ORIGINAL ARTICLE

Journal Section

Developing a ML pipeline for asthma and COPD: the case of a Dutch primary care service

Stefano Mariani[1] | Maarten M.H. Lahr[2] | Esther Metting[2] | Eloisa Vargiu[3] | Franco Zambonelli[1]

[1]Department of Sciences and Methods for Engineering, University of Modena and Reggio Emilia – Reggio Emilia, Italy

[2]Health Technology Assessment, Department of Epidemiology, University of Groningen – University Medical Center Groningen, the Netherlands

[3]EURECAT Technology Centre, Digital Health Unit – Barcelona, Spain

Correspondence
Stefano Mariani, Department of Sciences and Methods for Engineering, University of Modena and Reggio Emilia – Reggio Emilia, Italy
Email: stefano.mariani@unimore.it

Funding information
This work has been supported by the CONNECARE (Personalised Connected Care for Complex Chronic Patients) project (EU H2020-RIA) under contract no. 689802.

A complex combination of clinical, demographic, and lifestyle parameters determines the correct diagnosis and the most effective treatment for asthma and COPD patients. Artificial Intelligence techniques help clinicians in devising the correct diagnosis and designing the most suitable clinical pathway accordingly, tailored to the specific patient conditions. In the case of machine learning (ML) approaches, availability of real-world patient clinical data to train and evaluate the ML pipeline deputed to assist clinicians in their daily practice is crucial. However, it is common practice to exploit either synthetic datasets or heavily pre-processed collections cleaning and merging different data sources. In this paper, we describe an automated ML pipeline designed for a real-world dataset including patients from a Dutch primary care service, and provide a performance comparison of different prediction models for *(i)* assessing various clinical parameters, *(ii)* designing interventions, and *(iii)* defining the diagnosis.

KEYWORDS

Abbreviations: ML, Machine Learning; COPD, Chronic Obstructive Pulmonary Disease; AI, Artificial Intelligence; DSS, Decision Support System; CCQ, Clinical COPD Questionnaire; ROC, Receiving Operating Characteristic; AUC, Area Under the Curve; ACQ, Asthma Control Questionnaire.

[*]Study conception and design: S.M., E.M., E.V.; Acquisition of data: E.M., M.L.; Analysis and interpretation of data: S.M., E.V., E.M.; Drafting of manuscript: S.M., E.V., E.M.; Critical revision: E.V., M.L., F.Z; Software: S.M.; Validation, E.V., E.M., M.L.; Visualization: S.M.

asthma, COPD, diagnosis, machine learning, prediction, primary care, treatment

1 | INTRODUCTION

Asthma and Chronic Obstructive Pulmonary Disease (COPD) are prevalent chronic respiratory diseases. In total over 339 million people worldwide suffer from asthma and over 65 million people have moderate to severe COPD, which is responsible for 5% of all global deaths [1, 2]. Both diseases are treated by a combination of triggers avoidance (like smoking or allergens), healthy lifestyle habits, and usage of inhaled medications. For the diagnosis of asthma and COPD, spirometry is the gold standard [3]. The aim of asthma and COPD treatment is to reduce the impact of the disease on daily life. The severity of asthma may fluctuate over time but is not progressively worsening; on the contrary, COPD is a progressive and potentially deadly disease. The possibility of exacerbations is present in both asthma and COPD patients, and can significantly impact the patients' health status. Therefore, prevention of exacerbations is one of the pivotal goals of the treatment. Exacerbations are treated in primary care with oral antibiotics, corticosteroids, or with oxygen in case of the most severe episodes.

Current clinical practice is mostly based on personal professional skills acquired by training, past experience, and intuition, whereas digital support is often limited to *rule-based* clinical management systems configured to deliver pre-defined decision criteria depending on pre-established clinical parameters of the patient [4]. However, *predictive modelling tools* based on statistical models or Artificial Intelligence (AI) techniques such as machine learning (ML) are gaining traction as a innovative way to interpret the growing amount of patient data available, aimed at achieving more patient-centred and customised care plans [5, 6, 7]. In particular, within respiratory research the biggest area of expansion is application of machine learning within imaging [8]. However, the diagnosis of respiratory conditions such as asthma and COPD relies on much more than image analysis; it involves taking a patient's history, a physical examination together with pulmonary function tests, and possibly imaging (X-rays, CT scans, bronchoscopy) [9].

Personalised and patient-based health risk assessment is a common example of such a predictive digital tool, and is increasingly used by clinicians to support individual decisions. For example, it is used to predict clinical episodes like exacerbations, or by personalising clinical pathways to patient conditions [10]. Indeed, *clinical pathways definition* is another example of clinical task where usage of digital tools driven by AI techniques is steadily growing.

However, in order to build such predictive modelling tools patient data must be available, possibly accurately reflecting the same real-world conditions that they will be operating on, for instance regarding imbalanced representation of clinical attributes, missing data, or noisy and badly formatted data. Instead, as emphasised recently in the ML community [11], datasets used for advertising the most accurate prediction models are too often woefully out of touch with reality. For instance because the best ML models are usually measured against large and curated datasets that lack the noise typical of real-world deployments.

In this paper we contribute to the development of predictive modelling in the context of diagnosis and treatment of asthma and COPD patients, with particular attention to the *real-world setting* where the tool would be deployed:

- first, we describe the ML pipeline set up for the exploitation of data of asthma and COPD patients coming from a Dutch primary care service, with particular attention to its development aimed at obtaining an automated pipeline easy to deploy and integrate in existing software stacks
- then, we compare performance of different prediction models for various clinical parameters, automated diagnosis, and suggestion of medical interventions

- finally, we summarise the "lessons learnt" during our research for the benefit of researchers and clinicians willing to perform similar tasks

To the best of our knowledge, as summarised in Section 2.3, this is the first research work addressing both asthma and COPD for the comparison of multiple prediction models across different prediction targets, while relying on primary care data.

In the remainder of the paper, Section 2 gives context knowledge to health-risk prediction for asthma and COPD, then Section 3 presents the proposed ML pipeline detailing the techniques exploited for data pre-processing, and model training and scoring, whereas evaluation of the resulting prediction models is in Section 4. Section 5 summarises the "lessons learnt" during our research, and finally Section 6 provides for final remarks.

2 | BACKGROUND AND RELATED WORKS

In this section we provide the reader with the clinical and technological background knowledge necessary to fully understand the context, motivations, and goals of our research. Accordingly, we provide a brief account on the current state of art regarding usage of digital tools for clinical pathways definition, there including literature showing evidence of the effectiveness of ML-based DSS for predicting various asthma and COPD outcomes (in Section 2.2). Finally we overview what are the recent efforts in providing support to clinicians specifically through ML and in the case of asthma and COPD patients, and what are the main differences with our work.

2.1 | Digital tools for clinical pathways

Research performed on electronic (or, computerised) clinical pathways can be roughly divided in three macro-areas.

Clinical pathways analysis. Digital tools are created either to build an *in silico* representation of clinical pathways, enabling simulations and "what-if" analysis, or to define machine-readable specifications of constraints on pathways to be enforced in clinical practice. Both efforts share the goal of monitoring compliance or identifying bottlenecks and devise out reasons for under-performance of current care plans. In [12], for instance, *process mining* is used to mine electronic health records and build clinical pathways *post-hoc*. Others [13] have instead suggested an ontology-based approach for the definition of pathways in a workflow style.

Clinical pathways synthesis. The focus is on automated definition of clinical pathways, or enforcement and execution thereof, with the goal of practically assisting clinicians in the definition and execution of care plans. Both [14] and [15], for instance, propose a semantic *rule-engine* configured with domain expert knowledge and a suitable rule-set to suggest adaptations to care plans depending on the specific patient conditions or unexpected events.

DSS within clinical pathways. Here, emphasis is given to DSS that support both the pathways definition and their execution, but in specific tasks—as in our case. The work in [16], for instance, proposes a recommender system to suggest to clinicians the next steps of disease management depending on the evolving patients conditions.

In the following sections, we overview the issue of predicting clinical variables for asthma/COPD patients and the AI techniques that are increasingly used to support such a task, with a focus on ML approaches.

