Towards Advancing Cystic Fibrosis Care with Artificial Intelligence

Patient enrolment and sample/data collection

Entrants into this study were restricted to CF patients who possessed verifiable records confirming their CF diagnosis and who bore no history of undergoing lung transplantation. Ethical endorsement for this study was procured from the ethics committee of University Hospital of Cruces (Barakaldo, Spain) and the University of the Basque Country's ethics committee (Leioa, Spain). Preceding the incorporation into the study, patients and, if pertinent, their parents or guardians were comprehensively apprised of the study's nature, and their informed consent or assent, contingent on age, was duly procured.

Collection of pharyngeal swab and sputum specimens was undertaken, giving precedence to the latter where patient expectoration capacity allowed. The entire process of specimen procurement was entirely voluntary, with meticulous attention to the safeguarding of personal data and the utmost confidentiality of patient information. The initiative of sample procurement transpired over a duration of six months, characterised by a recurring temporal frequency of a singular sample per patient, with intervals of three months demarcating each collection episode.

Integral data encompassing variables such as age, gender, stature, CFTR mutation, antecedent pathologies including but not limited to pancreatitis, diabetes, and liver disease, the occurrence or non-occurrence of exacerbations, nutritional status, lung function indices (FEV1 and FVC or Forced Vital Capacity), as well as prior incidents of infections and antibiotic treatments have all contributed substantively to the scope of data acquisition.

Computational Methods

IFPTML model development

In this work, we use the IFPTML strategy to train/validate alternative models able to predict the outcome values v_{ij} for each i^{th} patient in different interviews ($I_j = I_0, I_1, I_2$). The model can classify the patients according to their personal subset of input continuous variables $\mathbf{c}_j$. See a detailed list of variables in Table 1. The model also considers the sub-set of personal discrete conditions/labels $\mathbf{s}_j$ of the patient. See Table 2.

As follows, we illustrate the linear form of the IFPTML models and their different cases. The strategy includes three different stages IF, PT, and AI/ML. The general linear form of the IFPTML model to be developed is the following:

Classic ML linear model (model of order zero).

In this case we use as input the original variables without re-grouping them in distances functions (r = 1, q = 1) and without scaling them with a moving average (α = 0). When we apply the previous constrains the equation of the model can be simplified as follow:

$$\textit{Hyper-parameters: } \alpha = 0,\ q = 1,\ r = 1,\ \textit{Parameters: } a_{c,s} = 0,\ a_{k,s} = a_k \in \boldsymbol{R}$$

$$f(v_{ij})_{calc} = a_0 + \sum_{k=1}^{kmax} a_k \cdot V_k \quad (1.1)$$

Table 1. Continuous variables partitions, partition names, variable names, levels, examples, and units.

Partition	Name	variables	n^0 levels	Examples	Units
$\mathbf{c}_I$	Medical and Personal data	v_1= Diagnosis Age (DA)	27 Ages	3,5; 41,9; 19; 6,5, etc.	(Yr)
		v_2=Age First Interview (AFI)	48 Interview	52,74; 7,22;48,9;25,73,etc.	(Yr)
		v_3=Initial Weight (Kg)	38 Weights	52,6; 49; 48,3;69;45,etc.	(Kg)
		v_4=Initial Eight (IE)	40 Eight	164;158,4;161;173,etc.	(m)
		v_5=Initial Body Mass Index (BMI)	46 BMI	19,56;19,53;18,63;23,05,etc.	(Kg/m)
		v_6= Sample-Inform Time lag (SITL)	32 Times	136;44;9;50,etc.	(dd)
$\mathbf{c}_{II}$	Treatment duration	v_7=months with modulator (MWM)	12-Month options	3,5;11,5;9,etc.	(mth)
		v_8= Days oral antibiotic 1 (DATBO01)	19 Days options	240; 249; 184; 87,etc.	(dd)
		v_9= Days oral antibiotic 2 (DATBO02)	24 Days options	42; 84; 260; 28,etc.	(dd)
		v_{10}= Days Intravenous Antibiotic (DATBiv)	7 Days options	15;21;41;42,etc.	(dd)
$\mathbf{c}_{III}$	Pharmacology drug cocktails	v_{11} =Inhalate antibiotic doses (ATBDinh01)	5 Options	2;1;300;75,etc.	(uu)
		v_{12} =Inhalate antibiotic doses (ATBDinh02)	4 Options	1;75;300,etc.	(uu)
		v_{13}=Inhalate Antibiotic doses (ATBDinh03)	4 Options	1;75;500,etc.	(uu)
		v_{14}= Oral antibiotic doses (ATBDoralc01)	6 Options	100;40;1000,etc.	(uu)
		v_{15}= Oral antibiotic doses (ATBDoralc02)	6 Options	30;100;1680;500;1000,etc.	(uu)
		v_{16}= Oral antibiotic doses (ATBDoral01)	13 Options	15;20;30;40,etc.	(uu)
		v_{17}= Oral antibiotic doses (ATBDoral02)	11 Options	10;15;20;40;600,etc.	(uu)
		v_{18}= Oral antibiotic doses (ATBDoral03)	6 Options	15;40;500;100,etc.	(uu)
		v_{19}= Oral antibiotic doses (ATBDoral04)	5 Options	15;500;600;1000,etc.	(uu)
		v_{20}= Intravenous Antibiotic doses (ATBDiv01)	4 Options	3,3;30;2000,etc.	(uu)
		v_{21}= Intravenus Antibiotic doses (ATBDiv02)	7 Options	2;30;500;2000;22500,etc.	(uu)
		v_{22}= Intravenus Antibiotic doses(ATBDiv03)	4 Options	1;2500;22500,etc.	(uu)
		v_{23}= Intravenus Antibiotic doses (ATDiv04)	2 Options	0.2	(uu)
$\mathbf{c}_{IV}$	Pharmacology drug regime	v_{24}=inhalate (inh01)	3 Options	0,12,8	(uu)
		v_{25}=inhalate (inh02)	3 Options	0,12,8	(uu)
		v_{26}=inhalate (inh03)	3 Options	0,12,8	(uu)
		v_{27}=continuous (c01)	3 Options	0,12,8	(uu)
		v_{28}=continuous (c02)	3 Options	0,12,8	(uu)
		v_{29}=oral (oral01)	4 Options	0,8,12,24	(uu)
		v_{30}=oral (oral02)	4 Options	0,8,12,24	(uu)
		v_{31}=oral (oral03)	4 Options	0,8,12,24	(uu)
		v_{32}=oral (oral04)	3 Options	0,8,24	(uu)
		v_{33}=intravenous (iv01)	3 Options	0,8,24	(uu)
		v_{34}=intravenous (iv02)	3 Options	0,8,12	(uu)
		v_{35}=intravenous (iv03)	3 Options	0,8,6	(uu)
		v_{36}=intravenous (iv04)	2 Options	0,8	(uu)

