



Privacy compliant health data as a service for AI development

Grant Agreement Number: 101095384

D2.5: PHASE IV AI Initial Technology Assessment and List

Deliverable Identifier:	D2.5
Deliverable Version:	v.0.6
Status	Final (F)
Work Package:	WP2 Requirements & Specifications
Task:	Task 2.5 PHASE IV AI - Technologies
Author(s) and Organisation:	Tunc Asuroglu (VTT), Juha Pajula (VTT)
Peer Reviewer(s):	Paula Subías-Beltrán (EUT), Rafael Redondo (EUT), Andrei Kazlouski (UTU), Antti Airola (UTU), Tapio Pahikkala (UTU)
Deliverable Due Date:	2024/05/31
Deliverable Submission Date:	2024/05/30
Dissemination Level:	PU: Public
Funding Authority:	European Commission
Funding Program:	Horizon Europe Health Work Programme 2021 – 2022
Topic:	HORIZON-HLTH-2022-IND-13-02
Rights:	PHASE-IV-AI Consortium

Document Control History

Version	Date	Edited by	Modification reason
v.0.1	2024/04/08	Tunc Asuroglu, VTT	1 st draft
v.0.2	2024/04/24	Juha Pajula, VTT	2 nd draft
v.0.3	2024/05/06	Paula Subías, EUT Rafael Redondo, EUT	Reviewed by EUT
v.0.4	2024/05/13	Andrei Kazlouski, UTU Antti Airola, UTU Tapio Pahikkala, UTU	Reviewed by UTU
v 0.5	2024/05/14	Tunc Asuroglu, VTT Juha Pajula, VTT	Revised according to EUT feedback
v 0.6	2024/05/17	Tunc Asuroglu, VTT Juha Pajula, VTT	Revised according to UTU feedback

Executive Summary

The objective of this deliverable is to outline foreseen technologies from partners across Work Packages 2, 3, 4, and 5 within the project. A structured table has been assembled, encompassing essential details such as purpose, description, expected inputs and outputs, and the likelihood of utilization for each technology. This table's architecture was initially conceived and organized internally at VTT, with subsequent refinement incorporating feedback from partners. Following collaborative discussions and iterative enhancements, the final version of the table was disseminated to all relevant partners via the project drive by INESC TEC (drive.inesctec.pt/). A transparent and inclusive review process was conducted, wherein partners were encouraged to provide input on technologies they are considering for implementation. An input gathering phase was facilitated through clear instructions and column descriptions provided to partners. This approach resulted in the acquisition of 46 inputs, contributing to the development of the initial technology list. 22 out of 46 technologies are likely to be utilized by the partners in the project. This compiled list of technologies functions as a strategic roadmap for the consortium when defining requirements and specifications of the project.

Disclaimer

The content of the publication herein is the sole responsibility of the publishers and it does not necessarily represent the views expressed by the European Commission or its services.

While the information contained in the documents is believed to be accurate, the authors(s) or any other participant in the PHASE-IV-AI consortium make no warranty of any kind with regard to this material including, but not limited to the implied warranties of merchantability and fitness for a particular purpose.

Neither the PHASE-IV-AI Consortium nor any of its members, their officers, employees or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein.

Without derogating from the generality of the foregoing neither the PHASE-IV-AI Consortium nor any of its members, their officers, employees or agents shall be liable for any direct or indirect or consequential loss or damage caused by or arising from any information advice or inaccuracy or omission herein.

Copyright message

©PHASE-IV-AI Consortium, 2023-2026. This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both. Reproduction is authorised provided the source is acknowledged.

Table of Contents

1. Introduction	6
1.1 Purpose of the document.....	6
1.2 Structure of the document.....	6
1.3 List of Acronyms.....	7
2. Initial Technology List	8
2.1 Initial Technology Review.....	8
2.2 Table Column Descriptions	8
2.3 Initial Technology Table.....	9
3. Conclusions	11
4. Annex A: Technology table template.....	12
5. Annex B: Technology table columns and descriptions.....	13
6. Annex C: PHASE IV AI initial technology list.....	14

Table of Figures

Figure 1: Example of inputs in the initial technology table 9

1. Introduction

PHASE IV AI aims to provide and validate a comprehensive set of scientific and technological results and make them accessible as services through the PHASE IV AI Health Data Hub. Beyond the mere aspects of research beyond the State of the Art (SotA) technology, the accessibility and applicability of the technological results are key aspects for competitiveness of the European Health industry, citizens, and health care providers. Health systems benefit from a swift uptake of innovative health technologies and services.

