DataTools4Heart

DataTools4Heart is funded by the European Union's Horizon Europe Framework Under Grant Agreement No. 101057849

What is DT4H?

DataTools4Heart (DT4H) is the first platform to combine real-world cardiology data from millions of people in multiple European countries.

Until now, cardiology data that could advance research and healthcare has largely remained unused in hospitals across Europe. This is due to data privacy requirements and variations in data formats and languages. To tackle these challenges, DT4H will extract, combine, harmonise, and reuse these data in a federated manner, i.e. without the information being shared with anyone or transferred out of the hospital.

Clinicians, researchers, and data scientists will have access to the data to answer pressing research questions in cardiovascular disease, ultimately improving diagnosis and treatment.

Collecting and combining data

DT4H will develop several tools to extract data from cardiology units across European regions and enable it to be combined for analysis. This tool will be created and validated in eight European hospitals. Interoperability will be tested in three artificial intelligence (AI) modelling scenarios.

The DT4H data pipeline will harmonise data from each hospital by converting it to a common data format (the common data model, Figure 1). Key features will be extracted to create datasets that can be used by other hospitals for AI modelling.

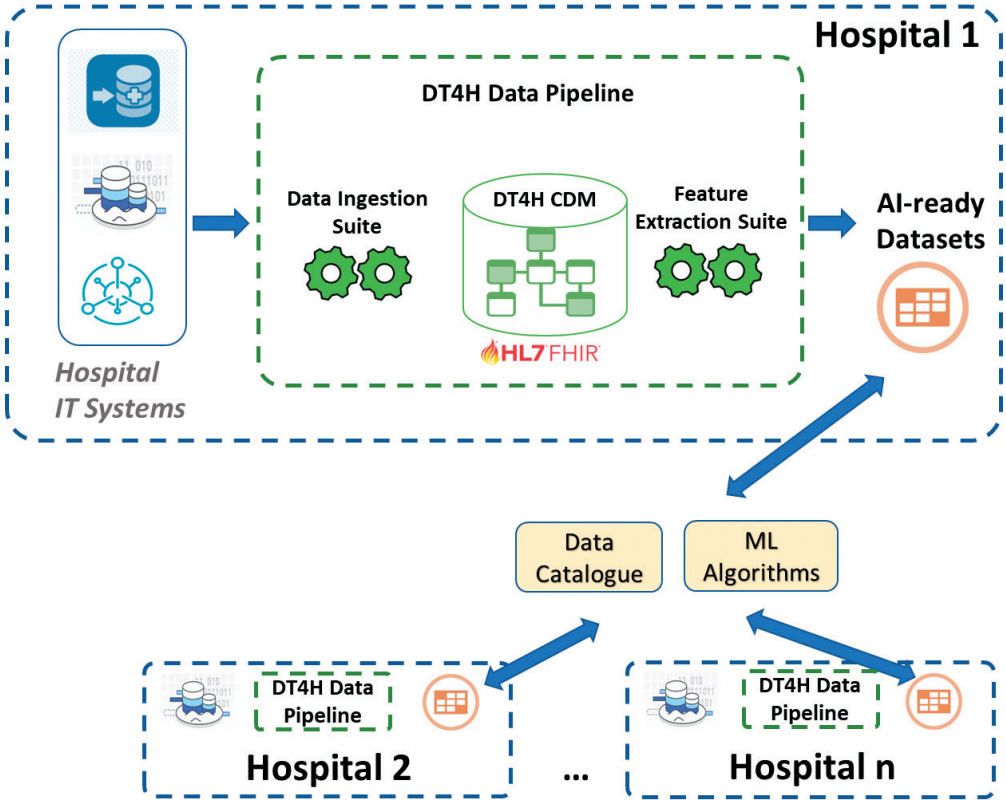

Figure 1: DT4H data pipeline

AI: artificial intelligence; CDM: common data model; FHIR: fast healthcare interoperability resources; HL7: Health Level Seven International; ML: machine learning; n: hospital numbers 3, 4, 5, etc.

To preserve patient privacy, the DT4H data pipeline will run inside each hospital's IT ecosystem and actual patient data will never leave the hospital. The pipeline will include three software modules:

- **Data ingestion suite**: imports heterogeneous patient data from multiple hospitals into an overarching data model.
- **Common data model**: stores the data securely in a standardised format on an open source data repository called onFHIR.io.
- **Feature extraction suite**: enables scientists to access the data they need to answer research questions by entering specific criteria.