Table 2. Discrete variables partitions, partition names, variable names, number of levels, and examples.

Partition	Name	variables	n^0 levels	Examples
s_I	Genetic and age group	s_1 = Mutation 1	12 Mutations	[delta]I507, [delta]F508, Y1092X, etc.
		s_2 = Mutation 2	24 Mutations	R334W, [delta]I507, R1066C, G85E, etc.
		s_3 = Group name	2 Age group	adults, children
		s_4 = Gender	2 Gender	Masculine (M), Femenine(F)
		s_5 = Neonatal screening	2 Options	NO= 0 , YES = 1
s_{II}	Disease conditions	s_6 = Pansreatis Insufficiency. 2020	2 Options	NO= 0 , YES = 1
		s_7 = Diabetes	2 Options	NO= 0 without insulin, YES =1 with insulin
		s_8 = Hepatis Disease	3 Options	NO = 0, without cirrhosis HTP= 2, Hepatic disease without cirrhosis=4
		s_9 = Exaserbation	4 Options	No (0), Mild (1), Moderate (2), Severe (3), High severity (4)
s_{III}	Microbiology lab data	s_{10} = Sample	2 Sample types	Sputum, Pharyngeal Smear.
		s_{11} = Microbiological culture 1	15 Microbiological culture	*Staphylococcus aureus MR, Pseudomonas aerugiNOsa,etc.*
		s_{12} = Microbiological culture1 note	6 Types of diagnosis	Moderate, abundant, scarce, too scarce,etc.
		s_{13} = Microbiological culture 2	15 Microbiological culture	*Staphylococcus aureus, Aspergillus fumigatus, etc.*
		s_{14} = Microbiological culture2 note	5 Types of diagnosis	Moderate, abundant, scarce, too scarce,etc.
		s_{15} = Microbiological culture 3	10 Microbiological culture	*Pseudomonas fluorescens, Exophiala dermatitidis,etc.*
		s_{16} = Microbiological culture3 note	3 Microbiological culture	scarce, moderate,etc.
s_{IV}	Medical & Pharm general support	s_{17} = Hypertonic Saline Solutions (HSS)	2 Options	NO= 0 , YES = 1
		s_{18} = Deoxyribonuclease (Dnasa)	2 Options	NO= 0 , YES = 1
		s_{19} = Azitromisine	2 Options	NO= 0 , YES = 1
		s_{20} = Bronchodilator	2 Options	NO= 0 , YES = 1
		s_{21} = corticoid-oral	2 Options	NO= 0 , YES = 1
		s_{22} = corticoid-inh	2 Options	NO= 0 , YES = 1
		s_{23} = Modulator	5 Options	No= 0, Ivacaftor = 1, Lumacaftor-Ivacaftor =2, Etc.
		s_{24} = O_2	2 Options	NO= 0 , YES = 1
		s_{25} = Non-invasive ventilation (NIV)	3 Options	NO= 0 , YES = 1
		s_{26} = Proton pump inhibitor (PPI)	2 Options	NO= 0 , YES = 1
s_V	Inhalate antibiotic (ATBinh)	s_{27} = Inhalate antibiotic (ATBinh01)	5 types	No= 0 , Colistin=1, Ceftazidime=4, Tobramycin=2,Etc.
		s_{28} = Inhalate antibiotic (ATBinh02)	3 types	No= 0 , Tobramycin=2, Ceftazidime=4
		s_{29} = Inhalate antibiotic (ATBinh03)	3 types	No= 0, Aztreonam=3,Vancomycin=7,Ceftazidime=4
s_{VI}	Oral antibiotic (ATBoral)	s_{30} = Oral antibiotic (ATBoral)	6 Types of antibiotics	Amoxicillin-Clavulanic Acid=9, Trimethoprim=10, Etc.
		s_{31} = Continuous antibiotic (ATBc)	6 Types of antibiotics	Trimethoprim =10, Cefaclor =17, Cefuroxime=19
		s_{32} = Oral antibiotic (ATBoral01)	10 Types of antibiotics	Amoxicillin=8, Trimethoprim=10, Ciprofloxacin=11, Etc.
		s_{33} = Oral antibiotic (ATBoral02)	9 Types of antibiotics	Trimethoprim=10, Ciprofloxacin=11, Etc.
		s_{34} = Oral antibiotic (ATBoral03)	6 Types of antibiotics	Minocycline=12, Levofloxacin=13,Etc.
		s_{35} = Oral antibiotic (ATBoral04)	5 Types of antibiotics	Levofloxacin=13, Fusidic Acid=16, Cefaclor =17
s_{VII}	Intravenous Antibiotic (ATBiv)	s_{36} = Intravenous Antibiotic (ATBiv01)	5 Types of antibiotics	Ceftazidime=4, Meropenem =6, Etc.
		s_{37} = Intravenous Antibiotic (ATBiv02)	5 Types of antibiotics	Colistin=1, Ceftazidime =4, Etc.
		s_{38} = Intravenous Antibiotic (ATBiv03)	5 Types of antibiotics	Colistin=1, Imipenem=22
		s_{39} = Intravenous Antibiotic (ATBiv04)	5 Types of antibiotics	Aztreonam=3