PHASE IV AI utilizes an agile approach within research by implementing results as Data as a Service (DaaS), WP3; Model as a Service (MaaS), WP4; and Health Data Hub, WP5. The DaaS concept provides innovative robust tools for privacy enhancing technologies including de-identification and data synthetization methods which are core technologies of PHASE IV AI. The MaaS concept in WP4 develops Machine Learning (ML) methods for privacy-preserving ML workflows using data residing in hospitals and private healthcare institutions utilizing the data and privacy enhancing technologies developed within the DaaS concept at WP3. The Health Data Hub is foreseen as the entry point for data search, request, and exchange, facilitating scaling up the DaaS and MaaS technologies developed at WP3 and 4 for project use cases managed at WP6.

Identifying potential SotA technologies from partners and literature provides the basis for the whole project and provides a good starting point for the whole consortium collaboration over work packages. To this end, foreseen technologies are gathered within overall processes of WP2, WP3, WP4, and WP5. The solutions developed with these technologies will be tested and validated within WP6 use cases.

1.1 Purpose of the document

The main purpose of this document is to present an initial assessment of potential SotA technologies that are relevant to the project. This assessment is facilitated by the inclusion of promising technologies identified through collaboration with partnering institutions. The goal is to establish a robust foundation for the consortium's collaborative efforts by pre-identifying core technologies to be integrated across the corresponding technical work packages (WP2 to 5) and ensure their alignment with the established WP6 use cases.

To achieve this objective, an initial technology list was compiled in the shape of a table to describe purpose, description, what is expected as input-output, likelihood of usage in the project, and reference of the initially identified technologies. This table will act as a reference document when the technological partners begin their implementations.

1.2 Structure of the document

The deliverable is structured as follows:

Chapter 2 provides information related to the initial technology list. It includes the technology collection phase, the gathered information, i.e., table columns and their descriptions, and an overall analysis. Chapter 3 summarizes key points and how the reported work will fit the project workflow. Chapter 4 provides an empty table template. Finally, chapter 5 includes technology table columns and descriptions as presented in the table template.

1.3 List of Acronyms

List of Acronyms	
SotA	State of the Art
DaaS	Data as a Service
MaaS	Model as a Service
ML	Machine Learning
ETL	Extract, Transform, and Load
GPU	Graphics Processing Unit
SQL	Structured Query Language
OHDSI	Observational Health Data Sciences and Informatics
EHR	Electronic Health Record
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
UC	Use Case

2. Initial Technology List

A table has been compiled summarizing foreseen technologies from WP2, WP3, WP4, and WP5 partners. The table includes details on the purpose, description, expected inputs and outputs, and the likelihood of each technology being used in the project, as well as links to official documentation.

2.1 Initial Technology Review

The initial task of T2.5 was to collect technologies which partners are aiming to use in PHASE IV AI implementations. For that we applied this simple technology review. For the review a data collection table was defined by VTT. Based on the initial discussions at WP2 bi-weekly meetings many partners were still waiting for the definitions of use cases from WP6 before they could have detailed information about used technology. For this reason and as this was the initial review, the collection titles were left on high level as definition of use cases and initial requirements of them were still open at the time of the collection. After initial formulation of the collection table the final titles were agreed in collaboration with WP2 partners with WP2 internal review and discussion at WP2 bi-weekly meetings. In addition, a separate support work sheet was created describing every column in the table.

The final version of the spreadsheet was uploaded to the project drive as a living document. WP2, WP3, WP4, and WP5 partners were informed about this table, and they were requested to provide input on it if they are developing any technology on the project. To achieve as complete input as possible additional dedicated instructions on how to fill the table were sent to the partners. The content was close to the same as what was in the support sheet. Additionally, VTT filled its own initial entries as an example to clarify the type of technology and inputs expected from the partners.

The columns of the collection table are described in detail in Section 2.2. The reasoning to include the specific topics in columns described in Section 2.2 was focusing to generate a wide initial view on technologies used at first phase of the project. This evaluation considered the project's objectives, the potential value of the input for achieving these objectives, and the feasibility of collecting the input at early stages of the project. Project had other similar types of collections ongoing in other WPs for which reason VTT decided to keep this collection as simple as possible and focused on basic input types. The aim is that the list is updated throughout the project and a detailed final technology review will be applied with it at the last phase of the project and will be reported with D2.6.

