* 2020;**21**(4):e180.
16. Zhou F, Yu T, du R, *et al*. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* 2020;**395**:1054–62.
17. *FDA-NIH_Biomarker_Working_Group in BEST (Biomarkers, EndpointS, and other Tools) Resource (Silver Spring (MD))*; (2016).
18. O'Connor JP, Aboagye EO, Adams JE, *et al*. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol* 2017;**14**:169–86.
19. Zhang L, Guo H. Biomarkers of COVID-19 and technologies to combat SARS-CoV-2. *Adv Biomark Sci Technol* 2020;**2**:1–23.
20. Hristova VA, Chan DW. Cancer biomarker discovery and translation: proteomics and beyond. *Expert Rev Proteomics* 2019;**16**:93–103.
21. Nature_editoriale. Early detection: a long road ahead. *Nat Rev Cancer* 2018;**18**:401.
22. Durinx C, McEntyre J, Appel R, *et al*. Identifying ELIXIR Core data resources. *F1000Res* 2016;**5**:2422.
23. Srivastava S, Rossi SC. Early detection research program at the NCI. *Int J Cancer* 1996;**69**:35–7.
24. Dingerdissen HM, Bastian F, Vijay-Shanker K, *et al*. OncoMX: a knowledgebase for exploring cancer biomarkers in the context of related cancer and healthy data. *JCO Clin Cancer Inform* 2020;**4**:210–20.
25. Information, N.C.F.B. Database resources of the National Center for biotechnology information. *Nucleic Acids Res* 2018;**46**:D8–d13.
26. Madeira F, Park Y, Lee J, *et al*. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* 2019;**47**:W636–41.
27. Alliance_of_Genome_Resources_Consortium. Alliance of genome resources portal: unified model organism research platform. *Nucleic Acids Res* 2019;**48**(D1):D650–D658.
28. de Coronado S, Wright LW, Fragoso G, *et al*. The NCI thesaurus quality assurance life cycle. *J Biomed Inform* 2009;**42**:530–9.
29. Haendel MA, Chute CG, Robinson PN. Classification, ontology, and precision medicine. *N Engl J Med* 2018;**379**:1452–62.
30. Köhler S, Carmody L, Vasilevsky N, *et al*. Expansion of the human phenotype ontology (HPO) knowledge base and resources. *Nucleic Acids Res* 2019;**47**:D1018–27.
31. Mungall CJ, Torniai C, Gkoutos GV, *et al*. Uberon, an integrative multi-species anatomy ontology. *Genome Biol* 2012;**13**:R5.
32. Munir K, Anjum MS. The use of ontologies for effective knowledge modelling and information retrieval. *Appl Comput Inform* 2018;**14**(2):116–26.
33. Sharma DK, Solbrig HR, Tao C, *et al*. Building a semantic web-based metadata repository for facilitating detailed clinical modeling in cancer genome studies. *J Biomed Semant* 2017;**8**:19.
34. Smith B, Arabandi S, Brochhausen M, *et al*. Biomedical imaging ontologies: a survey and proposal for future work. *J Pathol Inform* 2015;**6**:37.
35. UniProt_Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;**47**:D506–15.
36. Kim S, Chen J, Cheng T, *et al*. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* 2018;**47**:D1102–9.
37. Diehl AD, Meehan TF, Bradford YM, *et al*. The cell ontology 2016: enhanced content, modularization, and ontology interoperability. *J Biomed Semant* 2016;**7**:44.
38. Natale DA, Arighi CN, Blake JA, *et al*. Protein ontology (PRO): enhancing and scaling up the representation of protein entities. *Nucleic Acids Res* 2017;**45**:D339–d346.
39. Wu TJ, Schriml LM, Chen QR, *et al*. Generating a focused view of disease ontology cancer terms for pan-cancer data integration and analysis. *Database (Oxford)* 2015;**2015**:bav032.
40. Berman HM, Westbrook J, Feng Z, *et al*. The protein data Bank. *Nucleic Acids Res* 2000;**28**:235–42.
41. Sioutos N, Coronado S, Haber MW, *et al*. NCI thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform* 2007;**40**:30–43.
42. Hastings J, Owen G, Dekker A, *et al*. ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res* 2016;**44**:D1214–9.
43. Forrey AW, McDonald C, DeMoor G, *et al*. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clin Chem* 1996;**42**:81–90.
44. Chaturvedi AK, Caporaso NE, Katki HA, *et al*. C-reactive protein and risk of lung cancer. *J Clin Oncol* 2010;**28**:2719–26.
45. Wang L. C-reactive protein levels in the early stage of COVID-19. *Med Mal Infect* 2020;**50**:332–4.
46. Herold T, Jurinovic V, Arnreich C, *et al*. Elevated levels of IL-6 and CRP predict the need for mechanical ventilation in COVID-19. *J Allergy Clin Immunol* 2020;**146**:128–136.e124.
