Bonn-Aachen International Center for Information Technology

(B-IT)

University of Bonn

Master Program in Life Science Informatics

Master Thesis

**Landscaping of COVID-19 Clinical Trials for the Discovery of Insightful Patterns on Ethnicities**

Submitted by : Ram Kumar Ruppa Surulinathan

First Reviewer : Prof. Dr. Martin Hofmann-Apitius

Second Reviewer : Prof. Dr. Juliane Fluck

Internal Supervisor : Dr. Marc Jacobs

*A thesis submitted in fulfillment of the requirements for the degree of M.Sc. in Life Science Informatics in the*

Bonn-Aachen International Center for Information Technology (B-IT)

*in collaboration with*

Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)

August 2021

**ACKNOWLEDGEMENTS**

(draft)—natascha, meike, alpha, bruce, sepehr

I am extremely grateful to Prof. Dr. Martin Hofmann-Apitius, Head of Department of Bioinformatics at Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin for offering me a place in his group to perform my master thesis on COVID-19 and clinical trials.

I sincerely express my deep sense of gratitude to Dr. Marc Jacobs, Group leader of Software and Scientific Computing and Deputy Head of Department of Bioinformatics, for initiating and inspiring me into this research work, invaluable guidance, patience and encouragement throughout the period of the work.

Finally, we would like to offer our genuine gratitude to the God almighty for blessing us with this opportunity to explore the realms of science and last but not the least we would like to dedicate this to our parents who prayed for us and motivated us throughout our work and giving us the strength to prevail.

**A SMALL NOTE**

At this juncture, I am reminded of the renowned philosopher Ayyan Thiruvalluvar’s words, which is stated below:

*With rising flood the rising lotus flower its stem unwinds;*

*The dignity of men is measured by their minds.*

- Adapted from Thirukural Couplet No. 595

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| 1 | Introduction | 1 |
|  |  |  |
|  | Goals |  |
|  | Thesis Organization |  |
| 2 | Theoretical Background |  |
|  |  |  |
|  |  |  |
|  |  |  |
| 3 | Materials and Methods |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| 4 | Results |  |
|  |  |  |
|  |  |  |
|  |  |  |
| 5 | Conclusion |  |
|  |  |  |
|  |  |  |
| 6 | Scientific References |  |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure Number** | **Title** | **Page** |
|  | Structure of novel coronavirus SARS-CoV-2 |  |
|  | Manhattan plot representing a GWAS results |  |
|  | A simple directed labelled graph |  |
|  |  |  |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Table Number** | **Title** | **Page** |
|  | The list of various relationships and functions described by the BEL statements |  |
|  |  |  |
|  |  |  |
|  |  |  |

**LIST OF ABBREVIATIONS**

|  |  |
| --- | --- |
| SARS-CoV-2 | Severe Acute Respiratory Syndrome Coronavirus-2 |
| ICTRP | International Clinical Trials Registry Platform |
| WHO | World Health Organization |
| ECRIN | European Clinical Research Infrastructure Network |
| MDR | Metadata Repository |
| BEL | Biological Expression Language |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**ABSTRACT**

**INTRODUCTION**

**THEORETICAL BACKGROUND**

2.1 SARS-CoV-2 a novel coronavirus

The Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) a novel strain and fatal coronavirus was first identified in Wuhan city of China in December 2019. The common symptoms include increased body temperature, dry cough, nausea and body pains. The rapid spread of the virus posed a threat of life to the global environment. According to the records maintained by Worldometer, the top ten most affected nations include USA, India, Brazil, France, Russia, UK, Turkey, Italy, Spain and Germany (*COVID Live Update: Worldometer*, n.d.).

The coronaviruses generally is classified under the family Coronoviridae and subfamily Coronavirinae which is subdivided into four genera namely Alphacoronavirus, Betacoronavirus, Gammacoronavirus and Deltacoronavirus (Mittal et al., 2020). The SARS-CoV-2, a member of Betacoronavirus genera whose sequence is 96% homologous to the bat coronavirus. Its primary reservoir is considered to be bats and transmitted to human beings through an intermediate host called Pangolin (Zhao et al., 2020).