A one-and-a-half-month duration was given to partners to fill their input for the document at the initial technology review. A total of 46 inputs were gathered from the partners by the given deadline and those are the basis of the technology list described here.

2.2 Table Column Descriptions

Here are the descriptions of each column of the table (first sheet of the excel document):

- **Partner:** Name of the partner(s) using the technology.
- **Task:** Which task is the technology related to? For example, T3.1 or T4.3. Please note that task number also includes WP number.
- **Title:** Name of the tool/technology/solution, what is the title of the package, library, software, solution.
- **Purpose:** Purpose of the technology, e.g., Interacting with users, generating data, predicting disease outcome, or segmentation of images.
- **Language:** Programming language(s) or related code language, e.g., Python, R, or C++.
- **Source:** Public repository (if available). Link to source repository or homepage of the technology.

- **Input:** What the technology uses as input in any of its forms. For example, user interaction with UI, MRI images in nifty/DICOM format, tabular data, or specific data format like pandas/R data frames.
- **Output:** the format/style of output. For example, visualization on UI, single indicator value, table of probabilities, MRI volumetric segmentation, binary mask of detected area, or a specific file format.
- **Likelihood of usage:** This is very important information for T2.5 in order to define the initial approaches. Classification (1-4) of technology how probably it will be used:
 - 1: The best option, the first choice to start with.
 - 2: The second-best option.
 - 3: Possible but only after testing other options first.
 - 4: Plausible, might be used if Use Case (UC) and data fits.
- **Specifications:** Short description on how technology will be used at PHASE IV AI.
- **License:** License (if available) such as MIT, Apache, GNU, or CC BY-NC-ND.
- **Reference:** Homepage link or Publication DOI to the technology documentation (if exists).

2.3 Initial Technology Table

The complete table contains a total of 46 listed technologies collected from the technology partners within PHASE IV AI. The Figure 1 shows some of the provided entries.

Here we provide a summary analysis of the table content.

Partner	Task	Title	Purpose	Language	Source	Input	Output	Likelihood of usage	Specifications	License	Reference
UTU	T6.2 T6.3	Temporai	Survival analysis, treatment effects and time-series prediction	Python	https://github.com/vanderschaarlab/temporai	Tabular data: Static, Time series, Survival	Prediction, causal inference, and time-to-event analysis	3	Temporai will be used for time series and survival data analysis	Apache-2.0 license	https://arxiv.org/abs/2301.12260
EUT	T3.2, T3.3	Pytorch	Training AI models	Python	https://github.com/pytorch/pytorch	3D Volume data (images, masks) and potentially tabular data (including text)	Synthetic volumes	1	Framework to train, validate and deploy deep neural networks to generate synthetic CT volumes and predict lung cancer evolution in medical imaging. Generative DNN: Latent Diffusion Models, also backbone transformers (DiT FIT)	Modified BSD	https://pytorch.org

Figure 1: Example of inputs in the initial technology table.

Technology entries are dispersed across multiple work packages. The intended usage area of these technologies includes WP3, WP4, WP5, and WP6. It is reported that 27 technologies are planned to be used in WP3, 26 in WP5, 18 in WP4, and lastly, 12 in the WP6 workflow. This distribution aligns with the overall project workflow, because WP3, WP4, and WP5 are technical work packages that heavily rely on variety of technologies and tools.

The purposes of the technologies listed in the table, and the areas where these technologies are intended to be used, span a wide range of application areas. These technologies can be categorized into the following categories:

- Trustworthiness and quality assessment.
- ML and predictive models.
- Data privacy and security.

- Data infrastructure and management.
- Data synthesis and de-identification.
- Data standards and data vocabularies.
- Imaging and visualization of data.
- Database and extract-transform-load-tools (ETL).
- Parallel computing and graphics processing unit (GPU) acceleration.
- Secure processing environment.

As for technology implementation languages, Python is the most planned among the tools and technologies intended to be utilized by partners. C/C++ and Java programming languages follow Python. Solidity and Go programming languages are also present in the table. Finally, there are technologies that are implemented or make use of Structured Query Language (SQL), VHDL/Verilog, Bash shell scripting, and C# family languages.

30 out of 46 technologies reside in the GitHub repository, indicating that most of the technologies or tools are open source. 10 out of 46 of these technologies include web pages and documents corresponding to specifications or data model descriptions. Some of the available technologies have not been open-sourced, either because the research is still ongoing or because it is an internal tool.

