PRIMAGE

Medical imaging
Artificial intelligence
Childhood cancer research

D3.2 – Implementation of PRIMAGE database and mechanisms for archiving and storing imaging data, linkable to clinical data

Project Full Title: *PRedictive In-silico Multiscale Analytics to support cancer personalized diaGnosis and prognosis, Empowered by imaging biomarkers Business Place*
Project acronym: *PRIMAGE*
Project type: Horizon 2020 | RIA (Topic SC1-DTH-07-2018)
Grant agreement no: 826494

[redacted]

[redacted]

Introduction

The deliverable D3.2 (Implementation of PRIMAGE database and mechanisms for archiving and storing imaging data, linkable to clinical data) is part of the WP3, led by Universita di Pisa (UNIPI). It is mainly related to T3.3 (Data extraction and curation) pertaining to prepare and curate data previous to their upload to the PRIMAGE platform to construct a structured database that will be used in different work packages (WP4, WP5 and WP6). Three major actors are involved in the infrastructure:

- QUIBIM, which is developing the PRIMAGE web platform, a web application from where users have access to all the imaging studies and clinical variables included on the PRIMAGE platform.
- Medexprim which provides a software; MedexprimSuite™ and is developing curation tools to enable an automatic workflow for extraction and curation of clinical and imaging data to the PRIMAGE platform.
- Universitat Politécnica de Valencia (UPV-I3M), who provides the High-Performance Computing (HPC) infrastructure and middleware.

The D3.2 aims to define the design, the infrastructure and the technical implementation of PRIMAGE database as well as establishing the strategy regarding the data extraction and preparation.

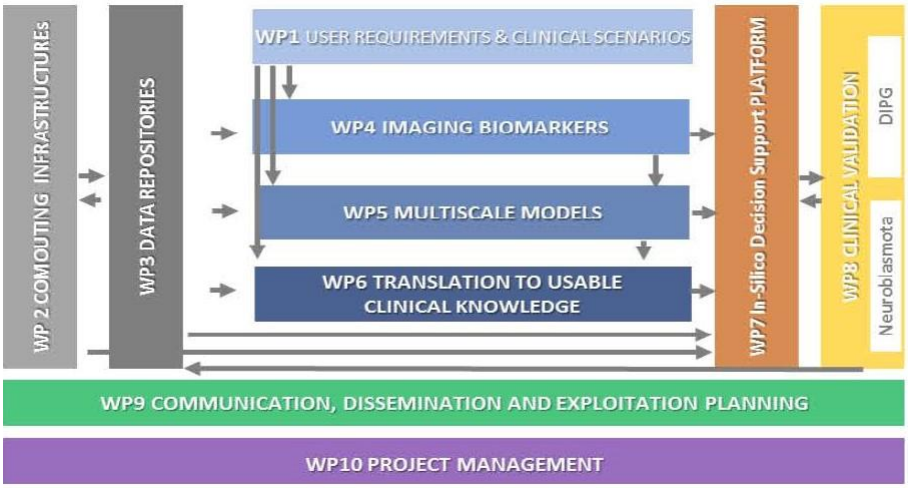

Figure 1: Global workflow of PRIMAGE's work packages.

As one of the main points of the PRIMAGE platform is to provide a persistent data storage service, a PRIMAGE database is developed and implemented. This point is detailed in section "*1. Implementation of PRIMAGE database*" and focuses on the data requirements of this database, the structured models developed and applied as well as the IT infrastructure on which this implementation is based.

This structured PRIMAGE database will store several types of data: medical images and their associated structured clinical information in electronic format through the e-form as defined in deliverable D3.1 (Definition of data requirements, e-forms, control procedures for data quality, and signed authorizations with different data holders to reuse existing datasets). The database implemented on PRIMAGE platform will be fed through different pipelines, sources, batch sizes and data types.