IFPTML model development

IFPTML model of order 1.

In this case, we use as input the original variables without re-grouping them into distances functions (r = 1, q = 1), but we do scaling (α = 1). In this scaling we transformed the original variables into first-order PTOs with the form of deviations $\Delta V_k(\mathbf{s}_j)$. These deviations $\Delta V_k(\mathbf{s}_j) = V_k - \langle V_k(\mathbf{s}_j) \rangle$ are calculated with a multi-conditional moving average $\langle V_k(\mathbf{s}_j) \rangle$. This multi-conditional moving average takes different values for different groups of patients depending on the sub-set selected $\mathbf{s}_j = \mathbf{s}_I$, $\mathbf{s}_{II}$, or $\mathbf{s}_{III}$, *etc*. Please, see Table 2. The different possible sub-sets of patients used according to (genetics, genre, treatments, etc.) When we apply the previous constraints, the equation of the model can be simplified as follows:

$$\textit{Hyper-parameters: } \alpha = 1,\ q = 1,\ r = 1,\ \textit{Parameters: } a_{c,s} = 1,\ a_{k,s} \in \boldsymbol{R}$$

$$f(v_{ij})_{calc} = a_0 + a_1 \cdot f(v_{ij})_{ref} + \sum_{s=1}^{smax} \sum_{k=1}^{kmax} a_{k,s} \cdot \left[V_k - \left(< V_k(\boldsymbol{s_j}) > \right) \right] \quad (1.2.1)$$

$$f(v_{ij})_{calc} = a_0 + a_1 \cdot f(v_{ij})_{ref} + \sum_{s=1}^{smax} \sum_{k=1}^{kmax} a_{k,s} \cdot \Delta V_k(\boldsymbol{s_j}) \quad (1.2.2)$$

IFPTML model of order 2.

This is an IFPTML model with PTOs of second order. In this case we use as input the original variables but doing firstly scaling (α = 1) with a multi-conditional moving average $\Delta V_k(\mathbf{s}_j)$ and next re-grouping them (r = 1/2, q = 2) into Euclidean Distances functions $\| \Delta V_k(\mathbf{s}_j) \|$. This distance was calculated as the square root (r = ½) of the sum of the power two (q = 2) of each deviation. When we apply the previous constraints, the equation of the model can be simplified as follows:

$$\textit{Hyper-parameters } \alpha = 1,\ q = 2,\ r = ½,\ \textit{Parameters: } a_{k,s} = 1$$

$$f(v_{ij})_{calc} = a_0 + a_1 \cdot \alpha \cdot f(v_{ij})_{ref} + \sum_{s=1}^{smax} a_{c,s} \cdot \left\{ \sum_{k=1}^{kmax} \left[V_k - \alpha \cdot \left(< V_k(\boldsymbol{c_j}) > \right) \right]^2 \right\}^{1/2} \quad (1.3.1)$$

$$f(v_{ij})_{calc} = a_0 + a_1 \cdot \alpha \cdot f(v_{ij})_{ref} + \sum_{s=1}^{smax} a_{c,s} \cdot \left\{ \sum_{k=1}^{kmax} \Delta V_k(\boldsymbol{s_j})^2 \right\}^{1/2} \quad (1.3.2)$$

$$f(v_{ij})_{calc} = a_0 + a_1 \cdot \alpha \cdot f(v_{ij})_{ref} + \sum_{s=1}^{smax} a_{c,s} \cdot \left\| \Delta V_k(\boldsymbol{s_j}) \right\|_{\mathbf{c_j}} \quad (1.3.3)$$

