RDA Plenary VP17
Sensitive Data Interest Group

Before, setting up for a sensitive data project

Lightning Talk

Rita Rb-Silva
MD, PhD

Department of Oncohematology
Portuguese Institute of Oncology of Porto

ICBAS – University of Porto

Porto, Portugal

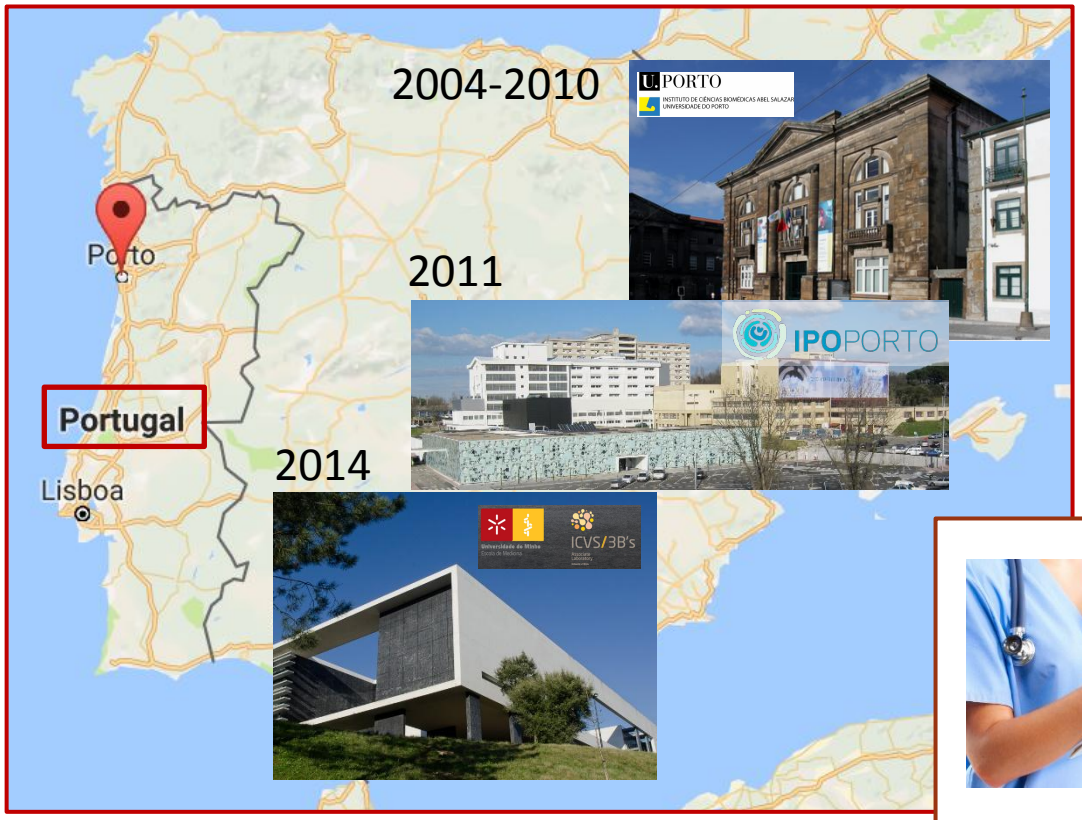
2021-04-22

IPOPORTO

U.PORTO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR
UNIVERSIDADE DO PORTO

RDA
RESEARCH DATA ALLIANCE

A little about myself



@ Calvin.Andrus DataScienceDisciplines.png
Case reports in pathology. 2015. 135684. 10.1155/2015/135684.

