Epidemiological Challenges in Pandemic Coronavirus Disease (COVID-19): Role of Artificial Intelligence

Abhijit Dasgupta1, Abhisek Bakshi2, Srijani Mukherjee1, Kuntal Das1, Soumyajeet Talukdar1, Pratyayee Chatterjee1, Sagnik Mondal1, Puspita Das1, Subhrojit Ghosh1, Archisman Som1, Pritha Roy1, Rima Kundu1, Akash Sarkar1, Arnab Biswas1, Karnelia Paul3, Sujit Basak4, Krishnendu Manna5, Chinmay Saha6, Satinath Mukhopadhyay7, Nitai P. Bhattacharyya7, and Rajat K. De8.

1 Department of Data Science, School of Interdisciplinary Studies, University of Kalyani, Kalyani, Nadia 741235, West Bengal, India; 2 Department of Information Technology, Bengal Institute of Technology, Basanti Highway, Kolkata 700150, West Bengal, India; 3 Department of Biotechnology, University of Calcutta, Kolkata 700019, West Bengal, India; 4 Department of Physiology and Biophysics, Stony Brook University, Stony Brook, New York 11794, USA; 5 Department of Food & Nutrition, University of Kalyani, Kalyani, Nadia 741235, West Bengal, India; 6 Department of Genome Science, School of Interdisciplinary Studies, University of Kalyani, Kalyani, Nadia 741235, West Bengal, India; 7 Department of Endocrinology and Metabolism, Institute of Post Graduate Medical Education & Research and Seth Sukhlal Karnani Memorial Hospital, Kolkata 700020, West Bengal, India; 8 Machine Intelligence Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700108, India.

Supplementary Material

* To be considered as ‘First Authors’.
* Correspondence should be addressed to Rajat K. De, Ph.D., Machine Intelligence Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700108, India, Email: rajat@isical.ac.in.
Databases for epidemiological studies

Here we are going to summarize some key characteristics of the publicly available data sets as follows.

**COVID-19 open research data set**

The Allen Institute for AI along with some leading research groups has shared the COVID-19 Open Research Data set (CORD-19). Such free resource comprises scholarly articles in order of thousands. The articles deal with the information on coronavirus family. One may apply natural language processing techniques to uncover the hidden relation among various findings. Such findings may help other researchers and doctors in assessing the outcome of some therapeutic procedure.

**WHO COVID-19 data**

World Health Organization (WHO) has already started publishing and updating information about the affected cases over the world in regular interval. The numbers of death and recovery results are provided, which convey the speed of spreading of coronavirus into different parts of the world. Besides, WHO has been sharing various reports related to the study on applying candidate vaccines and several drugs.

**ACAPS COVID-19**

Here various measures associated with coronavirus are integrated in a single platform. The data sets consider several issues, such as social distancing, movement restrictions, public health measures, social and economic measures and lock-down among others, for such measurement. Public health as well as socio-economic conditions are also considered here.

**World Bank indicators data set**

Presently, World Bank has taken an initiative to share data related to recent COVID-19 with the help of Humanitarian Data Exchange (HDX). HDX is an open platform that shares data across organizations during crises. HDX allows sharing data conveniently, using them for analysis.

---

1. https://pages.semanticscholar.org/coronavirus-research
2. https://allenai.org/
3. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/
4. https://data.humdata.org/dataset/acaps-covid19-government-measures-dataset
5. https://data.humdata.org/dataset/world-bank-indicators-of-interest-to-the-covid-19-outbreak
The data set can be broadly divided into three parts. Data are information related to health status of every individuals, basic hand sanitizing facility for the population with soap and water, and the population density with respect to a range of ages. Figures S1-S3 depict different entities and their attributes as well as relationship among them for the World Bank indicators database.

*Kaggle*

Kaggle, one of the largest data science community in the world, has tried to involve a number of scientists to visualize the pattern of such pandemic activity worldwide. In response this emergency, they have prepared a COVID-19 Open Research Dataset (CORD-19) by incorporating disease along with recovery/death related information in the form of tables. Moreover, they have developed a time series data set having the track of history of a large number of patients worldwide. Figure S4 depicts the entities and their attributes of a small portion of Kaggle database.

*Genetic sequence database*

Genetic sequence database is a compilation of all freely available annotated DNA sequences. DNA GenBank is the part of the International Nucleotide Sequence Database Collaboration, which comprises European Nucleotide Archive (ENA), GenBank at NCBI and DNA DataBank of Japan (DDBJ). These organizations exchange data among themselves regularly. The aim of GenBank database is to provide and promote access within the scientific community to the most recent and wide-ranging DNA sequence. National Center for Biotechnology Information (NCBI) has recently provided a set of SARS-CoV-2 sequences, accessible in the Sequence Read Archive (SRA) and GenBank. Currently, the repository contains 183 GeneBank sequences and 1 RefSeq sequence in Entrez Nucleotide, and the new NCBI Virus resource submitted from countries like China, Phillipines, Japan and Thailand.

*Genomic database*

Nextstrain provides a frequently updated view of publicly available data of COVID-19. In addition, it contributes a set of alongside powerful analytic and visualization tools that allow epidemiological understanding of the disease and empower the researchers for the solution. A genome database can be described as a storehouse of DNA sequences from different species of plants and animals.