47. de Lang A, Osterhaus AD, Haagmans BL. Interferon-gamma and interleukin-4 downregulate expression of the SARS coronavirus receptor ACE2 in Vero E6 cells. *Virology* 2006;**353**:474–81.
48. Mehta P, McAuley D, Brown M, *et al*. COVID-19: consider cytokine storm syndromes and immunosuppression. *Lancet* 2020;**395**:1033–4.
49. Yang AP, Liu JP, Tao WQ, *et al*. The diagnostic and predictive role of NLR, d-NLR and PLR in COVID-19 patients. *Int Immunopharmacol* 2020;**84**:106504.
50. Li H, Xiang X, Ren H, *et al*. Serum amyloid a is a biomarker of severe coronavirus disease and poor prognosis. *J Infect* 2020;**80**:646–55.

51. Uhlar CM, Whitehead AS. Serum amyloid a, the major vertebrate acute-phase reactant. *Eur J Biochem* 1999;**265**:501–23.
52. Adam SS, Key NS, Greenberg CS. D-dimer antigen: current concepts and future prospects. *Blood* 2009;**113**:2878–87.
53. Zhang L, Yan X, Fan Q, *et al*. D-dimer levels on admission to predict in-hospital mortality in patients with Covid-19. *J Thromb Haemost* 2020;**18**:1324–9.
54. Franchini M, Lippi G. The role of von Willebrand factor in hemorrhagic and thrombotic disorders. *Crit Rev Clin Lab Sci* 2007;**44**:115–49.
55. Aksenova AY. Von Willebrand factor and endothelial damage: a possible association with COVID-19. *EcoGen* 2020;**18**:135–8.
56. Escher R, Breakey N, Lämmle B. Severe COVID-19 infection associated with endothelial activation. *Thromb Res* 2020;**190**:62.
57. Connors JM, Levy JH. COVID-19 and its implications for thrombosis and anticoagulation. *Blood* 2020;**135**:2033–40.
58. Fan J, Wang H, Ye G, *et al*. Letter to the editor: low-density lipoprotein is a potential predictor of poor prognosis in patients with coronavirus disease 2019. *Metab Clin Exp* 2020;**107**:154243.
59. Klonoff D, Umpierrez G. COVID-19 in patients with diabetes: risk factors that increase morbidity. *Metabolism* 2020;**108**:154224.
60. Lou M, Luo P, Tang R, *et al*. Relationship between neutrophil-lymphocyte ratio and insulin resistance in newly diagnosed type 2 diabetes mellitus patients. *BMC Endocr Disord* 2015;**15**:9.
61. Mishra Y, Pathak BK, Mohakuda SS, *et al*. Relation of D-dimer levels of COVID-19 patients with diabetes mellitus. *Diabetes Metab Syndr* 2020;**14**:1927–30.
62. Srikanthan K, Feyh A, Visweshwar H, *et al*. Systematic review of metabolic syndrome biomarkers: a panel for early detection, management, and risk stratification in the west Virginian population. *Int J Med Sci* 2016;**13**:25–38.
63. Arelaki S, Arampatzioglou A, Kambas K, *et al*. Gradient infiltration of neutrophil extracellular traps in colon cancer and evidence for their involvement in tumour growth. *PLoS One* 2016;**11**:e0154484.
64. Thålin C, Lundström S, Seignez C, *et al*. Citrullinated histone H3 as a novel prognostic blood marker in patients with advanced cancer. *PLoS One* 2018;**13**:e0191231.
65. Skarstein Kolberg E. ACE2, COVID19 and serum ACE as a possible biomarker to predict severity of disease. *J Clin Virol* 2020;**126**:104350.
66. Pepe MS, Etzioni R, Feng Z, *et al*. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst* 2001;**93**:1054–61.
67. Srivastava S. Cancer biomarker discovery and development in gastrointestinal cancers: early detection research network-a collaborative approach. *Gastrointest Cancer Res* 2007;**1**:S60–3.
68. Wahlberg K, Høyer-Hansen G, Casslén B. Soluble receptor for urokinase plasminogen activator in both full-length and a cleaved form is present in high concentration in cystic fluid from ovarian cancer. *Cancer Res* 1998;**58**:3294–8.
69. Virogates suPAR_mongraph_v3. https://www.virogates.com/wp-content/uploads/2020/01/20191008_English_suPAR_mongraph_v3.pdf (2019).
70. York WS, Mazumder R, Ranzinger R, *et al*. GlyGen: computational and informatics resources for glycoscience. *Glycobiology* 2020;**30**:72–3.