RESEARCH DATA ALLIANCE

A little about my hospital

© Pedro Vidinha

Portuguese Institute of Oncology of Porto:

- 26 departments
- 11 outpatient clinics
- Approx. > 1.000 physicians and nurses
- 45.000 patients

Background

Problem: Obstacles for health data use and reuse

1) Doctors and researchers spend a lot of time finding and collecting the data they need

2) The majority of health data is not used for research purposes

3) Eletronic health records: data privacy, confidenciality and safety
   - Collaborations are not easy

4) Finding solutions:
   RDA (2013)
   OHDSI (2014)
   John Snow Labs (2015)
   Opaque systems (UC Berkeley, 2019)
   TEHDAS (2021)
   ...

https://www.ohdsi.org/data-standardization/the-common-data-model/
https://www.nlpsummit.org/ehr-question-answering/
https://www.nlpsummit.org/secure-collaborative-learning-using-the-mc2-platform/

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

SUMMARY

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.

KEYWORDS: Text mining, Clinical Research, Acute Myeloid Leukemia

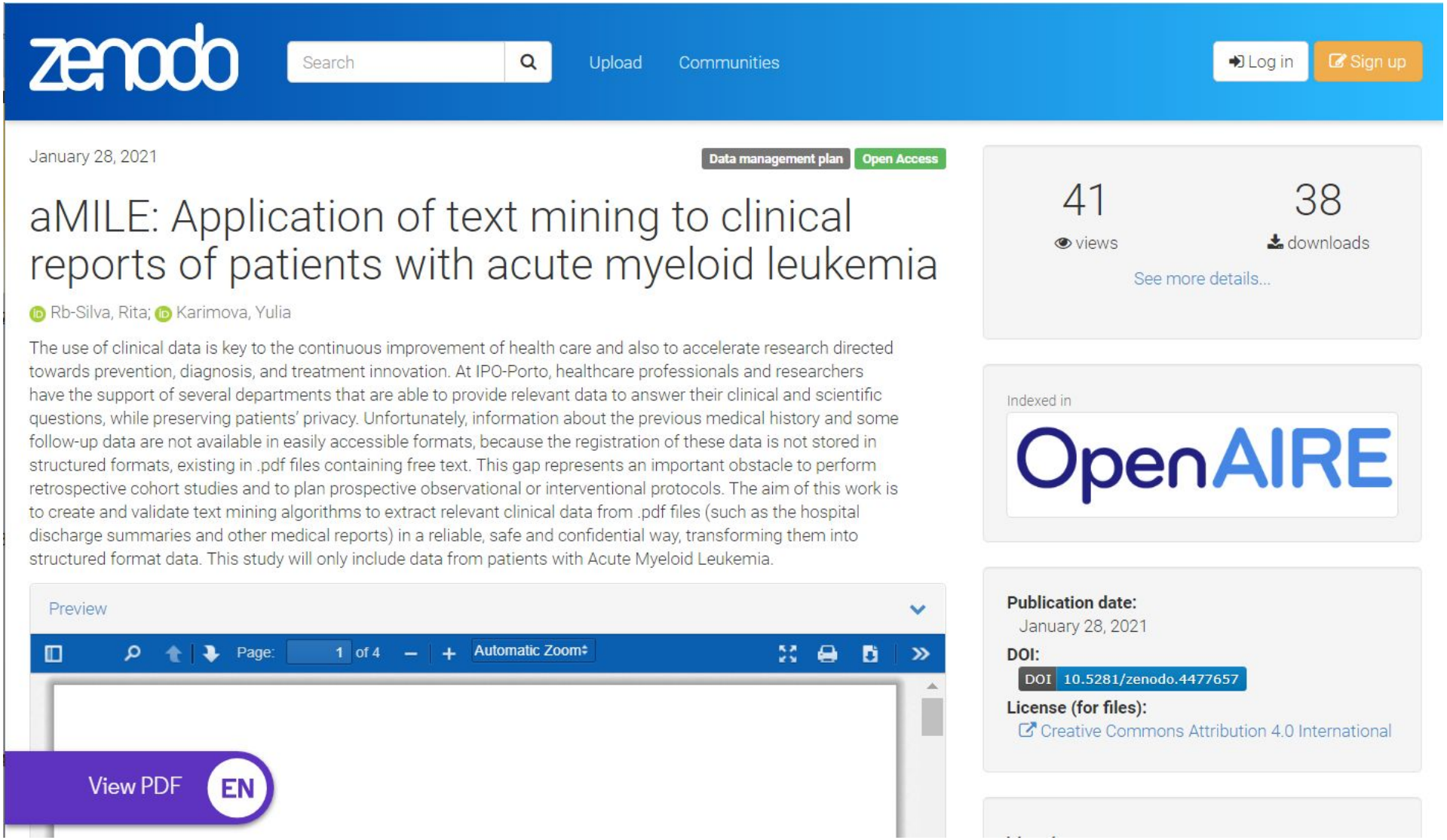
Measures for data privacy and security

- Disk encryption;
- No data transfer to third party services, such as cloud services;
