HEALTHYCLOUD
Health Research & Innovation Cloud

# D7.2 Functional requirement analysis report of Atrial Fibrillation Use Case Version 2.0

## Document Information

| | |
|---|---|
| Contract Number | 965345 |
| Project Website | http://www.healthycloud.eu/ |
| Contractual Deadline | M15, May 2022 |
| Dissemination Level | PU |
| Nature | R |
| Author(s) | Esmeralda Ruiz Pujadas (UB)<br>Karim Lekadir (UB) |
| Contributor(s) | Juan González-García (IACS) |
| Reviewer(s) | Helena Lodenius (CSC)<br>Michaela Th. Mayrhofer (BBMRI-ERIC) |
| Keywords | atrial fibrillation phenotypes, personalized treatment, hierarchical unsupervised machine learning and hierarchical models |

***Notice:*** *The HealthyCloud project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement №965345*

# Change Log

| Version | Author | Date | Description of Change |
|---|---|---|---|
| V0.1 | Juan González-García | 08/03/2022 | Initial Draft |
| V0.2 | Esmeralda Ruiz Pujadas | 11/04/2022 | 2nd revision round. Open to contributors |
| V0.3 | Esmeralda Ruiz Pujadas | 16/05/2022 | Updated with review requests |
| V0.4 | Esmeralda Ruiz Pujadas | 23/05/2022 | Updated with review requests |
| V1.0 | Juan González-García | 27/05/2022 | Updated styles, to be delivered |
| V2.0 | Esmeralda Ruiz-Pujadas | 20/01/2023 | Updates after rejection |
|  |  |  | (Final Change Log entries reserved for releases to the EC) |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Table of contents

# Executive Summary

Atrial fibrillation diagnosis, affects more than 14 million over-65s for which the European Society of Cardiology has risen need for urgent action[1]. As computational power increases, machine learning techniques are becoming more and more present and are integrated into many fields, especially in the healthcare sector. Machine learning has been used previously for both risk prediction and identification of the phenotypes of atrial fibrillation (AF). However, the resulting machine learning models were not externally validated or showed moderate predictive ability and high risk of bias in an external validation. Undoubtedly, there is room for improvement for the AF. We propose an unsupervised machine learning technique applied to an integrative model including clinical data, imaging data, electrocardiogram (ECG) signals and genetic variants. We will detect phenotypes within the population of AF using an initial cohort in a first stage, and then, extending the study to other cohorts to generalize the model in a federated learning scheme. In the context of the HealthyCloud, we will show the whole process from the data discoverability to the technical part and the development of the machine learning models. This will have a high impact in the other WPs, particularly in WP2 and WP5, considering a real case going through all the steps of implementation. With the inclusion of the functional requirements of the AF use case mapped as analysis requirements, WP5 can have a better understanding in order to perform a broad analysis of existing and planned computational solutions, in terms of both infrastructures for research and advanced data analysis. The legal barriers, we had to face to implement a federated learning, will be a potential source of information for WP2 that will incorporate by design the ethical and legal considerations. The discoverability of data is also a key point for other WPs which focus on how data is structured, organised, and accessed either individually (WP3), or through data hubs (WP4) and/or potentially discovered through the FAIR health data portal (WP6).

[1] https://www.escardio.org/The-ESC/Press-Office/Press-releases/Atrial-fibrillation-set-to-affect-more-than-14-million-over-65s-in-the-EU-by-2060

# 1. Background

## 1.1. Use case description

Atrial Fibrillation (AF) is the most frequently encountered cardiac arrhythmia in clinically practice[2]. It manifests itself as an irregular and often rapid heart rate that might cause the increase of risk of strokes, heart failure and even death. Atrial fibrillation diagnosis, affects more than 14 million over-65s for which the European Society of Cardiology has risen need for urgent action.

The main issue in AF is that some of the patients are asymptomatic and an early detection cannot be diagnosed. Patients who develop AF have a higher risk of thromboembolic events, in particular stroke, because of the pooling of blood in the left atrium and embolization to the brain. The risk of stroke is increased fivefold in individuals with AF[3].

For patients with undiagnosed AF, ischaemic stroke may be the first clinical manifestation of the condition. Only 10% of people who had an ischaemic stroke have been first diagnosed with AF. If it was possible to detect asymptomatic AF patients in an earlier stage, it would be possible to prevent strokes by offering anticoagulation treatments[3].

Approximately one-sixth of all strokes are attributed to AF. Patients with thromboembolic stroke from AF have a higher mortality and morbidity than patients with other stroke types. Moreover, the more AF progresses, the more the stroke risk increases. The presence of AF is also associated with an approximately twofold higher risk of future acute myocardial infarction. It is estimated that, of those with persistent AF, one-third will not have symptoms and therefore a first presentation of persistent AF might be a stroke. Patients with persistent AF are the ones who would benefit the most from anticoagulation therapy for stroke prevention[2].

Some studies have shown that vascular risk factors (VRF) such as age, hypertension, obesity and other cardiovascular diseases (CVDs) predispose to AF. However, those models were not externally validated or showed moderate predictive ability and high risk of bias in an external validation[4].

In our knowledge, an integrative model considering different modalities for incident AF has not been yet explored. We propose an unsupervised technique applied to an

[2] Welton, Nicky J. et al. "Screening strategies for atrial fibrillation: a systematic review and cost-effectiveness analysis." Health technology assessment vol.21 (2017): 29.

[3] Wolf, Philip A et al. "Atrial fibrillation as an independent risk factor for stroke: the Framingham Study." Stroke vol.22 (1991): 8.

[4] Nadarajah, R et al. "Prediction of incident atrial fibrillation in community-based electronic health records: a systematic review with meta-analysis." Heart (2021).

integrative model including clinical data, imaging data, electrocardiogram (ECG) signals and genetic variants. We will detect subgroups within the population of AF of the UK Biobank[5] (UKB) cohort in a first stage, and then, extended to other cohorts to generalize the model in a federated learning scheme.

## 1.2. Use case opportunities

This work will allow us a more early and precise AF diagnosis as well as a better personalised treatment for each patient. Moreover, it will help to have a better understanding of the complex cardiac structure and remodelling taking advantages of combining features of different modalities in an integrative hierarchical model. A similar study has not been analysed in literature, yet.

Apart from the clinical part, we will identify all the issues that we need to deal with from the initial stage of the study until we reach the technical part, starting from the discovery of potential data and fulfilment of the requirements for the ethical and legal regulations, up to the harmonization of the data from the different cohorts and creation of the models. This will have a high impact in the other WPs, particularly in WP2 and WP5, by considering a real case going through all the steps of implementation.

## 1.3. Use case challenges

The main challenge was to obtain sufficient incident AF events to build the models. Generally, research institutions start recording healthy participants. Some of these participants develop cardiovascular diseases and are tracked in a continuous and longitudinal follow-up. However, we found the number of incident AF cases very limited for each research centre, making it difficult to find powerful datasets to build the proposed models. In the case of the UKB, the main cohort that we used for our study, there are over half a million individuals recruited between 2006 and 2010. The incident CVDs are tracked using Hospital Episode Statistics (HES) and death registers to provide continuous tracking of the participants. In spite of the big number of participants included in the UKB, the number of patients who develop incident AF decreases up to 193 cases. The same occurred with the other research centres that we collaborated with. Hence, a multi-setting study approach will be considered. With the inclusion of different cohorts, problems of heterogeneities in the data and imbalance of the incident events distributions may be faced.

The use of the data from different cohorts will result in a robust model externally validated decreasing the high risk of bias in an external validation. We will directly

[5] Petersen, S. E et al. “The impact of cardiovascular risk factors on cardiac structure and function: Insights from the UK Biobank imaging enhancement study.” PLOS vol 12 (2017):10.

deal with the most important issues in literature and the main reason that previous machine learning models have not been introduced as a clinical tool for prediction of incident AF.

# 2. Data requirements

Initially, we identified the databases that contain the necessary information given the nature of the AF use case to conduct the proposed analysis. The data required includes health clinical data, biomarkers, genetic variants, imaging and ECGs. We identified possible cohorts in literature or via some connections of the group. From those, we selected the ones that had this data available and were willing to participate in this study.