2.2 | Health-risk prediction for asthma and COPD

Due to the large variation of characteristics in asthma and COPD patients, it can be extremely difficult even for experienced physicians to determine the most effective treatment for each patient.

In case of asthma, besides spirometry, the change in FEV_1 before and after administration of a large inhaled dosage of bronchodilator medications can confirm the diagnosis. However, absence of such a change does not necessarily mean absence of asthma, as many patients have normal lung function during diagnostic assessment. Hence, physicians often determine their diagnosis based on the clinical evaluation of symptoms, or via histamine provocation tests [17].

In case of COPD, the diagnosis is performed based on patient history including packyears and the presence of a fixed obstruction. An obstruction is determined using the Forced Expiratory Volume in 1 second (FEV_1) and Forced Vital Capacity (FVC). The severity of COPD is based on lung capacity measured with the proportion of predicted FEV_1. However, the burden that patients suffer in daily life activities is only partly determined by the level of obstruction. In addition, symptom questionnaires such as the Clinical COPD Questionnaire (CCQ) or the COPD Assessment Test (CAT) provide a more comprehensive assessment of the level of disease severity, which is divided in GOLD stages A (least severe), B, C, and D (most severe).

The optimal treatment is different for each patient and is based on disease severity, symptoms, patient characteristics, and individual risk factors. Personalised treatment has proven to be effective in improving patient outcomes, however, it is unlikely that physicians during their limited consultation times can accurately and comprehensively evaluate all of the many aspects that might affect worsening of the disease or triggering of an exacerbation. The human brain is simply not capable of recognising large amounts of longitudinal patterns and interactions between different predictors, as computers can [18]. For these reasons, AI can be promising especially when based on real-life large databases and when using machine learning techniques, for instance, by supporting healthcare providers in predicting the effect of specific treatment approaches [19]. In this way AI can be a valuable aid for healthcare professionals and reduces the risk of treatment side effects of patients [20].

The use of automatic clinical DSSs may improve the diagnosis and ongoing management of chronic diseases, which currently requires periodic visits to multiple health professionals, disease and medication monitoring, and modification of patient behaviour. The systematic literature review by Fathima et al. [21] brings evidence of effectiveness of clinical DSSs in the care of people with asthma. However, they did not found a clear evidence on using them for COPD.

Another systematic review of clinical DSSs was performed by Roshanov et al. [22], with the objective to determine if clinical DSSs improve the process of chronic care (in diagnosis, treatment, and monitoring) and associated patient outcomes. The authors identified 55 trials that measured and reported the impact of the clinical DSS on the process of care, and/or patient outcome. Out of the clinical DSSs that measured the impact on the process of care, 52% demonstrated a statistically significant improvement, and out of the trials that measured patient outcome 31% demonstrated benefits. Along the same line, Velickovski et al. [23] propose a clinical DSS in charge of delivering recommendations. Results show a high degree of accuracy to support COPD case-finding. Moreover, they demonstrate the integration into healthcare providers' workflow through the use of a modular design and service-oriented architecture that connects to existing health information systems already in use.

However, most of the DSSs considered in the aforementioned reviews do not focus on prediction of risk and suggestion of interventions after baseline assessment, as we do, and do not focus on interoperability and portability of the models embedded in the DSS, as we do. The ML pipeline we describe in Section 3 and evaluate in Section 4 is not a DSS *per se* but as a fully automated ML pipeline implemented in Python can be easily embedded into one. Indeed, such a pipeline could be easily served via a DSS such as the one described in [24].

2.3 | ML for asthma/COPD

The goal of best predicting clinical variables related to asthma/COPD with data-driven approaches is shared by many research works in state of art literature, and rightfully so, as the benefits that AI techniques may bring to the whole asthma/COPD management are widely recognised [25]. However, most of existing literature differ from our contribution in several aspects, such as data provenance and pre-processing, the approach adopted, and outcome of the research. For instance, many works we mention in the following perform heavy pre-processing of data to select the most likely predictors, or get data from specific clinical trials. Also, whereas they want to identify statistically meaningful predictors based on extensive statistical analysis (e.g. uni/multi-variate regression methods), we focus on performance of a *fully automated* ML pipeline autonomously building a slew of predictive models (also targeting suggestion of interventions) from unfiltered primary care data—that is, data fetched "as is" from a Dutch primary care centre, with no a-priori filtering or processing. Nevertheless, we here position with the most similar works to highlight strengths and differences.

Most of the related research works perform some form of *statistical analysis* on data collected at various scale (e.g. primary care vs. controlled trials) to evaluate the predictive value of specific attributes. For instance, in [26] the authors perform univariate analysis to identify likely predictors of COPD exacerbation events, then feed such predictors into a stepwise multivariable logistic regression model, and finally carry out sensitivity analysis with respect to asthma and smoking pack-years. Hence a first difference with our work is that they preselect a prediction model, whereas we compare a slew within an automated pipeline. They get data from a curated database and further select patients based on eligibility criteria meant to further remove confounding variables (e.g. of patients with others respiratory diseases, such as bronchiectasis). Also, they preselect relevant candidate predictors based on literature and expert knowledge. In contrast, in our work relies on primary care data "as is", and carries out minimal pre-processing automatically—with the purpose of being easily applicable without experts assistance. Finally, their work is aimed at finding the best predictors for COPD exacerbations, whereas our work as a twofold goal: find good prediction models for a few different asthma and COPD related variables, while also striving to develop a reusable and easy to deploy ML pipeline. For the work in [27] similar considerations can be done, as authors select candidate predictors from literature and through statistical analysis and then build an a-priori model to be evaluated using ROC AUC, the same scoring metric that we use. They rely on data from selected clinical trials and focus on asthma only. In [28] another statistical retrospective study deals with the issue of determining risk factors associated to asthma and COPD, applying univariate and multivariate logistic regression to determine predictors of future exacerbations. However, they only consider less than 400 patients specifically enrolled in the study. Although our work is considerably different in most aspects, such as sample size, data provenance, models comparison, and automation of our ML pipeline, it shares with [28] the inclusion of both asthma and COPD, which complicate the classification task as noted by authors themselves—even if they divided the population in subgroups. Nevertheless, our work provides for a much broader overview about predictive models for asthma and COPD, and it is immediately applicable to primary care datasets since our automated ML can deal with the data pre-processing steps reported in Figure 1 autonomously, as well as training and scoring models automatically as described in Section 3.

More focussed on automated ML is the recent work started in [29], that is similar to ours in both the goals (providing to clinicians a decision support tool) and the general approach (adopting ML techniques to learn prediction models from data), as authors aim to build a learning healthcare system exploiting a machine learning pipeline fed with primary care data to learn different asthma attack prediction models. However, they intend to consider asthma only and focus on a single task, that is, prediction of a single clinical variable (asthma attack). This notably simplifies the goal to be achieved with respect to our case, where we want to predict different outcomes, also give advises about

treatment plans, and finally consider COPD too—which complicates our goal as it is partially overlapped with asthma symptoms. Obviously, as the work is yet to be concluded and fully reported, we cannot compare results yet.

Another recent paper more focussed on ML, and with very similar objectives and methods to ours is [30], where authors compare a slew of different ML models to predict the risk of readmission for COPD patients. Authors compare logistic regression and its variants, random forests, linear support vector machine, gradient boosting, multi-layer perceptron, and also deep learning models including temporal features such as convolutional neural networks, recurrent neural networks, long-short term memory, and gated recurrent unit. The work is similar to ours in that authors build a mostly automated ML pipeline, with the objective of comparing which ML models could perform best for the target prediction task, which has strong reusability and modularity, hence appear to be easily adapted to slightly different tasks and easily integrated with existing DSSs for usage by clinicians in day to day practice—a goal we also have. Furthermore, they have a discussion section which is quite similar to our Section 5, although ours is a bit broader in scope as it covers both technical and non-technical aspects. However, reference [30] also differs under several aspects: first of all we do consider asthma too, and multiple prediction targets for the same primary care dataset, which notably complicates the learning task, whereas they focus on COPD and a single prediction target; second, they rely on data beyond primary care, and require 12 months of clinical data as inclusion criteria. Nevertheless, even given their more focussed scope, their results are mostly worse than ours, as their best reported performance has ROC AUC of .65—they do not report on precision-recall curves, that may report even worse results as described in Section 3.4.