Converting the data to common clinical codes

DT4H will enable scientists to combine clinical information currently held in different languages in countries throughout Europe. Natural language processing (NLP) will be used, a type of AI that allows computers to understand and process human languages.

The first step will be to identify the clinical information of interest – called semantic annotation – from electronic health records in English, Spanish, Italian, Romanian, Czech, Swedish, and Dutch (Figure 2).

The second step will be to categorise the type of information (e.g. disorder, procedure, or anatomical structure) and assign it a common code – called entity normalisation.

Each time new clinical text is used (e.g. type 2 diabetes mellitus), it will be compiled into libraries of terms (called corpora) in seven languages.

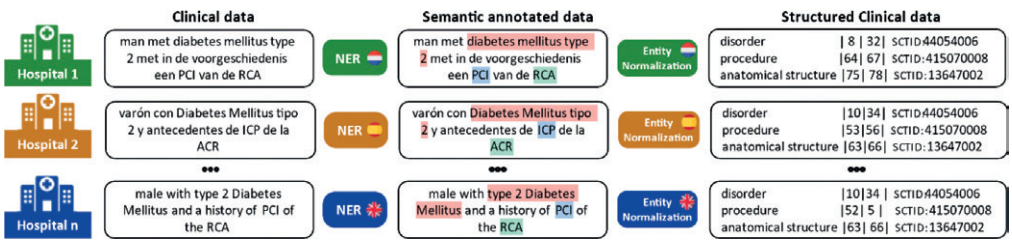

Figure 2: Using natural language processing to convert data to common clinical codes

n: hospital numbers 3, 4, 5, etc.; NER: named entity recognition; PCI: percutaneous coronary intervention; RCA: right coronary artery; SCTID: systematised nomenclature of medicine clinical terms (SNOMED CT) identifier.

Training the model to recognise clinical information

In the initial stages, clinicians will train the NLP model – a process called clinician-in-the-loop – so that it can eventually recognise clinical information and convert it into a common code without human assistance.

This process will occur in three phases (Figure 3), in which the NLP model performs automatic semantic annotation, the clinician manually corrects that annotation, and the model learns from the corrections.

The quality of the NLP model will largely depend on the amount of clinical information available for training. It is expected that the model will learn to interpret clinical information in Spanish or English more quickly than the other languages because a higher volume of data will be available.

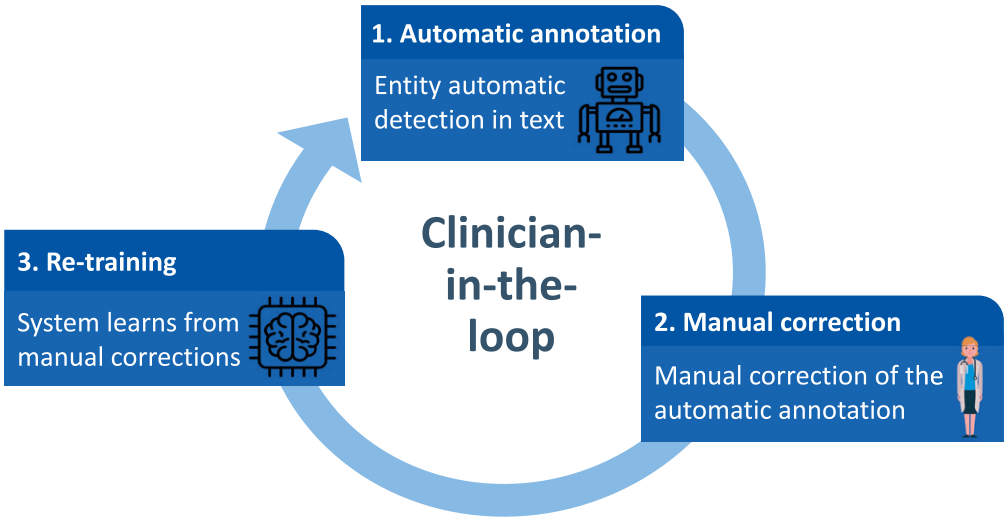

Figure 3: Clinicians will optimise performance of the NLP model in a three-step process

Preserving patient privacy

DT4H aims to securely combine clinical care data by employing innovative techniques such as federated learning (Figure 4) and data synthesis. Instead of centralising data collection from hospitals in one place to train the AI model for addressing specific research questions, the model itself will be sent to each hospital for onsite training using local data. Additionally, synthetic data will be generated to mimic the target population while ensuring that no individual information is retained. The synthetic data generation will be rigorously assessed over the course of the project to ensure its quality. The resultant models will then be aggregated at a central location. The updated model will be returned to each site to continue training and the process will be repeated until the model is fully trained. An open-source privacy-conscious synthetic dataset, CardioSynth, will be an important legacy of DT4H.