## 2.1. Existing data

The patient registries used to support open challenges in conferences organized by PhysioNet or MICCAI were firstly identified. However, only the ECGs were available and there was no possibility to obtain the clinical data from those subjects except from the UKB repositories. Then, the public patient registries were discarded for this project. To identify potential databases, we requested a minimum data requirement to the cohorts in order to participate in our study:

- Health and clinical data (AF related outcomes such as medication, interventions e.g. ablation, treatment response, hospitalisation, doctor visits, mortality, etc.);
- Imaging to: first, quantify cardiac structure, function and viability, thus to assess aetiology and associated cardiac comorbidities, e.g. heart failure; second, brain imaging to avoid side-effects and first choice therapy, as AF is a major risk factor for transient ischemic attack /stroke (due to emboli) and anticoagulation treatments may lead to a higher chance of brain haemorrhage.
- ECG to assess heart rhythm and electrical activity across individuals.
- Genetic variants (Single-nucleotide polymorphism) to estimate associated risks (including genome-wide association studies such as AFGEN but also CARDIOGRAMplusC4D and HERMES);
- Biomarkers (High-sensitive troponin, N-terminal B-type natriuretic peptide, and C-reactive protein) and lab results (e.g. blood pressure, glucose/insulin levels, etc.);
- Medical history, lifestyle information and family history (history of stroke);
- Data from research (rich phenotypes and omics data), clinical registries (longitudinal follow-up) and digital technologies (app-based follow-up of AF patients).

Only the cohorts fulfilling the minimum requirement are considered. We reached an agreement with the UKB to start our experiments. A second agreement was reached with the University of Greifswald where the contract is in process to be signed by the University of Barcelona. Finally, in a more initial stage, we found two more databases available from the University Medical Center Hamburg and McGill University Health Center, which are currently collecting all the data requested. Those databases are described in Table 1. In Annex I, the variables extracted are described. At this stage, only the UKB variables are available as we are waiting for the data extraction from the other cohorts.

| Participant No | Participant organisation name | Short name | Country |
|---|---|---|---|
| 1 | UK Biobank Team | UKB | United Kingdom |
| 2 | University of Greifswald | SHIP | Germany |
| 3 | University Medical Center Hamburg | HCHS | Germany |
| 4 | McGill University Health Centre | MUHC | Canada |

Table 1: The databases available for the atrial fibrillation use case.

### 2.2. Desired data

At this point, we have only available the UKB data. In the case of SHIP, we are still waiting for the signature from our home institution, University of Barcelona, before we can send the contract back to the research institution and move forward with the data request. In the case of the HCHS and the MUHC, the process just started. They are still extracting the data. We expect to have the extraction of the variables finished by the end of May, 2022. Unfortunately, HCHS does not fulfil all the requirements as the genetic variants are not available. The reason for this is that the procedure of data recording is in an initial stage and genetic variables are not usually collected. However, we will continue with the request procedure. In the future, we will discuss if it is possible to obtain the missing variables. If it is not the case, we will decide if the HCHS cohort or those variables will be included in the analysis.

### 2.3. Data access challenges

The procedure to request data is relatively slow both for the discoverability of the data and for the procedure to obtain the data. For discoverability, you need to identify, through Google Scholar or from collaborators' contacts, the research institutions who have data fulfilling the requirements, as there are not available any patient registries in Europe containing all the required information, except the UKB. Once the cohorts are identified, an initial meeting must be held to explain the research project proposal. If they are satisfied with the proposal, a formalization by writing the project proposal will be the next step. In a centralized system, you must include the following specifications:

- Background, objectives and methodology of the project.
- The required variables you want them to extract: ECGs, biomarkers, CMR images and so on.
- Sample size
- The server specifications and what levels of security are available to store the data, clarifying what methods for anonymization are going to be used and confirming that no datasets are going to be shared to third parties.
- Dissemination plan indicating the number of papers and conferences you expect to publish using the requested database.

Once, the project proposal is accepted, the research institution must extract all the variables available of the ones you requested in the proposal, and it can take up to four months to be collected. Mapping variables is a manual process and it usually takes several months. The research institution is usually in contact with the accountable researcher in case that some variables are not found and a solution is discussed between both institutions. Before the delivery of the datasets, a contract must be signed by both sides. Digital signatures are not usually allowed for security reasons and the contract cannot be sent via email if there is not a justification for that. The whole procedure can take up to one year.

In a federated scheme, that it is what we want finally to achieve, this process becomes even more complex and slower than a centralized system. The project proposal must also include the goal of the project, the clinical problem definition, sample size, hardware specifications and some explanation of the technology. The most significant difference of the project proposal compared with the one of a centralized system is that you need to specify the hardware requirements on their side as the training is happening locally in each research centre. They also need to be supplied with the necessary tools to run machine learning models as the data should not be exposed to anyone outside the research centre.

# 3. Analysis requirements

First, the feature extraction must be performed. Some of the variables can be directly used such as the diabetes or hypertension status, but some others must be extracted. This is the case for the CMR radiomics[6] which aim to extract a large number of quantitative features from medical images using data characterization algorithms. Radiomics features will be extracted from the CMR images and the corresponding contours using the open-source python-based PyRadiomics library (version 2.2.0) in end-diastole and end-systole. The features encode two phases: end-diastolic and end-systolic information of left ventricle, right ventricle and myocardium.

To compute the Radiomic features, we extract the relevant information present in the image by using three classes of features (Figure 1):

- First-Order Features: are histogram-based features related to the distribution of the grey level values in the tissue, without focusing on their spatial relationships.
- Shape Features: describe geometrical properties of the organ, such as volume, diameter, minor/major axis and sphericity.
- Texture Features are derived from images which encode the global texture information considering their spatial relationships.

For each chamber,16 shape, 19 first-order, and 73 texture features will be estimated. To reduce the number of features, an initial correlation analysis will be performed reducing the features that are highly correlated and keeping only one. The resulting radiomic features will be combined with the variables considered in our study such as medical history, ECG signals, biomarkers and genetic variants. The extraction of ECG features or the use of the whole signal will also be analysed.

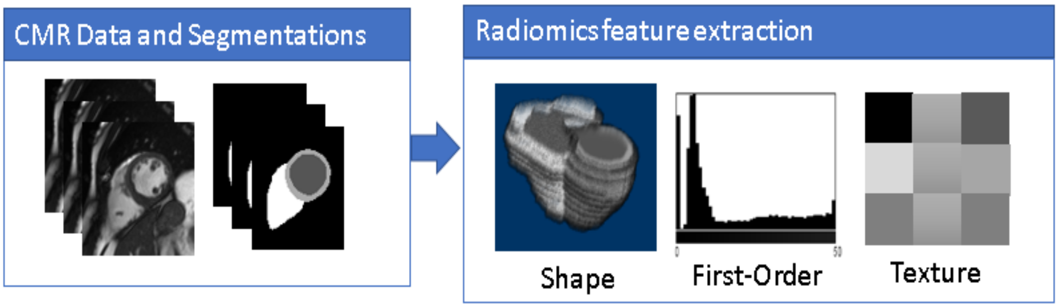

Figure 2: The CMR Radiomics extraction based on shape, first-order and texture features.

[6] Raisi-Estabragh, Z.et al. “Cardiac magnetic resonance radiomics: basic principles and clinical perspectives.” European heart journal vol 21 (2020):4.

## 3.1. Types of analysis envisaged

We aimed to identify different groups of patients with AF who shared common clinical phenotypes to evaluate the association between identified clusters. The hierarchical clustering technique will be considered due to its numerous properties. The main advantages of the clustering analysis algorithm are that the number of clusters are computed automatically and the resulting dendrogram gives a visualization of the hierarchical relationship between the identified groups. The clustering procedure starts by treating each observation as a separate cluster. Then, it iteratively identifies the two clusters that are closest together and merge the two most similar clusters. The process is repeated until all the clusters are merged together as shown in Figure 2.

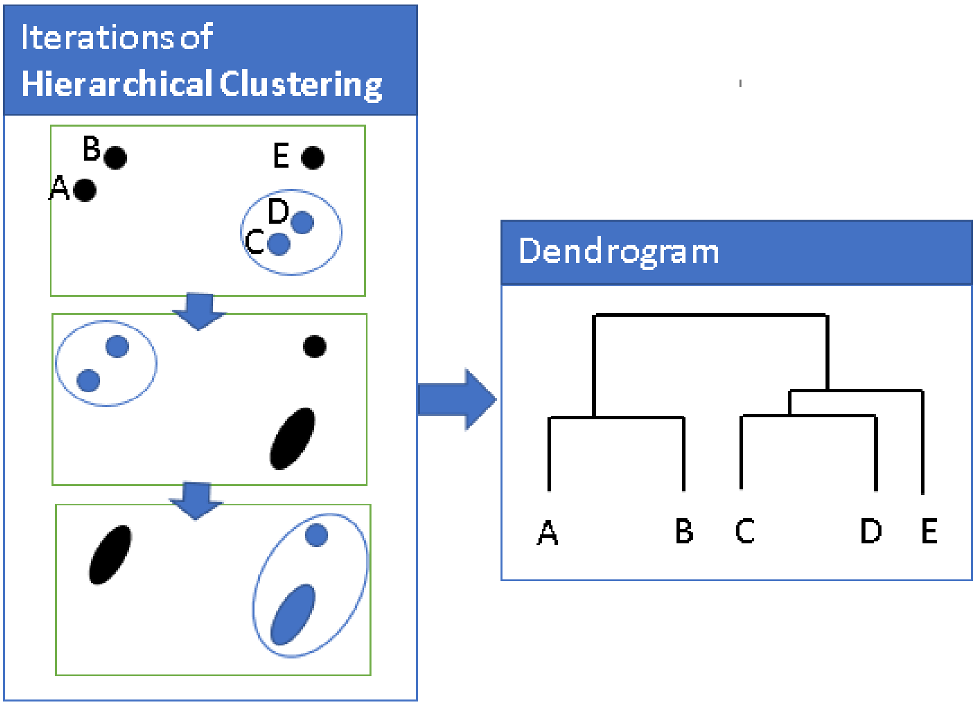

Figure 2: The process of the Hierarchical Clustering algorithm detecting the two clusters that are closest together and merging the two most similar clusters. The circles show the two selected clusters. The method results in a dendrogram showing the relationship among clusters.