The SARS-CoV-2 also called by COVID-19 is spherical in structure with positively stranded RNA genome packed inside the nucleocapsid protein (N) and enveloped by the membrane glycoprotein protein (M), envelope protein (E), and the spike protein (S). The typical virus lengths between 26.4 and 31.7 kb with the GC content ranging between 32% and 43% thus indicating to be the largest RNA virus (Mousavizadeh & Ghasemi, 2020).

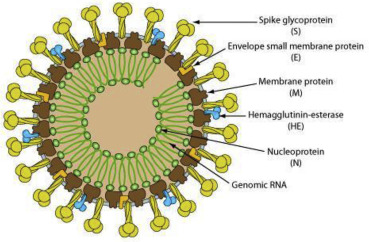


Figure 1 Structure of novel coronavirus SARS-CoV-2, Adapted from (Mousavizadeh & Ghasemi, 2020)

2.2 Clinical trials

Clinical trials are experimental studies performed using volunteers with the objective of examining different interventions like medical, surgical or behavioral ones. They can be classified into interventional studies and observational studies. These studies are led by a medical doctor assisted by other doctors, nurses carried out in hospitals, universities and research institutes (*Learn About Clinical Studies - ClinicalTrials.Gov*, n.d.).

The United States Food and Drug Administration (FDA) has defined the different stages of clinical trial phases to determine whether the drug could be employed for public usage. They are described as follows:

Phase I: During the initial phase, around 20 to 80 participants are enrolled having no underlying medical complications to evaluate the highest dosage levels that can be administrated without serious side effects.

Phase II: During this phase, around 100 to 300 participants are enrolled to evaluate the effectiveness of the medication along with short time side effects if occurred. This is carried out up to several years.

Phase III: During this phase, around 3000 participants are enrolled to evaluate the effectiveness of the medication across diverse population and varied dosage thus studying both the drug safety and efficacy. The rare and long time effects are observed during this phase. The medications are approved by the FDA if the trial results are positive.

Phase IV: During this final phase, long time side effects and efficacy of the approved medications are evaluated across the thousands of participants (*What Are Clinical Trials and Studies? | National Institute on Aging*, n.d.).

2.3 Clinical Trial Registries

The results of the clinical studies are recorded and published in a repository called Clinical Trial Registries. They are open source and their accession is available for scientific community and public. Each nation has their own trial registry systems centrally maintained by their Government or other approved institution. Some notable registries are described as follows:

2.3.1 ClinicalTrials.gov

ClinicalTrials.gov is an US based trial registry maintained by the National Library of Medicine (NLM) at the National Institute of Health (NIH), made available in February 2000. It holds both interventional and observational studies carried in 50 US states and 220 countries. The records contains information pertaining to the disease under investigation; the type of interventions employed; meta information like ethnicity of the participants, demographic data, inclusion/exclusion criteria for the study, location, list of comorbidities if any, clinical variables and laboratory variables. The results are included sometimes if the trials are subject to Section 801 of FDAAA (FDAAA 801) (*ClinicalTrials.Gov Background - ClinicalTrials.Gov*, n.d.).

The ClinicalTrials.gov registry could be accessed via <https://clinicaltrials.gov/>. The registry actually recorded 5,420 studies for the novel COVID 19 virus as of 21st April 2021.

2.3.2 ICTRP

International Clinical Trials Registry Platform (ICTRP) is a project of World Health Organization (WHO) is to ensure the registration of *WHO Trial Registration Data Set* and its accessibility to the public. In order for the clinical study to be considered as fully registered, it must contain atleast minimum amount of information called as Trial Registration Data Set (TRDS). Some of them include title of the study, disease conditions investigated, participant’s location, type of the study, duration of the study and outcome of the study etc. (*About ICTRP*, n.d.).

The ICTRP could be accessed via <https://apps.who.int/trialsearch/>. It was established in August 2005 in Geneva, Switzerland.