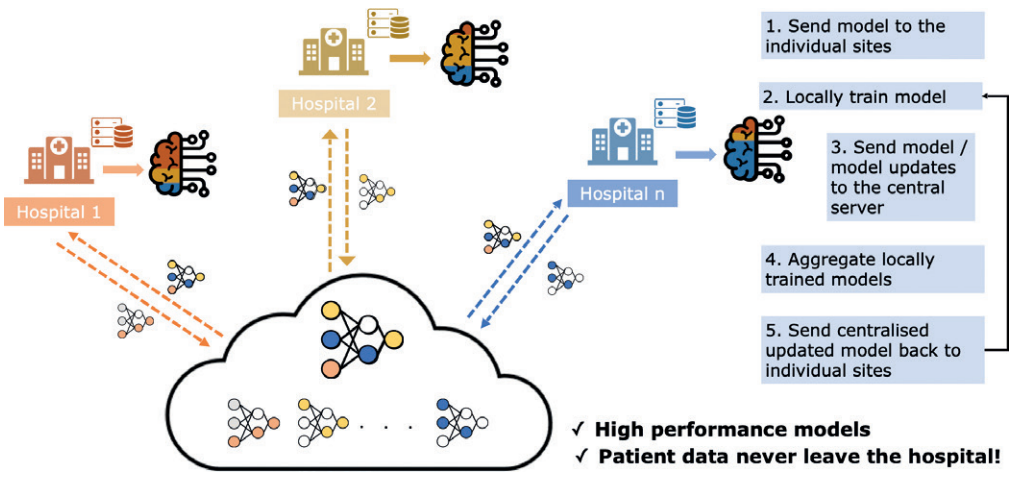

Figure 4: Training a model using federated learning

n: hospital numbers 3, 4, 5, etc.

Virtual assistants will help users navigate the platform

DT4H will develop an integrated platform incorporating all elements of the system, from collecting data, converting it to common clinical codes, and combining it anonymously to preserve patient privacy (Figure 5).

Users will access the system via the “DT4H entry point”. They will be able to input information into the data catalogue and extract information using search criteria such as “type 2 diabetes”.

AI-powered virtual assistants operating in seven languages will help researchers and clinicians to navigate the platform and access the information they need from the large-scale cardiology datasets. Ease of use will be examined in pilot studies conducted in eight European hospitals using real-world data.

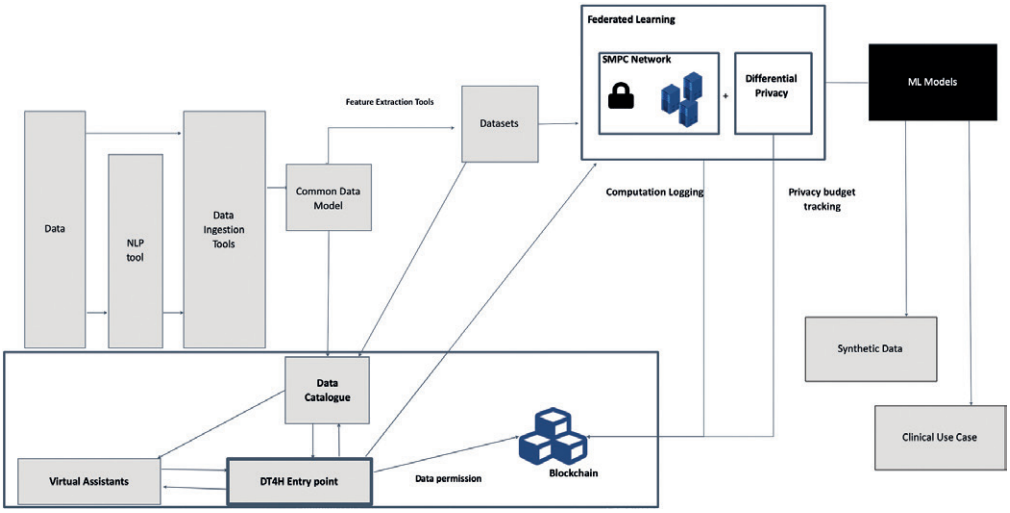

Figure 5: DT4H integrated platform

ML: machine learning; NLP: natural language processing; SMPC: secure multi-party computation.

info@datatools4heart.eu