We will also consider the distributed K-means and its variants. K-means algorithm is the most well-known unsupervised method. It partitions the data points into k clusters by minimizing the distances between each object and the centroid of the cluster. The original method is based on three simple steps (Figure 3):

- Step1: Initialization: The number of clusters K is defined. There are automatic procedures to find the optimum K.
- Step2: Assignment of data points to centroids: each data point is assigned to the closest centroid.

- Step3: Updating centroids: each centroid of each cluster is updated by averaging the data points of each group.

The second and third step is repeated until the centroids are not changed or are under certain threshold.

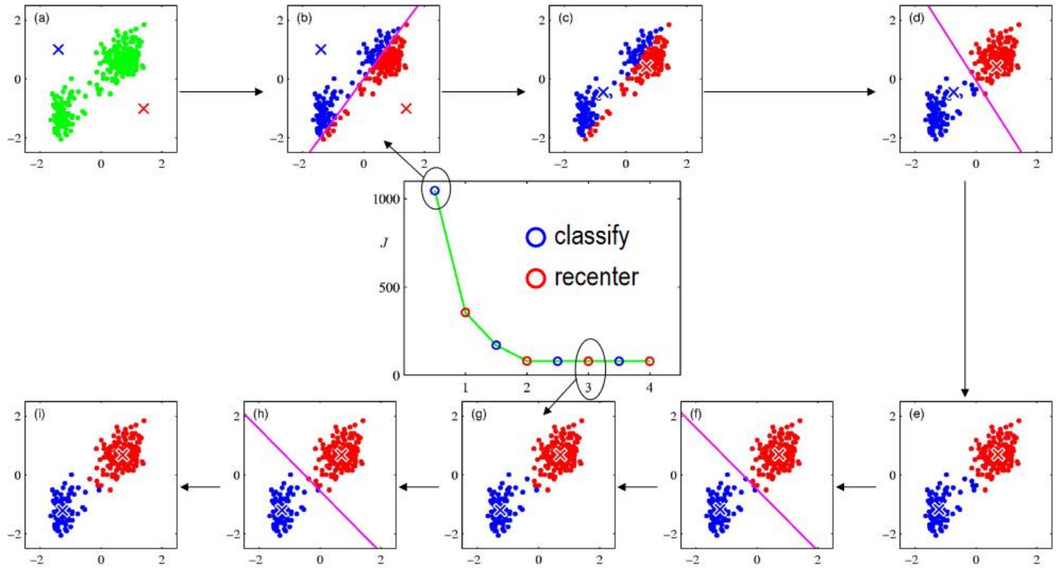

Figure 3: The process of the K-means algorithm: The K clusters are initialized and the data points are assigned to the closest cluster and the centroids are updated by averaging the data points in each group until the centroids do not change or are under a certain threshold *(Laura Igual, CVC-UAB, 2010).*

K -means has already been extended to the federated framework. One of the studies suggested to compute a weighted mean of local cluster centers in order to update global cluster centers[7], where the weights were given by the number of local data points assigned to the clusters. Additionally, some extensions have been proposed using fuzzy assignments of weights instead of hard assignments of number of data points.

Continuous variables will be reported as mean and standard deviations and categorical variables as percentages. For comparison among clusters, Chi-square test or Fisher’s exact test will be used for categorical variables, and Kruskal-Wallis test for continuous features. P-values of less than 0.05 will considered statistically significant.

Proportional hazards regression models will be computed to examine the differences in hazard ratios of the variables between the identified groups. While mortality is usually the primary event of interest for survival analysis, this type of analysis can also be used to assess treatment failure including outcomes such as

[7] Stallmann M. et al., “On a Framework for Federated Cluster Analysis” Appl. Sci. 12(20) (2022).

hospitalization or ablation. We will also examine interactions between variables for the binary outcomes related to the AF case (e.g., ablation, treatment response and hospitalisation) using classical machine learning classifiers.

## 3.2. Analysis development challenges

The main limitations come from moving from a centralized scheme to a federated learning framework[8]. Fortunately, the legal issues of the patient data in a federated learning is not something that we must deal with, as the patient data is not exposed outside the research centre and the legal regulations are performed within each research institution to collect the data. However as described in Section 2.3, we still need to write a project proposal for their own ethical committee. This process must be performed for each cohort you want to include in your study, so this makes the process relatively slow. Each institution usually has their own template. You can use the same information for all the cohorts but a different document must be filled in for each participant.

From a technical perspective, the process itself is also very challenging. We first initialize a global model on a central server that will be initially pre-trained with the UKB cohort. Then, the pre-trained model will be distributed across the research institutions. The initialization of each model must be the same in all the research centres in order to aggregate the information of each model. Each model will be trained in each client in the research centre. Each client computes the model performance on each cluster model and gets assigned to the cluster with the most fitting mode. The scheme of the federated learning for the AF use case is shown in Figure 4.

[8] Linardos, A. et al. “Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease.” Scientific Reports vol 12 (2022):1.

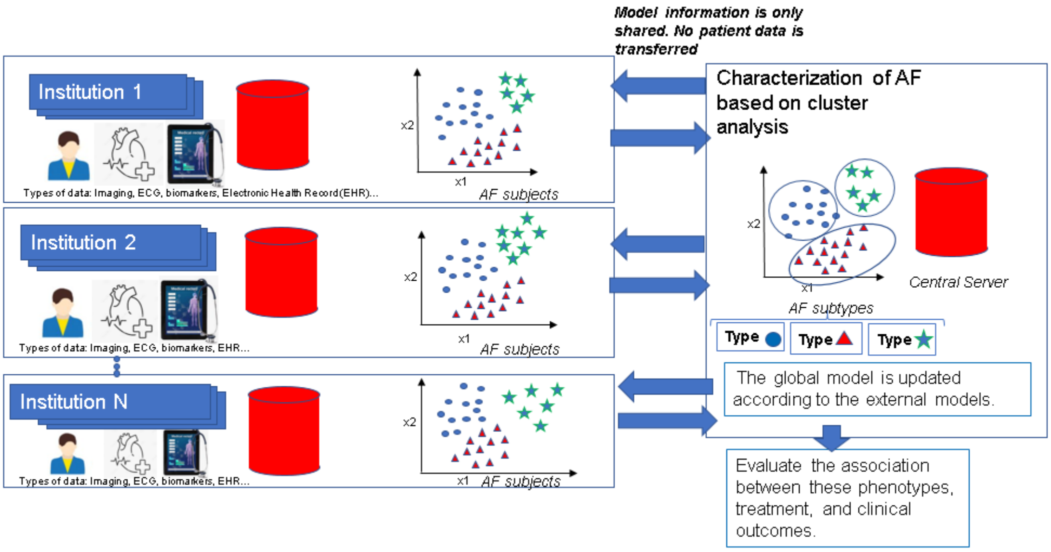
Model information is only shared. No patient data is transferred
Institution 1
Types of data: Imaging, ECG, biomarkers, Electronic Health Record(EHR)...
x2
x1
AF subjects
Institution 2
Types of data: Imaging, ECG, biomarkers, EHR...
x2
x1
AF subjects
Institution N
Types of data: Imaging, ECG, biomarkers, EHR...
x2
x1
AF subjects
Characterization of AF based on cluster analysis
x2
x1
AF subtypes
Central Server
Type ●
Type ▲
Type ★
The global model is updated according to the external models.
Evaluate the association between these phenotypes, treatment, and clinical outcomes.

Figure 4: The federated learning process which shows how a global model pre-trained in the central server is shared across the research institutions and the resulting information is sent back again to the central server whose global model is updated and sent again until certain convergence is reached.  No patient data is transferred or exposed in this scheme.

