REal-time data monitoring for **S**hared, **A**daptive, **M**ulti-domain and **P**ersonalised prediction and decision making for **L**ong-term Pulmonary care **E**cosystems

D3.3: Key features extraction

Dissemination level: PU
Document type: Report
Version: 1.0
Date: 31-08-2023

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 965315. This result reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Document Details

Reference No.	965315
Project title	**RE-SAMPLE** - **RE**al-time data monitoring for **S**hared, **A**daptive, **M**ulti-domain and **P**ersonalised prediction and decision making for **L**ong-term Pulmonary care **E**cosystems
Title of deliverable	Key features extraction
Due date deliverable	31-08-2023
Work Package	3
Document type	Report
Dissemination Level	PU: Public
Approved by	Coordinator
Authors	Serge Autexier, Jakob Lehmann, Gesa Wimberg (DFKI)
Reviewers	Danae Lekka (iSPRINT), Costas Lambrinoudakis (UPRC), Thrasyvoulos Giannakopoulos (UPRC), Christos Kalloniatis (UPRC)
Total No. of pages	45

Partners

Participant No	Participant organisation name (country)	Participant abbreviation
1 (Coordinator)	University of Twente (NL)	UT
2	Foundation Medisch Spectrum Twente (NL)	MST
3	University of Piraeus Research Center (GR)	UPRC
4	Foundation Tartu University Hospital (EE)	TUK
5	Foundation University Polyclinic Agostino Gemelli IRCCS (IT)	GEM
6	European Hospital and Healthcare Federation (BE)	HOPE
7	German Research Center for Artificial Intelligence GMBH (DE)	DFKI
8	ATOS IT Solutions and Services Iberia SL (ES)	ATOS
9	Roessingh Research and Development BV (NL)	RRD
10	Innovation Sprint (BE)	iSPRINT

Abstract

The goal of RE-SAMPLE is to use real-world data to empower patients with Chronic Obstructive Pulmonary Disease and complex chronic conditions to engage in self-care and to support their healthcare providers, by developing a virtual companionship programme. Within the virtual companionship programme, machine learning models will be used to provide predictions on disease progression and quality of life scores, accompanied with explanations. The machine learning models will also offer coaching suggestions and predictions for simulated future patient behaviour.

The objective of this deliverable is the documentation of the process of extracting the most important features from the RE-SAMPLE datasets containing data from the Hospital Information Systems and real-world data like environmental data, activity data and questionnaire scores. This enables the construction of a significantly smaller dataset that is well-suited for predictive machine learning models. Removing unnecessary or weak predictors reduces model complexity and can thereby improve performance and/or interpretability. Additionally, this approach is aligned with the data minimisation principle of the general data protection regulation, as it ensures that only necessary data is retained, reducing the potential privacy risks associated with handling large datasets. During the analysis, special attention is paid to the effort required for producing each variable, in terms of work for the healthcare professionals, cost to the hospitals and burden on the patients. The final dataset should minimise the effort while preserving the other objectives of the RE-SAMPLE project. Apart from the datasets utilised, the deliverable presents the methods used for feature extraction as well as the results of applying them to the available datasets.

Contents

List of Figures

List of Tables

Symbols, definitions, abbreviations, and acronyms

6MWT	6-minute walking test
ADO	Age, Dyspnoea, airflow Obstruction
AECOPD	Acute Exacerbations of Chronic Obstructive Pulmonary Disease
BMI	Body mass index
BODE	Body-mass index, airflow Obstruction, Dyspnea, and Exercise
CHF	Chronic Heart Failure
COPD	Chronic Obstructive Pulmonary Disease
D	Deliverable
EBM	Explainable Boosting Machine
EQ5D	EuroQol Group 5D Questionnaire
FEV_1	Forced Expiratory Volume in one second
FVC	Forced Vital Capacity
GDPR	General Data Protection Regulation
GOLD	Global Initiative for Chronic Obstructive Lung Disease
HIS	Hospital Information System(s)
ICS	Inhaled Corticosteroids
IHD	Ischaemic Heart Disease
IQR	Interquartile range
M	Month
ML	Machine Learning
mMRC	Modified Medical Research Council
QoL	Quality of Life
RWD	Real-World Data
SHAP	SHapley Additive exPlanations
WP	Work Package

1. Introduction

This deliverable (D) describes the process of key feature extraction from the RE-SAMPLE datasets. The extracted features are used for the predictive models in RE-SAMPLE. In D3.1 "Training of the predictive and simulation models" (Month (M) 24), datasets, first results and the procedure of training predictive models have been presented. D3.3 is the second deliverable in Work Package (WP) 3 that is concerned with the development of the predictive models for RE-SAMPLE and a prerequisite for D3.4 "Prediction and simulation model validation" (M38). Both, D3.3 "Key features extraction" and D3.4 "Prediction and simulation model validation" are part of task 3.2 "Validation and key feature extraction" (M18-M38).

The RE-SAMPLE dataset contains data from the Hospital Information Systems (HIS) from the three pilot sites such as blood test results and medication information. Moreover, it includes Real-World Data (RWD): activity data collected by a wearable device, daily weather and air quality information and additional information collected by the mobile app Healthentia, that is used by the patients and provided by iSPRINT. The results of the predictive models are shown in the clinical dashboard that is currently being developed and are therefore used during the shared decision-making process.

Feature selection and extraction is an important step in the design of Machine Learning (ML) models. The goal is to first identify features that are irrelevant for the prediction task, or redundant, with a high correlation to another feature being one such case. In addition, other features might only be weak predictors, providing little benefit while requiring a large amount of effort to collect. Using a reduced set of features in ML models offers several benefits, including diminished model complexity and mitigated overfitting risks. Moreover, reducing the number of features that have to be collected in the ongoing cohort study (Task 5.6 "Observational cohort for RWD collection", M1-M39) would reduce the burden on the patients and the workload for the clinicians in the hospital. A reduced number of features also improves the runtime of the ML model training and reduces the risk of errors in the data. It makes the ML models more precise, robust, and easier to interpret. Interpretability is important in RE-SAMPLE in order to support the shared decision making that will be implemented in the pilot sites of RE-SAMPLE, giving feedback and suggestions to healthcare professionals and patients.

The key feature extraction is also a step in ensuring that RE-SAMPLE complies with the data minimisation principle of the General Data Protection Regulation (GDPR). The GDPR is a comprehensive European Union law enacted to safeguard individuals' personal data privacy and provide them with greater control over their personal information. The data minimisation principle is one of the key principles of the GDPR, it mandates that organisations should collect only the absolutely necessary features or data for a specific purpose, thus reducing the potential risks associated with excessive data handling. More information on this can be found in D4.3 "GDPR related and security/privacy requirements" and D4.7 "Measure for organisational, legal and technical security and privacy requirements".

At the time of writing of this deliverable, there are very few patients enrolled sufficiently long enough such that they have already completed a follow-up. Therefore, useful predictive models cannot be trained, and feature extraction methods cannot be applied on the prospective data collected in the cohort study. For these reasons, the ML feature extraction methods presented in this deliverable have been applied on the retrospective data and the analysis will be repeated for the prospective data once the respective datasets are available in a suitable format.

This deliverable is structured as follows. After the introductory section, the objectives are presented in section 2. The description of the retrospective and prospective datasets is given as a summary from their description in D3.1 "Training of the predictive and simulation models" are presented in section 3. In section 4, the methods to perform the feature extraction are presented. The results of applying these methods on the data are described in section 5, and lastly, the next steps are described in section 6.

2. Objectives

The predictive models developed in RE-SAMPLE and particularly the results generated by them are a part of the virtual companionship programme. To design them in a professional manner that is aligned with the privacy-by-design approach of RE-SAMPLE, feature extraction is a necessary and important step.

The primary goals of the key feature extraction described in this deliverable encompass multiple aspects. Firstly, it aims to minimise the number of features collected, complying with the data minimisation principle mandated by GDPR. In addition to data minimisation, another objective is to alleviate the burden on RE-SAMPLE patients. By streamlining the feature set, the aim is to simplify the data collection process, making it less time-consuming and demanding for patients.

Furthermore, this deliverable seeks to reduce the workload for healthcare professionals, including clinicians, nurses, and others involved in the collection, analysis and interpretation of the data. By minimising the number of features, the task of producing, reviewing and analysing the data becomes more manageable, potentially freeing up valuable time and resources for healthcare professionals. While pursuing these efficiency improvements, it is crucial to maintain the performance and robustness of the ML models. Ensuring that the models continue to provide accurate and reliable results of the utmost importance.

Additionally, the objective is to enhance the interpretability and comprehensibility of the ML models. By simplifying the complexity of the models, their outputs become more transparent and easier to understand for both healthcare professionals and end-users. Lastly, an integral aspect is to aggregate features with high frequency. By consolidating commonly occurring features, the aim is to improve the efficiency of data analysis and potentially uncover valuable patterns or insights within the dataset.

Overall, this deliverable encompasses a comprehensive set of objectives, ranging from privacy compliance and patient convenience, to reducing the workload of healthcare professionals, maintaining model performance, improving interpretability, and optimising feature aggregation. Despite the of the fact that not enough prospective data are available, the objectives are fulfilled as much as possible at the time of writing the deliverable. The remaining objectives still to tackle in future work are described in section 6.

3. The datasets

There are two different base datasets used to train predictive ML models: a dataset that is made up of retrospective data provided by the pilot sites and a dataset produced by the ongoing RE-SAMPLE cohort study. From each base dataset, multiple training datasets are created that differ in number of follow-ups for prediction and target variable. Within the cohort study, patients are scheduled to have a follow-up visit every six months. A summary of the description of the datasets given in D3.1 "Training of the predictive and simulation models" is presented in this section. It is important to mention that currently, there are not enough patients enrolled in the cohort study long enough to apply common ML methods for feature extraction to the prospective data. Therefore, the feature extraction methods will be initially applied on the retrospective data, and the analysis on the prospective data will be conducted once enough patients are enrolled and have a suitable follow-up duration.

3.1 The retrospective dataset

The retrospective dataset in the RE-SAMPLE project contains a total of 2068 patients from the three pilot sites. There are 1138 patients from MST, 444 from GEM and 486 from TUK. In total, there are 256 features before pre-processing. The target variables in the dataset are the chance of survival over different time periods, number and occurrence of Chronic Obstructive Pulmonary Disease (COPD) exacerbations (moderate and/or severe) and different Quality of Life (QoL) scores. The main target variable is the presence of a COPD exacerbation within one year of follow-up. Since this feature is unbalanced within the three pilot sites, false relations to the target may be introduced if other features are also unbalanced. As an example, if a country has a higher occurrence of exacerbations while also having a population that is slightly taller than in the other countries, the ML models might assign a higher exacerbation risk to taller people in general, which would be false. For this reason, unbalanced features are removed. Moreover, if patients died or in the case of study patients dropped out of the study before the year of follow-up was completed, they are removed from the dataset for this target. There are many measurements of the Forced Expiratory Volume in 1 second (FEV_1) that need to be aggregated, resulting in several statistics for FEV_1 related features.

After this aggregation and dropping features with more than 50% missing values, as well as features like the *country* and *study* that should not be used as predictors, there are 33 predictors left, shown in Table 1 below.

Table 1: Names of the feature in the retrospective dataset and their description

Feature	Description
Age	Age of the patient
Height	Height of the patient
Weight	Weight of the patient
BMI	Body Mass Index (BMI) of the patient
Gender	Gender of the patient
Packyears	Years of active cigarette smoking multiplied by the packages smoked per day
FEV1_L_I	FEV_1 in litres at inclusion
FEV1_Per_I	FEV_1 value percentage of predicted at inclusion
FEV1_FVC_I	FEV_1 and Forced Vital Capacity (FVC) ratio at inclusion
GOLD	Global Initiative for Chronic Obstructive Lung Disease (GOLD) stage at baseline
GOLD_ABCD	GOLD ABCD status at baseline
BOD	BOD score – BODE score without 6 minutes walking test distance; considers BMI, dyspnoea and the FEV_1 value percentage of predicted
ADO	ADO score, considers age, dyspnoea, and airflow obstruction
Mod_AECOPD_Prev_Y	Number of moderate exacerbations in the previous year
Sev_AECOPD_Prev_Y	Number of severe exacerbations in the previous year
ICS	Inhaled corticosteroids use at inclusion
Pneu_vac	Pneumococcal vaccination status

Feature	Description
CHF	Presence of chronic heart failure
IHD	Presence of ischaemic heart disease
Diabetes	Presence of diabetes
mMRC	Modified medical research council questionnaire
EQ5D_I	Standardised measure for health related QoL, by EuroQol Group, measure in five dimensions at inclusion
Smoker_active	If the patient is an active smoker or not
FEV1_L_trend	Trend of the FEV_1 (in litres) values
FEV1_L_addit_max	Maximum of the follow-up measurements of the FEV_1 (in litres) values
FEV1_L_addit_min	Minimum of the follow-up measurements of the FEV_1 (in litres) values
FEV1_L_addit_mean	Mean of the follow-up measurements of the FEV_1 (in litres) values
FEV1_Per_addit_max	Maximum of the percentage of predicted of the FEV_1 follow-up measurements
FEV1_Per_addit_min	Minimum of the percentage of predicted of the FEV_1 follow-up measurements
FEV1_Per_addit_mean	Mean of the percentage of predicted of the FEV_1 follow-up measurements
FEV1_FVC_addit_max	Maximum of the FEV_1 FVC ratio of the follow-up measurements
FEV1_FVC_addit_min	Minimum of the FEV_1 FVC ratio of the follow-up measurements
FEV1_FVC_addit_mean	Mean of the FEV_1 FVC ratio of the follow-up measurements

The procedure for imputing the missing values is described in D3.1 “Training of the predictive and simulation models”.