2.3.3 EU Clinical Trials Register

The EU Clinical Trials Register records the interventional studies conducted in the European Union (EU) or the European Economic Area (EEA) started after May 01, 2004. In this registry, the description of phase II to phase IV along with its summary results are available. The summary results include the trial information, endpoints, adverse effects identified in patients if available with the additional information. The registry doesn’t contain any information on non-interventional studies, surgical procedures, medical devices and psychotherapeutic procedures (*About the EU Clinical Trials Register*, n.d.). The EU Clinical Trials Register could be accessed via <https://www.clinicaltrialsregister.eu/ctr-search/search>.

2.3.4 ECRIN-MDR

European Clinical Research Infrastructure Network (ECRIN) is an EU based non-profit organization built to facilitate multinational clinical research across twelve EU countries. To support the COVID-19 research, ECRIN developed the Metadata Repository (MDR). It standardizes the metadata about the clinical studies and thus it could be accessed by a web interface. This portal is an open source enabling researchers to access worldwide with the results directing to the open access journal article or a trial registry entry if the results are publicly available (*Clinical Research Metadata Repository | ECRIN*, n.d.). The ECRIN-MDR could be accessed via <https://ecrin.org/tools/clinical-research-metadata-repository>.

2.3.5 Global Coronavirus COVID-19 Clinical Trial Tracker

This is a specialized real time dashboard of clinical trial for COVID-19. The results are gathered from International Clinical Trials Registry Platform (ICTRP), Chinese Clinical Trial Registry (ChiCTR), ClinicalTrials.gov, EU Clinical Trials Register, Clinical Research Information Service – Republic of Korea (CRiS), Iranian Registry of Clinical Trials (IRCT), Japan Primary Registries Network (JPRN) and German Clinical Trials Register (DRKS) . To identify potential clinical studies Artificial Intelligence (AI) based methods are employed. COVID-19 trials are mapped based on geographical, patient and intervention characteristics whose results are visualized in a convincing plots (Thorlund et al., 2020). The real time dashboard could be accessed via <https://www.covid-trials.org/>.

2.4 Biological Expression Language (BEL)

The scientific knowledge available within a literature is mostly in the form of long free text and thus capturing of the available knowledge is cumbersome. This drawback is overcame by the Biological Expression Language (BEL) that represents the syntactical representation of biological relationships. Thus, the Biological Expression Language (BEL) originally designed by Selvanta is a language developed for the representation of knowledge in the life science domains thus paving the way for biocuration and enabling in identifying disease mechanisms, pathways, etc. in the literature (Madan et al., 2019).

In this language, the statements are transformed into subject, relation, object. The most common relationships are “increase” and “decrease”. The normalized namespaces for the subject and object is resulted from different databases entries like HGNC for human genes, MGI for mouse genes and ChEBI for chemical entities. The most common biological entities like gene is notated as g(), protein as p(), mRNA as r().



Figure : Structure of a BEL statement. Adapted from (Madan et al., 2019).

The various functions described by the BEL are listed below:

|  |  |
| --- | --- |
| **Function** | **Notation** |
| Abundance | a() |
| Biological process | bp() |
| Cell Secretion | sec() |
| Location | loc() |
| Degaradation | deg() |
| Fragment | frag() |
| Pathology | path() |
| Translocation | tloc() |
| Fusion | fus() |
| Variant | var() |

The various relationships described by the BEL are listed below:

|  |  |
| --- | --- |
| **Relationship** | **Meaning** |
| Increases | A indirectly increases B |
| Decreases | A indirectly decreases B |
| directlyIncreases | A directly increases B |
| directlyIncreases | A directly decreases B |
| isA | A is a subset of B |
| positiveCorrelation | A is positively correlated with B |
| negativeCorrelation | A is negatively correlated with B |
| biomarkerFor | A is a biomarker for B |
| transcribedTo | gene is transcribed to RNA |
| translatedTo | RNA is translated to Protein |

Table : The list of various relationships and functions described by the BEL statements. Adapted from (*Relations :: Documentation for BEL Language*, n.d.)