In order to implement this framework, some issues must be considered:

- The communication overhead between the clients and the centre server. In addition to that, there is certain risk of loss of transmission packets, limited network bandwidth or security/privacy breach. Then, some encryption and compression techniques must be adopted to reduce the model size and secure the privacy.  The encryption methods make the federated learning framework even more secure; although as discussed, no patient data is shown outside the research centre, not even to the programmer.

- Moreover, the models are trained in each research centre so a heterogeneous aggregation can produce a degradation of the performance. Different computational power and variety of data such as a different acquisition protocols, the variety of scanners, label imbalance and size might cause a drastically variation, making the trained client models hard to aggregate.

There are many tools to address the technical part of the federated framework for encryption and security that can be implemented. However, for the heterogeneity aggregation of the models, the solution remains open. We will discuss some measurements to alleviate those limitations in terms of the data pre-processing.

The intensity of the MRI is highly dependent on manufacturer, sequence, and acquisition parameters. Those variances may cause a weak reproducibility in multi-center studies. Standardizing the acquisition protocols is not a feasible solution, as some variables such as the pixel size must be individually adjusted to guarantee image quality. However, even if the acquisition protocol could be fixed, some studies have shown that even in this ideal case, some radiomic features would still be non- reproducible[9]. Then, a clear need of harmonization is required. In literature, many works have been developed in order to address those inter, intra- device heterogeneities but the problem remains open.

The standardization algorithms can be mainly grouped into two main categories[10]:

- Image domain whose procedures aim to correct the differences in acquisition before the extraction of features by standardization of protocols or application of image processing techniques.
- Feature domain which corrects unwanted variations after feature extraction by robust feature selection and batch effect correction.

The image-based procedures are not widely used as it is not a clear guideline to address the standardization of radiomic features. In the case of the feature domain, we can select robust image-derived biomarkers but there is certain loss of information by discarding the remaining features. This issue can be addressed by using Batch correction methods that allow standardization without loss of information. One well-known batch effect correction method is ComBat, that consists in dealing with the variability of parameters' distributions so they can be pooled together and makes the transformation to express all data in a common space. In a recent study, the authors showed that the feature-based harmonisation technique ComBat is able to remove the variability introduced by centre information from radiomic features, at the expense of slightly degrading classification performance. The authors suggested that piecewise linear histogram matching normalization was a better alternative as it gave features with greater generalisation ability for classification[11]. As we can observe, this is a challenging problem that will be explored.

In addition to the image-based features, there is a need to harmonize health records, health variables and questionnaires coming from different cohorts such as the 'alcohol intake', 'studies level' or risk factors such as diabetes or cholesterol and so on. In the real scenario, each cohort does not follow any standardization in order to collect patient´s data making the process tedious and slow. The variables needed for the study must be very well defined and identified in each cohort. But this process is not straightforward as some cohorts may not contain the same information and categories or might be different depending on the system of the country such as education. Moreover, we found private the patient tracking system

[9] Y. Nan .et al. "Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions". Information Fusion vol 82 (2022).
[10] H. Horng .et al. "Generalized ComBat harmonization methods for radiomic features with multi-modal distributions and multiple batch effects". Scientific Reports. 4493 (2022).
[11] Campello,V.M. et al. "Minimising multi-centre radiomics variability through image normalisation: a pilot study," Scientific Reports 12532 (2022).

in some cohorts making difficult the process of monitoring the progress of a person in terms of outcomes of the model (e.g., a healthy person at the baseline suffers atrial fibrillation in the future). We had to perform a "trick" to extract the evolution of the person such as checking the updated questionnaire or the medicine changes in the next follow-up of the patient of the research institution center after the baseline 5 years later. But this way, it is not the most efficient way to solve it. Then, there is a need to regularize the use of the follow-up patient records as well as a standardization of the data in order to process multi-center studies. Those regulations are not only in national level if not in the protocol of each research institutions which each one has their own regulations. One example of the difficulty of this task is the MORGAM cohort which aims to harmonize several databases from different research institution for 15 years. Then, it is clear that something must be done in order to automatize and speed up this process. There are some projects such as the DataTools4Heart project that are developing novel tools to automatize the harmonization. Their main goal is to implement a common data model (CDM) to make operable different data sources such as logical organization, terminologies, vocabularies and coding schemes. To standardize all the data sources in a comparable manner, they propose to transform data to the Observational Medical Outcomes Partnership (OMOP) CDM[12]. This project is in an initial stage, so it is need to wait until reaches a level of maturity.

Another problem to consider is the model imbalance. Fortunately, there are many techniques for this purpose such as upsampling or downsampling the minority class that will allow us to perform the model in a distributed manner.

Finally, federated learning inherits the hardware requirements of the model being distributed for deep learning models. The recommended setting proposed is to have available NVIDIA GPUs supporting CUDA and more than 12GB of RAM (e.g.,GPU-GeForce RTX 3090). This requirement is sometimes hard to meet in those health centers (in fact,  in one of the selected cohorts, they do not have this minimum hardware requirement and we are not allowed to extend the hardware). In this case, some solutions will need to be found, so as to facilitate the data processing offloading, for example using trusted third-parties computing facilities for sensitive data, as the ones described in Deliverable 5.1 and Deliverable 5.4.

[12] https://www.ohdsi.org/data-standardization/the-common-data-model/.

# 4. Summary

## 4.1. Detected data access challenges

The procedure for data request is relatively slow both for the data discoverability and for the procedure to obtain the data. For discoverability, you need to identify the research institutions which have available the required data via google scholar or from personal contacts and have an initial meeting to explain your project proposal. If they accept the proposal, a formalization by writing the project proposal will be the next step to address their ethical committee. In the case of federated learning, the project proposal must also include what technical requirements will be necessary in the research centre, making this process even slower as the research institution must provide the minimum technical specifications in order to run the necessary tools that we will provide to run the machine learning models.

## 4.2. Detected data analysis challenges

The main limitations come from the distribution issues of the federated learning framework. The legal issues of the patient data in a federated learning is not something that we must deal with, as the patient data is not exposed outside the research centre and the legal regulations are performed within each institution to collect the data. However, we still need to write a project proposal for their own ethical committee. This process must be performed for each cohort you want to include in your study so this makes the process relatively slow. Each institution usually has their own template. You can use the same information for all the cohorts but a different document must be filled in for each participant. The technical part is also very challenging. Some of the technical aspects, we need to face are the following: the communication overhead between the research institutions and the centre server, the risk of loss of transmission packets, limited network bandwidth or privacy breach. Then, some encryption and compression techniques must be adopted to reduce the model size and secure the privacy. The encryption methods make the federated learning framework even more secure; although as discussed, no patient data is shown outside the research centre, not even to the programmer. In addition to the transmission issues, the heterogeneities of each research centre (different acquisition protocols, the variety of scanners, label imbalance, size and different computational power) must also be dealt as heterogeneous aggregation can produce a degradation of the performance of the model making the trained client models hard to aggregate. In order to alleviate these issues some pre-processing techniques will be considered such as histogram matching from a reference data sample from the central server.

# Annex I: Extraction of Predictor Variables

| Name | UKB Field ID: code | Values | CovName | Notes |
|---|---|---|---|---|
| cov_age0 | 21003 | numeric | Age at baseline | No edits |
| cov_age2 | 21004 | numeric | Age at imaging | No edits |
| cov_sex | 31 | binary | Sex | No edits: 0= Female, 1= Male |
| cov_smoker | 1239 | binary | Current smoker | Combined categories 1 and 2 = "Current Smoker" |
| cov_deprivation | 189 | numeric | Townsend deprivation score | No edits |
| cov_bmi | 50, 21002 | numeric | BMI (kg/m2) | Derived from height and weight fields |
| cov_bsa | 50, 21003 | numeric | BSA | BSA by Dubois and Dubois: http://www.medcalc.com/body.html |
| cov_alcohol | 1558 | numeric | Alcohol intake | Reverse coded [0= Never, 1= Special occasions only, 2= One to three times a month, 3= Once or twice a week, 4= Three or four times a week, 5= Daily or almost daily, NA= Prefer not to answer or blank |
| cov_ipaq_tot | 864, 874, 884, 894, 904, 914 | numeric | IPAQ Score | As per. https://www.physio-pedia.com/images/c/c7/Quidelines_for_interpreting_the_IPAQ.pdf |
| cov_ipaq_group | 864, 874, 884, 894, 904, 914 | factor | IPAQ Group | 1= Less than 600, 2= 600-2999, 3= More than 3000 |
| cov_diabetes | 2443, 6153, 6177, 30750 | binary | Diabetes | Combination of any of: "yes" to 2443, "insulin" in medications, and HbA1c > 48 |
| cov_hypertens | 6153, 6177 | binary | Hypertension | "Blood pressure medication" present in either field |
| cov_highchol | 6133, 6177, 30690 | binary | High cholesterol | "Cholesterol lowering medication" in medication fields or cholesterol biochem > 7 |
| cov_educ | 6138, 845 | numeric | Education (years beyond age 14) | 0= Left school ≤14 years old without qualifications, 1= Left school ≤15 years old without qualifications, 2= High school diploma (eg. GCSE), 4= Sixth form qualification, 6= Professional qualification, 7= Higher education university degree |
| cov_ipaq_tot_log | Derived | numeric | Log IPAQ Score | ln(cov_ipaq_tot + 1) |
| cov_bmi_log | Derived | numeric | Log BMI | ln(cov_bmi) |
| cmr_LVEDV | 22421 | numeric | LV end diastolic volume | |
| cmr_LVESV | 22422 | numeric | LV end systolic volume | |
| cmr_LVSV | 22423 | numeric | LV stroke volume | |
| cmr_LVEF | 22420 | numeric | LV ejection fraction | |
| cmr_LVM | Derived | numeric | LV mass | |
| cmr_RVEDV | 24106 | numeric | RV end diastolic volume | |
| cmr_RVESV | 24107 | numeric | RV end systolic volume | |
| cmr_RVSV | 24108 | numeric | RV stroke volume | |
| cmr_RVEF | 24109 | numeric | RV ejection fraction | |