Well-known specific libraries and frameworks, such as OpenCV or CUDA libraries, are also present in the table. Finally, resources from the Observational Health Data Sciences and Informatics (OHDSI) are mentioned, referring to OHDSI's tools and standards in healthcare data management.

Inputs for initial technologies also differ variously. Most common input can be mentioned as tabular data which includes synthetic data. Some of the tabular data are static, while others are time series or survival data.

Clinical and Health Records Data are also present as input of technologies, which includes electronic health records (EHR) in general. Some technologies are also adapted to use 3-D Imaging data from computed tomography (CT) or magnetic resonance imaging (MRI) scans as input.

The outputs of the initial technologies consist of various types, which can be summarized into the following categories:

- Data analysis results, including uncertainty, novelty measures, and results of the analysis.
- Data transformations, which include de-identified text, image, and tabular data, as well as synthetic EHRs, synthetic tabular data, and transformed data.
- Data evaluation metrics, which include metrics for evaluating the distribution across dimensions, correlations among features, and the performance of the model.
- Imaging data processing, which includes image labeling, transformed files from DICOM to NifTI formats and vice versa, and harmonization of images.

The analysis of the likelihood of usage indicates that most technological partners have a best or second-best option technology in this project. 22 out of 46 technologies are indicated as the best option for a technology partner to use in the project, whereas 19 out of 46 technologies are indicated as the second-best option. 4 technologies are reported as possible use, but only after testing the other options. Finally, 1 technology is reported as plausible and might be used if the use cases and data fit.

In terms of licenses available in the initial technology list, the Apache License and MIT licenses appear multiple times. In addition, there are proprietary/commercially licensed technologies available. Finally, GNU and BSD licensed technologies, which have open-source licenses and commercial restrictions, are also present.

3. Conclusions

The timing of this deliverable is in the early stage of project solution development and various aspects are still under planning at the time of writing. Due to this, the technology list and assessment that are presented in this deliverable provide only an initial view of the technologies and expertise of the technology partners of the project. Thus, it provides foundation for the consortium's collaborative efforts and technological aspects to create new solutions to be validated with the three use cases of the project.

This deliverable describes an initial technology list utilized to collect the technologies with detailed breakdown of each technology, including its purpose, a concise description, the anticipated inputs and outputs, its projected applicability within the project's scope, and a corresponding reference. The collected table will serve as a valuable guide for technological partners as they initiate the implementation phase and when deciding which technologies to use and provide content for the Health Data Hub developed at WP5.

The technologies encompass various domains, such as trustworthiness and quality assessment, machine learning and predictive models, data privacy and security, data infrastructure and management, data synthesis and de-identification, data standards and data vocabularies, visualization of data, database and ETL tools, parallel computing, GPU acceleration, and lastly, secure processing environment. In terms of technology implementation languages, Python is the most prevalent, followed by C/C+ and Java. Most of the initial technologies reside in GitHub, indicating that most of the technologies are open source. This open-source nature of the technologies promotes transparency, collaboration, and continuous improvement, crucial for the project's success.

In conclusion, this technology list serves as a roadmap for the consortium, guiding the integration of promising technologies into the project. It is expected that the information and insights provided in this document will facilitate the successful implementation of the project's objectives and contribute to the achievement of a robust implementation phase.

4. Annex A: Technology table template

This annex is intended to show an empty technology template that was filled with inputs from technology partners.

Partner	Task	Title	Purpose	Language	Source	Input	Output	Likelihood of usage	Specifications	License	Reference

Annex A. Empty technology table template.

5. Annex B: Technology table columns and descriptions

This annex is intended to provide table column names and descriptions of the initial technology list.

Column Name	Description
Partner	Name of the partner(s) using the technology.
Task	Which task is the technology related? For example T3.1, T4.3 etc. Please note that task number includes also WP number.
Title	Name of the tool/technology/solution, what is the title of the package, library, software, solution.
Purpose	Purpose of the technology, where the technology is used: Interacting with users, generating data, predicting disease outcome, segmentation of images etc.
Language	Programming language(s): which code language is implemented or related, python, R, C++ etc.
Source	Public repository (If available). Link to source repository or homepage of the technology.
Input	What the technology uses as input: User interaction with UI, MRI images in nifty/DICOM format, tabular data, specific dataformat like pandas/R data frames etc.
Output	Output: what is the format/style of output: visualisation on UI, single indicator value, table of probabilities, volumetric segmentation of MRI images, binary mask of detected area, specific file format etc.
Likelihood of usage	This is very important detail for T2.5 so that initial approaches can be defined Classification (1-4) of technology how probably it will be used: 1: The best option, the first choice to start with, 2: The second best option, 3: Possible but only after testing other options first, 4: Plausible, might be used if UC and data fits.
Specifications	Description: short description how technology is used at PHASE IV AI.
License	License if available (MIT, Apache, Commercial, None, etc).
Reference	Homepage etc link or Publication DOI to reference describing related to the technology if exists. Publication in preparation phase.