3.2 The prospective dataset

Currently, due to delays in the recruitment process, there are 160 patients enrolled in the RE-SAMPLE observational cohort study. This number still falls short of the project target to enrol 675 patients among the three pilot sites, the process and problems on this are described in D5.4 “Mid-term recruitment report” (M22). Of the patients that are enrolled, only 81 of them have been included long enough such that models can be trained on them, because they must have completed at least one follow-up to have a target variable value recorded.

The RE-SAMPLE project uses an edge computing approach with edge nodes installed at the pilot sites and a central server that acts as an orchestrator in the training process. The full architecture is described in D2.6 “Architecture and technical specifications”. In this way the ML algorithms can benefit from the data available at all pilot sites without needing to centralize the data at a single central server. Currently, the edge nodes and in particular the connections to the respective HIS are in the process of being set-up.

The dataset contains data from the HIS, environmental data and data collected via the Healthentia app. The HIS data is collected during the regular follow-up visits every six months. There might be irregular emergency visits as well. The data collected are e.g., results of the six-minute walking test (6MWT), spirometry tests or blood test results. The environmental data contains weather and air quality information and is collected 4 times per day. The Healthentia data consists primarily of answers to custom and validated questionnaires by the patients about their health and activity data collected with a wearable device. In comparison to the hospital data, the Healthentia data can be collected multiple times a day, as is the case for the environmental data. Using the data as-is would lead to a very large and impractical dataset. Therefore, the variables that are collected very frequently should be aggregated in a way that preserves their usefulness. The procedure is described in more detail in D3.1 “Training of the predictive and simulation models”.

In ML, effective model training typically requires a large number of data samples when dealing with a substantial number of features (Köppen, 2000). Currently, there are very few samples available, making it challenging to train models effectively. As the data is collected from patients enrolled in a study, the number of samples will increase, albeit modestly. Thus, reducing the number of features should significantly

enhance the model training process given the limited dataset. In addition to this, it becomes easier to interpret a ML model when the number of features it is based on is reduced. Another reason to perform feature selection is an issue with correlated features. High correlation means that two features exhibit a strong statistical relationship or tendency to move together in a consistent manner. Positive correlation means that an increase in the first value implies a likely increase in the second value, while negative correlation means that they move in opposite directions, with an increase in the first value indicating a decrease in the second. An example for correlated features is the *weight* with *BMI* pair, where a high value in one typically indicates a high value in the other. If features are highly correlated, removing all but one of them can improve model performance and interpretability because they broadly provide the same information to a model. This issue affects both the retrospective and the prospective dataset. The workload for the clinicians to collect all these features is also high. So, it is beneficial for the end user as well to reduce the number of features. As an overview, all the feature subgroups and the number of features in each subgroup are listed below in Table 2. The questionnaires used to create a score, count as one feature as only their score is interesting to use. The number of medication features is quite high since the start and end date of 40 different types of medication are recorded. In total, there are 282 features collected.

Table 2: Feature subgroups in the prospective dataset and the number of features per subgroup

Feature subgroup	Number of features
Environmental data	16
Healthentia general info	11
Healthentia questionnaires	11
Healthentia questions	54
Garmin data	40
HIS general info	10
Spirometry	8
Hospitalisation	11
6MWT	29
Medication	80
Blood test	12
Total	282

4. Methods

This section describes the dataset examination methods that are applied to the RE-SAMPLE data in section 5. The goal of these methods is to help decide which features to keep and which to remove, resulting in a smaller dataset while also making sure that any features that are necessary to produce high-quality ML results are preserved. Once the methods are applied, the RE-SAMPLE datasets become a more compact version of themselves. At first, the terms *feature selection*, *feature extraction* and *feature engineering* are shortly explained. *Feature selection* refers to the task of selecting which features to include or exclude in a dataset intended for ML. During feature selection, the features are not transformed, aggregated or otherwise altered. During *feature extraction* on the other hand, multiple variables in the dataset are combined into one (new) feature or one feature is aggregated due to its frequency. *Feature engineering* requires domain knowledge and involves manually creating new features or transforming existing ones into a new feature. Generating the BMI out of weight and height from one patient is one example of feature engineering. All these tasks are performed before model training.

Another possibility to reduce the number of features is to look at the feature importance after training a predictive model and then to omit the features with low importance, as for example applied in (Khan, Madhav C, Negi, & Thaseen, 2020). This is an optional step during feature selection. Feature importance techniques are mainly used for debugging and understanding the models. Most features should be excluded based on the feature selection and extraction methods, but the feature importance can lead to a more hands-on decision.

It is important to carefully separate the tasks of just reducing the dimensionality of the feature space and interpreting the model, even if they can influence each other.

Different methods have been investigated for suitability for RE-SAMPLE. It is possible to automatically perform feature selection during the pre-processing pipeline together with the imputation of missing values, but an approach like this is more suitable for large datasets with hundreds or thousands of features. In RE-SAMPLE, we do not want to drop a feature without supervision. It could be that one specific feature is very important for clinicians, even if all patients until now have the same value. It would also be difficult to synchronise an automatic feature selection pipeline across the pilot sites for federated learning. So, unsupervised techniques are not applied.

In ML applications, it is best practice to:

- remove features that have almost only missing values,
- remove features with only one value,
- remove highly correlated features because it makes interpretability more difficult.

This is discussed in more detail in section 5, along with descriptions for its application and results.

In general, no feature should be dropped without the consent of all pilot sites. Even if the statistics for one feature led to the decision to drop it, maybe the clinicians would have a good reason to keep it.

In the following subsections, the common methods used in the ML domain that are still applicable are described as well as the alternative methods using expert knowledge.

4.1 Feature selection methods

Feature selection is a task that should be performed before model training. There are two different kinds of methods:

- Filter methods based on correlation or mutual information to internal model constraints, e.g.,
    - Information gain,
    - Fisher score,
    - Correlation matrix with heatmap,
- Wrapper methods that train the model with different subsets of features to select the best ones to optimise the performance, e.g.,

- Forward selection,
- Backward elimination.

Filter methods are model agnostic and usually not computationally expensive. They are based on the data's characteristics. It is a good first step that removes irrelevant features; constant and quasi constant features should be removed first. A threshold for a variance can be defined to decide which features should be dropped.

Using wrapper methods is computationally expensive because starting from the entire set of features, a subset is generated, and a predictive model trained on it. The model is evaluated, and another subset is generated. Repeating this several times, the best subset of features can be found.

Alternatively, some model types have embedded methods to select features while improving model performance using regularisation methods as additional constraints to the optimisation task. Examples are Lasso and ridge regression (Bonaccorso, 2017), but they are not discussed in more detail because the feature selection pipeline is standardised for all ML models.

4.1.1 Information gain

The feature selection method *information gain* is a filter method using the mathematical term for entropy, i.e., uncertainty, to evaluate how much information is gained for each feature regarding the prediction of the target variable. The mutual information between the target variable and one of the predictor features is estimated, originally described in (Kozachenko & Leonenko, 1987).

Features with little mutual information with the target could be dropped. More specifically, if two variables have high mutual information, knowing one would reduce the uncertainty about the other one. So, if one of our features does not reduce the uncertainty about our target variable, we can consider omitting it.

4.1.2 Fisher score

Another filter method is based on the widely used *Fisher score* (Gu, Li, & Han, 2012). It can only be applied for a classification task. The higher the Fisher score of a variable, the more important it is to predict the target variable, so features with a low Fisher score could be dropped.

The Fisher score F of one feature is defined as follows for a binary classification problem:

$$F = (\mu_1 - \mu_2)^2/(\sigma_1^2 + \sigma_2^2),$$

where μ_1 is the mean of the feature for all datapoints with a positive target and μ_2 the mean of the feature for the negative target, σ_1 and σ_2 are the standard deviations for the feature values for the positive and respectively negative class of the target.

4.1.3 Correlation matrix with heatmap

With correlation, the linear relationship between two or more variables can be measured. We are using the Pearson correlation coefficient. The most important things about correlation, regarding feature selection, are that features with high correlation with the target should be kept and predictors should not be highly correlated with each other because it affects the interpretability (Molnar, 2020). A heatmap of a correlation matrix is a method to easily get an impression if one of these two things occur. Most importantly, if in one column of the correlation matrix many values are high, then it is highly correlated with several features, which is not good for interpretability.

4.1.4 Forward selection

Forward selection is an iterative wrapper method that starts model training with one feature and increases the number of features used for training step-by-step. After each training step, the resulting model is evaluated and the performance is compared to the performance of the previous model trained on the smaller dataset. A feature is only selected for inclusion in the resulting dataset if it improves the model performance.

This method can therefore be computationally expensive. The implementation used in RE-SAMPLE is the SequentialFeatureSelector[1] from `scikit-learn` (Pedregosa, Varoquaux, Gramfort, & Michel, 2011).

4.1.5 Backward elimination

Backward elimination, a wrapper method and the reverse of forward selection, starts with all features and removes one in each step. Similarly, as with forward selection, the feature to be removed is the one whose exclusion diminishes model performance the least. The implementation used in RE-SAMPLE is the SequentialFeatureSelector[1] from `scikit-learn` (Pedregosa, Varoquaux, Gramfort, & Michel, 2011).

4.2 Feature importance methods

Feature importance should be studied after applying feature selection methods. It is usually used as part of model interpretation and explanation. But these methods can be used to validate the decisions, also on a context-related level.

4.2.1 Shapley additive explanations

One popular method for feature importance and ML explanations is to compute *Shapley additive explanations* (Molnar, 2020) (SHAP). In RE-SAMPLE the computed values are intended to help the clinicians and the patient to understand the predictions and therefore to evaluate the patient's behaviour and treatment. The intervention suggestions for the virtual coaching programme are planned to be based on feature importance and explanations that will be presented in D3.5 "Explainability of model predictions and simulations (M36).

4.2.2 Explainable Boosting Machines

An *Explainable Boosting Machine* (EBM) (Lou, Caruana, Gehrke, & Hooker, 2013) is a tree-based generalised additive model. The features are mainly modelled separately, with some limited interactions between pairs of features. Because the features are modelled separately, their impact on the target can be visualised as individual shape functions. EBMs are fully interpretable ML models, and they also provide feature importance values.

4.2.3 Permutation feature importance

The *permutation feature importance* (Molnar, 2020) is determined by shuffling the values of one feature and then compare the computed performance to the performance before shuffling. The value of the drop in performance is the importance of that feature.

4.3 Usage of expert knowledge

Apart from the commonly used methods described so far, an important step in the analysis is to include the expert knowledge of the members of the RE-SAMPLE consortium. A workshop about feature aggregation and extraction was done in the in-person meeting at GEM in March 2023 and online meetings were held afterwards. In the following results section, the collected expert knowledge taken into account.

[1] Forward selection and backwards elimination implementation: sklearn.feature_selection.SequentialFeatureSelector — scikit-learn 1.3.0 documentation

5. Results

This section first describes the results generated by applying the methods described in section 4 on the retrospective data. Afterwards, the features collected for the cohort study that could be included in the prospective dataset will be examined on a textual level without applying the ML methods for feature selection and extraction. These results will be extended and refined as for the retrospective data, once more data becomes available, this procedure is explained in section 6.

5.1 Results on the retrospective data

In this section, first filter methods are applied; the correlation heatmap described in section 4.1.3 is studied and the information gain from section 4.1.1 is calculated for the features left after dropping highly correlated features. Moreover, the Fisher score described in section 4.1.2 is calculated. Afterwards, wrapper methods, i.e., forward selection (section 4.1.4) and backwards elimination (section 4.1.5) are applied on the already reduced feature set. Different combinations of features are compared to determine the dataset that optimises performance of the EBM model and the logistic regression model with L2 regularisation and target *Any_AECOPD_FU_class*, the presence of any COPD exacerbation in one year of follow-up. These are the reference models and dataset that are most important after the analysis done in D3.1 "Training of the predictive and simulation models". For the target *EQ5D_12M*, the score of the EQ5D questionnaire after 12 months since the enrolment of the patient, a short analysis is done with the model ElasticNet, which was described in D3.1 "Training of the predictive and simulation models". Lastly, feature importance methods are applied to analyse the dataset, so a final feature set can be decided on.