2.5 Ontology:

Ontologies are developed to capture knowledge about a specific domain of interest like Genes, Clinical Trials, Parkinson Disease, COVID-19 etc. The ontology best describes the concepts within the specified domain along with the exploration of relationships existing between the concepts. An Ontology consisted of various components including Individuals, Classes, Attributes, Relations, Restrictions, Rules and Axioms. These ontologies are based on the *Open Biological and Biomedical Ontology* (OBO) Foundry principles in *Web Ontology Language* (OWL) representations.

Some applications of the Ontology includes extraction of information from various sources, improvement of communication/interoperability between people and organizations.

2.5.1 COVID-19 Ontology

The COVID-19 ontology is exclusively designed to facilitate COVID-19 research in its virology, epidemiology, clinical aspects, disease maps and thus widely applied in text mining and drug repurposing. For building the ontology, the concepts and entities were derived from several research articles, reviews, COVID-19 related websites which are assembled using Protégé ontology editor on the principles of *Open Biological and Biomedical Ontology* (OBO) (Sargsyan et al., 2021).

The metrics of the COVID-19 ontology are described as below:

|  |  |
| --- | --- |
| **Metrics** | **Total Number** |
| Classes | 2270 |
| Individuals | 6 |
| Properties | 10 |
| Maximum depth | 15 |
| Maximum number of children | 247 |
| Average number of children | 3 |
| Classes with single child | 447 |
| Classes with more than 25 children | 8 |
| Classes with no definition | 860 |

Table : Metrics of COVID-19 Ontology. Adapted from (*COVID-19 Ontology - Summary | NCBO BioPortal*, n.d.)

This ontology is made available for free to the scientific community in several platforms including BioPortal, a dedicated repository for biomedical ontologies. It is available in the web-link <https://bioportal.bioontology.org/ontologies/COVID-19>.

2.6 COVID-19 Literature mining

The primary step of analyzing the COVID-19 associated information is exploration of information from biomedical literature by semantic search and information retrieval system. The free search engine PubMed that contains citations related to biomedicine, life sciences and other associated fields uses Medical Subject Headings (MeSH) for the purpose of annotation of abstracts thus enabling the semantic search. Annotation of biological entities in the text corpus is not supported by the PubMed. The manual annotation of the large volumes of scientific literatures is practically infeasible.

Natural Language Processing (NLP) is an advanced field of artificial intelligence is a technique for the automatic extraction of knowledge from the unstructured data like the scientific literatures in this scenario The identification of different biological entities is known as “Named Entity Recognition (NER)” finds out the names of genes, proteins, chemicals, small molecules etc. The knowledge discovery process consisted of constructing network graphs using entities and relationships for the prediction of new function (Krallinger & Valencia, 2005).

SCAIView is a tool that integrates the knowledge discovery and semantic search.

2.6.1 SCAIView

The software SCAIView employs machine learning and Named Entity Recognition (NER) for the identification of biological entities from the scientific articles. Thus, the tool enables the researchers to answer the complex scientific queries in a simple and intuitive way as possible.

The key features of SCAIView includes user friendly query builder, accurate search results, visualization and ranking of the results and export of the results in a user preferred file format.

The SCAIView also supports the autocompletion features and the results are color coded thus representing different entities within the literature (*Introduction to SCAIView*, n.d.).

For the COVID-19 literature mining, there is a dedicated site called COPERIMOplus SCAIView at <https://coperimo.scaiview.com/> which is available for free to the scientific community.