Annex B. Technology table columns and descriptions.

6. Annex C: PHASE IV AI initial technology list

Annex C. PHASE IV AI initial technology list

Partner	Task	Title	Purpose	Language	Source	Input	Output	Likelihood of usage	Specifications	License	Reference
VIT	T4.3	MACAU	Assessing the trustworthiness of predictions	Python	https://github.com/merlinema/macau	Samples	Uncertainty and novelty measures for each sample	1	A wrapped for LightGBM to enable uncertainty and novelty estimation in inference in addition to providing linear trend extrapolation capabilities.	GPL-3.0	https://arxiv.org/pdf/1311.3190v1.pdf
VIT	T4.4	Synthetic Data Quality Metrics	Sample-level metrics for quantitatively assessing the quality of synthetic data	Python	Partially unpublished improved version of published work	Synthetic data	Precision, recall and authenticity values	1	A method and supporting implementation to assess the quality of synthetic data. The quality is evaluated from the following viewpoints: 1) precision (does the synth data come from the same distribution as the real data), 2) recall (does the synth data cover the whole distribution as the real one), and 3) authenticity (is the synthetic data authentic, i.e. not mere copies of the real data).	NA	https://proceedings.mlr.press/v62/alsaa22a/alsaa22a.pdf
VIT	T4.2	Diverse State Index FL	ML model for disease prediction	Python	FL version un-published	Formatted tabular data	Python variable with DSI model tree	2	ML model for DSI system with federated learning based training	Proprietary	Original DSI. https://csl.kuni.ku.jp/portals/ics/portals/1316142/matita.pdf
FIRE	T5.2 T5.3	Hyperledger Besu	Smart Contract Implementation and token mgmt for the Health Data Hub	Solidity	https://github.com/hyperledger/besu	Signed transactions	Executed smart contract	2	Permissioned EVM Blockchain Infrastructure for Smart Contract implementation and token mgmt	Apache License 2.0	
FIRE	T5.2 T5.3	OIDC and OpenID Connect standards	OIDC - Identity layer built on top of the OAuth 2.0 framework, for user authentication	NA	Inhouse implementation			1		NA	
FIRE	T5.2 T5.3	Gala X Federation services	Set up data space infrastructure	Multiple	https://github.com/edpco/efc			2	Tooling for setting up data space infrastructure: notarization service, trust framework (using SSI)	Apache License 2.0	
FIRE	T5.2 T5.3	Fame	Marketplace tooling for data hub to monetize data assets	NA	https://www.fame.foundation/			3	Federated data asset marketplace with embedded finance	NA	
FIRE	T5.2 T5.3	OAuth 2.0 open standard	Implementation for user authentication	GoLang	https://github.com/iplama/oauth2			1		BSD 2 Clause License	
FIRE	T5.2 T5.3	JSON Web Token (JWT) open standard	Communication between system components	JavaScript	https://jwt.io/introduction			1		MIT License	
FIRE	T5.2 T5.3	W3C standards for Decentralized Identifiers (DIDs) and Verifiable Credentials	Identity management implementation	NA	https://www.w3.org/TR/did-core/ https://www.w3.org/TR/vc-data-model/			1		W3C Patent Policy	
FIRE	T5.2 T5.3	ZeroDS	Core infra for Health Data Hub, allowing security and privacy enabling by design	Go	https://github.com/threefoldlabs/zerods			3	Autonomous decentralized operating system, abstracting provider side and consumer side, hereby enabling security and privacy by design. Tools implemented for security by design: encrypted overlay network using wireguard, quantum-secure file storage.	Apache License 2.0	
FIRE	T5.2 T5.3	EDC	Eclipse Data Connector, connector that allows to interconnect data space participants in a trusted way	Java	https://github.com/eclipse-edc/connector			2		Apache License 2.0	
UTU	T4.2 T4.3	Docker	Sending runnable code to the data curator, retrieving the results	Go	https://github.com/docker	Runnable data analytics program	Results of the analysis	2	Docker is a set of platform as a service products that use OS-level virtualization to deliver software in packages called containers	Apache License 2.0	https://www.docker.com/
UTU	T4.2	NVFlare	Training AI models in a distributed way	Python	https://github.com/NVIDIA/NVFlare	Tabular data	Result of the federated algorithm	2	NVIDIA FLARE is a domain-agnostic, open-source, and extensible SDK for Federated Learning	Apache-2.0 license	https://developer.nvidia.com/flare