A first pre-processing of the data dropped descriptive features that are unsuitable for training like the ZIP code and the country of the patients. Moreover, FEV_1 values are aggregated. This creates a feature set of 33 features, described in Table 1. As a first step, we create a correlation heatmap described in section 4.1.3 for these 33 features.

In the correlation heatmap shown in Figure 1, the actual variable names are omitted, instead the variables are numbered. The mapping between number and variable name can be found in Table 3 below.

Table 3: Mapping of variable numbers and names

Variable Number	Variable Name
0	*Gender*
1	*Age*
2	*Height*
3	*Weight*
4	*BMI*
5	*Packyears*
6	*FEV1_L_I*
7	*FEV1_Per_I*
8	*FEV1_FVC_I*
9	*GOLD*
10	*GOLD_ABCD*
11	*BOD*
12	*ADO*
13	*Mod_AECOPD_Prev_Y*
14	*Sev_AECOPD_Prev_Y*
15	*ICS*
16	*Pneu_vac*
17	*CHF*
18	*IHD*
19	*Diabetes*
20	*mMRC*
21	*EQ5D_I*

Variable Number	Variable Name
22	*Smoker_active*
23	*FEV1_L_trend*
24	*FEV1_L_addit_max*
25	*FEV1_L_addit_min*
26	*FEV1_L_addit_mean*
27	*FEV1_Per_addit_max*
28	*FEV1_Per_addit_min*
29	*FEV1_Per_addit_mean*
30	*FEV1_FVC_addit_max*
31	*FEV1_FVC_addit_min*
32	*FEV1_FVC_addit_mean*

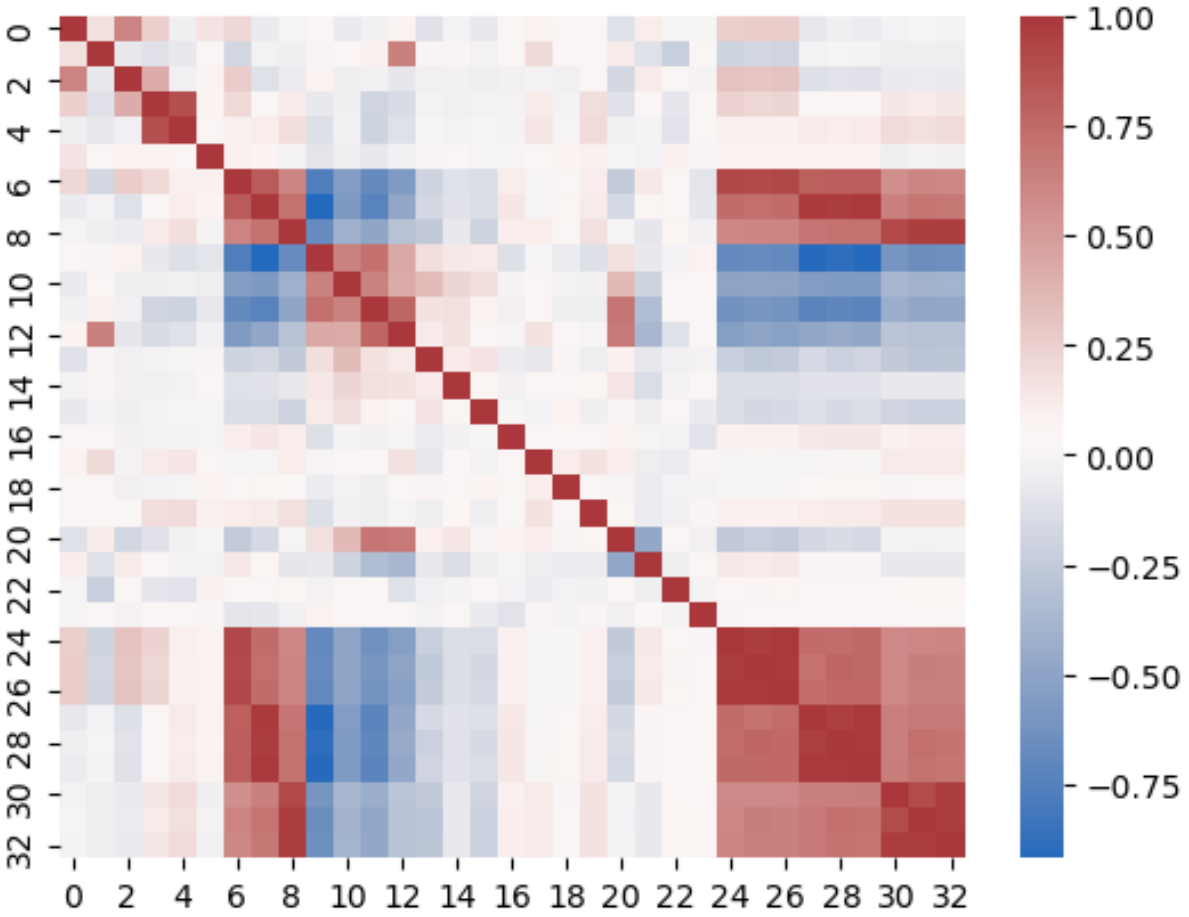

Figure 1: Correlation heatmap for the retrospective dataset

There are many highly correlated features, which is unfortunate due to the problems that arise from keeping highly correlated features, which are described in section 3.2. To have a closer look at these relations, the exact values for correlation of some features above an absolute value of 0.5 are shown in Table 4 below.

Table 4: Highly correlated features in the retrospective dataset

Variable 1	Variable 2	Correlation
FEV1_Per_I	*FEV1_L_I*	0.8228
FEV1_Per_I	*FEV1_FVC_I*	0.7095
FEV1_Per_I	*GOLD*	-0.9176
FEV1_Per_I	*GOLD_ABCD*	-0.5673
FEV1_Per_I	*BOD*	-0.7304
FEV1_Per_I	*ADO*	-0.4778

Variable 1	Variable 2	Correlation
FEV1_Per_I	*FEV1_L_addit_max*	0.7403
FEV1_Per_I	*FEV1_L_addit_min*	0.7303
FEV1_Per_I	*FEV1_L_addit_mean*	0.7417
FEV1_Per_I	*FEV1_Per_addit_max*	0.9837
FEV1_Per_I	*FEV1_Per_addit_min*	0.9732
FEV1_Per_I	*FEV1_Per_addit_mean*	0.9887
FEV1_Per_I	*FEV1_FVC_addit_max*	0.6396
FEV1_Per_I	*FEV1_FVC_addit_min*	0.6930
FEV1_Per_I	*FEV1_FVC_addit_mean*	0.6849
BMI	*Weight*	0.8854
Gender	*Height*	0.6195

As can be seen in the correlation heatmap in Figure 1, all FEV_1-related values are highly correlated with each other, like e.g., *FEV1_FVC_I* and *FEV1_Per_addit_min*, so it is not necessary to list all of them in Table 4. The bottom line is that from all the FEV_1-related values, only one should be kept, which is decided to be *FEV1_Per_I*. This value is preferable to the raw *FEV1_L_I* value because it already takes patient characteristics like the age, height and gender into account. It also does not depend on the presence of another value, like in the case of the FVC related values. Moreover, the *weight* is highly correlated with the *BMI*, so it is dropped. As well as the *height*, that is highly correlated with the *Gender*.

The number of comorbidities is correlated with the presence of some comorbidities. Since it is rather important to know if a specific comorbidity is present in some cases, for interpretability and the explanations, the number of comorbidities are dropped. For example, the prescription of specific medication is riskier if a heart-related comorbidity is present (Venkatesan, 2023).

Since the *BOD* and *ADO* score are using the *mMRC* to be calculated, the *mMRC* is kept and the scores are dropped. Additionally, *BOD* and *ADO* are more correlated with *FEV1_Per_I, mMRC* is not. Since *ADO* is slightly below the threshold of 0.5 that we picked, once the prospective dataset is available, it will be tested with the prospective dataset to drop *Age, mMRC* and *FEV1_Per_I* that are used to compute the *ADO* and to keep the *ADO* score.

The features left that are not overly correlated are the following:

- *Gender*,
- *Age*,
- *BMI*,
- *Packyears,*
- *FEV1_Per_I,*
- *Mod_AECOPD_Prev_Y,*
- *Sev_AECOPD_Prev_Y*,
- *ICS,*
- *Pneu_vac,*
- *CHF,*
- *IHD,*
- *Diabetes,*
- *EQ5D_I,*
- *Smoker_active,*
- *MMRC,*
- *FEV1_L_trend*.

The dataset containing only these 16 features (and the target) is now referred to as the reduced dataset.

The next step is to look at the other methods from section 4 to decide if the feature set should be reduced further than the list above. From now on, only these are considered. Next, the information gain method, explained in section 4.1.1, is applied with the results being listed in Table 5 below.

Table 5: Information gain of the uncorrelated features

Variable	Mutual information with *Any_AECOPD_FU_class*
Mod_AECOPD_Prev_Y	0.1430
EQ5D_I	0.1273
Age	0.0723
FEV1_L_trend	0.0689
Packyears	0.0517
ICS	0.0279
Sev_AECOPD_Prev_Y	0.0258
FEV1_Per_I	0.0207
Smoker_active	0.0203
CHF	0.0182
BMI	0.01542
Gender	0.0085
Pneu_vac	0.0019
IHD	0
Diabetes	0
mMRC	0

Because of the low mutual information with the *target Any_AECOPD_FU_class*, the features *IHD*, *diabetes* and *mMRC* can be considered to be dropped. It is not surprising that the feature representing the number of moderate exacerbations in the previous years has the highest value. The quality-of-life score *EQ5D_I* has also high mutual information with the target.

The computed Fisher score that is explained in section 4.1.2, ordered by decreasing values is shown in Table 6.

Table 6: Fisher score of the uncorrelated features

Variable	Fisher score
Mod_AECOPD_Prev_Y	0.4798
ICS	0.0897
FEV1_Per_I	0.0754
Sev_AECOPD_Prev_Y	0.0469
EQ5D_I	0.0243
Gender	0.0237
Pneu_vac	0.02067
CHF	0.0067
mMRC	0.0066
Smoker_active	0.0057
IHD	0.0045
Packyears	0.0031
FEV1_L_trend	0.0013
Diabetes	0.0010
BMI	0.0004
Age	0.00001

As was the case for the information gain method, *Mod_AECOPD_Prev_Y* is also the most important feature here, followed by *ICS* and then *FEV1_Per_I*. The features with the lowest values are *Age*, *BMI* and *Diabetes*. To have an idea which features result in the best performance, the EBM model is applied with the forward selection method explained in section 4.1.4, the results are in Table 7. The metrics considered are accuracy

and the F-beta score with β=2. The choice is explained in D3.1 "Training of the prediction and simulation models".