In order to seek the different cases of the IFPTML models mentioned above we need to carry out the following IFPMTL data pre-processing stages. In these stages we are going to explain how to reorganize the data (data engineering) and calculate the values for the objective function $f(v_{ij})_{obs}$, reference function $f(v_{ij})_{ref}$, the moving averages $\langle V_k(\mathbf{s}_j) \rangle$, deviations $\Delta V_k(\mathbf{s}_j)$ (first order PTOs), and Euclidean distances $\|\Delta V_k(\mathbf{s}_j)\|_{\mathbf{cj}}$. In the following sections we explain in more detail the calculation of the objection function and input parameters of the model, see also the general workflow in Figure 1. In all the previous cases, Linear Discriminant Analysis (LDA) could be the algorithm of election to seek this kind of IFPTML linear classification.

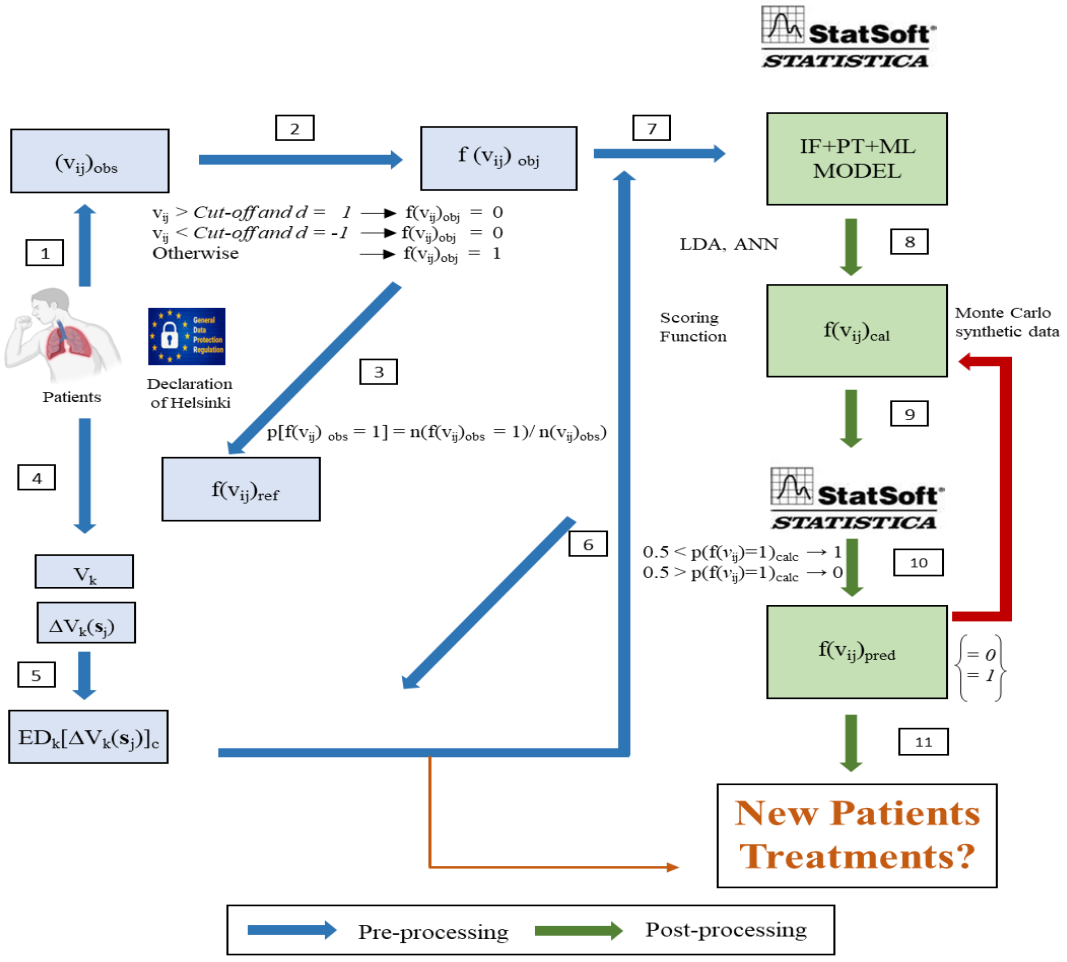

Figure 1. Workflow for variable pre and post-processing

IF pre-processing and data rearrangement

Firstly, we carried out an IF process for both patient clinical data and microbiological data. Patient clinical data (from clinical interviews) included personal data, pharmacological data, genetic data, etc. In contrast, microbiological data come from microbial cultures. Each row (case) corresponds to the i^{th} patient in the r^{th} visit/interview. These columns included four types of variables. The first class of variables v_{ij} includes the output values for each patient. The second class V_{ki} are input variables of patients (age, weight, drug doses,…,*etc*.). The third class s_j are input/output label variables of patients (gender, …,*etc*.).