INPH	T4.2	XOR Platform	Tool for privacy preserving data analytics and machine learning. Will be used to protect data privacy during the processing and sharing of clinical data.	Python 3 (Internal), C++ compute engine and Scala compiler	https://dev.inpher.io/xor/composy/	Tabular data	Encrypted or plaintext result (ex: trained model and/or predictions, statistics, scores, etc.)	1	The XOR Platform will be used to enable data analytics and ML across data coming from various data sources, while protecting the privacy of the plain data	Commercial	https://inpher.io/xor/secret-computing/
INPH	T4.2	THE	Open source library for fully homomorphic encryption. Can be used to protect data privacy during the aggregation and processing of clinical data.	C/C++	https://the-github.io/the/	Tabular data	Encrypted or plaintext result	2	The THE library allows to evaluate an arbitrary boolean circuit composed of binary gates, over encrypted data, without revealing any information on the data.	Apache 2.0	
UTU	T3.2, T3.3	OpenDP Library (SmartNoise SDK)	Generating differentially private (DP) data	Python	https://github.com/OpenDP/smart-noise-sdk	Tabular data	Depends on the task. It could be DP value, DP synthetic tabular data, DP data generator, DP model	2	The OpenDP library will be used to generate private data with DP guarantees.	MIT License	https://opendp.org/
UTU	T3.2	Microsoft Presidio SDK	De-identification	Python	https://github.com/microsoft/presidio	Text, image, tabular data	De-identified text, image, tabular data (depending on the input)	2	The MS Presidio SDK provides tools for data protection and de-identification	MIT License	https://microsoft.github.io/presidio/
UTU	T3.3	Halo (Epicent)	Synthesis of longitudinal data	Python	https://github.com/ethadonov/HALO	Longitudinal electronic health records (EHR) data	synthetic EHRs, Data generator	2	HALO library will be used to generate synthetic EHR data	Unknown	https://www.nature.com/articles/41467-023-41093-0
UTU	T3.2, T3.3	Synthity	Generating and evaluating synthetic tabular data, including Bayesian and GAN based Differentially private models	Python	https://github.com/vandercchaarlab/synthity	Tabular data: Static, Time series, Survival	synthetic tabular data and generator including no_DP and DP models. Evaluation metrics.	2	The Synthity library will be used to generate synthetic data with or without DP guarantees. Will be used for utility and privacy evaluation	Apache-2.0 license	https://arxiv.org/abs/2301.07573
UTU	T6.2, T6.3	Temporal	Survival analysis, treatment effects and time-series prediction	Python	https://github.com/vandercchaarlab/temporal	Tabular data: Static, Time series, Survival	Prediction, causal inference, and time-to-event analysis	3	Temporal will be used for time series and survival data analysis	Apache-2.0 license	https://arxiv.org/abs/2301.12260
EUT	T3.2, T3.3	Pytorch	Training AI models	Python	https://github.com/pytorch/pytorch	3D Volume data (images, masks) and potentially tabular data (including text)	Synthetic volumes	1	Framework to train, validate and deploy deep neural networks to generate synthetic CT volumes and predict lung cancer evolution in medical imaging. Generative DNN: Latent Diffusion Models, also backbone transformers (DLIT)	Modified BSD	https://pytorch.org
EUT	T3.2, T3.3	Diffusers Library	Training Generative Models	Python	https://github.com/huggingface/diffusers	3D Volume data (images, masks) and potentially tabular data (including text)	Synthetic volumes	2	Framework to train, validate and deploy diffusion models to generate synthetic CT volumes and predict lung cancer evolution in medical imaging. Generative DNN: Latent Diffusion Models, also backbone transformers (DLIT)	Apache-2.0 license	Apache-2.0 license
VARIA	T6.2, T6.3, T6.4	KVM & GEMU	Secure Processing Environment virtualization	C	https://linux-kvm.org/page/Code			1	Virtualization solution for Linux used to host secure processing environments	GNU GPL or LGPL	https://linux-kvm.org