- Authentication;
- Authorization to access;
- Collection of the minimum of personal data;
- De-identification and anonymization of data;
- Data storage security (Back-up);
- End of data use.

Ethical considerations

- Declaration of absence of conflict of interest
- GPDR compliance
- Data Protection Impact Assessment (DPIA)
- Informed consent
- Non-interventional study
- Purpose
- Public interest
- Data management plan
- Measures for data privacy and security

The aMILE data management plan

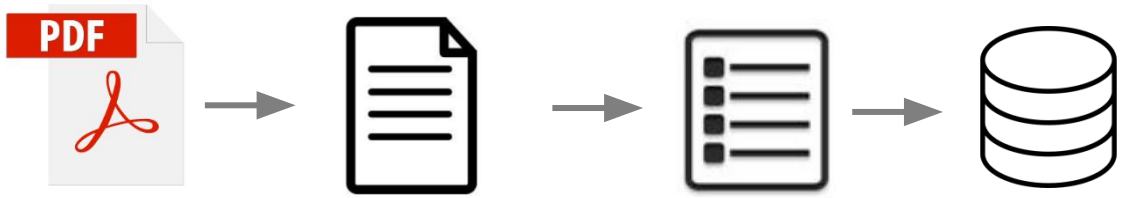


https://zenodo.org/record/4477657#.YH6rg-hKg2w

RDA RESEARCH DATA ALLIANCE

Contributing for a solution

- **aMILE**



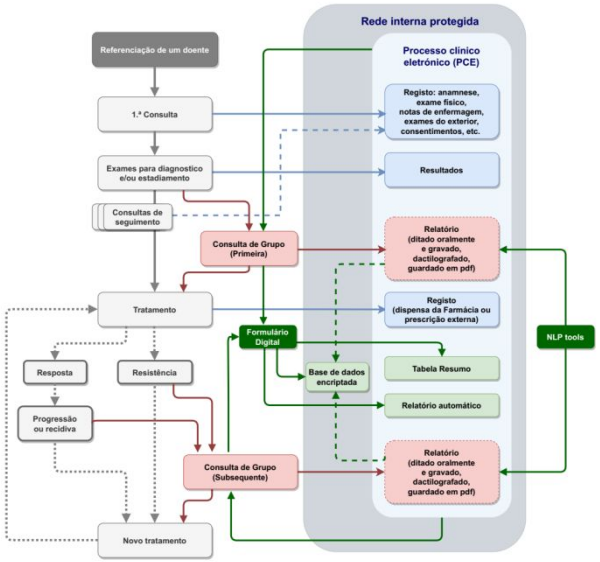
.pdf files
Hospital discharge summaries, Group Consultations and other clinical reports.
(3-10 files)

Raw text data

Primitive data

Validated Database

- **PCExtraHealth**



- Reliance on expert annotations
- Lack of resources

Thank you!

IPOPORTO
INSTITUTO PORTUGUÊS DE ONCOLOGIA DO PORTO FG, EPE

TEHDAS
Towards European Health Data Space

U.PORTO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR
UNIVERSIDADE DO PORTO

rita-rbsilva

r_rbsilva

Rb-Silva, Rita

RDA
RESEARCH DATA ALLIANCE