MUCOSA-PREDICT Clinical Data Reformatting for Analysis

Sebastian Van Blerk, Benjamin Lelouvier, Florence Servant

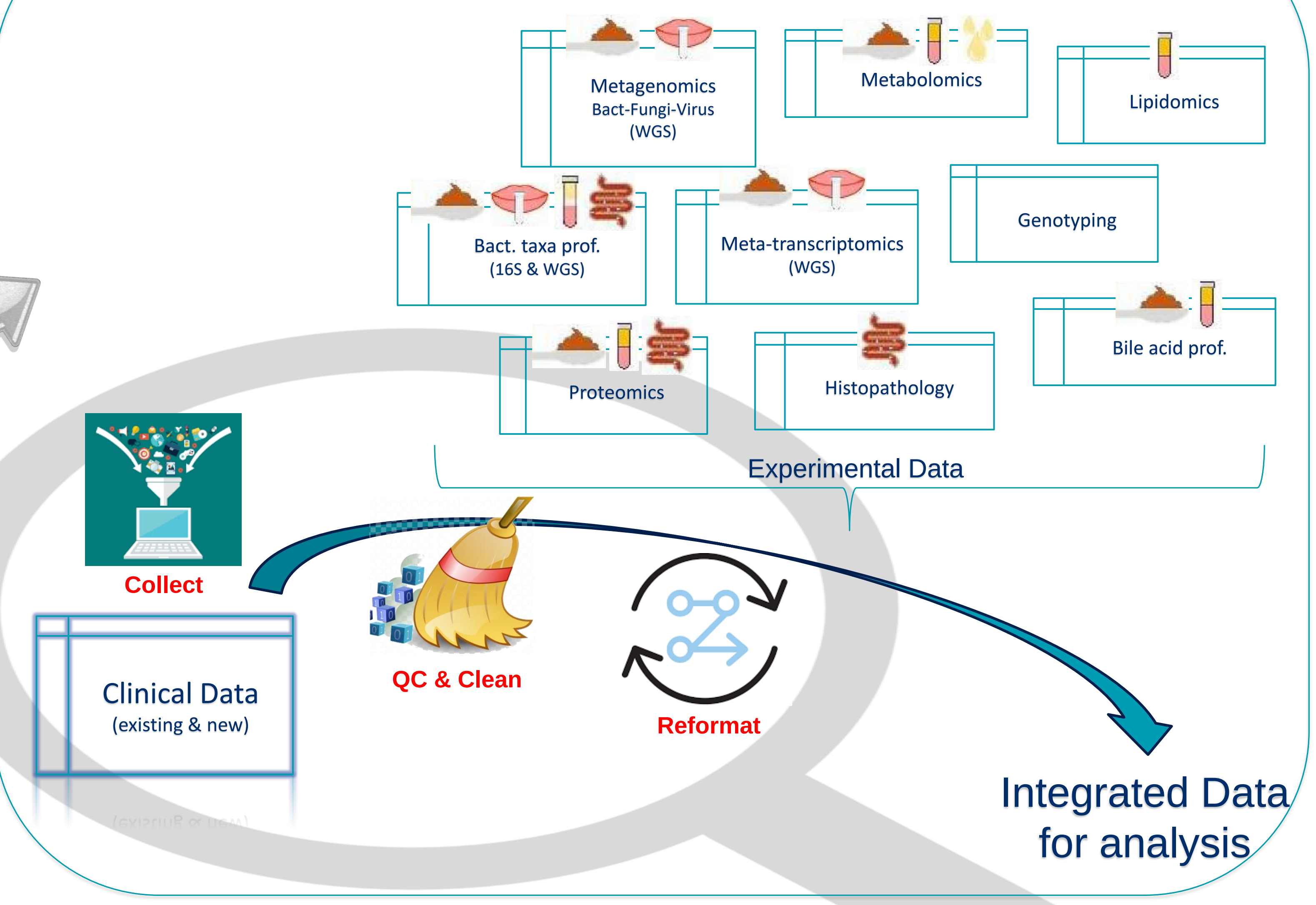
INTRODUCTION

Clinical data management is of major importance for the success of a project such Microb-Predict. Successful clinical data leveraging in R&D includes:

- Data collection using a common format for all centers and controlled vocabularies, addressed by WP1 "Clinical, genetic, expositional and geographic characterization of existing data“,
- Quality control with data sanity checks and data consolidation,
- Data reformatting to enable analysis with experimental data.

The goal of this work was to move the MUCOSA-PREDICT clinical data from a **patient visit oriented format** to a **format suitable for analysis**.