# Annex II: Disease Definitions

| Source | UKB Field ID: code | Description |
|---|---|---|
| **Myocardial infarction** | | |
| Self-report | 20002 | Heart attack/myocardial infarction |
| Algorithm | 42000 | Date of myocardial infarction |
| ICD10 | I21 | Acute myocardial infarction |
| | I22 | Subsequent myocardial infarction |
| | I23 | Certain current complications following acute myocardial infarction |
| First occurrences | 131298 | Acute myocardial infarction |
| | 131300 | Subsequent myocardial infarction |
| | 131302 | Certain current complications following acute myocardial infarction |
| Diagnosed by doctor | 6150: 1 | Heart attack |
| | 3894 | Age heart attack diagnosed |
| ICD9 | 410 | Acute myocardial infarction |
| | 411 | Other acute and subacute forms of ischaemic heart disease |
| | 412 | Old myocardial infarction |
| **Heart failure** | | |
| Self-report | 20002 | Heart failure/pulmonary odema |
| ICD10 | I500 | Congestive heart failure |
| | I501 | Left ventricular failure |
| | I509 | Heart failure, unspecified |
| First occurrences | 131354 | heart failure |
| **Atrial fibrillation** | | |
| Self-report | 20002 | Atrial fibrillation |
| ICD10 | I480 | Paroxysmal atrial fibrillation |
| | I481 | Persistent atrial fibrillation |
| | I482 | Chronic atrial fibrillation |
| **Stroke** | | |
| Self-report | 20002 | Stroke |
| | 20002 | Ischaemic stroke |
| | 20002 | Brain haemorrhage |
| Algorithm | 42006 | Date of stroke |
| | 42008 | Date of ischaemic stroke |
| | 42010 | Date of intracerebral haemorrhage |
| Diagnosed by doctor | 6150: 3 | Stroke |
| | 4056 | Age stroke diagnosed |
| ICD10 | I61 | Intracerebral haemorrhage |
| | I62 | Other nontraumatic intracranial haemorrhage |
| | I63 | Cerebral infarction |
| | I64 | Stroke, not specified as haemorrhage or infarction |
| ICD9 | 431 | Intracerebral haemorrhage |
| | 432 | Other and unspecified intracranial haemorrhage |
| | 434 | Occlusion of cerebral arteries |
| | 436 | Acute but ill-defined cerebrovascular disease |
| First occurrences | 131362 | Intracerebral haemorrhage |
| | 131364 | Other nontraumatic intracranial haemorrhage |
| | 131366 | Cerebral infarction |
| | 131368 | Stroke, not specified as haemorrhage or infarction |
| **Diabetes** | | |
| Diagnosed by doctor | 2443 | Diabetes diagnosed by doctor |
| | 2976 | Age diabetes diagnosed by doctor |
| Medications | 6177, 6153: 3 | Insulin |
| Biochemistry | 30750 | Glycated haemoglobin (HbA1c) > 48 mmol/mol |
| **High cholesterol** | | |
| Medications | 6177, 6153: 1 | Cholesterol lowering medication |
| Biochemistry | 30690 | Cholesterol > 7 mmol/L |
| **Hypertension** | | |
| Medications | 6177, 6153: 2 | Blood pressure medication |

# Annex III: AF-related Outcome Variables

| Name | Type of value | Label |
|---|---|---|
| **2296_x_x** | Textual | Touchscreen Health and medical history General health: Falls in the last year |
| **131342_x_x** | Date | related outcomes First occurrences Circulatory system disorders: Date I44 first reported (atrioventricular and left bundlebranch block) |
| **131343_x_x** | Textual | related outcomes First occurrences Circulatory system disorders: Source of report of I44 (atrioventricular and left bundlebranch block) |
| **131344_x_x** | Date | related outcomes First occurrences Circulatory system disorders: Date I45 first reported (other conduction disorders) |
| **131345_x_x** | Textual | related outcomes First occurrences Circulatory system disorders: Source of report of I45 (other conduction disorders) |
| **131346_x_x** | Date | related outcomes First occurrences Circulatory system disorders: Date I46 first reported (cardiac arrest) |
| **131347_x_x** | Textual | related outcomes First occurrences Circulatory system disorders: Source of report of I46 (cardiac arrest) |
| **131348_x_x** | Date | related outcomes First occurrences Circulatory system disorders: Date I47 first reported (paroxysmal tachycardia) |
| **131349_x_x** | Textual | related outcomes First occurrences Circulatory system disorders: Source of report of I47 (paroxysmal tachycardia) |
| **131350_x_x** | Date | related outcomes First occurrences Circulatory system disorders: Date I48 first reported (atrial fibrillation and flutter) |
| **131351_x_x** | Textual | related outcomes First occurrences Circulatory system disorders: Source of report of I48 (atrial fibrillation and flutter) |
| **131352_x_x** | Date | related outcomes First occurrences Circulatory system disorders: Date I49 first reported (other cardiac arrhythmias) |
| **131353_x_x** | Textual | related outcomes First occurrences Circulatory system disorders: Source of report of I49 (other cardiac arrhythmias) |
| **41200_x_x** | Textual | related outcomes Hospital inpatient Summary Operations: Operative procedures main OPCS4 |
| **41211_x_x** | Textual | related outcomes Hospital inpatient Summary Administration: Destinations on discharge from hospital (polymorphic) |
| **41231_x_x** | Textual | related outcomes Hospital inpatient Summary Administration: Hospital episode type |

| | | |
|---|---|---|
| **41232_x_x** | Textual | related outcomes Hospital inpatient Summary Administration: Administrative and legal statuses |
| **41233_x_x** | Textual | related outcomes Hospital inpatient Summary Administration: Sources of admission to hospital (polymorphic) |
| **41234_x_x** | Integer | related outcomes Hospital inpatient Recordlevel access: Records in HES inpatient diagnoses dataset |
| **41235_x_x** | Integer | related outcomes Hospital inpatient Summary Administration: Spells in hospital |
| **41244_x_x** | Textual | related outcomes Hospital inpatient Summary Administration: Intended management of patient (recoded) |
| **41248_x_x** | Textual | related outcomes Hospital inpatient Summary Administration: Destinations on discharge from hospital (recoded) |
| **41249_x_x** | Textual | related outcomes Hospital inpatient Summary Administration: Methods of admission to hospital (recoded) |
| **41250_1_1** | Textual | related outcomes Hospital inpatient Summary Administration: Methods of discharge from hospital (recoded) |
| **41251_x_x** | Textual | related outcomes Hospital inpatient Summary Administration: Sources of admission to hospital (recoded) |
| **41253_x_x** | Textual | related outcomes Hospital inpatient Summary Administration: Inpatient record format |
| **41256_x_x** | Textual | related outcomes Hospital inpatient Summary Operations: Operative procedures main OPCS3 |
| **41257_x_x** | Date | related outcomes Hospital inpatient Summary Operations: Date of first operative procedure main OPCS3 |
| **41258_x_x** | Textual | related outcomes Hospital inpatient Summary Operations: Operative procedures secondary OPCS3 |
| **41259_x_x** | Integer | related outcomes Hospital inpatient Recordlevel access: Records in HES inpatient main dataset |
| **41260_x_x** | Date | related outcomes Hospital inpatient Summary Operations: Date of first operative procedure main OPCS4 |
| **4501_3_0** | Textual | Touchscreen Family history: Nonaccidental death in close genetic family |
| **40000_x_x** | Date | related outcomes Death register: Date of death |
| **40001_x_x** | Textual | related outcomes Death register: Underlying (primary) cause of death: ICD10 |
| **40007_x_x** | Decimal | related outcomes Death register: Age at death |
| **40010_x_x** | Textual | related outcomes Death register: Description of cause of death |