Finally, two more related works are worth mentioning, despite their notable differences with the present paper. Reference [31] is one of the few works we found that considers asthma and COPD altogether. However, the learning task and methods are considerably different from ours: they want to discriminate asthma patients from COPD patients by analysing saliva samples. Furthermore, they rely exclusively on black-box ML models, that may be difficult to introduce in clinical practice due to their low transparency and understandability. However, one interesting aspect of reference [31] lies in usage of "few shots" (zero and one, for instance) learning models, which are able to learn from very few data samples. Reference [32], instead, is interesting for two reasons, even if the scope is quite different from ours, as they deal with prediction of individual asthma persistence upon clinical input for children under the age of 5 years with an incident asthma diagnosis. First, authors deal with highly imbalanced prediction classes, as for many of our prediction targets, by comparing different under-sampling techniques, such as the ENN method, that removes instances (of the majority class) whose class label differs from a majority of its k-nearest neighbours, R-ENN that repeats the ENN procedure until the majority of the k-nearest neighbours for every data point (of the majority class) have the same class label as the data point, and by removing majority class instances of Tomek links, that are pair of instances which are each other's nearest neighbour but are in different classes. Although these techniques may generally improve the prediction outcome, they do so by altering the dataset while artificially removing (or, at least, simplifying handling of) the most difficult instances to classify. For these reasons we chose not to exploit such techniques in this paper (as described in Section 3.1), leaving them as future work for comparison with our current results. Second, they use NPV-Specificity curves, which are similar to the precision-recall curves we use, as they too acknowledge ROC curves are poor for imbalanced classes.

Given the above analysis, our work is, to the best of our knowledge, the first one to consider both asthma and COPD patients for building an automated prediction pipeline based on primary care data.

3 | MATERIAL AND METHODS

To both develop the ML pipeline and validate prediction models, we exploited data coming from a Dutch primary care laboratory in the city of Groningen that receives approximately $2,000$ patients yearly with suspicion of asthma or COPD, that are referred for assessments and treatment advice. Patients receive assessment by a trained laboratory technician according to the American Thoracic Society and European Respiratory Society guidelines, there including respiratory testing with reversibility, medical history, smoking behaviour, Body Mass Index (BMI), medication, and inhaler technique evaluation. The primary care physician receives the advice from the pulmonologist directly in his/her electronic patient record. If the pulmonologist advises patients to change the medication regime, then patients are advised to have a follow-up assessment after three months to evaluate the effect of the new medication. Instead, if the pulmonologist advises to continue the current treatment policy, the patients is rescheduled for a yearly follow-up.

The primary care service shared a dataset storing the real-life observational data derived between 2007 and 2017. The dataset contains baseline assessment of the clinical conditions of $19,077$ patients. Attributes describe data such as age, gender, BMI, family history of the disease, lifestyle habits associated to the disease such as smoking, spirometry including FEV_1, FVC, and reversibility measured by a trained laboratory technician, common symptoms such as cough, wheeze, and dyspnea, information about medications including inhalation technique, and symptom questionnaires such as Asthma Control Questionnaire (ACQ) and Clinical COPD Questionnaire (CCQ). All mentioned data is assessed by a local pulmonologist though the internet. Diagnosis and treatment advice is send to the GP of the patient. Besides baseline, *follow-ups* are also available at different time points and only for some patients (more on this below). This results in $2,454$ attributes per patient, made of a set of ≈ 160 attributes repeated over time, for each potential follow-up. Amongst these, the primary care service focusses on:

- for *patient-based health risk assessment*, prediction of clinical variables relevant for diagnosis and prognosis of asthma, COPD, and asthma-COPD overlap syndrome (ACOS) patients. In particular, the aim is to build a predictive model for:
  - the amount of *exacerbations* at 1 year (0, 1, 2+), predicted after ACQ and CCQ assessment (usually done at baseline assessment)
  - the ACQ category at 3 and 12 months follow-ups after baseline assessment (controlled, partially controlled, uncontrolled)
  - the CCQ category at 3 and 12 months follow-ups after baseline assessment (stable, not entirely stable, unstable, very unstable)
- for *clinical pathways definition* in the form of suggestions for personalised intervention, the aim is to build a predictive model for:
  - advising usage of *Long-Acting Muscarinic Antagonists* (LAMA) after baseline assessment
  - advising usage of (low / high dosage of) *Inhaled Corticosteroids* (ICS), or *Long-Acting β-Agonists* (LABA), or *both* after baseline assessment
  - advising usage of β_2 bronchodilation after baseline assessment

  Also *automated diagnosis* has been explored, by trying to predict whether the patient has ACOS, COPD, or asthma after baseline assessment.

Since the dataset is extremely sparse and data is not "clean" (as real-life data), many pre-processing steps were necessary before starting with model training. The next section describes such steps, while Figure 1 depicts the whole machine learning pipeline described in this section.

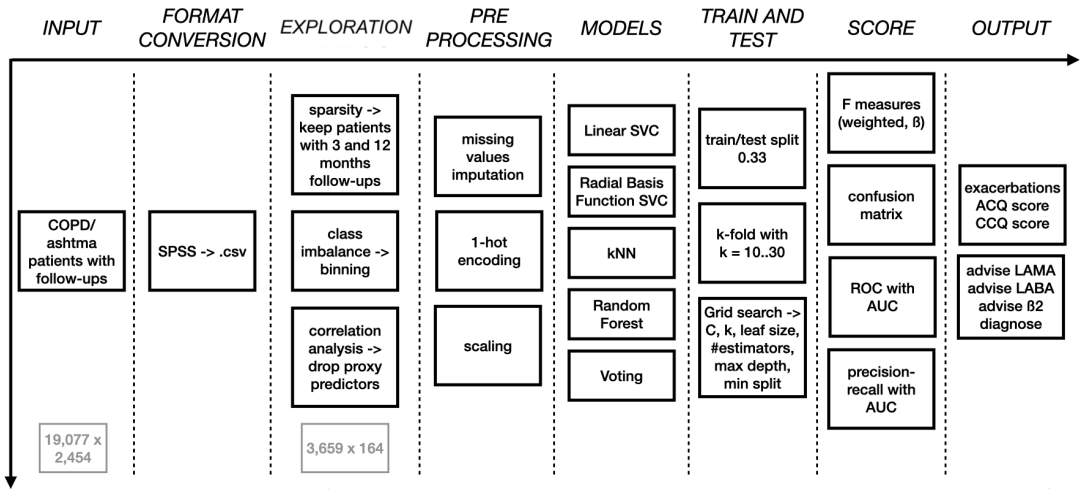
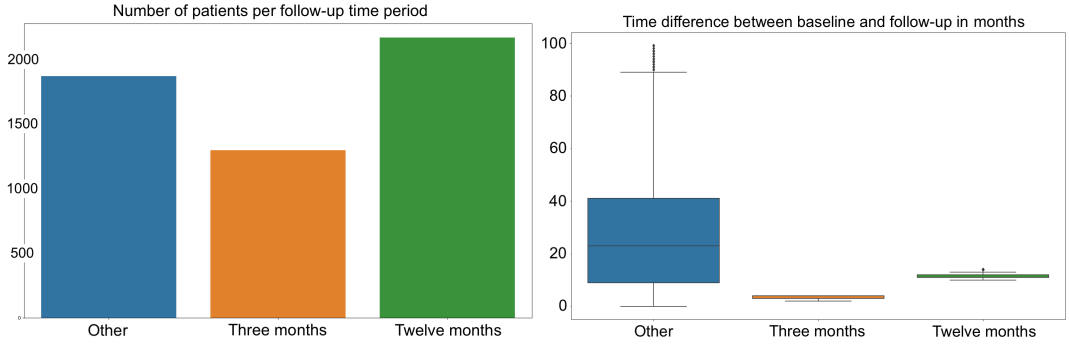


FIGURE 1 The machine learning pipeline. In light grey are the dataset dimensions (rows x columns).