The discrete labels s_j are of two types: the output variable label s_0 and input variable labels $s_{j>0}$. The output variable label get four different values s_0 =Forced Expiratory Volume in 1 second FEV_1(%), Forced Vital Capacities FVC(%), Hospitalization Days HD(dd), and days of intravenous antibiotics ATBiv days (dd). The different input discrete variables are organized in the form of a vector $\mathbf{s}_{j>0}$ = [s_1, s_2, ...s_{64}]. Some of these variables and their values are the following. The first input labeling variable is s_1 = Mutations 1, and get the values = ([delta]I507, [delta]F508, Y1092X, etc.) The second labeling variable is s_2 = Mutation 2, and get the values = (R334W, [delta]I507, R1066C, G85E, etc.)In total, we have $N_{patients}$ = 48, $N_{interviews}$ = 3, $N_{outputs}$ = 4 output continuous variables (v_{ij}), N_{vars} = 36 input continuous variables (V_{ki}), and N_{labels} = 39 output/input labeling variables (s_j). After the initial IF process, we carried out two additional IF and data rearrangement processes. The first was an IF horizontal process and the second and IF vertical process. The IF horizontal process applies to both input labels $\mathbf{s}_{j>0}$ and continuous input variables V_{ki}.

IF horizontal process for input labels

During the IF horizontal process for input labels, we carried out the following steps. Firstly, we carried out an IF and reorganization/regrouping of the 39 input labels $\mathbf{s}_{j>0}$ = [s_1, s_2, ...s_{39}] into 7 different subsets or partitions $\mathbf{s}_{j>0} = \mathbf{s}_I \& \mathbf{s}_{II} \& \mathbf{s}_{III} \ldots \& \mathbf{s}_{VII}$. Each one of these partitions is constructed as an IF or regrouping of different input labels, *i.e.*, $\mathbf{s}_I$ = [s_1, s_2, s_3, …, s_5], $\mathbf{s}_{II}$ = [s_6, s_7, s_8, s_9], *etc*. The IF of input labels $\mathbf{s}_j$>0 into partitions $\mathbf{s}_I$, $\mathbf{s}_{II}$, $\mathbf{s}_{III}$, … $\mathbf{s}_{XVII}$ was carried out according to expert criteria. In so doing, we grouped together labels of the same type or category of data, *i.e.*, $\mathbf{s}_I$ = Genetic and age group labels, $\mathbf{s}_{II}$ = Disease condition labels,$\mathbf{s}_{III}$ = Microbiology lab data labels, *etc*. Please, see the codes, names, and values of all the input labels and their partitions in Table 2.

IF horizontal process for continuous input

In this IF horizontal process, we performed two steps, see Figure 2. Firstly we carried out the IF of one continuous input variable V_{ki} at time *vs*. one partition of input labels $\mathbf{s}_I$, $\mathbf{s}_{II}$, …., or $\mathbf{s}_{III}$. This process leads to the calculation of the first order PTOs with the form of multi-condition deviations $\Delta V_{ki}(\mathbf{s}_j)$. The calculation of these PTO$_S$ is explained above. As we have N_{vars} = 36 input continuous variables (V_{ki}) and N_{labels} = 39 input labeling variables ($s_{j>0}$) grouped into 7 partitions $\mathbf{s}_j$ we could calculate theoretically a total of $N_{dev} = 36 \cdot 7 = 252$ operators $\Delta V_{ki}(\mathbf{s}_j)$. However, we used similar expert criteria to regroup all continuous input variables into partitions of continuous variables $\mathbf{c}_j$. Each one of these partitions $\mathbf{c}_j$ is constructed as an IF or regrouping of different input continuous variables V_{ki}, *i.e.*, $\mathbf{c}_I = [v_1, v_2, v_3, \ldots, v_6]$, $\mathbf{c}_{II} = [v_7, v_8, v_9]$, *etc*. The IF of input continuous variables into partitions $\mathbf{c}_I$, $\mathbf{c}_{II}$, $\mathbf{c}_{III}$, … $\mathbf{c}_{VII}$ was carried out according to similar expert criteria than those used for the input labels $\mathbf{s}_{j>0}$.In so doing, we grouped together labels of the same type or category of data, *i.e.*, $\mathbf{s}_I$ = Genetic and age group, $\mathbf{s}_{II}$ = Disease condition labels, $\mathbf{s}_{III}$ = Microbiology lab data labels, *etc*. After this, we were able to calculate the PTOs of second order in the form of Euclidean Distances $\|\Delta V_k(\mathbf{s}_j\|_{\mathbf{c}j}$. These parameters measure the distance between the values V_{ki} of the ith patient with respect to all patients we the same values of $\mathbf{s}_j$. Please, see the codes, names, and values of all the input variables and their partitions in Table 1 and Table 2).