| | | |
|---|---|---|
| **40018_x_x** | Textual | related outcomes Death register: Death record format |
| **40023_x_x** | Integer | related outcomes Death register: Records in death dataset |
| **136_x_x** | Integer | Verbal interview Operations: Number of operations, selfreported |
| **20004_x_x** | Textual | Verbal interview Operations: Operation code |
| **20010_x_x** | Decimal | Verbal interview Operations: Interpolated Year when operation took place |
| **20011_x_x** | Decimal | Verbal interview Operations: Interpolated Age of participant when operation took place |
| **2415_x_x** | Textual | Touchscreen Health and medical history Operations: Had major operations |
| **92_x_x** | Integer | Verbal interview Operations: Operation year age first occurred |
| **136_x_x** | Integer | Verbal interview Operations: Number of operations, selfreported |
| **137_x_x** | Integer | Verbal interview Medications: Number of treatments medications taken |
| **20003_x_x** | Textual | Verbal interview Medications: Treatment medication code |
| **2492_x_x** | Textual | Touchscreen Health and medical history Medication: Taking other prescription medications |
| **6153_x_x** | Textual | Touchscreen Health and medical history Medication: Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones |

# Annex IV: Dictionary of CMR Radiomics

| Name | Type | label:en |
|---|---|---|
| f_eid | integer | f_eid. Identifier of the patient. Each segmentation in short axis has three ROIs: left ventricle, right ventricle and myocardium in short axis and two ROIs in long axis: Right and left atrium for two frames (end diastole and end systole). |
| Volume_ROI_Frame | decimal | Volume_ROI_Frame. The volume of the ROI is approximated by multiplying the number of voxels in the ROI by the volume of a single voxel |
| SurfaceArea_ROI_Frame | decimal | SurfaceArea_ROI_Frame. Surface Area is an approximation of the ROI surface based on triangulation interpretation |
| SurfaceAreatoVolumRatio_ROI_Frame | decimal | SurfaceAreatoVolumeRatio_ROI_Frame. Lower values of this parameter indicate a sphere-like shape of the ROI |
| Sphericity_ROI_Frame | decimal | Sphericity_ROI_Frame. Sphericity is a measure of the roundness of the ROI relative to a sphere |
| Max3Ddiameter_ROI_Frame | decimal | Max3Ddiameter_ROI_Frame. The largest pairwise Euclidean distance between ROI surface voxels |
| Max2DdiameterSlice_ROI_Frame | decimal | Max2DdiameterSlice_ROI_Frame. The largest pairwise Euclidean distance between ROI surface voxels of specific axial slice |
| Max2DdiameterColumn_ROI_Frame | decimal | Max2DdiameterColumn_ROI_Frame. The largest pairwise Euclidean distance between ROI surface voxels of specific coronal slice |
| Max2DdiameterRow_ROI_Frame | decimal | Max2DdiameterRow_ROI_Frame. The largest pairwise Euclidean distance between ROI surface voxels of specific sagittal slice |
| MajorAxis_ROI_Frame | decimal | MajorAxis_ROI_Frame. A feature derived from the principal component analysis proportional to the square root of length of the largest principal component axes |
| MinorAxis_ROI_Frame | decimal | MinorAxis_ROI_Frame. A feature derived from the principal component analysis proportional to the square root of length of the second largest principal component axes |
| LeastAxis_ROI_Frame | decimal | LeastAxis_ROI_Frame. A feature derived from the principal component analysis proportional to the square root of length of the smallest largest principal component axes |
| Elongation_ROI_Frame | decimal | Elongation_ROI_Frame. A feature derived from the principal component analysis proportional to the ratio of lengths of the second largest and the largest principal component axes |
| Flatness_ROI_Frame | decimal | Flatness_ROI_Frame. A feature derived from the principal component analysis proportional to the ratio of lengths of the smallest and the largest principal component axes |
| Energy_ROI_Frame | decimal | Energy_ROI_Frame. Energy is a measure of the magnitude of voxel values in an image |
| TotalEnergy_ROI_Frame | decimal | TotalEnergy_ROI_Frame. Total Energy is the value of Energy feature scaled by the volume of the voxel in cubic mm |
| Entropy_ROI_Frame | decimal | Entropy_ROI_Frame. Entropy specifies the uncertainty or randomness in the image values. It measures the average amount of information required to encode the image values. |

| | | |
|---|---|---|
| Minimum_ROI_Frame | decimal | Minimum_ROI_Frame. Minimum intensity value present in the ROI |
| Percentile10_ROI_Frame | decimal | Percentile10_ROI_Frame. Value below which 10% of the intensities may be found in the histogram of the ROI |
| Percentile90_ROI_Frame | decimal | Percentile90_ROI_Frame. Value below which 90% of the intensities may be found in the histogram of the ROI |
| Maximum_ROI_Frame | decimal | Maximum_ROI_Frame. Maximum grey level intensity found in the ROI |
| Mean_ROI_Frame | decimal | Mean_ROI_Frame. Mean gray level intensity found in the ROI |
| Median_ROI_Frame | decimal | Median_ROI_Frame. Median grey level intensity found in the ROI |
| InterquartileRange_ROI_Frame | decimal | InterquartileRange_ROI_Frame. The difference between the 25th and 75th percentile of ROI |
| Range_ROI_Frame | decimal | Range_ROI_Frame. Difference between the maximum and minimum gray tone present in the ROI |
| MeanAbsoluteDeviation_ROI_Frame | decimal | MeanAbsoluteDeviation_ROI_Frame. Mean Absolute Deviation is the mean distance of all intensity values from the Mean Value present in the ROI |
| RobustMeanAbsDeviation_ROI_Frame | decimal | RobustMeanAbsoluteDeviation_ROI_Frame. Robust Mean Absolute Deviation is a modification of Mean Absolute Deviation that takes into account only ROI intensities present in between 10th and 90th percentile which helps to avoid noise impact |
| RootMeanSquared_ROI_Frame | decimal | RootMeanSquared_ROI_Frame. Root Mean Squared is the square-root of the mean of all the intensity values squared. Characterizes the magnitude of the image gray tone |
| Skewness_ROI_Frame | decimal | Skewness_ROI_Frame. Skewness measures the asymmetry of the distribution of values around the Mean value |
| Kurtosis_ROI_Frame | decimal | Kurtosis_ROI_Frame. Kurtosis measures the peakedness of the values distribution in the image ROI |
| Variance_ROI_Frame | decimal | Variance_ROI_Frame. Variance is the he mean of the squared distances of each intensity value from the Mean value |
| Uniformity_ROI_Frame | decimal | Uniformity_ROI_Frame. Uniformity is a measure of the sum of the squares of each intensity value. This is a measure of the heterogeneity of the ROI |
| Autocorrelation_glcm_ROI_Frame | decimal | Autocorrelation_glcm_ROI_Frame. Autocorrelation detects repetitive patterns present in the ROI. Intends to measure the magnitude of the fineness and coarseness of texture |
| JointAverage_glcm_ROI_Frame | decimal | JointAverage_glcm_ROI_Frame. Joint Average returns the mean gray level intensity of the i distribution |
| ClusterProminence_glcm_ROI_Frame | decimal | ClusterProminence_glcm_ROI_Frame. Cluster Prominence is a measure of the skewness and asymmetry of the GLCM |
| ClusterShade_glcm_ROI_Frame | decimal | ClusterShade_glcm_ROI_Frame. Cluster Shade is a measure of the skewness and uniformity of the GLCM. Somewhat similar to Prominense |
| ClusterTendency_glcm_ROI_Frame | decimal | ClusterTendency_glcm_ROI_Frame. Cluster Tendency is a measure of groupings of voxels within the ROI with similar gray-level values |
| Contrast_glcm_ROI_Frame | decimal | Contrast_glcm_ROI_Frame. Contrast is a measure of the local intensity variation, favoring values away from the diagonal (i=j) of the GLCM(diagonal elements represent the co-occurrence of the same intensities between compared pixels). |