VARIA	T6.2, T6.3, T6.4	Guacamole	Secure Processing Environment remote access	C, Java, JavaScript	https://github.com/guacamole/guacamole-server			1	Clientless remote desktop gateway used to access secure processing environments	Apache 2.0	https://guacamole.apache.org
SU	T4.2	OpenHE	Open source library for homomorphic encryption.	C/C++ , Python 3.x, HDLC, Verilog/System C, Bash Shell	https://github.com/openheor/openhe-development	Binary data	Binary data	1	Acceleration of homomorphic encryption primitives	BSD 2-Clause	https://www.openhe.org/
SU	T4.2	Microsoft SEAL	Open source library for homomorphic encryption.	C/C++ , Python 3.x, HDLC, Verilog/System C, Bash Shell	https://github.com/microsoft/SEAL	Binary data	Binary data	1	Acceleration of homomorphic encryption primitives	MIT license	https://github.com/microsoft/SEAL
SU	T4.2	NVIDIA CUDA	Parallel computing program and API that allows for GPU acceleration	C/C++ , Python 3.x, HDLC, Verilog/System C, Bash Shell	https://developer.nvidia.com/cuda-zone	Binary data	Binary data	1	Acceleration of homomorphic encryption primitives	GNU GPL	https://developer.nvidia.com/cuda-zone
LKS	T3.5 T4.4 T5.4	LeanScale DB	Distributed Relational Database	C, Java	Not Available	SQL statements submitted using well known standards (i.e. JDBC/ODBC) or popular frameworks (i.e. Apache Spark)	relational data	4	As a database technology that provides real time analytics	under LKS proprietary license	https://www.leanscale.com
KUL	T3.4	Synthetic Data Quality Metrics	The framework established to assess fidelity, utility, privacy and ensure the quality of synthetic medical data.	Python	https://github.com/Vicomtech/SDGS-evaluation-metrics	Tabular real data and synthetic data	Metrics for evaluating the distribution across dimensions, correlations among columns, and the performance of the model	1	Standardized Metrics and Methods for Synthetic Tabular Data Evaluation	MIT license	https://www.himem-connect.de/products/qsuaml/abstract/10.1055/a-0442-176024/
KUL	T3.4	Synthetic Data Quality Metrics	The framework aims to create synthetic datasets using GANs and evaluates various GAN models by assigning different weights to suit specific objectives.	Python	https://github.com/yulinba/synthetic-dataset-benchmarking	Tabular real data	GAN based synthetic data and multifaceted assessment - utility, privacy	2	Benchmark Generative Adversarial Networks(GAN) based approaches for generating synthetic electronic health record (EHR) data	MIT License	https://arxiv.org/abs/2208.01230
KUL	T3.4	Synthetic Data Quality Metrics	From the latent representation, calculate the support coverage between real data and synthetic data	Python	https://github.com/ahmedmohamada/evaluating-generative-models	Tabular real data and synthetic data	a-Precision, β -Recall, Authenticity test	2	Introduce Latent EHR - OneClass Embedding approach, then measuring the adapted metrics of Precision, Recall, and Authenticity to evaluate its performance. The code is in progress.	MIT license	https://arxiv.org/abs/2102.08921
KUL	T6.2, T6.3	Survival analysis for time series	Survival analysis, treatment effects and time-series prediction	Python	https://github.com/vandercchaarlab/temporal	Tabular data: Static, Time series, Survival	Prediction, causal inference, and time-to-event analysis	3	Temporal will be used for time series and survival data analysis	Apache-2.0 license	https://arxiv.org/abs/2301.12260
KUL	T6.2, T6.3, T3.2, T3.3	Synthetic Data Vault	Generation of synthetic data	Python	https://github.com/rds-dev/SDV	Tabular real data	Synthetic Data	2	State-of-the-art methods for generation of synthetic data	MIT license	https://repositorio.icec.org/document/1794926
AIN	T3.1, T4.1, T5.1	OMOP CDM	Common Data Model (CDM) is an open community data standard, designed to standardize the structure and content of observational data and to enable efficient analyses that can produce reliable evidence. A central component of the OMOP CDM is the OHDSI standardized vocabularies.	CDM v5.4	https://ohdsi.github.io/CommonDataModel/	Clinical Data	Structured & Standardized to OMOP CDM Clinical Data	1	OMOP CDM will be used as a data structure for clinical data across the project	NA	https://ohdsi.github.io/CommonDataModel/