2.7 COVID-19 Risk genes and Host genetics initiative

The human genome consisted of nearly 3 billion base pairs namely Adenine (A), Cytosine (C), Thymine (T) and Guanine (G). The genetic sequences between any two individuals are identical nearly up to 99.9% and the remaining 0.01% difference that makes each unique individual is called genetic variation. The “Genome Wide Association Studies (GWAS)” is performed to analyze whether the genetic variation in the regions of the genome are associated with the particular disease. In this study, there are two groups of participants namely the group having the disease under investigation (disease group) and the group not having the disease (normal group). The researchers identify for the variants that occur frequently thus indicating to be the risk effect for the disease group while protective effect on the control group. The results of the GWAS studies are considered to the risk signs that could be prominent for a specified disease in a group of population (*Explainer: Genome-Wide Association Studies | Broad Institute*, n.d.).

Concerning the COVID-19, COVID-19 Host Genetics Initiative (COVID-19 hg) was developed by Dr. Andrea Ganna from Institute of Molecular Medicine in Finland (FIMM) to understand the genetic variation and disease susceptibility/severity in COVID-19 patients. The COVID-19 hg was developed with the objective of generating, sharing and analyzing data related to COVID-19 genetic determinants. The knowledge could help the researchers for the drug repurposing and contribute to the global knowledge of COVID-19 (Andrea Ganna, 2020).

There are many challenges associated with the study of genetic variants:

1. Availability of large cohort size of population
2. Inability of the GWAS statistical methods to deal with the rare variants.
3. Identification of the rare variants and associating it with an exact biological etiology

The summary of the results are provided in visualization as Manhattan plot. It associates the COVID-19 traits with the genetic variants across the entire genome.

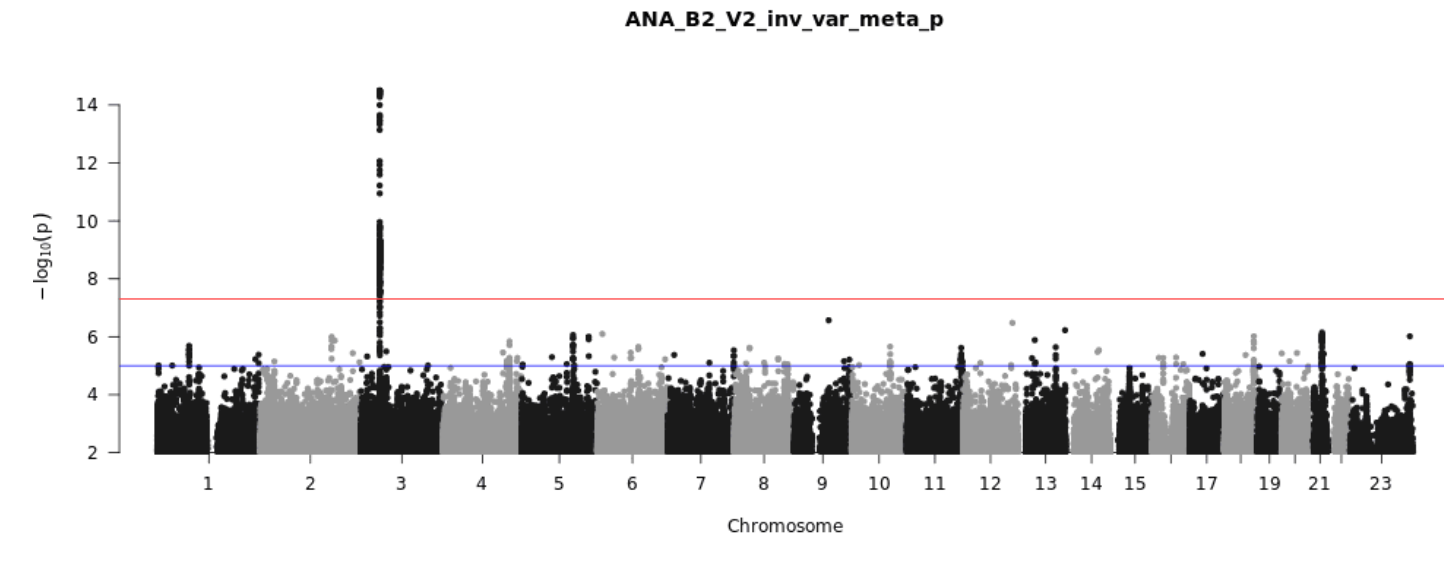


Figure: Manhattan plot representing a GWAS results. Adapted from (*COVID-19 HGI Results for Data Freeze 3 (July 2020) | Blog*, n.d.)