| | | |
|---|---|---|
| | | A larger value correlates with a greater disparity in intensity values among neighboring voxels |
| Correlation_glcm_ROI_Frame | decimal | Correlation_glcm_ROI_Frame. Correlation is a value between 0 (uncorrelated) and 1 (perfectly correlated) showing the linear dependency of gray level values to their respective voxels in the GLCM |
| DifferenceAverage_glcm_ROI_Frame | decimal | DifferenceAverage_glcm_ROI_Frame. Difference Average measures the relationship between occurrences of pairs with similar intensity values(closer to diagonal of the GLCM) and occurrences of pairs with differing intensity values in GLCM |
| DifferenceEntropy_glcm_ROI_Frame | decimal | DifferenceEntropy_glcm_ROI_Frame. Difference Entropy is a measure of the randomness/variability in neighborhood intensity value differences |
| DifferenceVariance_glcm_ROI_Frame | decimal | DifferenceVariance_glcm_ROI_Frame. Difference Variance is a measure of heterogeneity that places higher weights on differing intensity level pairs that deviate more from the mean |
| JointEnergy_glcm_ROI_Frame | decimal | JointEnergy_glcm_ROI_Frame. Joint Energy is a measure of how homogeneous are the patterns in the ROI |
| JointEntropy_glcm_ROI_Frame | decimal | JointEntropy_glcm_ROI_Frame. Joint entropy is a measure of the randomness/variability in neighborhood intensity values |
| InformalMeasOfCorr1_glcm_ROI_Frame | decimal | InformalMeasureofCorrelation1_glcm_ROI_Frame. Alternative definition of Correlation based on ratio of entropy dependencies to the maximum entropy |
| InformalMeasOfCorr2_glcm_ROI_Frame | decimal | InformalMeasureofCorrelation2_glcm_ROI_Frame. Alternative definition of Correlation based on entropy dependencies. Uses square root of entropies difference instead of the max |
| InverseDiffMoment_glcm_ROI_Frame | decimal | InverseDifferenceMoment_glcm_ROI_Frame. Inverse Difference Moment (IDM) is a measure of the local homogeneity of an image |
| InverDiffMomentNorm_glcm_ROI_Frame | decimal | InverseDifferenceMomentNormalized_glcm_ROI_Frame. Normalization of Inverse Difference Moment. It normalizes the square of the difference between neighboring intensity values by dividing over the square of the total number of discrete intensity values |
| InverseDifference_glcm_ROI_Frame | decimal | InverseDifference_glcm_ROI_Frame. Inverse Difference is a measure of the local homogeneity of an image |
| InverDifferenceNorm_glcm_ROI_Frame | decimal | InverseDifferenceNormalized_glcm_ROI_Frame. Inverse Difference Normalized (IDN) normalizes the difference between the neighboring intensity values by dividing over the total number of discrete intensity values |
| InverseVariance_glcm_ROI_Frame | decimal | InverseVariance_glcm_ROI_Frame. Inverse of the variance sums up the elements of the GLCM matrix while decreasing the values which lay further from the diagonal proportional to the distance |
| MaximumProbability_glcm_ROI_Frame | decimal | MaximumProbability_glcm_ROI_Frame. Maximum Probability is the occurrence of the most predominant pair of neighboring intensity values |
| SumAverage_glcm_ROI_Frame | decimal | SumAverage_glcm_ROI_Frame. Sum Average measures the relationship between occurrences of pairs with lower intensity values and occurrences of pairs with higher intensity values |
| SumEntropy_glcm_ROI_Frame | decimal | SumEntropy_glcm_ROI_Frame. Sum Entropy is a sum of neighborhood intensity value differences |

| | | |
|---|---|---|
| SumofSquares_glcm_ROI_Frame | decimal | SumofSquares_glcm_ROI_Frame. Sum of Squares is a measure in the distribution of neighboring intensity level pairs about the mean intensity level in the GLCM |
| SmallAreaEmphasis_glszm_ROI_Frame | decimal | SmallAreaEmphasis_glszm_ROI_Frame. Small area emphasis (SAE) measures how many small regions with the same intensity value(fine texture) are present in the ROI opposed to big regions with same intensity value(homogeneous texture).A greater value of this feature indicates the presence of more fine textures within the ROI |
| LargeAreaEmphasis_glszm_ROI_Frame | decimal | LargeAreaEmphasis_glszm_ROI_Frame. Large Area Emphasis (LAE) measures how many big regions with same intensity value(homogeneous texture) are present in the ROI opposed to the small regions with the same intensity value(fine texture). A greater value of this feature indicates the presence of more coarse textures within the ROI |
| GrayLevelNonUnifor_glszm_ROI_Frame | decimal | GrayLevelNonUniformity_glszm_ROI_Frame. Gray Level Non-Uniformity (GLN) measures the variability of gray-level intensity values in the image, with a lower value indicating more homogeneity in intensity values and higher value indicating the presence of fine texture |
| GrayLevlNonUniNorm_glszm_ROI_Frame | decimal | GrayLevelNonUniformityNormalized_glszm_ROI_Frame. Normalized version of Gray Level Non-Uniformity which takes into account the number of zones with the same intensity present within the ROI |
| SizeZoneNonUniform_glszm_ROI_Frame | decimal | SizeZoneNonUniformity_glszm_ROI_Frame. Size-Zone Non-Uniformity (SZN) measures the variability of the size zone volumes(regions with the same intensity) in the image, with a lower value indicating that ROI has even size zones volumes |
| SizeZoneNonUniNorm_glszm_ROI_Frame | decimal | SizeZoneNonUniformityNormalized_glszm_ROI_Frame. Normalized version of Size-Zone Non-Uniformity which takes into account the number of zones with the same intensity present within the ROI |
| ZonePercentage_glszm_ROI_Frame | decimal | ZonePercentage_glszm_ROI_Frame. Zone Percentage (ZP) measures the coarseness of the texture by taking the ratio of number of zones with the same intensity and number of voxels in the ROI |
| GrayLevelVariance_glszm_ROI_Frame | decimal | GrayLevelVariance_glszm_ROI_Frame. Gray Level Variance measures the variance in gray level intensities for the zones (regions with same intensity) |
| ZoneVariance_glszm_ROI_Frame | decimal | ZoneVariance_glszm_ROI_Frame. Zone Variance measures the variance in zone (region with the same intensity) size |
| ZoneEntropy_glszm_ROI_Frame | decimal | ZoneEntropy_glszm_ROI_Frame. Zone Entropy measures the uncertainty/randomness in the distribution of zone sizes and gray levels. A higher value indicates more heterogeneous texture patterns |
| LowGrayLevlZoneEmp_glszm_ROI_Frame | decimal | LowGrayLevelZoneEmphasis_glszm_ROI_Frame. Low Gray Level Zone Emphasis measures the distribution of lower gray-level size zones, with a higher value indicating a greater proportion of lower gray-level values and size zones in the image |
| HighGrayLvlZoneEmp_glszm_ROI_Frame | decimal | HighGrayLevelZoneEmphasis_glszm_ROI_Frame. High Gray Level Zone Emphasis measures the distribution of the higher gray-level values, with a higher value indicating a greater proportion of both higher gray-level values and size zones in the image |
| SmallAreaLowGraEmp_glszm_ROI_Frame | decimal | SmallAreaLowGrayLevelEmphasis_glszm_ROI_Frame. Small area low gray level emphasis measures the proportion in |

| | | |
|---|---|---|
| | | the image of the joint distribution of smaller size zones with lower gray-level values |
| SmallAreaHighGrEmp_glszm_ROI_Frame | decimal | SmallAreaHighGrayLevelEmphasis_glszm_ROI_Frame. Small area high gray level emphasis measures the proportion in the image of the joint distribution of smaller size zones with higher gray-level values |
| LargeAreaLowGraEmp_glszm_ROI_Frame | decimal | LargeAreaLowGrayLevelEmphasis_glszm_ROI_Frame. Large area low gray level emphasis measures the proportion in the image of the joint distribution of larger size zones with lower gray-level values |
| LargeAreaHighGrEmp_glszm_ROI_Frame | decimal | LargeAreaHighGrayLevelEmphasis_glszm_ROI_Frame. Large area high gray level emphasis measures the proportion in the image of the joint distribution of larger size zones with higher gray-level values |
| ShortRunEmphasis_glrlm_ROI_Frame | decimal | ShortRunEmphasis_glrlm_ROI_Frame. Short run emphasis is a measure of the distribution of short run lengths, with a greater value indicative of shorter run lengths and more fine textural textures |
| LongRunEmphasis_glrlm_ROI_Frame | decimal | LongRunEmphasis_glrlm_ROI_Frame. Long run emphasis is a measure of the distribution of long run lengths, with a greater value indicative of longer run lengths and more coarse structural textures |
| GrayLevlNonUniform_glrlm_ROI_Frame | decimal | GrayLevelNonUniformity_glrlm_ROI_Frame. Gray level non-uniformity (GLN) measures the variability of gray-level intensity values in the image, with a lower value indicating more homogeneity in intensity values and higher value Gray level nonuniformity (GLN) measures the similarity of gray-level intensity values in the image, where a lower GLN value correlates with a greater similarity in intensity values |
| GrayLevlNonUniNorm_glrlm_ROI_Frame | decimal | GrayLevelNonUniformityNormalized_glrlm_ROI_Frame. Gray level non-uniformity normalized (GLNN) measures the similarity of gray-level intensity values in the image, where a lower GLNN value correlates with a greater similarity in intensity values. This is the normalized version of the GLN formula |
| RunLengtNonUniform_glrlm_ROI_Frame | decimal | RunLengthNonUniformity_glrlm_ROI_Frame. Run length non-uniformity (RLN) measures the similarity of run lengths throughout the image, with a lower value indicating more homogeneity among run lengths in the image |
| RunLengtNonUniNorm_glrlm_ROI_Frame | decimal | RunLengthNonUniformityNormalized_glrlm_ROI_Frame. Run length non-uniformity normalized (RLNN) measures the similarity of run lengths throughout the image, with a lower value indicating more homogeneity among run lengths in the image. This is the normalized version of the RLN formula |
| RunPercentage_glrlm_ROI_Frame | decimal | RunPercentage_glrlm_ROI_Frame. Run percentage measures the coarseness of the texture by taking the ratio of number of runs and number of voxels in the ROI |
| GrayLevelVariance_glrlm_ROI_Frame | decimal | GrayLevelVariance_glrlm_ROI_Frame. Gray level variance measures the variance in gray level intensity for the runs |
| RunVariance_glrlm_ROI_Frame | decimal | RunVariance_glrlm_ROI_Frame. Run variance is a measure of the variance in runs for the run lengths |
| RunEntropy_glrlm_ROI_Frame | decimal | RunEntropy_glrlm_ROI_Frame. Run entropy measures the uncertainty/randomness in the distribution of run lengths and gray levels. A higher value indicates more heterogeneity in the texture patterns |
| LowGrayLevelRunEmp_glrlm_ROI_Frame | decimal | LowGrayLevelRunEmphasis_glrlm_ROI_Frame. Low gray level run emphasis measures the distribution of low gray-level |

