Optimizing outbreak data management tools using epidemiological and genomic data collected during the COVID-19 pandemic, and integrating artificial intelligence tools to enable monitoring the evolution of an outbreak in real time

A Data Management Plan created using DMPonline.be

Creator: Jonathan Thibaut

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Grant number / URL: 1130423N

ID: 201224

Start date: 31-10-2022

End date: 30-10-2026

Project abstract:

With 6 million deaths caused by COVID-19, and 30 Ebola virus disease outbreaks identified since the late 1970s, emerging diseases are an important and continuous cause for concern worldwide, as our ability to rapidly contain outbreaks emerging can rapidly have a global impact. A key element of disease control lies in the ability to rapidly understand transmission patterns and measure the impact of (non-)pharmaceutical interventions. In real life conditions, the data collection and analysis surrounding the early phase of an outbreak can have a critical impact on further epidemiological evolution. Therefore, it is important to ensure that on-site teams have direct access to tools which allow them to input the information collected by testing & tracing, and to identify and update risk factors for infection. Since September 2020, the Laboratory of Clinical Microbiology at the KU Leuven has gathered unique sets of epidemiological and genomic data from a 30.000+ student community. From this data, I will first perform a joint epidemiological and phylogenetic analysis to assess the correctness of contact tracing data. Secondly, I will use this data to develop a decision model capable of identifying behavioral, social, immunological and chronological risk factors for SARS-CoV-2 infection. Based on this experience, and in synergy with the World Health Organization, we will perform on-sites studies to integrate analytical modules in an outbreak investigation tool.

Last modified: 31-07-2023

Optimizing outbreak data management tools using epidemiological and genomic data collected during the COVID-19 pandemic, and integrating artificial intelligence tools to enable monitoring the evolution of an outbreak in real time

FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .cvs,.pdf, .txt, .rtf, .dwg, .gml, ... • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
Contact tracing	COVID-19 contact tracing data, involving vaccination status, disease test results, description of high-risk contacts and high-risk events, ...	Reuse existing data	Digital	Compiled/aggregated data	.csv	<1GB	
Genomic sequencing	RNA sequences corresponding to the test results stored in contact tracing database.	Reuse existing data	Digital	Observational	.fasta	<1GB	
Phylogenetic pipeline	Scripts exploiting both contact tracing and genomic sequencing dataset to assess the correctness of contact tracing data.	Generate new data	Digital	Software	.py, .nwk	<10GB	
Decision model capable of identifying risk factors for SARS-CoV-2 infection.	Machine Learning model trained based on data contained in the contact tracing dataset.	Generate new data	Digital	Software	.py files (source code) .h5 (model's weights)	<1GB	
Contact tracing application	Web application providing an epidemics visualization tool.	Generate new data	Digital	Software	.py, .sqlite, .html, .css, .js	<10GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

Contact tracing: Non publicly available data stored on a KU Leuven internal server, and gathered under study protocol S64919. There is no persistent identifier for this dataset yet.
Genomic sequencing: Dataset available on GISAID (https://gisaid.org/).

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes, human subject data

EC approval:
Contact tracing: S64919
Software development is covered by S64919.

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes

Contact tracing: Approved study protocol S64919.
This dataset contains personal information from the student population in Leuven, namely:
- Demographic information: first name, surname, addresses of domicile and student residence, phone number, email address, recent activities and contacts.
- Social interactions: close contacts links between patients, textual and structured description of the transmission event including the corresponding location and date.
- Patients' clinical information: SARS-CoV-2 PCR test result, date of diagnosis, symptoms onset dates, vaccination status.

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

Contact tracing: The data stored in this data set was inputted via clear questionnaire forms. Those forms are available to all researchers inside the team. The (anonymized) data that will be shared in the future in the context of scientific publication will be available on a public Git repository with a corresponding README.txt file describing this data subset. A README.md file will also be created to describe the variables from the complete dataset, for the research team.
Genomic sequencing: The RNA sequences are anonymous and are publicly available and encoded within the format required by GISAID (https://gisaid.org/).
The three software applications will be stored on distinct public GitHub repositories (https://github.com/), including source files, compiled models and analysis graphs. Those GIT repositories will all be accompanied with a README.md file describing the software, its functionalities, and the corresponding requirements for anyone wishing to reuse or enhance the applications. Moreover, I will directly input comments in source code to describe the role of each written function.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

Contact tracing: The anonymized data that will be shared in the future in the context of scientific publication will be available in a GitLab repository. The whole contact tracing dataset (restricted access) is currently stored on a private KU Leuven's GitLab repository. It will also be stored in the central KU Leuven data centers. DataCite will be used to describe the dataset.
This data was gathered through structured web forms inputted by patients themselves and by contact tracers during the contact tracing procedure. This data is structured within a database software that allows for data export into a '.csv' file. Each file correspond to one form. The headers correspond to the fields through the form and each entry corresponds to one submitted form.
Genomic sequencing: Each RNA sequence is identified via a unique GISAID identifier (https://gisaid.org/).
The three software applications will be stored on distinct public GitHub repositories (https://github.com/). Each repository will be accompanied with one README file and in-file documentation. Links to repositories will be reported within their corresponding published articles.