3.1 | From format conversion to pre-processing

The dataset has been exported from IBM SPSS proprietary software, which is not natively compatible with Python, our language of choice fro the ML pipeline; then it has been suitably converted as CSV preserving SPSS metadata. Inspection of the resulting dataset has been required to confirm correctness, for instance with respect to data types congruence (e.g. datetime format, categoricals) and missing values preservation (e.g. custom missing values).

Then, *sparsity* has been addressed by looking at which follow-ups are the most common amongst the 19,077 patients: Figure 2a shows the results of such analysis, where time points between 2 and 4 months have been grouped in the "Three months" category, those between 10 and 14 in the "Twelve months" one, and all the rest in the "Other". As confirmed by Figure 2b, which shows the number of months between baseline and follow-up for the three groups above, three and twelve months follow-ups are the most common—y axis indicates the number of months elapsed between follow-up visits.



(a) 3 and 12 months follow-ups are most common.

(b) Other follow-ups are too sparse to consider.

FIGURE 2 Analysis of patients follow-ups periods to perform grouping.

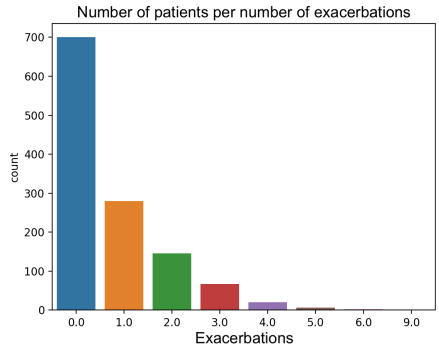


FIGURE 3 Exacerbations at 1 year is extremely imbalanced.

Based on these groupings, patients with missing values for 3^{rd} and 12^{th} month follow-ups have been removed from the dataset, as well as attributes with missing values for the corresponding measurements. This resulted in a restricted dataset of 3,659 patients for 164 attributes, with only $\approx 5\%$ of missing values. On this dataset we performed univariate analysis to identify *imbalanced* attributes, that is, categorical attributes whose classes are not equally represented hence can skew prediction performance negatively, and multi-variate correlation analysis to detect *proxy* predictors, that is, independent variables having high correlation with dependent ones hence could skew prediction performance positively.

For instance, one of the predicted variables for patient-based health risk assessment is the total number of exacerbations at 1 year, whose class distribution is shown in Figure 3: it is extremely imbalanced, hence techniques such as down/over-sampling are necessary to improve prediction performance. However, application of such techniques is challenging, as downsampling further reduces the size of the dataset, hindering the learning task, and oversampling reduces accuracy of the dataset in representing the real-world situation. For these reasons, we decided not to perform any of the two, but instead to group together categories in 3 bins: no exacerbations, 1, 2 or more.

For multi-variate analysis, instead, we looked at various forms of *correlation* amongst dependent, independent, and both dependent and independent variables. For instance, Figure 4 examines the co-occurrences of exacerbations and ACQ/CCQ categories, by looking at the percentage of patients with at least one exacerbation per ACQ/CCQ category. The few peaks at $\approx 20\%$ are deemed not solid enough to justify elimination of one of the variables.

A similar analysis has been conducted systematically on the entire set of independent variables, both between each other and against dependent ones, so as to identify proxy predictors and opportunities for dimensionality reduction (when independent variables have high correlations, it may be sufficient to keep only one). For instance, Figure 5 shows the heatmap built out of the correlations between a handful of independent variables chosen at random just to exemplify the heterogeneous situations which can be found in the dataset: whereas FEV and FVC related measurements have a high correlation, the others are mostly unrelated.

In summary, given the analysis described above, all the 164 attributes have been kept for further pre-processing stages required by the predictive models described in Subsection 3.2. Such stages include:

- *imputation* of missing values: for numerical variables the median value has been propagated, whereas for categorical variables a random one has been sampled according to class distribution frequency, that is, more represented classes have higher chances of being drawn. Although we are aware that this choice does nothing for mitigating the class imbalance problem, it preserves distribution of the original dataset;

- one-hot *encoding* for categorical variables, which is a required step for all the learning algorithms exploited;
- *scaling* of numerical variables to normal distribution (mean 1, standard deviation 0), which is a requirement for the learning algorithms exploited.

Numerical and categorical variables where already defined as such in the source dataset. It is worth emphasising that all the pre-processing steps described above have been *automated* through Python programming, so as to *(i)* provide a reusable and configurable pipeline to data scientists, and *(ii)* be ready for deployment on any target platform (e.g. as a web service).

3.2 | Predictive models

Based on the above described reduced dataset, we set up a model *training pipeline* to compare performance of the following predictive models, all implemented by the well-known `scikit-learn` Python module (to which we refer the reader for further technical information not reported here)—we refer the reader to the referenced literature for a thorough description of the models:

Linear SVC. Support Vector Classification (SVC), that is, classification based on Support Vector Machines (SVM) [33], with a linear kernel. SVM are a set of supervised learning methods particularly versatile and powerful for high dimensional spaces, especially thanks to the notion of *kernel* enabling to plug into the SVM different decision functions depending on the problem at hand (e.g. linear for Linear SVC).

RBF SVC. SVC with Radial Basis Function kernel, the default kernel in `scikit-learn`, notably performing well on average [34].

k-Nearest Neighbours. Neighbours-based classifiers do not attempt to build a model of the data, rather, store instances of the training data and compute predictions on unknown data based on majority voting amongst known data points. k-Nearest Neighbours (kNN) [35] takes into account the k data points *nearest* to the one to be assigned a class (according to a configurable metric), where k is a integer value that can be set as a learning parameter.

Random Forest. Ensemble methods combine predictions from several classifiers so as to enhance robustness over

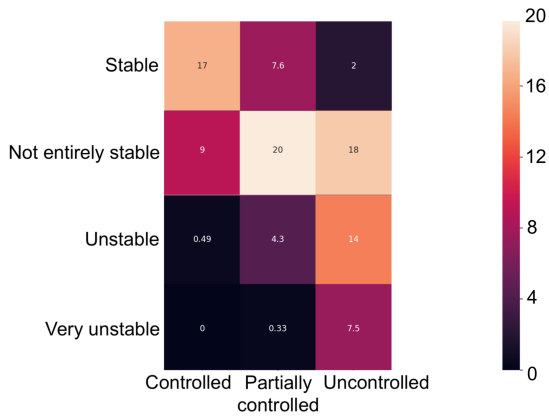


FIGURE 4 Percentage of patients with at least one exacerbation per ACQ (x-axis) and CCQ category (y-axis).

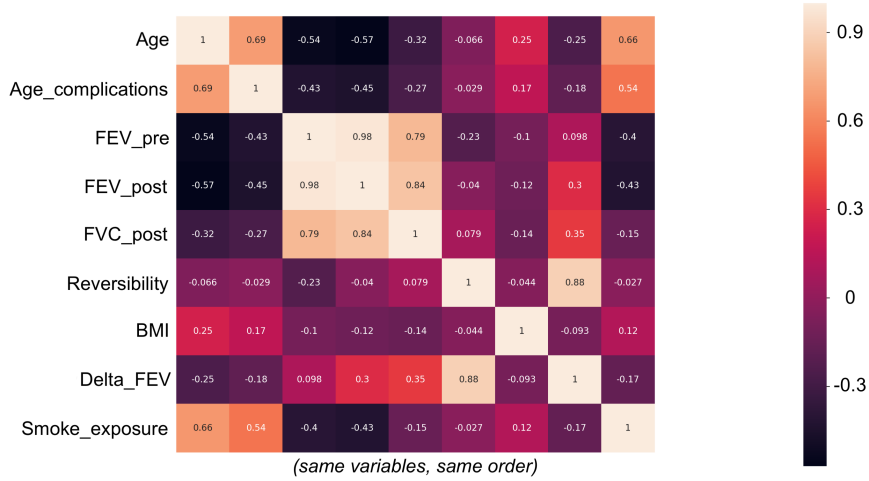


FIGURE 5 Excerpt of correlation analysis.

a single model. *Averaging methods* build several independent models over subsets of data and average out their predictions in a single one, whereas boosting methods build models sequentially trying to reduce bias of the whole sequence. The Random Forest [36] exploits an averaging method with two sources of non-determinism, meant to reduce overfitting: the subset of data for training is chosen at random, and a random subset of features is chosen when splitting each node of the decision tree.