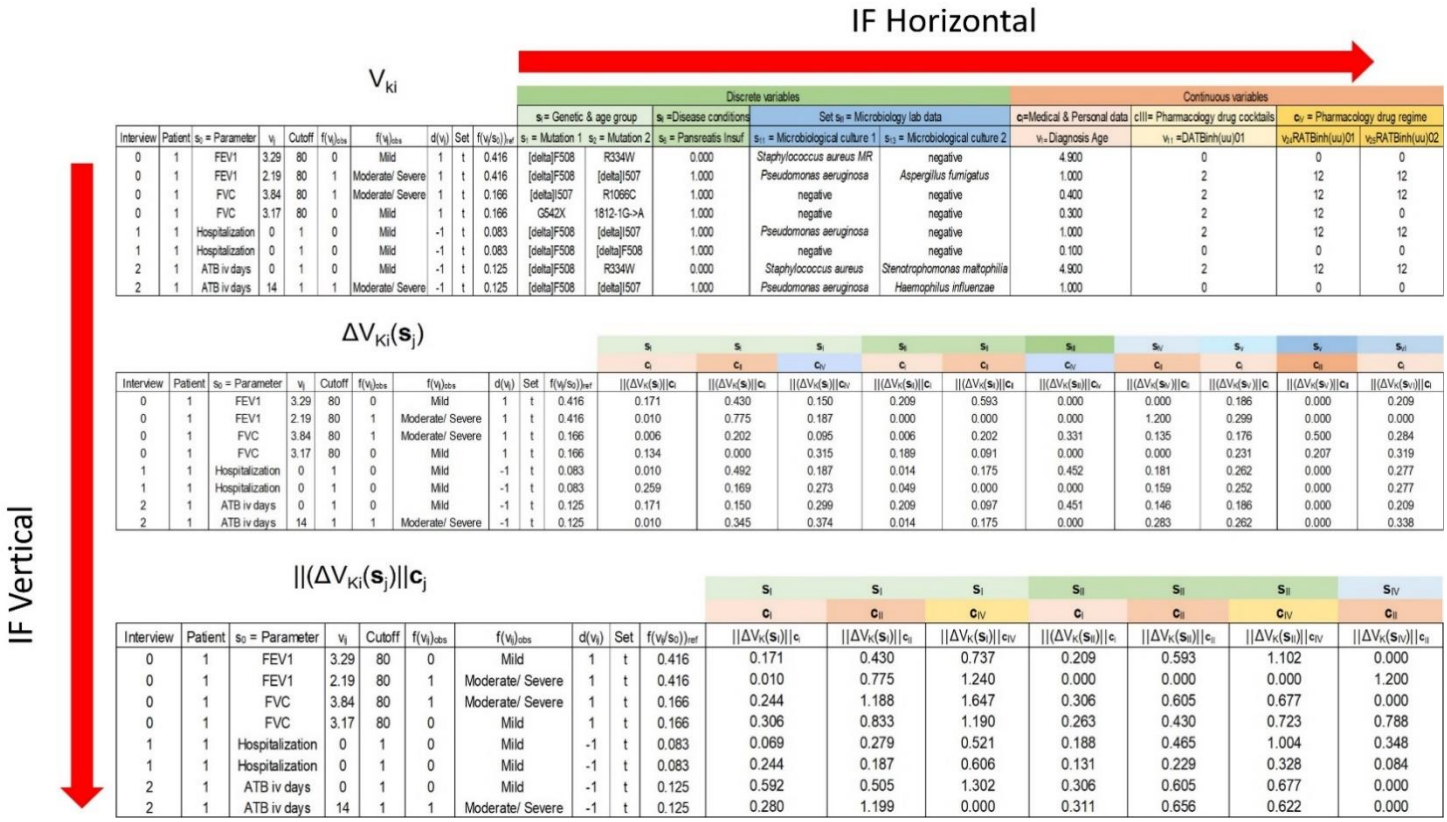

Figure 2. Information Fusion (IF) processes illustration

IF vertical process and outputs preprocessing

The vertical IF process is illustrated also in Figure 2. In the original dataset compiled with have a total of N_{output} = 4 outcome values by patient and interview. These output variables have the values v_{ij} and the labels or names s_0= FEV_1, FVC%, HD(dd) and ATBiv days. In the case we treat each one of these variables as a column we shall have 4 objective variables. As consequence, we will need to train 4 different models (one model for each objective variable) and/or use multi-objective techniques allowing more than one output variable. One alternative is to carry out an IF process grouping all the 4 output variables into only one by means of a vertical rearrangement all these values into on column. In so doing, we need to use an output labeling variable s_0 to indicate to what variable belongs each outcome value v_{ij}. This allow the IFPTML algorithm to process multiple output variables at the same time by training multi-output models with a single equation. Once we assembled all outcome values v_{ij} into one objective variable (column) we proceeded to the preprocessing of the values in order to put all of them into the same scale. In this work we transformed the original values v_{obsij} into the objective function with values $f(v_{ij})_{obj}$ = 1 => patient with moderate to severe disease symptoms or $f(v_{ij})_{obj}$ = 0 => patient with mild disease symptoms. The values of the objective function are calculated as follow:

$$f(v_{ij})_{obj} = IF\left(AND(v_{ij} > cutoff_j, d_j = 1), 1, IF\left(AND(v_{ij} < cutoff_j, d_j = -1), 1,0\right)\right) \quad (2)$$