AIN	T3.1, T4.1, T5.1	OHDSI Vocabularies	OHDSI Standardized Vocabularies are used for: Mapping data to OMOP CDM. Querying of the transformed data, Interpreting the meanings of the data	R/SQL	https://github.com/OHDSI/Vocabulary-v5.0			1	OHDSI Standardized Vocabularies will be used in order to map all the clinical data to standardized codes, enabling reproducible and accurate research	Free SW	https://github.com/OHDSI/Vocabulary-v5.0/wiki
AIN	T3.1, T4.1, T5.1	PostgreSQL	Open source object-relational database system	SQL	https://www.postgresql.org/about/	Clinical Data	Transformed Data	2	SQL will be used for ETL processes	MIT License	https://www.postgresql.org/about/
AIN	T3.1, T4.1, T5.1	ATHENA	Search and load standardized vocabularies	NA	https://athena.ohdsi.org/	Search clinical terms	CDM code	1	Athena will be used to search clinical terms and map the to OMOP CDM codes	NA	https://github.com/OHDSI/Athena
AIN	T3.1, T4.1, T5.1	WHITERABBIT	Scans data and creates a report containing all the information necessary to begin designing the ETL	GUI	https://ohdsi.github.io/WhiteRabbit/WhiteRabbit.html	Clinical Data	A report containing all the information necessary to begin designing the ETL	1	ETL preparation	NA	https://ohdsi.github.io/TheBookOfOhdsi/ExtractTransformLoad.html#white-rabbit
AIN	T3.1, T4.1, T5.1	RABBIT IN A HAT	Perform a scan of the source data, providing detailed information on the tables, fields, and values that appear in a field	GUI	https://ohdsi.github.io/WhiteRabbit/WhiteRabbit.html#hat	White Rabbit scan report	Rabbit In a Hat uses White Rabbit report and through a graphical user interface to allow a user to connect source data to tables and columns within the CDM	1	ETL design	NA	https://ohdsi.github.io/TheBookOfOhdsi/ExtractTransformLoad.html#rabbit-in-a-hat
AIN	T3.1, T4.1, T5.1	USAGI	Is a tool to aid the manual process of creating a code mapping. It can make suggested mappings based on textual similarity of code descriptions.	GUI	https://ohdsi.github.io/usagi/	Source code	Vocabulary concept mapping	1	Map non-standard codes to standard. Map codes that needs translation	NA	https://ohdsi.github.io/TheBookOfOhdsi/ExtractTransformLoad.html#usagi
AIN	T3.1, T4.1, T5.1	MIRCOGL	View 2D slices and renderings of your brain imaging data, allows you to draw regions of interest which can aid lesion mapping	Python	https://www.mirc.org/projects/mircng/	Neuroimages	Neuroimage labeling, transformed files from DICOM to NIfTI and vice versa	2	Easily adapt and harmonize the provided labels, Transform DICOM to NIfTI, Change of Coordinate system	BSD 3 Clause License	
AIN	T3.1, T4.1, T5.1	Opencv	Computer vision and machine learning software library designed for real-time image processing and analysis	Python	https://docs.opencv.org/4.x/dh/dhgg_tutorial.html	CT / MRI	Harmonized Images	1	Medical Imaging Harmonization including: Resolution, orientation, space, file format	Apache 2.0	https://github.com/opencv/opencv-python
AIN	T3.1, T4.1, T5.1	Nilearn	Analyses of brain volumes	Python	https://nilearn.github.io/stable/index.html	Neuroimages	Neuroimage labeling, transformed files from DICOM to NIfTI and vice versa	2	Easily adapt and harmonize the provided labels, Transform DICOM to NIfTI, Change of Coordinate system	BSD 3 Clause License	https://nilearn.github.io/stable/introduction.html
ENG	T3.3, T3.4	ALIDA	Data Science and Machine Learning Platform for Big Data Analytics.	Python/Java	https://github.com/ai-lab/ai-lab-internal-temporary-data-services	Data/Models	Data/Models	1	Ingesting data, training models.	NA	https://home.aidalab.io/
ENG	T3.3	MONAI	Project MONAI is an initiative started initially by NVIDIA and King's College London to establish an inclusive community of AI researchers to develop and exchange best practices for AI in healthcare imaging across academia and enterprise researchers.	Python/C#	https://github.com/Project-MONAI	Data	Trained Models	2	The library main focus is generating medical data. Its usage will be evaluated according to T3.3 requirements.	Apache 2.0	https://monai.io/