The set of chosen models represents well the most commonly used models for prediction of categorical variables: the Linear SVC model for its simplicity and ease of interpretation, the RBF kernel has been reported to perform well on average, independently of the specific characteristics of the dataset, the kNN serves as comparison with purely data-driven methods not forcing observations into a pre-defined model, and the Random Forest is the best performing ensemble method on average. It is worth noting that we explicitly avoided *opaque* models with limited *explainability*, such as neural networks, as the clinicians of the primary care service expressed interest in working with easily interpretable models, for which they can precisely track due to what a given prediction or suggestion has been delivered by the software.

3.3 | Training and testing models

Each of the aforementioned models exposes several parameters influencing the underlying learning algorithm.

All models have been trained and tested using a train / test split ratio of 0.33, hence 33% of the dataset has been left out of training to be used for evaluating the produced prediction models. Also, k-fold cross validation has been used to assess performance of the models, with k ranging from 10 to 30 depending on the specific model (for some, more than 10 validation rounds were impractical either for overflowing memory or for taking too long to complete).

Grid search has been exploited for automatically tuning hyper-parameters of the learning algorithms, in particular: the C regularisation factor for SVC models (both linear and RBF kernels); the k number of neighbours parameter of the k-NN model, and the "leaf size" property (regulating the trade-off between efficiency and greed while building the model); the number of estimators (models), maximum tree depth, and minimum samples split at nodes, for the

Random Forest. For further information regarding each parameter meaning, we refer the interested reader to the technical documentation available starting from `scikit-learn` interactive map: `https://scikit-learn.org/stable/tutorial/machine\_learning\_map/index.html`.

3.4 | Scoring of models

Finally, the following scoring techniques for assessment of the best models have been used: F-measure and its variations (e.g. weighted, F-beta) [37], confusion matrix, and Receiving Operating Characteristic (ROC), also with area under the curve. Literature claims that the ROC should be the standard tool for assessing performance of clinical risk prediction models [38]. However, it also warns about the misleading results it can provide for imbalanced datasets [39]. Hence, in the following we first report both the ROC and *precision-recall* curves, which are known to overcome some ROC-related issues, then we stick to the latter.

4 | RESULTS

The whole pre-processing pipeline described in Subsection 3.1 as well as the predictive models described in Subsection 3.2 have been applied to the dataset subject to our investigation, for both patient-based health risk assessment, and clinical pathways definition. The following sections report on the best performing models for each.

4.1 | Patient-based health-risk assessment

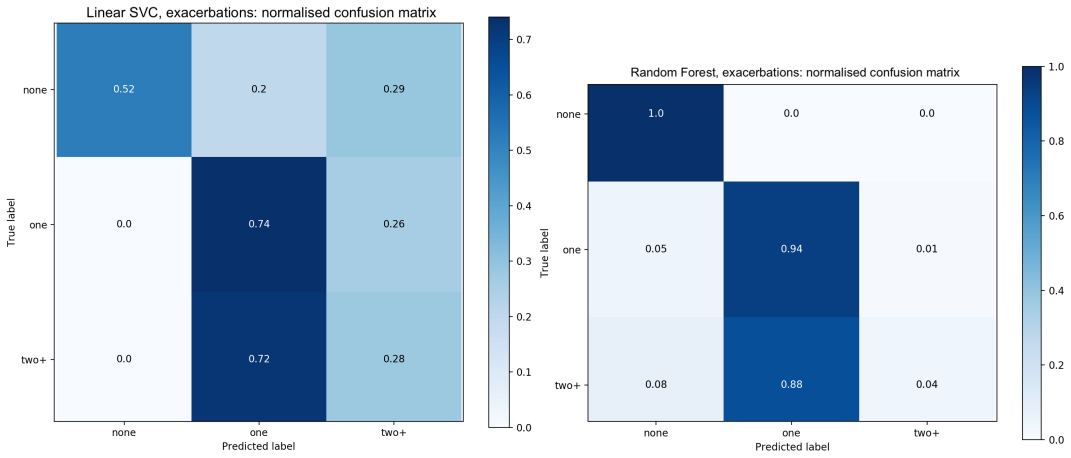
With the aim of predicting health-risk, three variables have been focussed by the primary care service: exacerbation, ACQ category, and CCQ category.

Exacerbations.
We have to deal with the extremely imbalanced multi-class problem of predicting the amount of exacerbations at 1 year amongst `0`, `1`, and `2 or more`. Our experiments shows that in this case the best performing models are the Linear SVC and the Random Forest.

Figure 6a depicts the confusion matrix comparing true classes (y-axis) against predicted ones (x-axis) for the Linear SVC classifier. As shown in the graph, the Linear SVC model behaves well in predicting cases with one exacerbation (the centre square). On the other hand, for cases with no exacerbation it works only slightly better than a random predictor, while for cases with 2 or more exacerbations, it mostly fails by predicting only 1.

Figure 6b depicts the confusion matrix of the Random Forest classifier. The best performing model has maximum depth of each tree set to 40, minimum samples for splitting to 30% of the population, and the number of trees to generate to 40 (hyper-parameters automatically set through Grid search). The model behaves extremely well for both cases with no exacerbation and with 1, but not for 2 or more exacerbations, incorrectly predicted as 1.

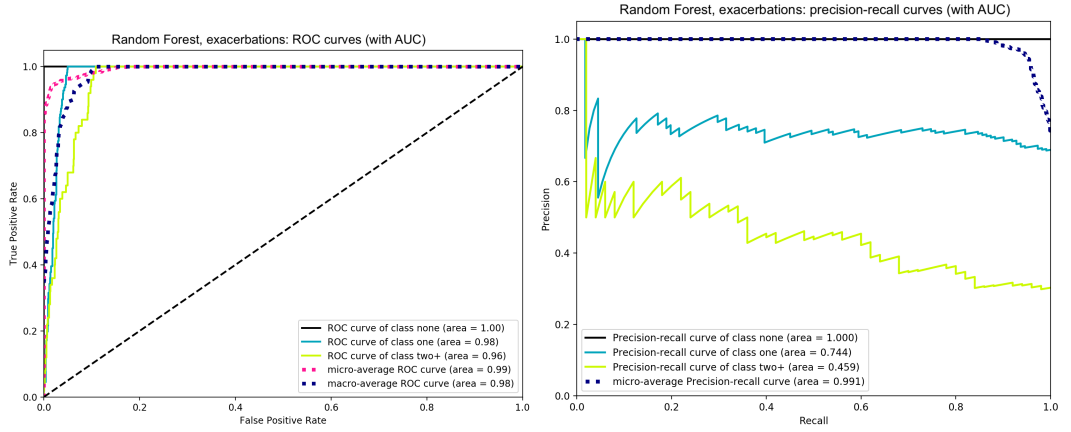
As the Random Forest model natively provides probability estimates in `scikit-learn`, we also report on ROC curves, depicted in Figure 7a. Such curves indicate the true positives rate in the y-axis, and the false positives rate in the x-axis, hence their ideal shape is a curve with a steep elbow on the top-left corner. Those curves represent an excellent model, as they follow the mentioned elbow, and the AUC is always far superior to 0.9. Nevertheless, ROC curves disregard information about *baseline probabilities*, that is, the relative proportion of the different classes in the dataset. In other words, if the model misclassifies a low represented class, it still scores high according to ROC.



(a) The Linear SVC works well in case of one exacerbation. Most of the errors are in attributing one exacerbation to cases with 2 or more.

(b) The Random Forest model is extremely good for patients with 0 and 1 exacerbation. On the contrary, it fails in assigning 1 exacerbation to most patients who actually experience 2 or more.

FIGURE 6 Confusion matrices of best Linear SVC vs. best Random Forest.