Analysis with Microb-Predict Clinical Data



METHODS AND RESULTS

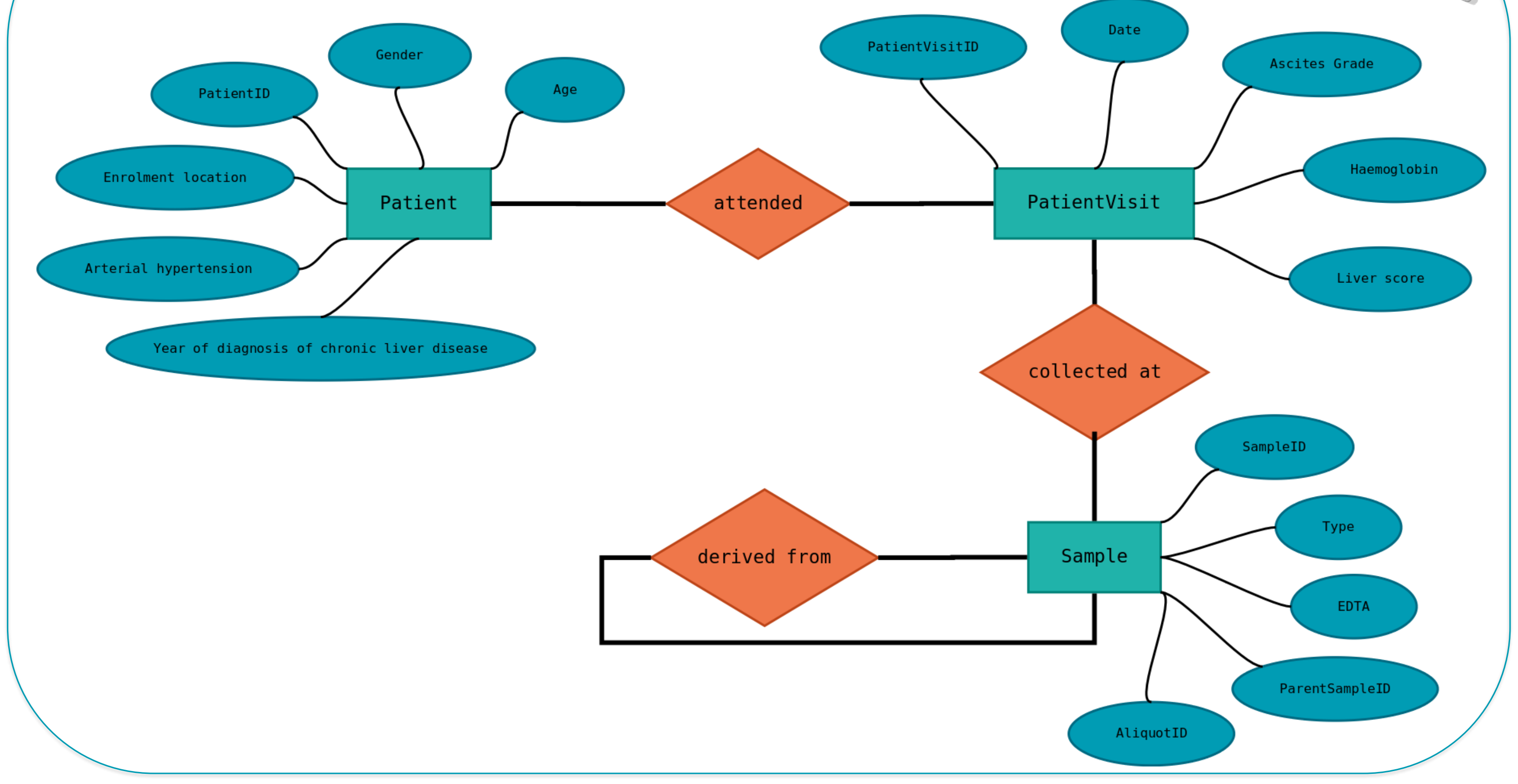
1. Data Modeling

In data management, an Entity Relationship model, also called “ER model” is composed of 3 main components:

- **entities**: « things of interest
- **attributes** for each entity: « features »
- **relationships** between entities

A simple Entity Relationship modeling allowed the identification of 3 main entities (🟩): patient, patient visit and sample.
Each clinical parameter was mapped to 1 entity to become an entity attribute (●).
Relationships between entities (◆) were determined as follow: a patient attends at least 1 patient visit; samples can be collected from the patient (feces, blood, saliva, etc.) at a patient visit; a sample can derive from another sample to produce aliquots.