The $cutoff_j$ values used for the different properties were $FEV_1(\%)_{cutoff}$ = $FVC(\%)_{cutoff}$ = 80 % and $HD(dd)_{cutoff}$ = $IAAD(dd)_{cutoff}$ = 1 day. The parameters desirability d_j indicates whether the values v_{ij} is desired to be higher than cutoff d_j = 1 or not d_j = -1. The reference function $f(v_{ij})_{ref}$ is the first variable that enters the model. It is equal to the expected probability p[$f(v_{ij})$ = 1] with which the parameter (v_{ij}) of a sample at the end of the process falls within the specified limits, $f(v_{ij})_{obs}$=1. It depends on the number of patients with moderate to severe disease $n(f(v_{ij})_{obj}$= 1) and on the total number of patients $n(f(v_{ij})_{obs})$ with the value of the parameter $n(v_{ij})_{obs}$ for the first interview I_0 (see equation 3). In Table 3 we show the values of the $cutoff_j$, d_j, number of patients n_j, function of reference, *etc*., for the first interview $I_{0.}$

$$f(v_{ij})_{ref} = p[f(v_{ij})_{obj} = 1] = n(f(v_{ij})_{obj} = 1)/\, n(v_{ij})_{obj}) \quad (3)$$

Table 3. Values of cutoff, desirability, overall statistics, and function of reference for outcomes

s_0 = Observed	General statistics						$f(v_{ij})_{ref}$ =
parameters	$Cutoff_j$	d_j	Avg.	Desvt	n_j	$n(f(v_{ij})=1)$	$p(f(v_{ij})=1/I_0)$
FEV1	80	1	79.7	24.62	48	20	0.417
FVC	80	1	94.1	14.93	48	8	0.167
Hospital days	1	-1	0.5	1.9	48	4	0.083
ATB iv days	1	-1	0.9	2.6	48	6	0.125

Output Variable and posterior probabilities calculation

Classification models the output variable $f(v_{ij})_{calc}$ is a real-valued scoring function related to the probability $p[f(v_{ij})_{pred} = 1]$ with which the experimental value of each property measured is predicted v_{ij} to fall within limits specified by the $cutoff_j$ values. In these cases, the output variable $f(v_{ij})_{calc}$ was discretized to obtain the predicted classification of each case ($f(v_{ij})_{pred} = 1$) or not ($f(v_{ij})_{pred} = 0$), see next section. This Boolean variable $f(v_{ij})_{calc}$ should be compared to the observed classification $f(v_{ij}/s_0)_{obs}$ for each sample in order to calculate the Specificity (Sp) and Sensitivity (Sn) (Hill and Lewicki, 2006).

Posterior probabilities calculation

When the IFPTML model is a classification model generated by LDA, ANN, *etc*., we can obtain the values of $f(v_{ij})_{calc}$ by introducing in the model the input values. However, $f(v_{ij})_{calc}$ is a real value variable (scoring function) and we need to undergo a discretization process in order to classify the samples as $f(v_{ij})_{pred} = 1$ or $f(v_{ij})_{pred} = 0$. In so doing, we used a sigmoid function to calculate the probabilities a posteriori $p[f(v_{ij})_{pred}=1]$ with which a sample is classified as $f(v_{ij})_{pred} = 1$. This sigmoid function uses as input a priori probabilities (π_0 and π_1). These prior probabilities are used in the Bayesian method to calculate the posterior probabilities. The equation of the sigmoid function of the posterior probabilities is the following (Hill and Lewicki, 2006).

$$p\left[f(V_{ij})_{pred} = 1\right] = \frac{1}{1 + \left(\frac{\pi_0}{\pi_1}\right) \cdot e^{-f(vij)_{calc}}} \quad (4)$$

Monte Carlo Synthetic Data (MCSD) generation

Monte Carlo (MC) like algorithms allow to generate a large number of new cases known as Synthetic Data (SD) presenting "small" perturbations in the values of the input variables (Harrison et al.,2022). In this work, we used the following procedure to create new SD cases. The method supposes that sufficiently small yet not trivial perturbations in the input values of an experimental case of reference will create a new plausible SD case. This new SD case is expected to have approximately the same values of the observed parameterv_{ij} or at least the same value of the objective function $f(v_{ij})_{obs}$ = 1 or 0. In order to create new SD cases with this method the perturbations can be done over the original input variables V_{ki}, their first order PTOs like $\Delta V_k(\mathbf{s}_j)$, or second order PTOs like Euclidean distance $\|\Delta V_k(\mathbf{s}_j)\|_{\mathbf{c}j}$. In this work we out the changes only in continuous input variables $\|\Delta V_k(\mathbf{s}_j)\|_{\mathbf{c}j}$ with a demonstrated linear and additive relationship with the output variable $f(v_{ij})_{calc}$. Firstly, the values of minimum $\|\Delta V_k(\mathbf{s}_j)\|_{\mathbf{c}jmax}$ and maximum $\|\Delta V_k(\mathbf{s}_j)\|_{\mathbf{c}jmin}$ for all the input variables were calculated. Next, a MC model based on the following system of equations was used to generate the new SD cases.