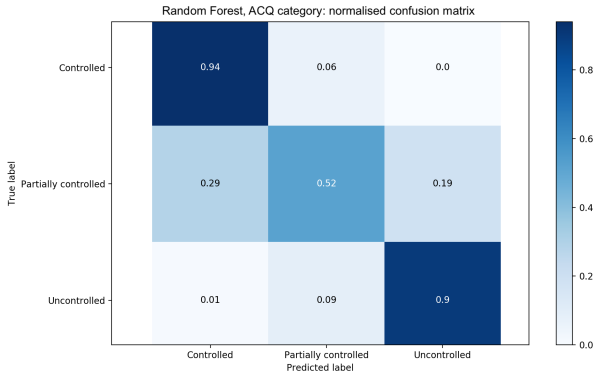


(a) ROC curves are excellent, but they do not take into account the class imbalance problem (see Figure 7b).

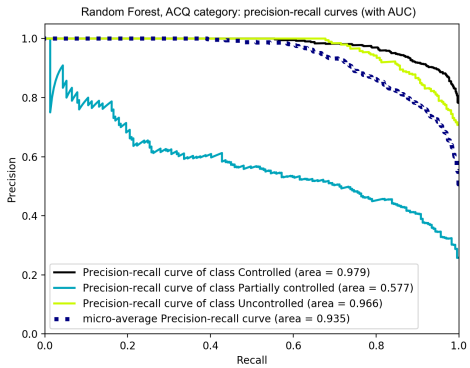
(b) Precision-recall curves better emphasise the weaknesses of the model, too, by considering class imbalance.

FIGURE 7 ROC curves vs. precision-recall curves (both with AUC) of best Random Forest.

For this reason we also report the precision-recall curves, depicted in Figure 7b. They indicate the precision along the y-axis and the recall along the x-axis, hence, the ideal shape is a curve with a steep elbow on the top-right corner this time. Here, it is more evident how the model sometimes misclassifies low represented classes such as `2 or more` exacerbations. Indeed, the precision-recall curve of the `2 or more` class in particular is far from the ideal shape just described. However, AUC is above 0.7 for two out of three classes despite class imbalance, and the class with the



(a) The Random Forest model is the best performing one, and only fails for a category gathering "edge" cases.



(b) Precision-recall curves for `controlled` and `uncontrolled` categories are extremely good.

FIGURE 8 Confusion matrix and precision-recall curves (with AUC) of best Random Forest.

most errors is also the least represented amongst samples; hence, results are still arguably good.

ACQ category.

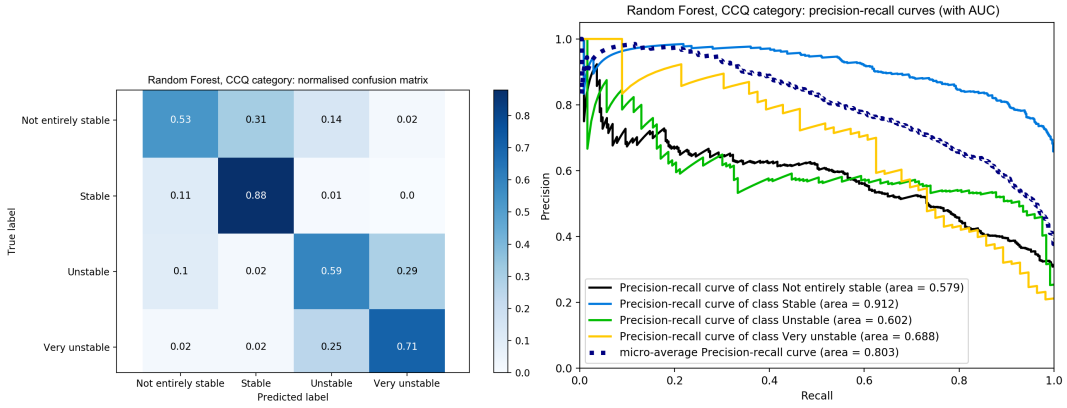
We have a multi-class classification problem. In fact, we are interested in predicting the ACQ category of patients at 3 and 12 months amongst `controlled`, `partially controlled`, and `uncontrolled`. Our results show that the best performing model is the Random Forest, trained with maximum depth of each tree set to 30, minimum samples for splitting to 10% of the population, and number of trees to generate to 40 (obtained through Grid search). Figure 8a shows the confusion matrix, while Figure 8b depicts the precision-recall curves. They confirm that the model behaves very well, as both the `controlled` and `uncontrolled` categories have over 0.9 correct predictions and over 0.95 AUC. The only category with results comparable to a random classifier is `partially controlled`.

It is worth noting, however, that such a prediction is complicated by the nature itself of the class: it represents "edge" cases with no clinical variable clearly hinting at either one of the other two categories, and corresponds to situations difficult to assess even for experienced clinicians. All the other models mentioned in Subsection 3.2 have similar performances but score lower in every class. Also, predictions at 12 months follow similar patterns but with slightly degraded performance, hence are not reported.

CCQ category.

We have to address the highly imbalanced classification problem of predicting the CCQ category at 3 and 12 months amongst `stable`, `not entirely stable`, `unstable`, and `very unstable`. In fact, category `not entirely stable` and `stable` are represented 3 to 7 times more than others.

Also in this case the best performing model is the Random Forest, trained with fully developed trees, minimum samples for splitting set to 10% of the population, and number of trees to generate to 90 (again, obtained through Grid search). Figure 9a and Figure 9b show, respectively, the confusion matrix and the precision-recall curves with AUC. Predicting `stable` cases gives the best results, followed by `very unstable` ones, whereas cases `unstable` and `not entirely stable` are only slightly better than a random classifier. However, it is worth noting that most misclassifications happen by incorrectly attributing cases to the `not entirely stable` class; this should not surprise: similarly to the case of ACQ, this class represents the most uncertain cases, difficult to assess also for experienced



(a) The Random Forest model is the best performing one, and achieves excellent results (≈ 0.9) for `stable` cases.

(b) Precision-recall curves confirm that class `stable` is the best performing one, followed by `very unstable` (AUC ≈ 0.7). Others are slightly better than a random classifier.

FIGURE 9 Confusion matrix and precision-recall curves (with AUC) of best Random Forest.

clinicians.

Predictions at 12 months show similar results, again with slightly degraded performance, hence are not reported.

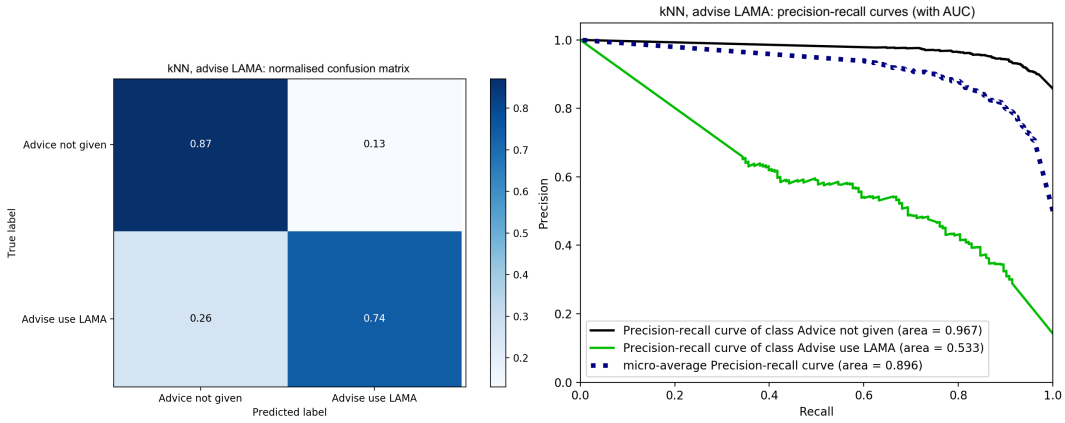
Discussion.

For each of the three prediction targets Random Forest is the approach that gives the best performance. To better understand our results from a clinical perspective, we have to consider the underlying operational context.

The training dataset after pre-processing has limited size ($3,659$ samples for 164 attributes, 5% missing values) because most GPs referred patients for one single assessment. Moreover, most tasks involved highly imbalanced class distributions, which complicates the learning task. Finally, as the whole data processing pipeline, from pre-processing to application of models, is conceived to be as *automated* as possible so as to be easily reusable and deployable on different software platforms with minimal effort, undertaking specific operations for specific tasks and on specific portions of the training dataset is not always possible. For instance, inspection of data for exploratory analysis and manual pre-processing, or fine-tuning of the learning algorithms under specific conditions, "by hand", is not supported by our Python pipeline at the moment. This makes easier to adapt the pipeline to slightly different data (e.g. coming from similar primary care services), at the cost of possibly sacrificing a bit of accuracy.