Entity Relationship data model



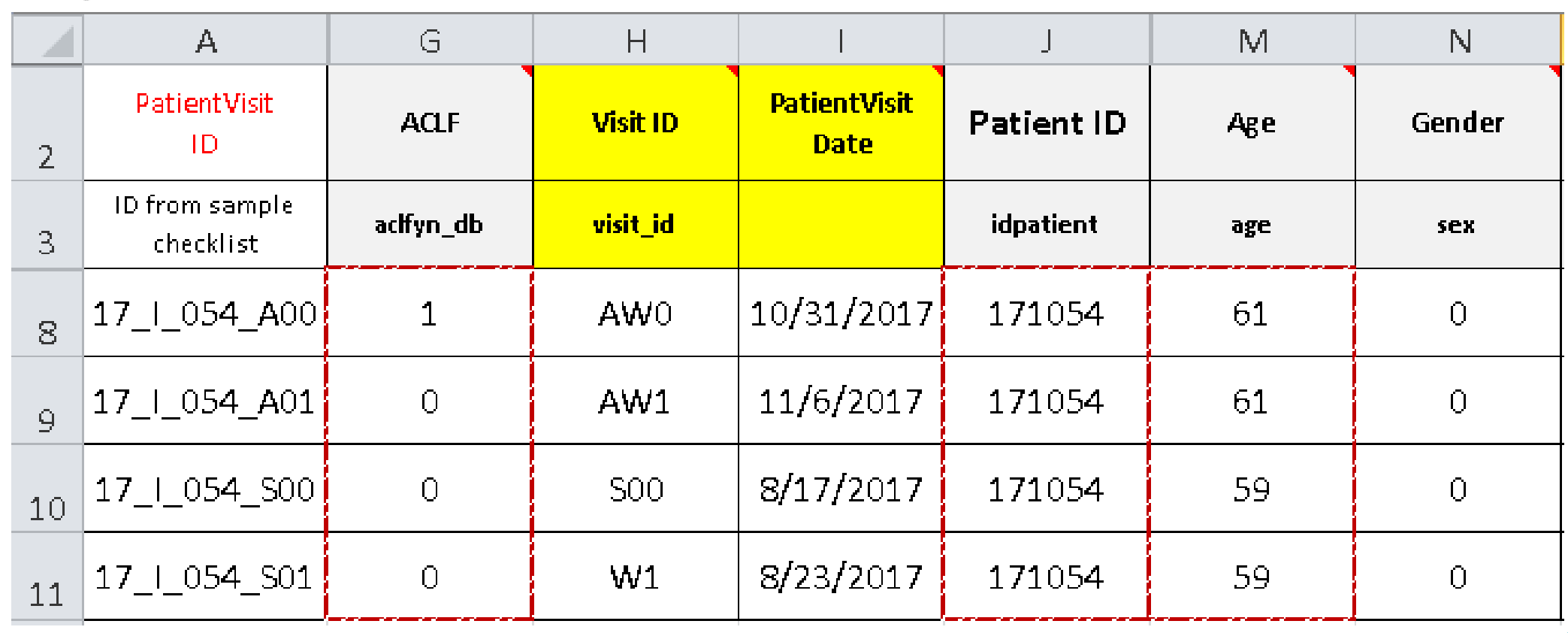
Reformatting process automation

original single
sheet format
python
3 entity
sheet format

2. Data Cleanup

With the ER model, a sample inherits from its patient visit information, which inherits from its patient information. This reduces information redundancy and has helped to implement a program that performs data sanity checks for entity attributes.
Inconsistencies detected this way were then checked and corrected in the original database.

Data cleanup example



	A	G	H	I	J	M	N
2	PatientVisit ID	ACLF	Visit ID	PatientVisit Date	Patient ID	Age	Gender
3	ID from sample checklist	aclfyn_db	visit_id		idpatient	age	sex
8	17_I_054_A00	1	AW0	10/31/2017	171054	61	0
9	17_I_054_A01	0	AW1	11/6/2017	171054	61	0
10	17_I_054_S00	0	S00	8/17/2017	171054	59	0
11	17_I_054_S01	0	W1	8/23/2017	171054	59	0

3. Data Reformatting

The cleaned data were finally reformatted in an Excel format handy for data analysis. The new format has 3 sheets corresponding to the 3 entities. Each column of an entity sheet corresponds to a clinical parameter (i.e. attribute).

The cleanup and reformatting steps were automated using python scripts and will be run on future clinical data versions.

CONCLUSIONS

- This modelling, cleaning and reformatting of the clinical data has allowed to:
  - ✓ Reduce redundancy: final Excel file size is ¼ of original Excel file size (300Kb vs 1.2Mb),
  - ✓ Improve data quality: 9 inconsistencies have been detected, manually checked and corrected in the original Excel file,
  - ✓ Facilitate data integration: linking analysed samples to the sample entity will allow retrieving both patient visit and patient related clinical data.
- **Both formats contain the exact same information and will co-exist for the consortium partners**, the choice of format is based on convenience.
- Additional data sanity checks could be implemented when appropriate to ensure the highest quality data for the Microb-Predict partners.
- There is still some work to link samples to data. Interested partners can contact us to complete this effort.

ACKNOWLEDGEMENTS

We would like to thank Esau Moreno and Jonel Trebicka for their help, support and responsiveness on this work.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 825694. This reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.

Follow MICROB-PREDICT on twitter:
www.twitter.com/MicrobPredict

MICROB-PREDICT online:
www.microb-predict.eu