$$\left\|\Delta V_{ki}(\boldsymbol{s}_j)\right\|_{cjnew} = \left\|\Delta V_{ki}(\boldsymbol{s}_j)\right\|_{\boldsymbol{c}_j} + \left[(Rnd(0{,}100)/100)\cdot\left\|\Delta V_{ki}(\boldsymbol{s}_j)\right\|_{\boldsymbol{c}_{jmin}}\right] \quad (5)$$

$$\left\|\Delta V_{ki}(\boldsymbol{s}_j)\right\|_{\boldsymbol{c}_{jnew}} = if\left[\left\|\Delta V_{ki}(\boldsymbol{s}_j)\right\|_{\boldsymbol{c}_{jrnd}} > \left\|\Delta V_{ki}(\boldsymbol{s}_j)\right\|_{\boldsymbol{c}_{jmax}}; \left\|\Delta V_{ki}(\boldsymbol{s}_j)\right\|_{\boldsymbol{c}_jmax}; \left\|\Delta V_{ki}(\boldsymbol{s}_j)\right\|_{\boldsymbol{c}_{jrnd}}\right] (6)$$

Firstly, the first equation was used to generate new putative random values V_{kirnd} starting from the original minimum value V_{kmin}. Next, using the second equation ($V_{kmin} \leq V_{kinew} \leq V_{kimax}$), we obtained the new synthetic data value V_{kinew} after applying a boundary condition. This boundary condition keeps the synthetic values V_{kinew} within the range [V_{kimin}, V_{kimax}]. It means that the new synthetic data values are equal to $\|\Delta V_{ki}(\mathbf{s}_j)\|_{\mathbf{c}jnew} = \|\Delta V_{ki}(\mathbf{s}_j)\|_{\mathbf{c}j} + [(\text{Rnd}(0, 100)/100)\cdot\|\Delta V_{ki}(\mathbf{s}_j)\|_{\mathbf{c}jmin}]$ iff (if and only if) they are lower than $\|\Delta V_{ki}(\mathbf{s}_j)\|_{\mathbf{c}jmax}$; otherwise, they are equal to $\|\Delta V_k(\mathbf{s}_j)\|_{\mathbf{c}jmax}$. In the particular case when the function Rnd (0, 100) = 0 the original value$\|\Delta V_{ki}(\mathbf{s}_j)\|_{\mathbf{c}jnew} = \|\Delta V_k(\mathbf{s}_j)\|_{\mathbf{c}j}$ remains unaltered. The function Rnd (0, 100) is a generator of pseudo-random natural numbers (n = 0, 1, 2, … 100) based on Mersenne-Twister MC algorithm (MT19937) (Hongo et al., 2010; Ghersi et al., 2017).

Cumulative Validation CDM IFPTML-LDA model.

To identify how the model behaves when new patients are added cumulatively, an additional validation was performed. This validation considered 131 cases from the primary validation dataset to predict the outcome values v_{ij} for each i_{th} patient in different interviews ($I_j = I_0, I_1, I_2$). Then, 64 new data points (cases) were added to perform the cumulative validation. These cases come from 16 new CF patients that were added, obtaining 64 data points = [16 patients x 1 interview x 4 output variables]. It resulted in a total of 195 new cases for the cumulative validation study.

To use the IFPTML methodology, the same input continuous variables c_j and the set of discrete variables s_j used in the CDM IFPTML-LDA model were used, but in this case, they were derived from the new patient data. In the same way, the value of $f(v_{ij})_{obs}$ was determined for each output variable (FEV1 FVC%, hospitalization days (HD), and days of intravenous antibiotics during visits to the clinic (ATB-iv)), taking into account the same cutoffs established in the **IFPTML-OD** model. The function of reference $f(v_{ij})_{ref}$ used was the same as that obtained for the **IFPTML-OD** model, without any modification.

Additionally, for the multi-conditional moving average: $\Delta V_{ki}(\mathbf{s}_j/\mathbf{s}_{0j}) = V_{ki}(\mathbf{s}_j) - \langle V_k(\mathbf{s}_{0j})\rangle$, the V_k of the continuous variables from the new patients were used. However, *as per* model definition, the average values of reference $\langle V_k(\mathbf{s}_{0j})\rangle$ were taken from. These are the averages values for the patients in the interview of reference ($I_j = I_0$). In order to predict a new patient or same patient in different interview we used the $\langle V_k(\mathbf{s}_{0j})\rangle$ values from patients of reference with same $\mathbf{s}_j = \mathbf{s}_{0j}$ multi-conditional variables. In the case that non patient in the reference interview I_0 have the same conditions $\mathbf{s}_j \neq \mathbf{s}_{0j}$ we used the values of reference $\langle V_k(\mathbf{s}_{0j})\rangle$ values of the more similar conditions $\mathbf{s}_j \approx \mathbf{s}_{0j}$.

Finally, the same type of Euclidean distances $\|\Delta V_{ki}(\mathbf{s}_j/\mathbf{s}_{0j})\|$ were calculated as in the original model, obtaining all the necessary data to use the IFPTML methodology.

The machine learning (ML) process was carried out. Predictions for new data were made in Python using the scikit-learn library. The previously trained **IFPTML-OD** model was applied to the new set of patients, producing the predictions $f(v_{ij})$pred.