Our results show that the proposed models could be introduced in clinical practice as for each problem at least one model produces excellent results in at least one target category. A boosting approach based on Linear SVC and Random Forest can be use to predict cases with no exacerbation or 1. A Random Forest approach can be adopted to predict `controlled` and `uncontrolled` ACQ categories. Similarly, a Random Forest model can be used to identify `stable` cases in CCQ category prediction.

Our Python pipeline can be integrated with Electronic Health Records by delivering its predictions along with the degree of confidence (or even the precision-recall curves as a whole): this is crucial to ensure that clinicians are informed about the confidence of the prediction, and we argue that is also fundamental to boost adoption of this form of AI-driven support. For some models, also the relative importance of the different independent variables in contributing to the prediction may be available, and will be surely precious in further informing the clinician exploiting



(a) The *k*NN behaves very well in both cases.

(b) Precision-recall curves emphasise the error, as the AUC for `advise use LAMA` class is ≈ 0.5. They also confirm excellence of the model in the complementary case.

FIGURE 10 Confusion matrix and precision-recall curves (with AUC) of best *k*NN.

the predictors. As such, including this aspect in our analysis is already intended as a future work.

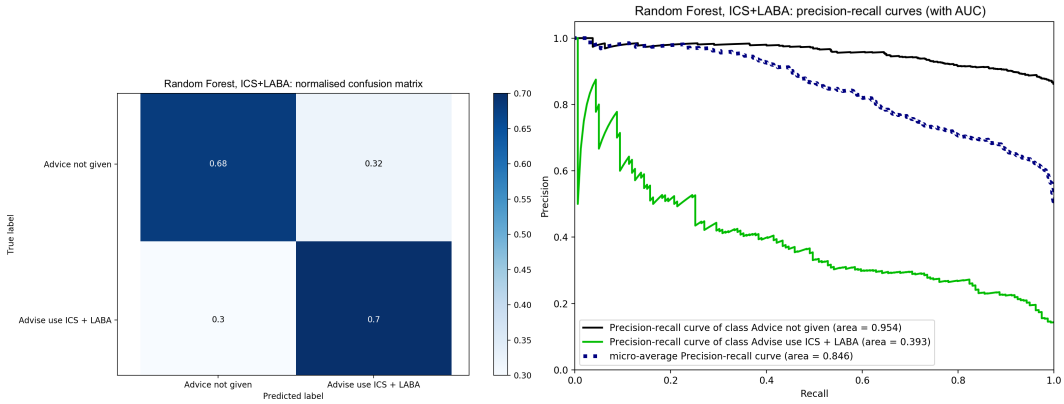
4.2 | Clinical pathways definition

As regards clinical pathways definition, and in particular suggestion of interventions, the primary care service wanted to tackle 4 classification problems: whether to advise usage of LAMA, or not (binary); whether to advise usage of both ICS and LABA, or not (binary); whether to advise usage of β_2 bronchodilation, or not (binary); suggesting the diagnosis of a patient amongst `asthma`, `COPD`, `asthma/COPD overlap`, and `diagnosis unclear` (multi-class). It is worth mentioning that we decided to adopt a supervised approach because our dataset is actually labelled, as it contains information about outcomes, hence, could base predictions on whether the suggestion led to a positive outcome or not. By doing so, a more accurate solution is likely to be obtained with respect to adopting an unsupervised approach as, for instance, *association rules mining*.

Advise LAMA.

The *k*NN classifier is the best performing model. Figure 10a shows its confusion matrix on test data. Performance is very good for both the positive and negative classes, even if some samples (cases) are incorrectly not provided with suggestion to use LAMA when instead due. The model uses weighted distance amongst samples to label classes, and leaf size and number of neighbours hyper-parameters set to 10 and 4, respectively, auto-tuned through Grid search. Weighted F1 score [37] is used as scoring metric. Both the excellent performance in one case the and misclassifications in the other are confirmed by the precision-recall curves depicted in Figure 10b.

It is worth mentioning that the Random Forest slightly improves performance when correctly not giving the advise, but at the cost of sensibly degrading performance when giving it, hence the kNN is chosen as the best model.



(a) Both classes exhibit good performance.

(b) Precision-recall curves reveal the misclassifications when advising usage of ICS and LABA.

FIGURE 11 Confusion matrix and precision-recall curves (with AUC) of best Random Forest.

Advise ICS+LABA.

The best model is the Random Forest.

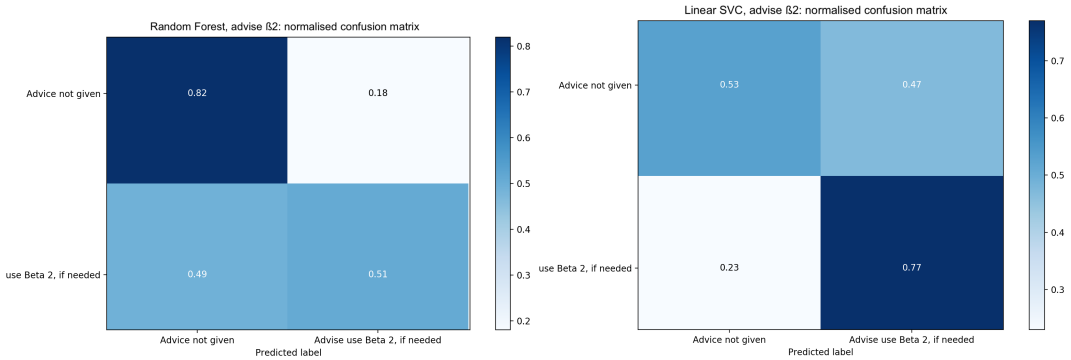
Figure 11a shows the confusion matrix obtained by applying the Random Forest model on test data, configured with balanced class weights, fully developed trees, and an auto-tuned (through Grid search) number of estimators (80) and minimum split samples (10% of population). The model does a good job in both delivering the advice when due and refraining to do so when unnecessary. The precision-recall curves reported in Figure 11b reveal that most errors are made while predicting the `advise use ICS+LABA` class.

Achieving better results across classes is complicated by the fact that class representation is highly imbalanced: for ICS and LABA usage we have a ratio of 1:5. Undersampling and oversampling are not deemed to be viable solutions for seeking improvement: the former would leave too few samples to train the model, whereas the latter would overfit the model to artificial data. As such, we prefer to stick with good results actually reflecting real-world datasets as truthfully as possible (e.g. in reference to the amount and quality of data that clinicians have at their disposal).

Advise β_2.

The two best performing models are a Random Forest and a Linear SVC, whose confusion matrices are shown respectively in Figure 12a and Figure 12b. The two mentioned models are complementary in what they are good at suggesting: the Random Forest is almost excellent when deciding not to advise usage of β_2 bronchodilation, whereas fails half the times when deciding to deliver the advice. The Linear SVC, instead, is very good when suggesting to advise usage, but fails half the times when deciding not to do so.

For these complementarity, a further attempt at improving classification results has been done with a Voting classifier, that is a kind of ensemble method (like the Random Forest) with a very simple idea at its core: combine different machine learning classifiers and use a majority vote (or the average predicted probabilities) to predict the class labels. Unfortunately, no sensible improvement has been achieved.



(a) The Random Forest is almost excellent in deciding when to not deliver the advice.

(b) The Linear SVC does the opposite, being very good in deciding when to advise usage of β_2.