Table 7: Forward selection results using the EBM model

Number of variables	List of variables	Accuracy	F-beta
1	*Mod_AECOPD_Prev_Y*	0.7566	0.7280
2	*Mod_AECOPD_Prev_Y* *Sev_AECOPD_Prev_Y*	0.7678	0.7484
3	*Mod_AECOPD_Prev_Y,* *Sev_AECOPD_Prev_Y, CHF*	0.7640	0.7416
4	*Mod_AECOPD_Prev_Y,* *Sev_AECOPD_Prev_Y, CHF, Smoker_active*	0.7678	0.7484
5	*Mod_AECOPD_Prev_Y,* *Sev_AECOPD_Prev_Y, CHF,* *Smoker_active, Diabetes*	0.7640	0.7416
6	*Mod_AECOPD_Prev_Y,* *Sev_AECOPD_Prev_Y, CHF,* *Smoker_active, Diabetes, ICS*	0.7566	0.7050
7	*Mod_AECOPD_Prev_Y,* *Sev_AECOPD_Prev_Y, CHF,* *Smoker_active, Diabetes, ICS, BMI*	0.7640	0.7189
8	*Mod_AECOPD_Prev_Y,* *Sev_AECOPD_Prev_Y, CHF,* *Smoker_active, Diabetes, ICS, BMI, IHD*	0.7640	0.7246
9	*Mod_AECOPD_Prev_Y,* *Sev_AECOPD_Prev_Y, CHF,* *Smoker_active, Diabetes, ICS, BMI, IHD,* *Gender*	0.7528	0.7097
10	*Mod_AECOPD_Prev_Y,* *Sev_AECOPD_Prev_Y, CHF,* *Smoker_active, Diabetes, ICS, BMI, IHD,* *Gender, Age*	0.7790	0.7351
11	*Mod_AECOPD_Prev_Y,* *Sev_AECOPD_Prev_Y, CHF,* *Smoker_active, Diabetes, ICS, BMI, IHD,* *Gender, Age, FEV1_Per_I*	0.7640	0.7073
12	*Mod_AECOPD_Prev_Y,* *Sev_AECOPD_Prev_Y, CHF,* *Smoker_active, Diabetes, ICS, BMI, IHD,* *Gender, Age, FEV1_Per_I, FEV1_L_trend*	0.7640	0.7189
13	*Mod_AECOPD_Prev_Y,* *Sev_AECOPD_Prev_Y, CHF,* *Smoker_active, Diabetes, ICS, BMI, IHD,* *Gender, Age, FEV1_Per_I, FEV1_L_trend,* *Packyears*	0.7603	0.7177
14	*Mod_AECOPD_Prev_Y,* *Sev_AECOPD_Prev_Y, CHF,* *Smoker_active, Diabetes, ICS, BMI, IHD,* *Gender, Age, FEV1_Per_I, FEV1_L_trend,* *Packyears, EQ5D_I*	0.7491	0.7085
15	*Mod_AECOPD_Prev_Y,* *Sev_AECOPD_Prev_Y, CHF,* *Smoker_active, Diabetes, ICS, BMI, IHD,* *Gender, Age, FEV1_Per_I, FEV1_L_trend,* *Packyears, EQ5D_I, Pneu_vac*	0.7528	0.7154

Number of variables	List of variables	Accuracy	F-beta
16	*Mod_AECOPD_Prev_Y, Sev_AECOPD_Prev_Y, CHF, Smoker_active, Diabetes, ICS, BMI, IHD, Gender, Age, FEV1_Per_I, FEV1_L_trend, Packyears, EQ5D_I, Pneu_vac, mMRC*	0.7528	0.7097

As was confirmed by Table 5 and Table 6, the feature *Mod_AECOPD_Prev_Y* is the most important feature. The performance of the EBM model is best for 4 and 10 features, but the differences are small. The minimum value for accuracy is 0.7528 and the maximum value is 0.7790, for the F-beta the minimum value is 0.7050 and the maximum value is 0.7484.

In the following Table 8, the forward selection is applied to the same dataset but with the logistic regression model with L2 regularisation.

Table 8: Forward selection results using the logistic regression model with L2 regularisation

Number of variables	List of variables	Accuracy	F-beta
1	*Mod_AECOPD_Prev_Y*	0.7566	0.7280
2	*Mod_AECOPD_Prev_Y, Gender*	0.7566	0.7280
3	*Mod_AECOPD_Prev_Y Gender, Age*	0.7640	0.7360
4	*Mod_AECOPD_Prev_Y Gender, Age, ICS*	0.7715	0.7212
5	*Mod_AECOPD_Prev_Y Gender, Age, ICS, Smoker_active*	0.7715	0.7212
6	*Mod_AECOPD_Prev_Y Gender, Age, ICS, Smoker_active, mMRC*	0.7715	0.7212
7	*Mod_AECOPD_Prev_Y Gender, Age, ICS, Smoker_active, mMRC, IHD*	0.7715	0.7212
8	*Mod_AECOPD_Prev_Y Gender, Age, ICS, Smoker_active, mMRC, IHD, Pneu_vac*	0.7640	0.7131
9	*Mod_AECOPD_Prev_Y Gender, Age, ICS, Smoker_active, mMRC, IHD, Pneu_vac, BMI*	0.7566	0.7108
10	*Mod_AECOPD_Prev_Y Gender, Age, ICS, Smoker_active, mMRC, IHD, Pneu_vac, BMI, FEV1_L_trend*	0.7640	0.7131
11	*Mod_AECOPD_Prev_Y Gender, Age, ICS, Smoker_active, mMRC, IHD, Pneu_vac, BMI, FEV1_L_trend, Sev_AECOPD_Prev_Y*	0.7566	0.7108
12	*Mod_AECOPD_Prev_Y Gender, Age, ICS, Smoker_active, mMRC, IHD, Pneu_vac, BMI, FEV1_L_trend, Sev_AECOPD_Prev_Y, CHF*	0.7416	0.7120
13	*Mod_AECOPD_Prev_Y Gender, Age, ICS, Smoker_active, mMRC, IHD, Pneu_vac, BMI, FEV1_L_trend, Sev_AECOPD_Prev_Y, CHF, Diabetes*	0.7528	0.7097
14	*Mod_AECOPD_Prev_Y Gender, Age, ICS, Smoker_active, mMRC, IHD, Pneu_vac, BMI, FEV1_L_trend, Sev_AECOPD_Prev_Y, CHF, Diabetes, EQ5D_I*	0.7715	0.7327

Number of variables	List of variables	Accuracy	F-beta
15	*Mod_AECOPD_Prev_Y Gender, Age, ICS, Smoker_active, mMRC, IHD, Pneu_vac, BMI, FEV1_L_trend, Sev_AECOPD_Prev_Y, CHF, Diabetes, EQ5D_I, FEV1_Per_I*	0.7378	0.7051
16	*Mod_AECOPD_Prev_Y Gender, Age, ICS, Smoker_active, mMRC, IHD, Pneu_vac, BMI, FEV1_L_trend, Sev_AECOPD_Prev_Y, CHF, Diabetes, EQ5D_I, FEV1_Per_I, Packyears*	0.7828	0.7362

Again, the feature *Mod_AECOPD_Prev_Y* is the first to be selected and the model performs surprisingly well on only one feature. As for the EBM model, the performance varies very little by changing the number of used features.

In the following Table 9, the results of applying the backwards elimination method (explained in section 4.1.5) on the EBM model are shown.

Table 9: Backwards elimination results using the EBM model

Number of variables	List of variables	Accuracy	F-beta
1	*Mod_AECOPD_Prev_Y*	0.7566	0.7280
2	*Mod_AECOPD_Prev_Y, FEV1_Per_I*	0.7640	0.7246
3	*Mod_AECOPD_Prev_Y, FEV1_Per_I, FEV1_L_trend*	0.7603	0.6944
4	*Mod_AECOPD_Prev_Y, FEV1_Per_I, FEV1_L_trend, ICS*	0.7640	0.7189
5	*Mod_AECOPD_Prev_Y, FEV1_Per_I, FEV1_L_trend, ICS, Sev_AECOPD_Prev_Y*	0.7603	0.6944
6	*Mod_AECOPD_Prev_Y, FEV1_Per_I, FEV1_L_trend, ICS, Sev_AECOPD_Prev_Y, Age*	0.7640	0.7131
7	*Mod_AECOPD_Prev_Y, FEV1_Per_I, FEV1_L_trend, ICS, Sev_AECOPD_Prev_Y, Age, Packyears*	0.7640	0.7131
8	*Mod_AECOPD_Prev_Y, FEV1_Per_I, FEV1_L_trend, ICS, Sev_AECOPD_Prev_Y, Age, Packyears, Gender*	0.7603	0.7120
9	*Mod_AECOPD_Prev_Y, FEV1_Per_I, FEV1_L_trend, ICS, Sev_AECOPD_Prev_Y, Age, Packyears, Gender, mMRC*	0.7603	0.7062
10	*Mod_AECOPD_Prev_Y, FEV1_Per_I, FEV1_L_trend, ICS, Sev_AECOPD_Prev_Y, Age, Packyears, Gender, mMRC, Diabetes*	0.7566	0.7050
11	*Mod_AECOPD_Prev_Y, FEV1_Per_I, FEV1_L_trend, ICS, Sev_AECOPD_Prev_Y, Age, Packyears, Gender, mMRC, Diabetes, EQ5D_I*	0.7416	0.6947
12	*Mod_AECOPD_Prev_Y, FEV1_Per_I, FEV1_L_trend, ICS, Sev_AECOPD_Prev_Y, Age, Packyears, Gender, mMRC, Diabetes, EQ5D_I, Pneu_vac*	0.7491	0.7085
13	*Mod_AECOPD_Prev_Y, FEV1_Per_I, FEV1_L_trend, ICS, Sev_AECOPD_Prev_Y, Age, Packyears, Gender, mMRC, Diabetes, EQ5D_I, Pneu_vac, IHD*	0.7491	0.7085

Number of variables	List of variables	Accuracy	F-beta
14	*Mod_AECOPD_Prev_Y, FEV1_Per_I, FEV1_L_trend, ICS, Sev_AECOPD_Prev_Y, Age, Packyears, Gender, mMRC, Diabetes, EQ5D_I, Pneu_vac, IHD, CHF*	0.7491	0.7085
15	*Mod_AECOPD_Prev_Y, FEV1_Per_I, FEV1_L_trend, ICS, Sev_AECOPD_Prev_Y, Age, Packyears, Gender, mMRC, Diabetes, EQ5D_I, Pneu_vac, IHD, CHF, Smoker_active*	0.7453	0.6958
16	*Mod_AECOPD_Prev_Y, FEV1_Per_I, FEV1_L_trend, ICS, Sev_AECOPD_Prev_Y, Age, Packyears, Gender, mMRC, Diabetes, EQ5D_I, Pneu_vac, IHD, Smoker_active, CHF, BMI*	0.7528	0.7097

We observe a very similar behaviour compared to Table 7. In Table 10 below, the backwards elimination method is used with the logistic regression, L2 regularisation.

Table 10: Backwards elimination results using the logistic regression model with L2 regularisation

Number of variables	List of variables	Accuracy	F-beta
1	*Mod_AECOPD_Prev_Y*	0.7566	0.7280
2	*Mod_AECOPD_Prev_Y, EQ5D_I*	0.7566	0.7280
3	*Mod_AECOPD_Prev_Y, EQ5D_I, ICS*	0.7678	0.7258
4	*Mod_AECOPD_Prev_Y, EQ5D_I, ICS, Diabetes*	0.7678	0.7201
5	*Mod_AECOPD_Prev_Y, EQ5D_I, ICS, Diabetes, mMRC*	0.7715	0.7270
6	*Mod_AECOPD_Prev_Y, EQ5D_I, ICS, Diabetes, mMRC, Gender*	0.7753	0.7282
7	*Mod_AECOPD_Prev_Y, EQ5D_I, ICS, Diabetes, mMRC, Gender, CHF*	0.7753	0.7339
8	*Mod_AECOPD_Prev_Y, EQ5D_I, ICS, Diabetes, mMRC, Gender, CHF, BMI*	0.7790	0.7351
9	*Mod_AECOPD_Prev_Y, EQ5D_I, ICS, Diabetes, mMRC, Gender, CHF, BMI, FEV1_Per_I*	0.7865	0.7374
10	*Mod_AECOPD_Prev_Y, EQ5D_I, ICS, Diabetes, mMRC, Gender, CHF, BMI, FEV1_Per_I, Pneu_vac*	0.7865	0.7374
11	*Mod_AECOPD_Prev_Y, EQ5D_I, ICS, Diabetes, mMRC, Gender, CHF, BMI, FEV1_Per_I, Pneu_vac, Packyears*	0.7828	0.7305
12	*Mod_AECOPD_Prev_Y, EQ5D_I, ICS, Diabetes, mMRC, Gender, CHF, BMI, FEV1_Per_I, Pneu_vac, Packyears, Smoker_active*	0.7828	0.7305
13	*Mod_AECOPD_Prev_Y, EQ5D_I, ICS, Diabetes, mMRC, Gender, CHF, BMI, FEV1_Per_I, Pneu_vac, Packyears, Smoker_active, Sev_AECOPD_Prev_Y*	0.7828	0.7248
14	*Mod_AECOPD_Prev_Y, EQ5D_I, ICS, Diabetes, mMRC, Gender, CHF, BMI, FEV1_Per_I, Pneu_vac, Packyears, Smoker_active, Sev_AECOPD_Prev_Y, Age*	0.7828	0.7305

Number of variables	List of variables	Accuracy	F-beta
15	*Mod_AECOPD_Prev_Y, EQ5D_I, ICS, Diabetes, mMRC, Gender, CHF, BMI, FEV1_Per_I, Pneu_vac, Packyears, Smoker_active, Sev_AECOPD_Prev_Y, Age, IHD*	0.7828	0.7305
16	*Mod_AECOPD_Prev_Y, EQ5D_I, ICS, Diabetes, mMRC, Gender, CHF, BMI, FEV1_Per_I, Pneu_vac, Packyears, Smoker_active, Sev_AECOPD_Prev_Y, Age, IHD, FEV1_L_trend*	0.7828	0.7305

Again, a very similar behaviour to Table 8 can be observed. It seems like the only very important feature is *Mod_AECOPD_Prev_Y* but adding the other variables does not worsen the performance of the two reference models EBM and logistic regression with L2 regularisation much. Since 16 features is not a high number for a dataset with around 2000 samples, they can all be kept at this stage of the analysis. Of course, future results on the prospective dataset can lead to other decisions and the explainability task can also influence the choice of the final feature set; it will be documented in D3.5 "Explainability of model predictions and simulations" (M36).