It contains the chromosome numbers along the X-axis and the log transformed p-value along the Y-axis. The points in the plot denotes the p-value between the variant and the disease under study. The most associated genetic variant is the variant that has the highest p-value (*COVID-19 HGI Results for Data Freeze 3 (July 2020) | Blog*, n.d.).

The COVID-19 hg results could be directly accessed from the weblink at <https://app.covid19hg.org/>.

2.8 Knowledge graph

For the mining of information from the large volumes of scientific literatures for answering complex queries, artificial intelligence technology called Knowledge graph (KG) is employed.

A knowledge graph is a directed labelled graph consisting of nodes representing the entities connected by edges representing the relationship between the nodes.

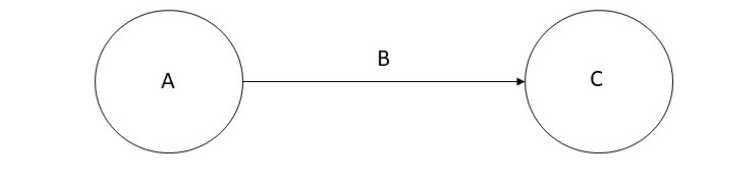


Figure : A simple Directed labelled graph. Adapted from (*What Is a Knowledge Graph?*, n.d.)

In the field of life sciences, the knowledge graphs are employed in the literature searching for the specific query provided by the user, drug repurposing for identifying drugs effective for the disease etc. (Chatterjee et al., 2021).

For supporting the COVID-19 research, a knowledge graph specific to COVID-19 drug-target interactions, mechanism of infection, pathophysiology etc. was developed. It consisted of 4,016 nodes (genes, proteins, drugs etc.) and 10,232 relationships as edges (increases, decreases, has\_component etc.). The information available within the knowledge graph, most are pertaining to the proteins and biological processes associated with COVID-19.

The graph was developed from the corpus of 160 scientific literatures from various resources including PubMed, LitCovid etc. Then those information were manually encoded in Biological Expression Language (BEL) which resulted in source, relationship, target patterned statements.

An User Interface to explore the knowledge graph and query was developed using the Python Django and OrientDB as Biological Knowledge Miner (BikMi). It is also supported by an API. The web application can be accessed directly for free at <https://bikmi.covid19-knowledgespace.de/>.

An interesting application of the knowledge graph is the identification of drug candidates for the drug repurposing. The most prominent drugs for the treatment of HIV, Ebola, Malaria like Lopinavir/Ritonavir combination, Remdesivir, Hydroxychloroquine respectively are identified within the knowledge graph (Domingo-Fernández et al., 2020).

**SCIENTIFIC REFERENCES**

*About ICTRP*. (n.d.). Retrieved April 22, 2021, from https://www.who.int/clinical-trials-registry-platform/about

*About the EU Clinical Trials Register*. (n.d.). Retrieved April 22, 2021, from https://www.clinicaltrialsregister.eu/about.html

Andrea Ganna. (2020). The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *European Journal of Human Genetics*, *28*(6), 715–718. https://doi.org/10.1038/s41431-020-0636-6

Chatterjee, A., Nardi, C., Oberije, C., & Lambin, P. (2021). *Knowledge Graphs for COVID-19 : An Exploratory Review of the Current Landscape*.

