







MELLODDY

# Federated machine-learning of clinical and omics data

### **Yves Moreau**

University of Leuven – ESAT-STADIUS



### **Federated analytics**

"Bring the computation to the data"



### Federated analytics & the GDPR



### Buyer beware: No free lunch!

Federated analytics facilitates GDPR compliance

- It does not provide an escape from responsibility
- Processing from personal data to aggregate data must be legitimate, which depends on foreseen/foreseeable uses
- Consent to participate in research is independent of GDPR
- Consent to process personal data is not the same as consent to participate in research study
- GDPR mandates data minimization
  - Do not process personal beyond necessary
  - Not only pseudonymization
  - Federated analytics is a key data minimization technique



### Attacks against aggregate data

- Aggregate data or models might sometimes be susceptible to differential privacy and model inversion attacks.
  - From aggregate data and part of a person's data, it might be possible to know whether someone's data was used in the calculation of the aggregate data or model.
  - From a model, it might be possible to recover some information about the underlying personal data.
- Risk analysis must be carried out for each specific aggregation scheme.
  - In case of plausible attack, mitigation strategies might be possible.



### Horizontally vs. vertically partitioned data

Vertical

Horizontal C1 C2 C3



- Horizontal = patient (= item) aggregation
- Vertical = attribute (= variable) aggregation



### Multiple Sclerosis Data Alliance: COVID & MS



## Over 80 countries are participating in this initiative



### A large number of partners were involved



A L L I A N C E acting under the umbrella of the European Charcot Foundation









# Associations of DMT therapies with COVID-19 severity in multiple sclerosis

Data From 28 countries aggregated (including three level data acquisition)

Demographic (age, sex), and clinical (MS phenotype, disability), and DMT covariates were queried, alongside COVID-19 clinical severity outcomes, hospitalization, admission to ICU, requiring artificial ventilation, and death.

Characteristics of outcomes were assessed in patients with suspected/confirmed COVID-19 using multilevel mixed-effects logistic regression, adjusted for age, sex, MS phenotype, and EDSS.

**KU LEU** 

Using the largest cohort of people with MS and COVID-19 available, we demonstrated consistent associations of rituximab with increased risk of hospitalization, ICU admission, and requiring artificial ventilation, and ocrelizumab with hospitalization and ICU admission.

### Federated Learning 4 Everyone





### With Data Custodian User Perspective i **Study Center Repository Center** Model Center **Data Center** Study Catalogue Create a Study: Data Custodians can create Model Center: · Scripts Repository: Data Dictionary: · Data Custodians can download the scripts · In Model Center, data Custodians can use a · Data Custodians can upload and share an a study in a study catalogue according to attached to an analysis. By downloading the trained model; this model is crossempty data frame according to the data they the data they have. referenced to the study catalogue. They can scripts they can go over through script line have, this Data Scientist can use this data o Observe a Study: Data Custodianscan by line to have the sanity and privacy use the trained model to test their own data dictionary to develop a cleaning function as observes different studies created by the in the local. incorporating this trained model checked based on their governance. these well as an analysis plan. this Data Dictionary other stakeholders and, based on the with test data can be developed as a DSS scripts should be cross-referenced to the should be cross-referenced with the study authorization level of the study, can step study center. and shared in the model center. center. Server-Side further into that study. Models Repository: · Raw Data: Analysis Center · This module can be used as a shortcut for o data custodians who prefer to share their o Join an Analysis: Data Custodians can join a the data custodian to access the existing patient-level data can use this module in the federated analysis. model being shared by the other data center to share their data with the stakeholders. these models are crisis central platform, cleaning functions always referenced to the model center. make sure that data is never duplicated and take over the common errors and ensure the quality of the uploaded data.

**Mounting Data** 

· Data Custodians who want to participate in a

their local with the docker container.

federated analysis can benefit from a user-friendly

docker container interface in their local. Using this

module, they can mount the patient-level data in

Local-Side

### Sharing Scripts

 Data custodians have to upload the scripts they have downloaded and want to participate in analysis from the central platform in this part.

### Executing Analysis

 After mounting data with the container and sharing the scripts, all the information data custodians need is the IP and Port address of the data scientist machine to start the federated learning analysis. For broader detail, a logger has been bounded to the container; using this logger data custodian can follow the different sections of the learning.





MSDA infrastructure V:COV-2.1



### Data heterogeneity

- Multicentric analyses will reveal substantial heterogeneity across partners
  - Harder to obtain reproducible results
  - Need to correct for biases
  - Need for research on better machine learning approaches
  - More robust results



### MSBase – Data size imbalance



Proportion of Each Dataset (Sorted)

**122 Clinics** 









Individual Performances in AUC-PR vs Dataset Size





Experiment time with Error Bars

### New collaborative models



## Value proposition

- Hospitals
  - Hub model
    - Protects patient privacy (GDPR)
    - Hospitals maintain control over own data at all times
  - Predictive model
    - Support therapeutic decisions
  - Generate revenue from industry
    - Access to modeling results
    - Patient recruitment
  - Fund IT costs related to disenclaving/curating data

- Public health authorities
  - Use of the platform would be valuable for public health policy
    - Disease burden
    - Drug effectiveness research

**KU LEU** 

- ...
- To obtain governmental cooperation, public health authorities could be given access to the platform

# Value proposition

- Pharma
  - Modeling of patient trajectories
  - Better stratification of patients
  - Improved design of clinical trials
  - Improved patient recruitment
- Pay-for-cure model
  - Incentive for all pharma companies to give the right drug to the right patient
  - Support better therapeutic decisions at hospitals



## Acknowledgments

- KU Leuven ESAT-STADIUS
  - Kevin Bardool
  - Ashkan Pirmani
  - Adam Arany
  - Martijn Oldenhof
  - Edward De Brouwer
  - Amin Davani (alumn)
  - Jaak Simm (alumn)
  - Haleh Chizari (alumn)
  - Nasim Shabani (alumn)
  - Regis Loeb (alumn)

- UHasselt
  - Liesbet Peeters
- > UAntwerp
  - Geert Vandeweyer
  - Nishkala Sattanathan
- Janssen Pharma
  - Hugo Ceulemans
  - Joerg Wegner
  - Bart Vannieuwenhuyse
  - ➢ Tinne Lewi



## MELLODDY





Flanders Al