Apart from the target variable of the occurrence of an exacerbation within 12 months of follow-up (*Any_AECOPD_FU_class*), QoL questionnaire scores are also predicted as previously described. Since the EQ5D is also available in the prospective dataset, the feature selection methods are tested for predicting *EQ5D_12M*. The same dataset is prepared in the same way as for the target *Any_AECOPD_FU_class*, but since only MST patients have this target available, the columns that are dropped because of more than 50% missing values are a bit different. Afterwards, the correlation heatmap is also studied as presented in section 5.1, i.e., Figure 1. The precise values for the correlations are not shown for this target, i.e., we omit showing detailed results as in Table 4. The result is the following list of features left after dropping the correlated features in the dataset for predicting the EQ5D score 12 months after inclusion:

- *Gender*,
- *Age*,
- *BMI*,
- *Packyears*,
- *Occ_stat*,
- *Civil_stat*,
- *FEV1_Per_I*,
- *Sev_AECOPD_Prev_Y*,
- *Flu_vac*,
- *CHF*,
- *IHD*,
- *Diabetes,*
- *Number_comorb* (the number of present comorbidities),
- *MMRC*,
- *6MWD_I* (the distance walked in the 6MWT at inclusion),
- *EQ5D_I*,
- *Mod_AECOPD_Prev_Y*,
- *CVD* (presence of cardiovascular disease),
- *Smoker_active*,
- *CRQ_I* (score of the chronic respiratory disease questionnaire at inclusion).

Only these features will be used to generate the following results.

Now, forward selection and backwards elimination are applied as for the retrospective dataset, the metric examined is the coefficient of determination R^2. The results on forward selection using the ElasticNet regression are shown in Table 11.

Table 11: Forward selection results using the ElasticNet model

Number of variables	List of variables	R^2
1	*EQ5D_I*	0.4625
2	*EQ5D_I, Mod_AECOPD_Prev_Y*	0.5107
	EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb	0.5518
4	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF*	0.5513
5	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD*	0.5629
6	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD*	0.5149
7	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Civil_stat*	0.5130
8	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Civil_stat, Sev_AECOPD_Prev_Y*	0.5539
9	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Civil_stat, Sev_AECOPD_Prev_Y, Flu_vac*	0.5465
10	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Civil_stat, Sev_AECOPD_Prev_Y, Flu_vac, CRQ_I*	0.5394
11	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Civil_stat, Sev_AECOPD_Prev_Y, Flu_vac, CRQ_I, Occ_stat*	0.5641
12	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Civil_stat, Sev_AECOPD_Prev_Y, Flu_vac, CRQ_I, Occ_stat, Diabetes*	0.5630
13	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Civil_stat, Sev_AECOPD_Prev_Y, Flu_vac, CRQ_I, Occ_stat, Diabetes, mMRC*	0.5407
14	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Civil_stat, Sev_AECOPD_Prev_Y, Flu_vac, CRQ_I, Occ_stat, Diabetes, mMRC, Packyears*	0.5380
15	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Civil_stat, Sev_AECOPD_Prev_Y, Flu_vac, CRQ_I, Occ_stat, Diabetes, mMRC, Packyears, FEV1_Per_I*	0.5375
16	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Civil_stat, Sev_AECOPD_Prev_Y, Flu_vac, CRQ_I, Occ_stat, Diabetes, mMRC, Packyears, FEV1_Per_I, Age*	0.5317
17	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Civil_stat, Sev_AECOPD_Prev_Y, Flu_vac, CRQ_I, Occ_stat, Diabetes, mMRC, Packyears, FEV1_Per_I, Age, Gender*	0.5276
18	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Civil_stat,*	0.5410

Number of variables	List of variables	R^2
	Sev_AECOPD_Prev_Y, Flu_vac, CRQ_I, Occ_stat, Diabetes, mMRC, Packyears, FEV1_Per_I, Age, Gender, 6MWD_I	
19	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Civil_stat, Sev_AECOPD_Prev_Y, Flu_vac, CRQ_I, Occ_stat, Diabetes, mMRC, Packyears, FEV1_Per_I, Age, Gender, 6MWD_I, Smoker_active*	0.5364
20	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Civil_stat, Sev_AECOPD_Prev_Y, Flu_vac, CRQ_I, Occ_stat, Diabetes, mMRC, Packyears, FEV1_Per_I, Age, Gender, 6MWD_I, Smoker_active, BMI*	0.5397

As observed in the previous tables for the exacerbation target showing the results of forward selection, Table 7 and Table 8, the performances do not vary much by changing the number of features. The first feature selected is *EQ5D_I* which is not surprising, the second is *Mod_AECOPD_Prev_Y* that is the most important feature predicting the target *Any_AECOPD_FU_class*. Overall, the values of R^2 are not very good, which might be due to the small number of samples (156 patients).

In Table 12, the results for backwards elimination are shown.

Table 12: Backwards elimination results using the ElasticNet model

Number of variables	List of variables	R^2
1	*EQ5D_I*	0.4625
2	*EQ5D_I, Mod_AECOPD_Prev_Y*	0.5107
3	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb*	0.5518
4	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF*	0.5513
5	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD*	0.5629
6	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD*	0.5149
7	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Packyears*	0.5229
8	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Packyears, FEV1_Per_I*	0.5243
9	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Packyears, FEV1_Per_I, Age*	0.5504
10	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Packyears, FEV1_Per_I, Age, Flu_vac*	0.5403
11	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Packyears, FEV1_Per_I, Age, Flu_vac, 6MWD_I*	0.5627
12	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Packyears, FEV1_Per_I, Age, Flu_vac, 6MWD_I, Gender*	0.5597
13	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Packyears,*	0.5665

Number of variables	List of variables	R^2
	FEV1_Per_I, Age, Flu_vac, 6MWD_I, Gender, Occ_stat	
14	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Packyears, FEV1_Per_I, Age, Flu_vac, 6MWD_I, Gender, Occ_stat, Diabetes*	0.5598
15	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Packyears, FEV1_Per_I, Age, Flu_vac, 6MWD_I, Gender, Occ_stat, Diabetes, Sev_AECOPD_Prev_Y*	0.5588
16	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Packyears, FEV1_Per_I, Age, Flu_vac, 6MWD_I, Gender, Occ_stat, Diabetes, Sev_AECOPD_Prev_Y, Smoker_active*	0.5373
17	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Packyears, FEV1_Per_I, Age, Flu_vac, 6MWD_I, Gender, Occ_stat, Diabetes, Sev_AECOPD_Prev_Y, Smoker_active, Civil_stat*	0.5364
18	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Packyears, FEV1_Per_I, Age, Flu_vac, 6MWD_I, Gender, Occ_stat, Diabetes, Sev_AECOPD_Prev_Y, Smoker_active, Civil_stat, mMRC*	0.5397
19	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Packyears, FEV1_Per_I, Age, Flu_vac, 6MWD_I, Gender, Occ_stat, Diabetes, Sev_AECOPD_Prev_Y, Smoker_active, Civil_stat, mMRC, CRQ_I*	0.5674
20	*EQ5D_I, Mod_AECOPD_Prev_Y, Number_comorb, CHF, IHD, CVD, Packyears, FEV1_Per_I, Age, Flu_vac, 6MWD_I, Gender, Occ_stat, Diabetes, Sev_AECOPD_Prev_Y, Smoker_active, Civil_stat, mMRC, CRQ_I, BMI*	0.5713

The results are similar to the ones of Table 11. Further analysis needs to be done on the prospective data to come to a decision, so, the feature set for the target *EQ5D_12M* does not need to be reduced further so far.

An additional step towards the explanations that will be worked out in D3.5 "Explainability of model predictions and simulations" (M36) is to check the feature importance of the trained models, the EBM model and the logistic regression model with L2 regularisation. The methods were explained in section 4.2.

A plot of the overall importance of the features is shown in Figure 2 below, all features listed in Table 3 are included, including the correlated features.

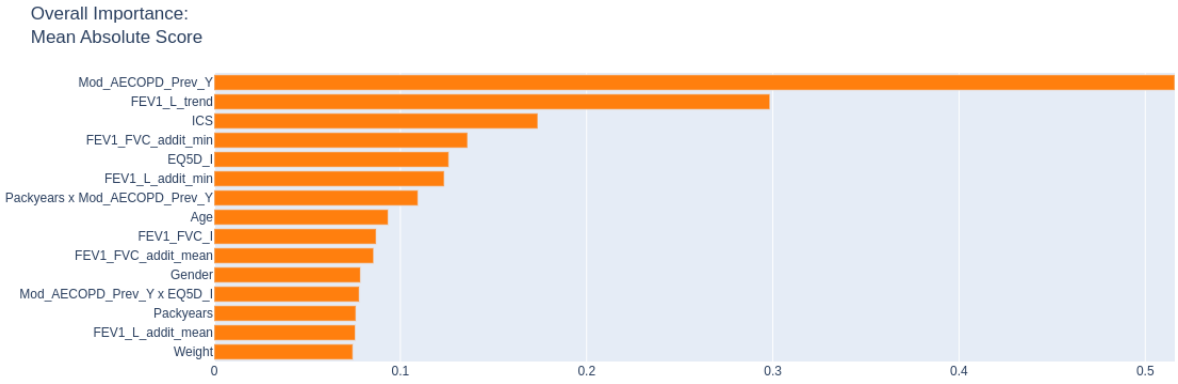

Figure 2: Feature importance computed by the EBM model for all retrospective features

The features with highest importance are *Mod_AECOPD_Prev_Y*, *FEV1_L_trend* and *ICS*. There are several other FEV_1 related features, which are all correlated. There are two interactions, *Packyears x Mod_AECOPD_Prev_Y* and *Mod_AECOPD_Prev_Y x EQ5D_I*, and we have seen in previous results that these features seemed to be important and improve the performance of the model.

Next, the SHAP feature importance values are computed for the features including the correlated ones, the results are shown in Figure 3. Since the ML task is a binary classification problem, the values for each feature are identical for the target having either 0 or 1 as values. We only look at the absolute value because we are not interested in the features having an increasing or decreasing effect on the exacerbation risk.

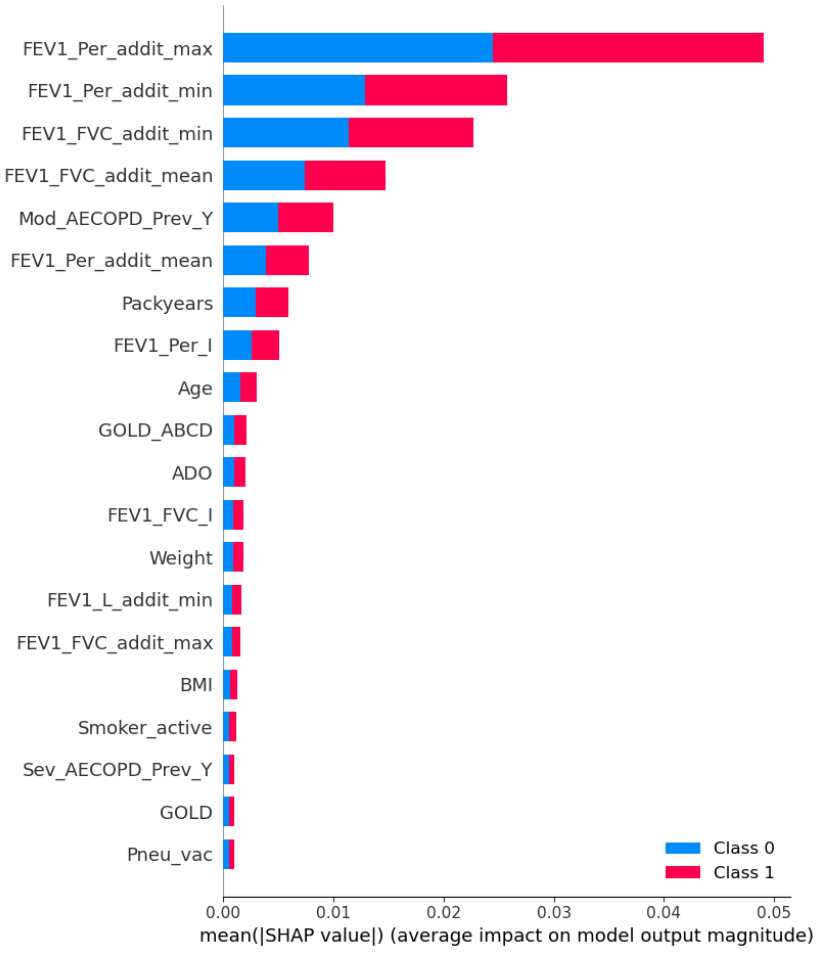

Figure 3: SHAP feature importance for the logistic regression model for all retrospective features

For the logistic regression model, the first four most important features are FEV_1-related again, then there is *Mod_AECOPD_Prev_Y*, another two FEV_1-related features and *Packyears*. For the EBM, the results are shown in Figure 4.