*Clinical Research Metadata Repository | ECRIN*. (n.d.). Retrieved April 22, 2021, from https://ecrin.org/clinical-research-metadata-repository

*ClinicalTrials.gov Background - ClinicalTrials.gov*. (n.d.). Retrieved April 21, 2021, from https://clinicaltrials.gov/ct2/about-site/background

*COVID-19 HGI Results for Data Freeze 3 (July 2020) | Blog*. (n.d.). Retrieved May 8, 2021, from https://www.covid19hg.org/blog/2020-09-24-freeze-3-results/

*COVID-19 Ontology - Summary | NCBO BioPortal*. (n.d.). Retrieved April 29, 2021, from https://bioportal.bioontology.org/ontologies/COVID-19

*COVID Live Update: Worldometer*. (n.d.). Retrieved April 21, 2021, from https://www.worldometers.info/coronavirus/

Domingo-Fernández, D., Baksi, S., Schultz, B., Gadiya, Y., Karki, R., Raschka, T., Ebeling, C., Hofmann-Apitius, M., & Kodamullil, A. T. (2020). COVID-19 knowledge graph: A computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. *BioRxiv*, 19–21. https://doi.org/10.1101/2020.04.14.040667

*Explainer: Genome-Wide Association Studies | Broad Institute*. (n.d.). Retrieved May 8, 2021, from https://www.broadinstitute.org/visuals/explainer-genome-wide-association-studies

*Introduction to SCAIView*. (n.d.). Retrieved May 7, 2021, from https://www.scaiview.com/en/introduction.html

Krallinger, M., & Valencia, A. (2005). Text-mining and information-retrieval services for molecular biology. In *Genome Biology* (Vol. 6, Issue 7, pp. 1–8). BioMed Central. https://doi.org/10.1186/gb-2005-6-7-224

*Learn About Clinical Studies - ClinicalTrials.gov*. (n.d.). Retrieved April 21, 2021, from https://www.clinicaltrials.gov/ct2/about-studies/learn

Madan, S., Szostak, J., Komandur Elayavilli, R., Tsai, R. T. H., Ali, M., Qian, L., Rastegar-Mojarad, M., Hoeng, J., & Fluck, J. (2019). The extraction of complex relationships and their conversion to biological expression language (BEL) overview of the BioCreative VI (2017) BEL track. *Database*, *2019*(1). https://doi.org/10.1093/database/baz084

Mittal, A., Manjunath, K., Kumar, R., Id, R., Kaushikid, S., Kumar, S., & Vermaid, V. (2020). *COVID-19 pandemic: Insights into structure, function, and hACE2 receptor recognition by SARS-CoV-2*. https://doi.org/10.1371/journal.ppat.1008762

Mousavizadeh, L., & Ghasemi, S. (2020). Genotype and phenotype of COVID-19: Their roles in pathogenesis. In *Journal of Microbiology, Immunology and Infection*. Elsevier Ltd. https://doi.org/10.1016/j.jmii.2020.03.022

*Relations :: Documentation for BEL Language*. (n.d.). Retrieved April 28, 2021, from https://language.bel.bio/language/reference/2.1.0/relations/

Sargsyan, A., Kodamullil, A. T., Baksi, S., Darms, J., Madan, S., Gebel, S., Keminer, O., Jose, G. M., Balabin, H., DeLong, L. N., Kohler, M., Jacobs, M., & Hofmann-Apitius, M. (2021). The COVID-19 Ontology. *Bioinformatics*, *36*(24), 5703–5705. https://doi.org/10.1093/bioinformatics/btaa1057

Thorlund, K., Dron, L., Park, J., Hsu, G., Forrest, J. I., & Mills, E. J. (2020). A real-time dashboard of clinical trials for COVID-19. *The Lancet Digital Health*, *2*(6), e286–e287. https://doi.org/10.1016/S2589-7500(20)30086-8

*What Are Clinical Trials and Studies? | National Institute on Aging*. (n.d.). Retrieved April 21, 2021, from https://www.nia.nih.gov/health/what-are-clinical-trials-and-studies

*What is a Knowledge Graph?* (n.d.). Retrieved May 8, 2021, from https://web.stanford.edu/class/cs520/2020/notes/What\_is\_a\_Knowledge\_Graph.html

Zhao, J., Cui, W., & Tian, B. (2020). The Potential Intermediate Hosts for SARS-CoV-2. *Frontiers in Microbiology*, *11*, 2400. https://doi.org/10.3389/fmicb.2020.580137