---

6https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
7https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
8https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/
9https://nextstrain.org/ncov
Drug Database

In order to support the prediction based on different AI models, many drug-related databases have been developed to contain several types of drug-target interaction (DTI) information. Simultaneously, drug related databases are vital resources for DTI predictions in silico. Based on the content of databases, it can be subdivided into four categories, drug centered or target centered databases, DTI databases, DTI affinity databases, and other supporting databases. In the class of “drug centered or target centered databases”, seven databases are generally used, such as, BRENDA, PubChem, SuperDRUG2, DrugCentral, PDIT, Pharos, and ECOdrug. On the other hand, DTI database has been developed for collecting and validating the DTI and related information. ChEMBL, ChemProt 3.0, DGIdb 3.0, DrugBank, GtoPdb, KEGG, LINCS, PROMISCUOUS, STITCH, SuperTarget, and TTD have commonly used databases in this category. The aim of “DTI affinity databases” is to focus on the various collection of binding affinity among drugs, some related molecules and target proteins. In this category, PDBBind, BindingDB, and PDSP Ki have been frequently used as a repository of a more significant number of binding affinity data. Few other databases like FAERS, SIDER, and JAPIC are being used to obtain the information related to country-specific adverse drug reaction (ADR) reports and medication error reports.
Supplementary Figures
Figure S1: The figure illustrates the entity set of World Bank indicators database about COVID-19 and its attributes along with the relation among them (Part 1)
Figure S2: The figure illustrates the entity set of World Bank indicators database about COVID-19 and its attributes along with the relation among them (Part 2)

Figure S3: The figure illustrates the entity set of World Bank indicators database about COVID-19 and its attributes along with the relation among them (Part 3)
**Figure S4:** The figure illustrates the entity set of Kaggle database about COVID-19 and its attributes along with the relation among them.
References

[1] Bagherian, M. et al. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Briefings in Bioinformatics* (2019).

[2] Jeske, L., Placzek, S., Schomburg, I., Chang, A. & Schomburg, D. Brenda in 2019: a european elixir core data resource. *Nucleic acids research* **47**, D542–D549 (2019).

[3] Kim, S. et al. Pubchem substance and compound databases. *Nucleic acids research* **44**, D1202–D1213 (2016).

[4] Siramshetty, V. B. et al. Superdrug2: a one stop resource for approved/marketed drugs. *Nucleic acids research* **46**, D1137–D1143 (2018).

[5] Ursu, O. et al. Drugcentral 2018: an update. *Nucleic acids research* **47**, D963–D970 (2019).

[6] Wang, C. et al. Pdid: database of molecular-level putative protein–drug interactions in the structural human proteome. *Bioinformatics* **32**, 579–586 (2016).

[7] Nguyen, D.-T. et al. Pharos: collating protein information to shed light on the druggable genome. *Nucleic acids research* **45**, D995–D1002 (2017).

[8] Verbruggen, B. et al. Ecodrug: a database connecting drugs and conservation of their targets across species. *Nucleic acids research* **46**, D930–D936 (2018).

[9] Bento, A. P. et al. The chembl bioactivity database: an update. *Nucleic acids research* **42**, D1083–D1090 (2014).

[10] Kringelum, J. et al. Chemprot-3.0: a global chemical biology diseases mapping. *Database* **2016** (2016).

[11] Wagner, A. H. et al. Dgiddb 2.0: mining clinically relevant drug–gene interactions. *Nucleic acids research* **44**, D1036–D1044 (2016).

[12] Armstrong, J. F. et al. The iuphar/bps guide to pharmacology in 2020: extending immunopharmacology content and introducing the iuphar/mmv guide to malaria pharmacology. *Nucleic acids research* **48**, D1006–D1021 (2020).

[13] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research* **45**, D353–D361 (2017).

[14] Stathias, V. et al. Lincs data portal 2.0: next generation access point for perturbation-response signatures. *Nucleic acids research* **48**, D431–D439 (2020).
[15] Von Eichborn, J. et al. Promiscuous: a database for network-based drug-repositioning. *Nucleic acids research* **39**, D1060–D1066 (2010).

[16] Zhu, D., Vaishampayan, P. A., Venkateswaran, K. & Fox, G. E. Stitch: Algorithm to splice, trim, identify, track, and capture the uniqueness of 16s rRNAs sequence pairs using public or in-house database. *Microbial ecology* **61**, 669–675 (2011).

[17] Günther, S. et al. Supertarget and matador: resources for exploring drug-target relationships. *Nucleic acids research* **36**, D919–D922 (2007).

[18] Chen, X., Ji, Z. L. & Chen, Y. Z. Ttd: therapeutic target database. *Nucleic acids research* **30**, 412–415 (2002).

[19] Liu, Z. et al. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics* **31**, 405–412 (2015).

[20] Gilson, M. K. et al. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research* **44**, D1045–D1053 (2016).

[21] Roth, B. L., Lopez, E., Patel, S. & Kroeze, W. K. The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches? *The Neuroscientist* **6**, 252–262 (2000).

[22] Moore, T. J., Morrow, R. L., Dormuth, C. R. & Mintzes, B. Us food and drug administration safety advisories and reporting to the adverse event reporting system (faers). *Pharmaceutical Medicine* 1–6 (2020).

[23] Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J. & Bork, P. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology* **6** (2010).

[24] Murakami, T. Drug information and japic (japan pharmaceutical information center). *Nihon yakurigaku zasshi. Folia pharmacologica Japonica* **140**, 135 (2012).