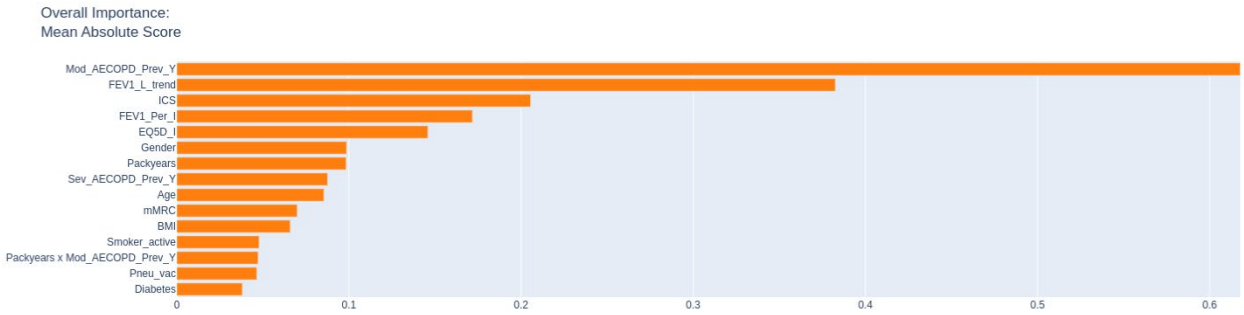

Figure 4: Feature importance computed by the EBM model for the reduced dataset

In Figure 4, the feature importance computed by the EBM model is shown as in Figure 2 but with the reduced dataset, i.e., without the correlated features. The results are as expected from the previous analysis.

In Figure 5, the SHAP feature importance is repeated on the reduced dataset.

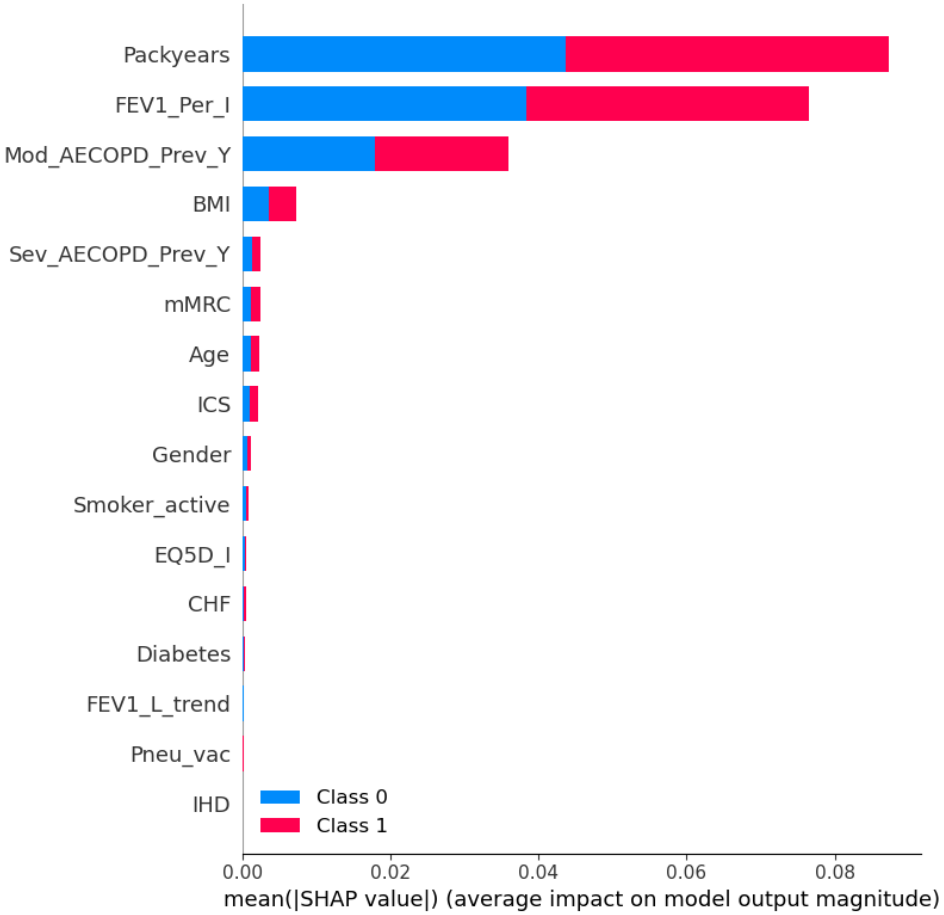

Figure 5: SHAP feature importance for the logistic regression model for the reduced dataset

Figure 5 shows a high importance for the feature *Packyears* – the number of daily cigarette packs smoked by a patient multiplied by the number of years that the patient smoked this amount – something that was not observed in the feature selection methods. Upon further inspection, it becomes apparent that the feature *Packyears* has an outlier section in the retrospective data between the values 62 and 87. For the total dataset, about 47% of patients had an exacerbation within one year of follow-up. In contrast, almost 70% of patients

with *Packyears* between 62 and 87 had an exacerbation in this timeframe, while only 37.5% of patients with *Packyears* of more than 87 had an exacerbation.

This non-linear association between *Packyears* and exacerbations does not seem to make medical sense, since a higher value for this feature either means that a patient has smoked more or longer. Therefore, it is to be expected that an increase in *Packyears* is correlated with an increase in exacerbation risk and prevalence. This behaviour makes the *Packyears* feature unsuitable as a predictor in the retrospective data. It will therefore be removed from any datasets enriched with retrospective data. Since the effect is unlikely to reoccur for prospective patients, the variable is kept in prospective datasets.

The SHAP plot from Figure 5 is created once again with *Packyears* removed from the dataset and the result is shown in Figure 6.

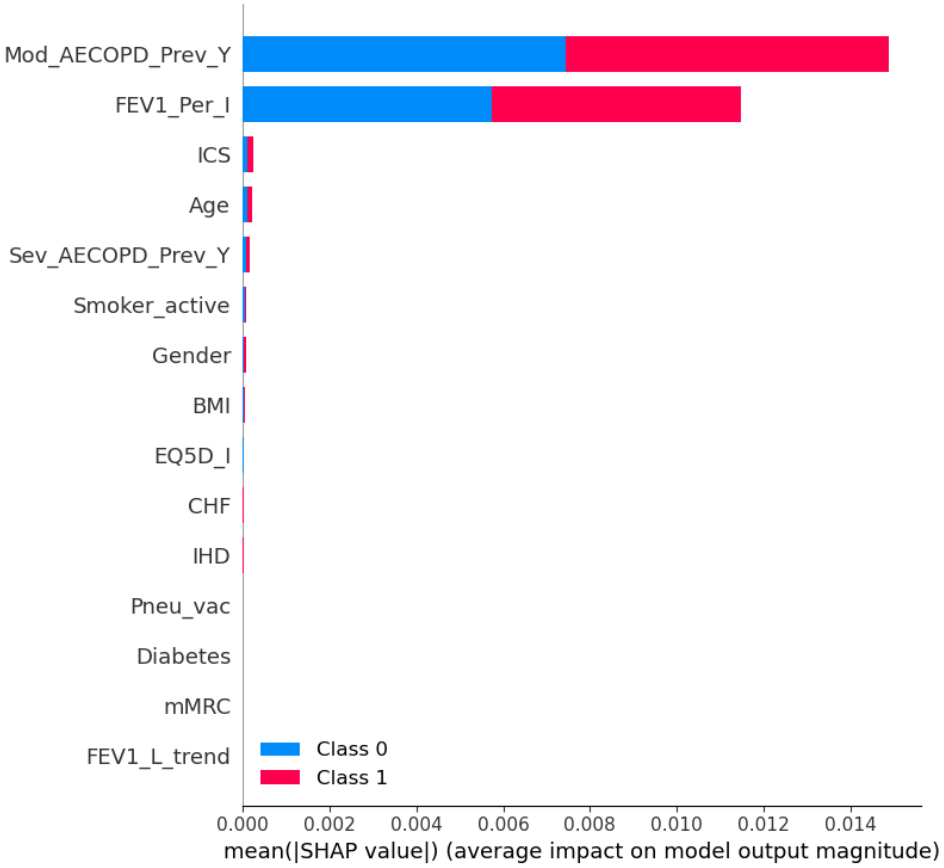

Figure 6: SHAP feature importance for the logistic regression model without *Packyears*

Without *Packyears*, the only features with an importance that can be mentioned are *Mod_AECOPD_Prev_Y* and *FEV1_Per_I*. Since we have seen that a dataset containing only very few features have almost the same performance as the whole reduced dataset this is not surprising.

Lastly, the permutation feature importance from section 4.2.3 is computed for the features in the reduced dataset, the results are in Table 13.

Table 13: Permutation feature importance for logistic regression model with L2 regularisation

Variable	Permutation feature importance
FEV1_Per_I	0.0371 ± 0.0117
Mod_AECOPD_Prev_Y	0.0262 ± 0.0070
Sev_AECOPD_Prev_Y	0.0030 ± 0.0020
BMI	0.0009 ± 0.0048
Smoker_active	0.00090 ± 0.0016

Variable	Permutation feature importance
Diabetes	0.0002 ± 0.0009
Pneu_vac	0.0001 ± 0.0012
Age	0 ± 0.0036
CHF	0 ± 0
IHD	0 ± 0
FEV1_L_trend	0 ± 0
EQ5D_I	-0.0009 ± 0.0021
ICS	-0.010 ± 0.0025
Gender	-0.0012 ± 0.0028
mMRC	-0.0024 ± 0.0020
Packyears	-0.0027 ± 0.0157

The features with highest (absolute value) of permutation feature importance are *FEV1_Per_I* and *Mod_AECOPD_Prev_Y*. Interestingly, *FEV1_L_trend* seems to have no influence, while the EBM model seemed to use it more as can be seen in Figure 4.

To sum up, with the methods that were applied on the retrospective dataset, it could be seen that apart from *Mod_AECOPD_Prev_Y*, there is no good predictor, and the models tend to overfit on some other features, i.e., *Packyears*. During the design of the predictive model on the prospective data, this must be carefully studied. But since including the other features also does not worsen the performance of the models, the dataset is not reduced drastically.

Thus, the result of this section is the following Table 14 whereas the starting point was Table 3. Compared to Table 1, the features that were dropped are explained in the very beginning of this section 5.1.

Table 14: Final features to keep and omit for the retrospective data

Features to keep	Features to omit
Age	*Number_comorb*
BMI	*Height*
FEV1_Per_I	*Weight*
Gender	*GOLD*
Mod_AECOPD_Prev_Y	*GOLD_ABCD*
Sev_AECOPD_Prev_Y	*ADO*
Diabetes	*BOD*
CHF	*FEV1_L_I*
IHD	*FEV1_FVC_I*
MMRC	*FEV1_L_addit_max*
EQ5D_I	*FEV1_L_addit_min*
FEV1_L_trend	*FEV1_L_addit_mean*
Smoker_active	*FEV1__Per_addit_max*
mMRC	*FEV1_Per_addit_min*
ICS	*FEV1_Per_addit_mean*
Pneu_vac	*FEV1_FVC_addit_max*
	FEV1_FVC_addit_min
	FEV1_FVC_addit_mean
	Packyears

5.2 Results on the prospective data

Due to the problem with the low number of patient enrolments and since not all edge nodes are fully connected to the HIS at the time of writing this deliverable, the methods described in section 4 cannot be applied yet to the prospective dataset. What can be done by knowing the structure of the prospective dataset is to study the features regarding their suitability for creating features that are combining several features,

this is done in section 5.2.1. Moreover, the aggregation of the longitudinal data can be determined, this is the topic of section 5.2.2. In section 5.2.3 the preliminary results are summarised, and a list of the reduced feature set is given.

First, there are some variables that can be dropped for the ML dataset from the general info of the Healthentia dataset:

- Dates like *Withdrawal date, inclusion date,*
- *Education,*
- *Country,*
- *Zip code,*
- *Weight,*
- *Height,*
- *COPD presence*.

The weight and height are dropped as discussed in section 5.1. The presence of COPD is the same value for every patient. The country and zip code are not used as predictors for the target variables. The dates are also not useful for predictions.

5.2.1 Description of similar features

In the complex dataset that is collected on the one hand via the Healthentia app and on the other hand via the hospitals, there are some features that contain similar or redundant information. To improve the ML models and facilitate their interpretation, we group them in Table 15 to either create a new feature combining their information or decide which of these should be used.

