aMILE: Application of text mining to clinical reports of patients with acute myeloid leukemia

A Data Management Plan created using DMPonline

Creators: Rita Rb-Silva, Yulia Karimova

Affiliation: Other

Template: DCC Template

ORCID iD: 0000-0002-1422-0974

Project abstract:

The use of clinical data is key to the continuous improvement of health care and also to accelerate research directed towards prevention, diagnosis, and treatment innovation. At IPO-Porto, healthcare professionals and researchers have the support of several departments that are able to provide relevant data to answer their clinical and scientific questions, while preserving patients' privacy. Unfortunately, information about the previous medical history and some follow-up data are not available in easily accessible formats, because the registration of these data is not stored in structured formats, existing in .pdf files containing free text. This gap represents an important obstacle to perform retrospective cohort studies and to plan prospective observational or interventional protocols. The aim of this work is to create and validate text mining algorithms to extract relevant clinical data from .pdf files (such as the hospital discharge summaries and other medical reports) in a reliable, safe and confidential way, transforming them into structured format data. This study will only include data from patients with Acute Myeloid Leukemia.

ID: 57790

Last modified: 27-03-2021

Copyright information:

aMILE: Application of text mining to clinical reports of patients with acute myeloid leukemia

Data Collection

What data will you collect or create?

In this project, we will collect all the text data contained in .pdf files corresponding to the hospital discharge summaries, the Multidisciplinary Consultation report and other medical reports of patients with the diagnosis of acute myeloid leukemia from a single tertiary healthcare center.

The free text contained in the .pdf files (raw data) will be transformed into structured format data (processed data). All the unnecessary personal data included in the raw data will be excluded during the data processing and will be absent in the processed data.

How will the data be collected or created?

A list of identification codes provided by the Epidemiology Department of IPO-Porto corresponding to all the patients with the diagnosis of acute myeloid leukemia will be used. All the .pdf files corresponding to the codes in that list containing text data will be manually downloaded and saved by the Principal Investigator, in a specific folder (Folder A) in a computer with an encrypted disk and other privacy and safety measures, in alignment with General Data Protection Regulation (GDPR). The folder will only be accessible to the PI. More details regarding the data storage, back-up, selection and preservation are available below.

Structured format data will be extracted from the raw data using text mining algorithms in Python.

Documentation and Metadata

What documentation and metadata will accompany the data?

- A list of the identification codes of patients with acute myeloid leukemia will be stored in an excel file with a password, together with the raw text data, in a specific folder (Folder B) in a computer with an encrypted disk, in alignment with GDPR. More details are available below on this DMP.
- A DPIA (Data Privacy Impact Assessment) was evaluated by the Data Privacy Officer of IPO-Porto, and received a favourable opinion after confirmation of its compliance with the European Union data protection law. This document will be stored in a

local paper folder at EPOP, and it will not accompany the data.
- The project was analysed by the Ethical Committee for Health (Comissão de Ética para a Saúde, CES) and received a favourable opinion(Ref. CES. 167/020). A .pdf version of this document will be stored in Folder B.
- An agreement between the Principal Investigator and the other project collaborators regarding Intellectual Property Rights (IPR) ownership issues will be written and signed. A .pdf version of this document will be stored in Folder B.
- Other documents relevant to the project may be created and information about them will be added during monitorization and actualization of the DMP (at least every 6 months).
- Dublin Core schema will be used in the description of processed data. DDI will be used for the description of specific topics (e.g. instrument names).
- More details will be added in an updated version of this document.

Ethics and Legal Compliance

How will you manage any ethical issues?

1. Any ethical issues will be discussed and managed by all the authors from the two collaborating institutions (IPO-Porto and CHU HSJ).
2. All the procedures of the project will be evaluated by the Local Ethical Committee for Health and will only be performed after ethical approval.
3. The Principal Investigator will be responsible for the management of all data collected and processed in this project.
4. It is not necessary that patients sign an informed consent statement because this study is observational, not interventive, and data privacy and safety will be ensured by several well-established procedures, in alignment with GDPR. More details will be added in an updated version of this document.
5. The lawfulness of the treatment of the data collected in this study is based on the need of processing for the performance of a task carried out in the public interest, in accordance with Article 6 (1) (e) of the European Union (EU) General Data Protection Regulation (GDPR).
6. The .pdf files and their contents will be permanently deleted at the end of the work, and the collected data (raw text data) will be anonymized and codified. In this way, the processing of personal data will be restricted to the specific purpose that is intended to be achieved with this study, in accordance with Resolution 1704/2015 of the National Data Protection Commission.