Data may come from two **different sources**:

Past clinical trials / Registries:	Databases of clinical data, where images may either be at the main investigator site or spread across participating hospitals. International Society of Paediatric Oncology European Neuroblastoma Research Network (SIOPEN-R-NET) at St. Anna Kinderkrebsforschung – Children's Cancer Research Institute (CCRI), Gesellschaft für Pädiatrische Onkologie und Hämatologie (GPOH) at Klinikum der Universitaet zu Koeln (UKOELN) and the Society of Paediatric Oncology European registry for Diffuse Intrinsic Pontine Glioma (SIOPE-DIPG) at Hospital Universitario y Politécnico La Fe (HULAFE).
Hospitals:	Patients that have not been included in clinical trials or registries. Routine care data. For example, from HULAFE and UNIPI.

These data may be composed of **different data types**:

Clinical trials data:	Structured database from past clinical trials
Medical images:	CT, PET, MRI, mIBG images obtained from either routine care data or from clinical trials
Unstructured data:	Various data formats, as unstructured text report, genomic information etc.

To assure as much as possible an intuitive and easy-to-use extraction process for the users, **different ways of sending information** to the PRIMAGE platform are provided:

Manual feeding:	Pseudonymize, upload pseudonymized images and fill structured report (e-form) patient by patient on a web interface → PRIMAGE web platform provided by QUIBIM.
Automatic feeding:	Extract from PACS and other sources of information, pseudonymize, upload pseudonymized batch of patients to the platform and automatically fill corresponding structured report (e-form) from unstructured data using curation tools → Solutions provided by Medexprim.

Hence, the strategy developed to perform the task T3.3 (Data extraction and curation) must be compliant with all these aspects. The Figure 2 below is an overview of the process of how the data is included into the PRIMAGE platform. This point is detailed in section *"2. Extraction strategy – Technical workflow".*