As already explained for the retrospective data, the *BMI* is kept and the *weight* that is used to calculate it, will be dropped. Additionally, the *Gender* is kept, and the *height* will be dropped, because it is used to calculate the *BMI* and strongly correlated to the *Gender*.

The features related to the 6MWT are described in the section 5.2.2 about aggregation of longitudinal variables.

Table 15: Description of similar features

Content	Similar features	Comments
Lifestyle	*Occupation*	The *living situation* is most interesting for the targets to predict. The other features can be neglected.
	Marital status	
	Social role	
	Living situation	
FEV_1-related	*FEV_1*	*Predicted percentage of FEV_1* is not biased by the *gender*, *Age* or *height* of the patient and thus a better predictor. The 4 features are additionally collected post short-acting bronchodilators for comparison. The clinical partners advised that we should not use the post short-acting bronchodilators spirometry feature because they might not be collected during every follow-up.
	Predicted Percentage FEV_1	
	FVC	
	FEV_1/FVC	
	Port short-acting bronchodilators spirometry - FEV_1	
	Port short-acting bronchodilators spirometry - Predicted Percentage FEV_1	
	Port short-acting bronchodilators spirometry - FVC	
	Port short-acting bronchodilators spirometry - FEV_1/FVC	
Number of moderate/severe exacerbations	*Number of exacerbations in the year before last year*	Usually, the last year is more meaningful than the year before

Content	Similar features	Comments
	Number of hospitalisations in the year before last year	that, maybe the sum of the two could be used to generate a new feature. A comparison in importance and performance will be done on the dataset once it is available.
	Number of exacerbations in the last year	
	Number of hospitalisations in the last year	
Information about white blood cells	*Thrombocytes*	These features describe information about the white blood cells of the patient. The clinical partners advised that *Eosinophils* are meaningful for COPD patients. They help to decide if prednisolone during an acute exacerbation of COPD should be prescribed. Moreover, *Neutrophils* to *Lymphocytes* ration is an important predictor for incidence of exacerbation of COPD. *Neutrophils* are key mediators of the inflammatory changes in the airways.
	Leukocytes	
	Eosinophils	
	Basophils	
	Neutrophils	
	Lymphocytes	
	Monocytes	
Medication to widen the airways	*SAMA*	Since these medications have very similar purposes, the information about the four could be combined to get a new feature.
	SABA	
	LAMA	
	LABA	
Corticosteroids	*Inhaled corticosteroids*	*Inhaled corticosteroids* are available in the retrospective dataset. Maybe we can combine the two to get a new feature. Both versions should be tested.
	Oral corticosteroids	
Treatment for heart-related diseases	*ACE-inhibitors*	A new feature could be generated to indicate that the patient received medication treating symptoms of their heart-related comorbidity.
	ARB	
	Beta blocker	
	Digoxin	
	Ivabradine	
Treatment for diabetes	*SGLT2-inhibitors*	A new feature could be generated to indicate that the patient received medication treating their comorbidity diabetes.
	Insulin	
	Metformin	
	Sulfonylureumderivates	
	Glinidines	
	GLP-1-analogs	
	DPP-4-inhibitors	
	Acarbose	

Content	Similar features	Comments
Treatment for mental disorders	*Benzodiazepines*	A new feature could be generated to indicate that the patient received medication treating their mental disorder.
	Selective serotonin reuptake inhibitors	
	Noradrenaline and dopamine reuptake inhibitors	
	Tricyclic antidepressants	
	Z-products	
	MAO inhibitors	
	Lithium	
	Quetiapine	
Treatment to lower cholesterol	*Statins*	A new feature could be generated to indicate that the patient received medication treating their high cholesterol.
	Ezetimibe	

This preliminary discussion facilitates future analysis of the prospective dataset and reduces the number of features already at this point.

5.2.2 Feature aggregation of longitudinal data

In this section, we take a closer look at the longitudinal data that is available in the prospective RE-SAMPLE dataset. There is data collected in the 6MWT, even if it is only performed every 6 months like the other data collected at the HIS, there is data for every minute for some of these features and they might be aggregated over the 6 minutes. Moreover, some data is collected even daily and this needs to be aggregated to be able to generate a training dataset with all the combined data. Below we provide several tables, grouped by their content, containing the variable name, the frequency with which it is collected, how it should be aggregated and if it should be dropped. Only if the variable is not dropped, the aggregation type is of interest. We based the decision on which features to drop and which to keep on discussions with several RE-SAMPLE partners, i.e., technical partners with experience working on longitudinal data and clinical partners and the expected availability of the data.

In general, we decided to aggregate over a period of 2 months and to calculate the median, trend an interquartile range (IQR) as statistics for each variable that is to be aggregated. The trend is the slope of a fitted linear regression.

What has to be taken into account apart from that is that in case of a moderate exacerbation, questionnaires might be asked once more, so we would use the most recent score. Lastly, we use the most recent value if there are, e.g., blood samples updated during hospitalisation.

The following tables, starting with Table 16 about activity, heart rate, sleep and exercise data are subsets of the data model described in D4.1 “Representation of Multi-Modal Data and Disease Progression Monitoring Features”.

Table 16: Aggregation of activity, heart, sleep and exercise data

Variable	Frequency	Aggregation type	Drop?
Did you have more symptoms than usual during the last 24 hours?	daily	Count consecutive days over two months	no
Daily Activity - Steps walked	daily	Median, IQR, trend over two months	no
Daily Activity - Distance travelled	daily		yes

Variable	Frequency	Aggregation type	Drop?
Daily Activity - Calories burned	daily	Median, IQR, trend over two months	no
Daily Activity - Floors climbed	daily		yes
Daily Activity - Lightly active minutes	daily		yes
Daily Activity - Moderately active minutes	daily		yes
Daily Activity - Highly active minutes	daily		yes
Heart - Min heart rate	daily		yes
Heart - Max heart rate	daily		yes
Heart - Out of range minutes	daily		yes
Heart - Fat burn minutes	daily		yes
Heart - Cardio minutes	daily		yes
Heart - Peak minutes	daily		yes
Sleep - Sleep start (hours relative to midnight)	daily		yes
Sleep - Sleep end (hours relative to midnight)	daily		yes
Sleep - REM minutes	daily		yes
Sleep - Light minutes	daily		yes
Sleep - Deep minutes	daily		yes
Sleep - Awake minutes	daily	Median, IQR, trend	no
Sleep - Total minutes	daily		no
Exercise - Start Time	daily		yes
Exercise - Duration	daily		yes
Exercise - Active Duration	daily		yes
Exercise - Calories	daily		yes
Exercise - Steps	daily		yes
Exercise - Distance	daily		yes
Exercise - Average Heart Rate	daily		yes
Exercise - Fat Burn Minutes	daily		yes
Exercise - Cardio Minutes	daily		yes
Exercise - Peak Minutes	daily		yes
Exercise - Sedentary Minutes	daily		yes
Exercise - Lightly Active Minutes	daily		yes
Exercise - Fairly Active Minutes	daily		yes
Exercise – Very Active Minutes	daily		yes

Most of the features above can be dropped, only the *steps walked* and *calories burned* are of interest for the shared-decision making task of the RE-SAMPLE project. The patients are not told how to use the exercise functionality of the Garmin device, so this functionality will likely not be used. The daily heart rate information and the sleep information is not considered as particularly important to predict the target variables we are focusing on. What is kept about the sleep is the *total minutes* and the *awake minutes* that are particularly important for patients with the comorbidity obstructive sleep apnoea syndrome.

The following Table 17 is about the 6MWT data collected every 6 months during the follow-up at the hospital.

Table 17: Aggregation of 6MWT data

Variable	Frequency	Aggregation type	Drop?
Six-minute walking test - Medication			yes
Six-minute walking test - Walking aid			yes
Six-minute walking test - Oxygen use			yes
Six-minute walking test - Oxygen used			yes
Six-minute walking test - Systolic pressure before test			yes
Six-minute walking test - Diastolic pressure before test			yes
Six-minute walking test - Walked distance			no
Six-minute walking test - Theoretical walked distance base on BMI and Age			yes
Six-minute walking test - If the patient has stopped			no
Six-minute walking test - Oxygen saturation at baseline			yes
Six-minute walking test - Oxygen saturation in min 1			yes
Six-minute walking test - Oxygen saturation in min 2			yes
Six-minute walking test - Oxygen saturation in min 3			yes
Six-minute walking test - Oxygen saturation in min 4			yes
Six-minute walking test - Oxygen saturation in min 5			yes

Variable	Frequency	Aggregation type	Drop?
Six-minute walking test - Oxygen saturation in min 6			yes
Six-minute walking test - Minimum Oxygen saturation during the test			yes
Six-minute walking test - Percentage of time that patient has SP02 below 85%			no
Six-minute walking test - Heart rate at baseline			no
Six-minute walking test - Heart rate in min 1		Trend over the six minutes	no
Six-minute walking test - Heart rate in min 2			no
Six-minute walking test - Heart rate in min 3			no
Six-minute walking test - Heart rate in min 4			no
Six-minute walking test - Heart rate in min 5			no
Six-minute walking test - Heart rate in min 6			no
Six-minute walking test - Borg score dyspnea before test			yes
Six-minute walking test - Borg score dyspnea after test			yes
Six-minute walking test - Borg score fatigue before test			yes
Six-minute walking test - Borg score fatigue after test			yes

After discussing with clinicians, the only features left are the *distance walked* and the *trend over six minutes for the heart rate* as well as the *heart rate at baseline*. Lastly, the *percentage of time that patient has SP02 below 85%* is kept because it indicates a dangerous situation for patients with COPD.

In Table 18, the aggregation of the environmental data that is collected 4 times daily is summarised.

Table 18: Aggregation of the environmental data

Variable	Frequency	Aggregation type	Drop?
Air Quality Index	4 times per day	Median, IQR, trend over 2 months	no
Carbon monoxide	4 times per day		yes
Nitrogen monoxide	4 times per day		yes
Nitrogen dioxide	4 times per day		yes

Variable	Frequency	Aggregation type	Drop?
Ozone	4 times per day		yes
Sulfur dioxide	4 times per day		yes
Ammonia	4 times per day		yes
PM2,5	4 times per day		yes
PM10	4 times per day		yes
Temperature	4 times per day	Count very hot/cold days (thresholds: below 5 degrees, above 25 degrees) in the last two months	no
Feels_like	4 times per day		yes
Temp_min	4 times per day		yes
Temp_max	4 times per day		yes
Pressure	4 times per day		yes
Humidity	4 times per day	Count very dry days in the last two months, threshold: 30%	no
Wind_speed	4 times per day		yes

Since bad air quality is dangerous for COPD patients (Li, et al., 2016), (Hansel, McCormack, & Kim, 2016), this information should be kept in the ML dataset. The air quality index is already combining several air quality features and therefore used and aggregated over two months computing the median, IQR and the trend. The other features are dropped. Regarding the weather information, very hot and very cold days are problematic for COPD patients (Hansel, McCormack, & Kim, 2016). Clinicians involved in RE-SAMPLE meetings about the feature extraction also named that dry weather can be dangerous. Thus, there are thresholds defined for temperature and humidity and days above and below that are counted over the past two months.

5.2.3 Final results' discussion and preliminary list

This section presents the preliminary list of features that will be utilised, even though they will be further studied in future analysis by applying the methods described in section 4. For every subgroup of the features, the main results are outlined.

The key points from the analysis of the retrospective data, in section 5.1, are very important for our future work on the prospective data due to the small number of patients enrolled in the cohort study. The retrospective data will be used to enhance the training dataset where possible, so the features available in the retrospective data are important. We have seen that the number of exacerbations in the previous year is an important predictor, as for other COPD exacerbation prediction models, e.g., (Adibi, et al., 2020). Using only this feature leads to a quite good performance of the ML models. However, adding the other features that are not highly correlated with each other to the training data, does not worsen performance. The only critical feature is *Packyears* which might have a false relation with the target variables in the retrospective dataset leading to an improvement of performance that is not representative. This behaviour was already observed with other features in the analysis performed in D3.1 "Training of the predictive and simulation

models". That is why we drop *Packyears* and keep all the other 15 uncorrelated retrospective features until further analysis is possible.

In the retrospective dataset, only approximately 17% of the patients were hospitalised. This would mean at the current number of patients, that only 20 of the RE-SAMPLE patients might be hospitalised. There are 9 additional features collected during hospitalisation that are mainly used for the clinician dashboard. They are therefore likely to have a high rate of missing values, so they will be dropped for ML usage. In addition to this, 20 patients are not enough to train a separate model only for patients that were hospitalised and the patients in the retrospective dataset do not have these 9 features available. In case it is decided to train a model on the patients that were hospitalised to e.g., predict mortality, the *presence of pneumonia* and *mechanical ventilation* are important predictors.