More details will be added in an updated version of this document.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

The raw text data correspond to health information owned by the respective patients and can not be licensed for reuse. Processed anonymized and codified data belong to the project group. Intellectual Property Rights (IPR) ownership issues will be covered by a

written agreement between the Principal Investigator and the other project collaborators.

Storage and Backup

How will the data be stored and backed up during the research?

The data will be stored in a specific folder in a computer owned by the PI, with an encrypted disk using the XTS-AES-128 algorithm at rest and a strong encryption key.

A back-up will be saved by the PI, at least every 6 months, in a protected external hard drive owned by the PI. Only the last two back-up versions will be preserved and they will only be accessible to the PI. Older versions of the back-up will be deleted.

More details will be added in an updated version of this document.

How will you manage access and security?

ACCESS:

Only the Principal Investigator and two project collaborators (José Mário Mariz and Isabel Oliveira) will have full access to the raw data, because they are physicians at the Department where patients were diagnosed, treated and/or followed. They all already have access to that raw data. The other project collaborator (Tiago Taveira-Gomes) will only have access to processed data without any personal information.

The processed anonymized and codified data and their metadata will be saved in a specific folder (Folder C), accessible to the PI and all the project collaborators. The files in this folder will be shared at the end of the project in an open research data repository with a defined license without any access restriction.

The specific research data repository will be selected in the future and its description will be included in an updated version of this DMP.

SECURITY:

The data transfer will be carried out through a removable flash memory disk that will be formatted before the extraction process starts and will be formatted again at the end of it. This device will always be in the possession of the Principal Investigator and will have no other purpose nor will it be used on computers other than those of the IPO and the computer where the analysis will be performed.

There will be no data transfer to third party services, namely cloud services.

The computer where the raw data will be stored has an encrypted disk using the XTS-AES-128 algorithm at rest with a strong encryption key. Only the Principal Investigator will be able to access the device by combining a unique username and strong password, which will be required automatically after 5 minutes of inactivity.

The antivirus software on this computer will be licensed and constantly updated.

The security of backed up data will be assured by similar strategies.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

Raw data (corresponding to all the .pdf files) and raw data text will be definitely erased at the end of the project.

Processed data will be preserved because it will be necessary to validate the research findings and it includes data that cannot easily be recreated or produced and/or is costly to reproduce.

Processed anonymized and codified data will be preserved without time limit in Folder C.

The files contained in Folders A and B will only be shared with José Mariz and Isabel Oliveira, both authors of the project and physicians at the Department of Onco-Hematology.

What is the long-term preservation plan for the dataset?

The long-term preservation plan will be defined according to the requirements of the research data repository in which the processed data will be stored, preserved and shared.

The specific research data repository will be selected in the future and its description will be included in an updated version of this DMP.

Data Sharing

How will you share the data?

The processed anonymized and codified data and their metadata will be shared in an open research data repository without any restriction (Creative Commons BY).

Are any restrictions on data sharing required?

The sharing of processed data will have no restriction, without any embargo period.

Responsibilities and Resources

Who will be responsible for data management?

Responsible for DMP creation: Rita Rb-Silva (Principal Investigator, ORCID: 0000-0002-1422-0974) e Yulia Karimova (ORCID: 0000-0002-1015-6709).

Responsible for the collection, processing and preservation of the raw data: Rita Rb-Silva.

Responsible for the sharing of the processed data: Rita Rb-Silva.

Rita Rb-Silva (MD, PhD) is a resident doctor of Onco-Hematology at IPO-Porto, an invited assistant at the University of Porto and a researcher of the Population Health Research Domain of the School of Medicine of the University of Minho.

What resources will you require to deliver your plan?

During the project the following assets will be used:

- Hardware/devices: work desktop and laptop computers, personal laptop computers, USB flash drives, external disks, institutional servers, smartphones, tablets.
- Software: Windows, Linux, Excel, Word, Python, SQL, Adobe Acrobat Reader.
- Message exchanges: entities' email services, Google "GMail" emails, Skype.
- Paper transmission channels: notes related to teams, meetings, participants names, etc.
- Access to the selected research data repository.