| | | |
|---|---|---|
| | | values, with a higher value indicating a greater concentration of low graylevel values in the image |
| HighGrayLevlRunEmp_glrlm_ROI_Frame | decimal | HighGrayLevelRunEmphasis_glrlm_ROI_Frame. High gray level run emphasis measures the distribution of the higher gray-level values, with a higher value indicating a greater concentration of high gray-level values in the image |
| ShortRunLowGrayEmp_glrlm_ROI_Frame | decimal | ShortRunLowGrayLevelEmphasis_glrlm_ROI_Frame. Short run low gray level emphasis measures the joint distribution of shorter run lengths with lower gray-level values |
| ShortRunHighGrEmp_glrlm_ROI_Frame | decimal | ShortRunHighGrayLevelEmphasis_glrlm_ROI_Frame. Short run high gray level emphasis measures the joint distribution of shorter run lengths with higher gray-level values |
| LongRunLowGrayEmp_glrlm_ROI_Frame | decimal | LongRunLowGrayLevelEmphasis_glrlm_ROI_Frame. Long run low gray level emphasis measures the joint distribution of long run lengths with higher gray-level values |
| LongRunHighGrayEmp_glrlm_ROI_Frame | decimal | LongRunHighGrayLevelEmphasis_glrlm_ROI_Frame. Long run high gray level emphasis measures the proportion in the image of the joint distribution of larger size zones with higher gray-level values |
| Coarseness_ngtdm_ROI_Frame | decimal | Coarseness_ngtdm_ROI_Frame. Coarseness is a measure of average difference between the center voxel and its neighbourhood and is an indication of the spatial rate of change. A higher value indicates a lower spatial change rate and a locally more uniform texture |
| Contrast_ngtdm_ROI_Frame | decimal | Contrast_ngtdm_ROI_Frame. Contrast is a measure of the spatial intensity change, but is also dependent on the overall gray level dynamic range. Contrast is high when both the dynamic range and the spatial change rate are high, i.e. an image with a large range of gray levels, with large changes between voxels and their neighborhood |
| Busyness_ngtdm_ROI_Frame | decimal | Busyness_ngtdm_ROI_Frame. Busyness is a measure of the change from a pixel to its neighbor. A high value for busyness indicates a busy image, with rapid changes of intensity between pixels and its neighborhood |
| Complexity_ngtdm_ROI_Frame | decimal | Complexity_ngtdm_ROI_Frame. An image is considered complex when there are many primitive components in the image, i.e. the image is non-uniform and there are many rapid changes in gray level intensity |
| Strength_ngtdm_ROI_Frame | decimal | Strength_ngtdm_ROI_Frame. Strength is a measure of the primitives in an image. Its value is high when the primitives are easily defined and visible, i.e. an image with slow change in intensity but more large coarse differences in gray level intensities |
| SmallDepEmphasis_gldm_ROI_Frame | decimal | SmallDependenceEmphasis_gldm_ROI_Frame. Small dependence emphasis measures how many small dependencies are present in ROI. Greater values represents smaller dependence and less homogeneous texture |
| LargeDepEmphasis_gldm_ROI_Frame | decimal | LargeDependenceEmphasis_gldm_ROI_Frame. Large dependence emphasis measures how many large dependencies are present in ROI. Greater value indicates larger dependence and more homogeneous texture |
| GrayLevelNonUniform_gldm_ROI_Frame | decimal | GrayLevelNonUniformity_gldm_ROI_Frame. Gray level non-uniformity measures the similarity of gray-level intensity values in the image. Higher value indicates smaller similarity whereas lower value indicates higher similarity in gray level intensity values. |

| | | |
|---|---|---|
| DepNonUniform_gldm_ROI_Frame | decimal | DependenceNonUniformity_gldm_ROI_Frame. Dependence non-uniformity measures the similarity of dependence throughout the image, with a lower value indicating more homogeneity among dependencies in the image |
| DepNonUniformNorm_gldm_ROI_Frame | decimal | DependenceNonUniformityNormalized_gldm_ROI_Frame. Dependence non-uniformity normalized measures the similarity of dependence in the image, with a lower value indicating more homogeneity among dependencies in the image. This is the normalized version of the dependence non-uniformity formula |
| GrayLevelVariance_gldm_ROI_Frame | decimal | GrayLevelVariance_gldm_ROI_Frame. Gray level variance measures the variance in grey level in the image |
| DepVariance_gldm_ROI_Frame | decimal | DependenceVariance_gldm_ROI_Frame. Dependence variance measures the variance in gray level dependence size in the image |
| DepEntropy_gldm_ROI_Frame | decimal | DependenceEntropy_gldm_ROI_Frame. Dependence entropy measures the randomness in the gray level dependencies and gray levels |
| LowGrayLevelEmphasi_gldm_ROI_Frame | decimal | LowGrayLevelEmphasis_gldm_ROI_Frame. Low gray level emphasis measures the distribution of low gray-level values, with a higher value indicating a greater concentration of low gray-level values in the image |

# Annex V: Dictionary of Electrocardiogram waveforms

| Name of Features | Type | Description |
|---|---|---|
| **Ventricular rate** | decimal | Ventricular rate indicated by the frequency of the QRS complex |
| **QRS duration** | decimal | QRS duration time for ventricular depolarization. |
| **QT interval Time** | decimal | QT interval Time taken for ventricular depolarisation and repolarisation |
| **Corrected QT** | decimal | Corrected QT Correction of the QT interval for heart rate extremes |
| **R-R Interval R** | decimal | R-R Interval R intervals between successive heartbeats |
| **P-P Interval Distance** | decimal | P-P Interval Distance between consecutive P waves due to atrial depolarization |
| **R- wave axis** | decimal | R- wave axis deviation Rotation of the R wave in the frontal plane |
| **T- wave axis** | decimal | T- wave axis deviation Rotation of the T wave in the frontal plane |
| **QRS Number** | decimal | Number of QRS complexes |
| **T Offset** | decimal | T offset of the T wave |
| **P Onset** | decimal | P onset of the P wave |
| **P Offset** | decimal | P offset of the P wave |

# Acronyms and Abbreviations

- AF – Atrial Fibrillation
- BMI – Body Mass Index
- BSA – Body Surface Area
- CA – Consortium Agreement
- CHD – Coronary Heart Disease
- CVD – Cardiovascular Disease
- D – deliverable
- DoA – Description of Action (Annex 1 of the Grant Agreement)
- EB – Executive Board
- EC – European Commission
- ECG – Electrocardiogram
- ED – End Diastole
- ES – End Systole
- GA – General Assembly / Grant Agreement
- HPC – High Performance Computing
- IPAQ – International Physical Activity Questionnaires
- IPR – Intellectual Property Right
- KPI – Key Performance Indicator
- LV – Left Ventricle
- LVM – Left Ventricular Mass
- M – Month
- MS – Milestones
- PM – Person month / Project manager
- ROI – Region of Interest
- RV – Right Ventricle
- UKB – UK Biobank
- WP – Work Package
- WPL – Work Package Leader