Some blood test variables are described in Table 15 have similar features and are summarised in a newly created feature. We have to test the feature importance and model performance of the others to decide which ones to keep but based on clinician's opinion, it is foreseen that only few of them might be of use as good predictors, for example *NT-proBNP, Eosinophils, Neutrophils* and *Lymphocytes*.

Most features collected during the 6MWT will be dropped as they are not good predictors; they are listed in Table 17. We keep the *distance walked*, the information *if the patient stopped* and *the percentage of time that patient has SP02 below 85%*.

The selection of the questionnaires was intensely discussed in WP5, so no further selection will be done at this point. If the information of single answers to questions or the questionnaire scores is improving the prediction quality or connected to the target variables will be tested during the analysis. From the single questions, the *living situation* is kept, the information about the age via *Birth date* and the comorbidity and risk factor information, which will be one-hot encoded and thus leads to 10 single features instead of one answered question. Moreover, the daily question "Did you have more symptoms than usual?" is aggregated over two months to a symptom score.

Even after grouping the medication as much as possible – see Table 15 about similar features – there are 16 features left about medication of which some are probably not good indicators for exacerbation risk prediction or QoL prediction. The number should be reduced during an analysis. It was mentioned by clinicians that a high number of different medications can cause dangerous side-effects, so it can be studied if the number of medications taken would be a good predictor. The prescription of antibiotics is particularly important if it is related to pneumonia, but often it is prescribed too inconsiderately by the doctors. So, it has to be studied if this information can be misleading or prescription of antibiotics is only mentioned if related to pneumonia. Moreover, what is very important about the medication is their adherence and the use of inhalers.

As mentioned in the discussion about aggregating the environmental data, Table 18, only *air quality index* is used as well as *temperature* and *humidity*.

Table 19 below summarises which features of the prospective data are kept as is, which features are omitted and features that are created anew from available features, aggregating or combining them.

Table 19: Features to keep and to omit for the prospective data

Features to keep	Features to omit	New features created
Diabetes	*Birth date*	*Age*
Anxiety	*Inclusion date*	*Diabetes treatment*
Depression	*Withdrawal date*	*Treatment for heart related diseases*
OSAS	*Country*	*Treatment for mental disorders*
IHD	*Zip code*	*Treatment to widen the airways*
Paroxysmal atrial fibrillation	*COPD presence*	*Corticosteroids*
CHF	*Civil status*	*Treatment to lower cholesterol*

Features to keep	Features to omit	New features created
Hypertension	*Marital status*	*Information about white blood cells*
Hypercholesterolemia	*Education level*	*6MWT – heart rate trend over the 6 minutes*
Kidney failure	*Occupational status*	*Number of very dry days below threshold over two months*
Smoking status	*Social role*	*Number of very hot days over threshold over two months*
Packyears	*Inhaled corticosteroids*	*Number of very cold days below threshold over two months*
Hemoglobin	*Oral corticosteroids*	*Air quality index: median over two months*
Hematocrit	*ACE-inhibitors*	*Air quality index: IQR over two months*
NT-proBNP	*ARB*	*Air quality index: trend over two months*
HbA1c	*Beta blocker*	*Sleep - Total time: median over two months*
Predicted percentage FEV1	*Digoxin*	*Sleep - Total time: IQR last two months*
Living situation	*Ivabradine*	*Sleep- Total time: trend over two months*
BMI	*SGLT2-inhibitors*	*Sleep - Awake time: median over two months*
Sex	*Insulin*	*Sleep - Awake time: IQR last two months*
MMSE	*Metformin*	*Sleep - Awake time: trend over two months*
MMRC	*Sulfonylureumderivates*	*Daily activity - steps walked: median over two months*
6MWT - walked distance	*Glinidines*	*Daily activity - steps walked: IQR over two months*
6MWT – heart rate at baseline	*GLP-1-analogs*	*Daily activity - steps walked: trend over two months*
6MWT – if the patient has stopped	*DPP-4-inhibitors*	*Daily activity – calories burned: median over two months*
6MWT - Percentage of time that patient has SP02 below 85%	*Acarbose*	*Daily activity – calories burned: IQR over two months*
Number of exacerbations in the year before last year	*Benzodiazepines*	*Daily activity – calories burned: trend over two months*
Number of hospitalisations in the year before last year	*Selective serotonin reuptake inhibitors*	*Did you have more symptoms than usual? – Symptom score*
Number of exacerbations in the last year	*Noradrenaline and dopamine reuptake inhibitors*	*Hospitalisation after x days (where x is a variable number of days depending on the target)*
Number of hospitalisations in the last year	*Tricyclic antidepressants*	
Number of hospitalisations in the year before last year	*Z-products*	
Antibiotics	*MAO inhibitors*	
PDE4-inhibitor	*Lithium*	
Diuretics	*Quetiapine*	
Digoxin	*Statins*	

Features to keep	Features to omit	New features created
Neprilysin-inhibitors	*Ezetimib*	
Nitrate	*LABA*	
Calcium antagonists	*LAMA*	
Antiplatelets	*SABA*	
Anticoagulants	*SAMA*	
Anti-epileptic drugs	*FEV1 in 1 second*	
RAND36 score	*FVC*	
EQ5D	*FEV1/FVC*	
FACIT-Fatigue SF	*Post short-acting bronchodilators spirometry - FEV1*	
Brief illness perception questionnaire	*Post short-acting bronchodilators spirometry - Predicted Percentage FEV1*	
Test of adherence to inhalers	*Post short-acting bronchodilators spirometry - FVC*	
Health literacy	*Post short-acting bronchodilators spirometry - FEV1 /FVC*	
International physical activity questionnaire	*Weight*	
Willingness to change	*Height*	
E-Health usability benchmarking instrument	*6MWT - medication*	
UX1month	*6MWT - walking aid*	
COPD assessment test	*6MWT - oxygen use*	
Hospital anxiety and depression scale	*6MWT - oxygen used*	
	6MWT - systolic pressure before test	
	6MWT - diastolic pressure before test	
	6MWT - Theoretical walked distance base on BMI and Age	
	6MWT - Oxygen saturation at baseline	
	6MWT - Oxygen saturation in min 1	
	6MWT - Oxygen saturation in min 2	
	6MWT - Oxygen saturation in min 3	
	6MWT - Oxygen saturation in min 4	
	6MWT - Oxygen saturation in min 5	
	6MWT - Oxygen saturation in min 6	
	6MWT - Minimum Oxygen saturation during the test	
	6MWT - Borg score dyspnea before test	

Features to keep	Features to omit	New features created
	6MWT - Borg score dyspnea after test	
	6MWT - Borg score fatigue before test	
	6MWT - Borg score fatigue after test	
	Hospitalisation - Admission date	
	Hospitalisation - Discharge date	
	Hospitalisation - Oxygen use	
	Hospitalisation - Mechanical ventilation	
	Hospitalisation - Presence of pneumonia	
	Hospitalisation - Blood pH level	
	Hospitalisation - Partial pressure of carbon dioxide	
	Hospitalisation - Bicarbonate	
	Hospitalisation - Base Excess	
	Hospitalisation - Partial pressure of oxygen	
	Hospitalisation - Oxygen saturation	
	Carbon monoxide	
	Nitrogen monoxide	
	Nitrogen dioxide	
	Ozone	
	Sulfur dioxide	
	Ammonia	
	PM2,5	
	PM10	
	Feels_like	
	Temp_min	
	Temp_max	
	Pressure	
	Wind_speed	
	Daily Activity - Distance travelled	
	Daily Activity - Floors climbed	
	Daily Activity - Lightly active minutes	
	Daily Activity - Moderately active minutes	
	Daily Activity - Highly active minutes	
	Heart - Min heart rate	
	Heart - Max heart rate	
	Heart - Out of range minutes	
	Heart - Fat burn minutes	
	Heart - Cardio minutes	
	Heart - Peak minutes	
	Sleep - Sleep start (hours relative to midnight)	

Features to keep	Features to omit	New features created
	Sleep - Sleep end (hours relative to midnight)	
	Sleep - REM minutes	
	Sleep - Light minutes	
	Sleep - Deep minutes	
	Exercise - Start Time	
	Exercise - Duration	
	Exercise - Active Duration	
	Exercise - Calories	
	Exercise - Steps	
	Exercise - Distance	
	Exercise - Average Heart Rate	
	Exercise - Fat Burn Minutes	
	Exercise - Cardio Minutes	
	Exercise - Peak Minutes	
	Exercise - Sedentary Minutes	
	Exercise - Lightly Active Minutes	
	Exercise - Fairly Active Minutes	
	Exercise – Very Active Minutes	
	Wellbeing	
	Body temperature	
	Provide feedback	
	Follow-up questions if more symptoms than usual	
	NYHA question in case of CHF	

There are 135 fields of features that are dropped, some of them containing information which is not completely lost but processed in new features e.g., *Insulin* is used to generate the feature *Diabetes treatment*. Moreover, not all follow-up questions that are asked are listed if there are more daily symptoms than usual, which are a maximum of 21 additional questions. This means that the number of features so far is 81 but that will be further reduced through analysis to be performed on the data. The starting point of the prospective data was Table 2, using Table 19 we can summarise the features used per subgroup in Table 20.

Table 20: Number of features available and used per subgroup

Feature subgroup	Number of features available	Number of features used
Environmental data	16	6
Healthentia general info	11	4
Healthentia questionnaires	11	10
Healthentia questions	54	13
Garmin data	40	12
HIS general info	10	8
Spirometry	8	1
Hospitalisation	11	1
6MWT	29	5
Medication	80	16
Blood test	12	5
Total	282	81

All in all, this preliminary analysis already reduced the number of features a lot, so the methods that were applied to the retrospective data can be applied in the same way as presented in section 5.1 to the pre-processed prospective dataset. The number of features can be reduced from 282 to 81 features with concrete ideas of how to reduce these further.

6. Conclusion and next steps

The main contribution of D3.3 “Key features extraction” is the identification of the methods adopted to select and extract the final features used for ML prediction. The methods are described and tested on the retrospective data. The main result of this analysis is to keep the features that are not overly correlated and do not overfit the model. The prospective dataset is as intensely as possible, the dataset could be reduced from 282 to 81 features and there is a clear plan on how to reduce this number further as the edge nodes become fully connected to the HIS. The final decision on the set of features is going to be presented in D3.5 “Explainability of model predictions and simulations” (M36).

The medical experts that are part of the RE-SAMPLE consortium have been involved in all decisions and will be involved in future work on the dataset. In this way, interpretability of the predictive ML models can be ensured, and the models can be robust despite the small number of patients enrolled in the cohort study.

The next step to be done regarding feature extraction is mainly to apply ML methods on prospective data, i.e., the features that are kept and the new ones created from Table 19 above. There were some concrete ideas mentioned about features that are likely to be dropped and about which ones are likely to be good predictors. The results of this will be documented in D3.5 “Explainability of model predictions and simulations” (M36).

References

Adibi, A., Sin, D. D., Safari, A., Johnson, K. M., Aaron, S. D., FitzGerald, J. M., & Sadatsafavi, M. (2020). The acute COPD exacerbation prediction tool (ACCEPT): a modelling study. *The Lancet Respiratory Medicine 8.10*, 1013-1021.

Bonaccorso, G. (2017). *Machine learning algorithms.* Packt Publishing Ltd.

Gu, Q., Li, Z., & Han, J. (2012). Generalized fisher score for feature selection. *arXiv preprint arXiv:1202.3725*.

Hansel, N. N., McCormack, M. C., & Kim, V. (2016). The effects of air pollution and temperature on COPD. *COPD: Journal of Chronic Obstructive Pulmonary Disease 13.3*, 372-379.

Khan, N., Madhav C, N., Negi, A., & Thaseen, I. (2020). Analysis on Improving the Performance of Machine Learning Models Using Feature Selection Technique. *Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) held in Vellore, India, December 6-8, 2018, Volume 2, Springer International Publishing*, 69-77.

Köppen, M. (2000). The curse of dimensionality. *5th online world conference on soft computing in industrial applications (WSC5)*, 4-8.

Kozachenko, L. F., & Leonenko, N. N. (1987). Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 9-16.

Li, J., Sun, S., Tang, R., Qiu, H., Huang, Q., Mason, T. G., & Tian, L. (2016). Major air pollutants and risk of COPD exacerbations: a systematic review and meta-analysis. *International journal of chronic obstructive pulmonary disease*, 3079-3091.

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.*

Molnar, C. (2020). *Interpretable machine learning.* Independently published (28 February 2022).

Pedregosa, F., Varoquaux, G., Gramfort, A., & Michel, V. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2825-2830.

Venkatesan, P. (2023). *GOLD COPD report: 2023 update.* The Lancet Respiratory Medicine, 11(1), 18.