3. Data storage & back-up during the research project

Where will the data be stored?

Contact tracing: The data is currently stored on server from the KU Leuven, structured via the Go.Data software, and accessible via the Go.Data API (https://www.who.int/tools/godata). It will also be stored in the central KU Leuven data centers.
Genomic sequencing: The data is stored and publicly accessible on GISAID.

Both contact tracing and genomic sequencing data are also stored on a private GIT repository stored on the KU Leuven GitLab server. This data is therefore unavailable to users outside the research team.

The source code of the three software applications will be stored on distinct GIT repositories stored on the public GitHub.com server. Software will be made available to anyone outside the research team or outside KU Leuven's organization. Moreover, a copy of the source code and corresponding datasets that the code analyses will be uploaded on KU Leuven's GitLab servers, and therefore on KU Leuven's datacenters.

How will the data be backed up?

The data will be stored on the KU Leuven's central servers with automatic back-up procedures.
The three software applications that will be uploaded on GitHub will also be stored on KU Leuven's central servers with automatic back-up procedures.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

Ample storage on KU Leuven group biomedical sciences' servers.
Ample storage on KU Leuven GitLab servers.
Ample storage on GitHub servers.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Complex data security setup including password access (with differential access for data entry, alterations, visualisation, export), access logs to database system, data discretion agreement for all users, regular audits of access logs, access controls to buildings.
The three software applications will be open-source, i.e. their source code will be publicly available.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

The storage of 2 GB in the central KU Leuven data centers managed by the group biomedical sciences costs around 2€ per year.
Storage on KU Leuven's GitLab servers is free.
Storage on GitHub servers is free.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

Contact tracing: We will preserve data for 10 years. Specific requirements are in place in agreement with the KU Leuven data protection officer. Non pseudonimised information can be kept but can only remain accessible to members of our research team. Others can only access encrypted data as soon as the epidemiological situation of the person involved allows for encryption to take place without compromising their follow up in the light of COVID-19 exposure.
Genomic sequencing: Publicly available data on GISAID. We do not need to preserve this data.
Software applications: We will preserve the source code in three distinct publicly available GitHub repositories. Moreover, we will preserve back-ups for 10 years on KU Leuven data centers.

Where will these data be archived (stored and curated for the long-term)?

Contact tracing: We will preserve this data on the central KU Leuven data centers.
Genomic sequencing: Publicly available data on GISAID. We do not need to preserve this data.
Software applications: Publicly available on GitHub repositories. A back-up will also be preserved on the central KU Leuven data centers.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

Contact tracing: 2€/year for 2GB.
Genomic sequencing: Publicly available data on GISAID. We do not need to preserve this data.
Software applications:

- GitHub: Free
- Backup on KU Leuven's data servers: Contact tracing: 2€/year for 2GB

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in a restricted access repository (after approval, institutional access only, ...)

- Yes, in an Open Access repository

Contact tracing data: Yes, in a restricted access repository (after approval, institutional access only, ...). This repository is stored on internal KU Leuven's server.
Genomic sequencing data: Yes, in an Open Access repository. This data is stored on GISAID and is therefore publicly available.
Software applications: Yes, in an Open Access repository. This data will be made available on public repositories stored on GitHub servers.

If access is restricted, please specify who will be able to access the data and under what conditions.

Contact tracing: Non pseudonimised information can be kept but can only remain accessible to members of our research team. Others can only access encrypted data as soon as the epidemiological situation of the person involved allows for encryption to take place without compromising their follow up in the light of COVID-19 exposure.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Privacy aspects

Contact tracing: This dataset contains personal data. No data will be shared with individuals outside of our research group without the prior consent of the KU Leuven / UZ Leuven ethics committee.
When sharing the data for scientific publication purpose, the data will be anonymized.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

Contact tracing data will be available on request after signing a data sharing agreement. EC approval will always be gained before sharing data.
Source code of software applications will be made available on public GitHub repositories. Those repositories are not available yet, and will be made available upon publication, once the code is finalized, commented, reviewed, and therefore publishable.

When will the data be made available?

Anomymized data will be made available upon acceptance of publication in a journal via a GitLab repository with the corresponding analyses script.
Software applications: After code peer-reviewing and as soon as a related publication is submitted to a scientific journal.

Which data usage licenses are you going to provide? If none, please explain why.

Contact tracing - anonymized data: I will provide the license CC-BY-NC-4.0. The material therefore cannot be used for commercial purpose.
Software: The software applications developed in the context of this PhD will be open source. I will therefore provide the AGPL-3.0-or-later license.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- No

What are the expected costs for data sharing? How will these costs be covered?

Data sharing is free of charge: Anonymized data and software applications will be available on GitHub.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Myself and my supervisors.

Who will manage data storage and backup during the research project?

The KU Leuven group biomedical sciences IT department

Who will manage data preservation and sharing?

Myself and my supervisors.

Who will update and implement this DMP?

The PI (Emmanuel André) bears the end responsibility of updating and implementing this DMP.