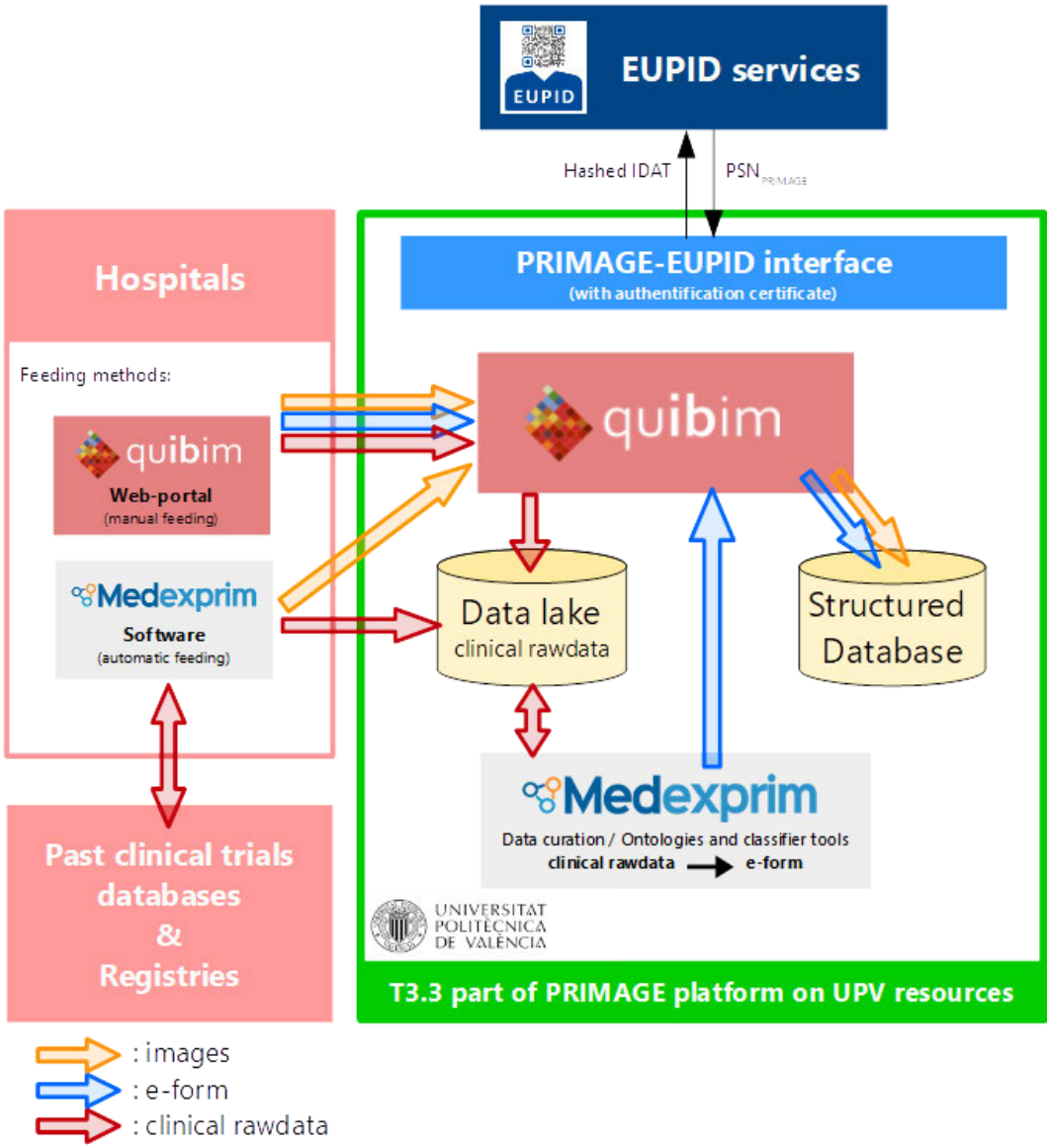

Figure 2: Overview of the global workflow to feed data to the structured database inside PRIMAGE platform. Each feeding approach will be detailed below in the deliverable. Hospitals can either manually feed data to the platform through the PRIMAGE web platform or automatically using Medexprim's software.

Regarding the pseudonymisation process, the European Patient Identity Management (EUPID) service will be used in the PRIMAGE project. EUPID services are introduced and detailed in section *"3.1. Integration evaluation of EUPID Services for pseudonymisation*". This point has been widely discussed between clinical and technical partners during different meetings and it has been decided this pseudonymisation process will be beneficial for the project.

In addition to the data extraction and pseudonymisation, a strategy regarding the curation and extraction of clinical and biological features from unstructured clinical data is needed to be defined. At the current state of the project, access to sample or example of these data has not yet been granted. Hence, this

deliverable focus on the anticipation of potential technological locks and gather different sources of relevant resources and technologies. This point is detailed in section "*4. Curation strategy"*.

Extraction and pseudonymisation strategy for each clinical partner is detailed in section "*5.* Strategy per clinical partner*".*

6. Conclusion

This deliverable aims to define the design, implementation and infrastructure of PRIMAGE database, to present the planned strategies for the extraction, pseudonymisation and curation of imaging and clinical data for the PRIMAGE project.

The implementation of the PRIMAGE database presented is based on data requirements established in D3.1, using the platform architecture defined in D1.2 and is built on the infrastructure described in D2.2 and D2.3.

The extraction strategy presented allows imaging and clinical data from different sources (past clinical trials, registries, routine care data) to be fed separately to the PRIMAGE platform using either a manual or automatic method according to the user's preferences.

The strategy presented in this deliverable is based on EUPID services, which provide patient pseudonymized identifier and ensure the same patient within different hospitals or clinical trials get the same identifier, hence avoiding duplicate studies on the PRIMAGE platform.

This deliverable includes a description of different sources of relevant technologies and resources for text feature extraction and data structuration & curation. We focused on the selection and extraction of relevant clinical and biological endpoints based on ontologies, decision trees, natural language processing methods etc. The curation strategy is focused on identifying the potential technological locks and challenges that will be faced during the development of curation tools, once access to data samples is granted.