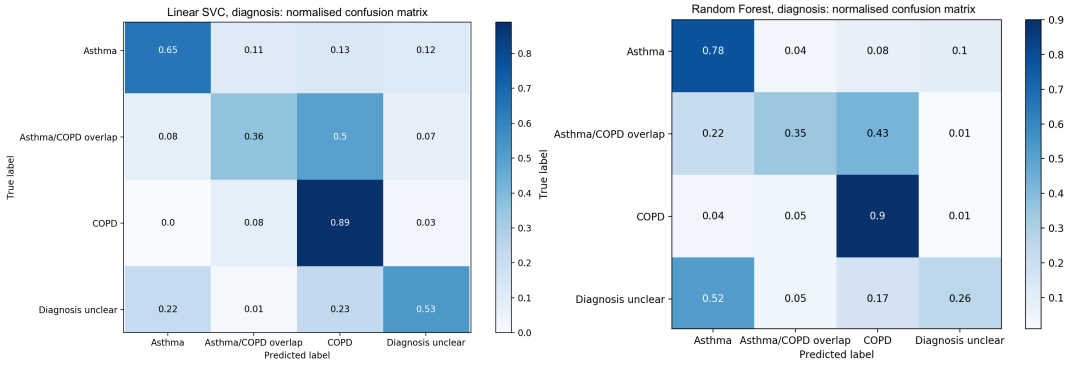
FIGURE 12 Confusion matrices of best Random Forest vs. best Linear SVC.

Diagnose.

Besides the pre-processing presented in Subsection 3.1, additional filtering of features is necessary for automated diagnosis. In particular, as we are interested in diagnosing COPD, asthma, and ACOS, we remove from the training data the features obtained as a consequence of the diagnosis itself (such as repetition of questionnaires to track evolution of the condition), as they would inflate results—being essentially a "proxy" for the predicted variable.

The two best performing models are a Linear SVC and a Random Forest, shown respectively in Figure 13a and Figure 13b. The Linear SVC is excellent in diagnosing COPD. Extensive search through the hyper-parameters plane to find the best regularisation factor (which helps especially in the case of unbalanced problems, as this one) does not help in improving classification recall. The Random Forest improves where the Linear SVC is already good (asthma and COPD) but does not help where the Linear SVC fails.

This is particularly unlucky, because otherwise a Voting classifier may have helped, by combining the complemen-



(a) Diagnosis based on Linear SVC is excellent at spotting COPD cases, and good at spotting asthma cases.

(b) The Random Forest improves the Linear SVC.

FIGURE 13 Confusion matrices of best Linear SVC vs. best Random Forest.

tary classifications of different models.

Discussion.

Given that the same considerations discussed in Section 4.1 still hold here, regarding both the limits of the used dataset and our focus on pipeline automation, we have to take into account an additional complication: in the case of clinical pathways, establishing when to give the above advices and which diagnosis to make may quickly become very difficult even for experienced clinicians, depending on the specific patient's conditions. This means not only that the task is intrinsically difficult as no combination of independent variable clearly determines the outcome, but also and most importantly that the input dataset used for training has some "intrinsic error" that inevitably biases the generated prediction models, as the labelled outcomes assigned by clinicians cannot be assumed to be always correct.

That being said, we have good results all across the different tasks. Advising usage of LAMA is accomplished with very good results by a kNN classifier. Advising usage of ICS and LABA is accomplished by a Random Forest classifier, exhibiting good results despite having to deal with an highly imbalanced problem. Advising usage of β_2 is jointly accomplished by a Random Forest and a Linear SVC: the former is very good in advising usage, whereas the later is good in the opposite case (advice not due). Finally, a Linear SVC is excellent in diagnosing COPD, and good in diagnosing asthma. Once again then, we have at least one model able to provide useful support to clinicians in their daily practice when dealing with asthma and COPD patients.

4.3 | Summary of key results

Table 1 summarises the key results we presented and discussed. For each ML task, we report: whether the classification problem is imbalanced ("Imbal.?"), whether it is multi-class ("Multi?"), the best model and its hyper-parameters, the ROC AUC, and the precision-recall curves AUC.

The Random Forest performs well across tasks, as it is the most represented model. However, its best parameters vary greatly from task to task, hence care and time must be dedicated to explore the hyper-parameters space (e.g. through automated Grid Search, as we have done). It is interesting to note that for binary classification problems where class representation is balanced (LAMA and β_2 tasks), in our case, an alternative exist: kNN for the former, which outperforms other classifiers, Linear SVC for the latter, that is more or less on par with the Random Forest. Finally, it is worth emphasising that ROC is very good across tasks, as most of the prediction errors happen in the least represented classes.

In particular, our results show that: *(i)* for predicting the number of exacerbations at 1 year, the ACQ category at 3 and 12 months, and the CCQ category at 3 and 12 months, the best performing model is a Random Forest; *(ii)* for delivering suggestions about usage of LAMA a kNN model performs best, while for ICS+LABA and β_2 a Random Forest; *(iii)* for automated diagnosis a Linear SVC performs best. For each task, at least one predictive model exhibits actionable results, that is, good performance according to the clinicians collaborating to the research, and feasibility of deployment according to the data actually available in the primary care centres taken as reference.

In [30], authors compare the following ML models to predict the risk of readmission for COPD patients: logistic regression and its variants, random forests, linear support vector machine, gradient boosting, multi-layer perceptron, and also deep learning models including temporal features such as convolutional neural networks, recurrent neural networks, long-short term memory, and gated recurrent unit. Although they do consider COPD solely, and focus on readmission risk, as noted in Section 2.3 it is the most similar work to ours, hence the only one amenable of a detailed performance comparison. Their best performing white-box model is a gradient boosting decision tree achieving mean .643 ROC AUC, whereas the best deep learning model they obtained is a gated recurrent unit, that achieves mean .65

TABLE 1 Summary of models performance results.

Task	Imbal.?	Multi?	Best model	Best params	ROC AUC	prec-rec AUC
Exacerbations	yes	yes	Random Forest	max tree depth = 40 min split = 30% trees = 40	none = 1.00 one = .98 two+ = .96	none = 1.00 one = .74 two+ = .46
ACQ	no	yes	Random Forest	max tree depth = 30 min split = 10% trees = 40	controlled = .97 partially = .85 uncontrolled = .99	controlled = .98 partially = .58 uncontrolled = .97
CCQ	yes	yes	Random Forest	max tree depth = ∞ min split = 10% trees = 90	not entirely = .77 stable = .91 unstable = .94 very = .97	not entirely = .58 stable = .92 unstable = .60 very = .69
LAMA	no	no	kNN	leaf size = 10 neighbours = 4	advise = .87 no advise = .87	advise = .97 no advise = .53
ICS+LABA	yes	no	Random Forest	max tree depth = ∞ min split = 10% trees = 80	advise = .75 no advise = .75	advise = .96 no advise = .39
β_2	no	no	Random Forest Linear SVC	max tree depth = min split = % trees = dual = true class = balanced C = $1.29e^{-05}$	advise = .73 no advise = .73 —	advise = .92 no advise = .42 —
Diagnosis	no	yes	Random Forest	max tree depth = ∞ min split = 43% trees = 70	asthma = .94 ACOS = .76 COPD = .96 unclear = .86	asthma = .90 ACOS = .34 COPD = .93 unclear = .25

ROC AUC. In our case, as reported in Table 1, all of our models have mean ROC AUC above .73, peaking at mean .98 for the number of exacerbations, .937 for ACQ category, .898 for CCQ category, .87 for advising LAMA, .75 for advising ICS+LABA, .73 for advising β_2, and .88 for diagnosis. Looking at the precision-recall curves, the performance of our models degrades as the class imbalance problem makes prediction some classes extremely complicated. In fact, the best models whose ROC AUC has been been just described now achieve mean .73 for the number of exacerbations, .843 for ACQ category, .698 for CCQ category, .75 for advising LAMA, .675 for advising ICS+LABA, .67 for advising β_2, and .605 for diagnosis. It is worth noting then that even in the case of precision-recall AUC, performance is always better than random choice, and in all cases but for diagnosis, outperforms the state of art.

5 | LIMITATIONS & LESSONS LEARNT

Although our study has shown good generalisation capabilities during validation, as demonstrated by the results described in previous section, there are limitations. Training data is limited in size (66% of 3,659), and only represent patients from a specific Dutch region, hence our results may not hold still for a different population. Also, validation data comes from the same primary care centre, hence data distribution is the same of training data, which may hinder

generalisation capabilities of models. Another limitation of our study stems from our attention to automatisation of the ML pipeline (depicted in Figure 1): since one of our aims is to provide a software package easy to integrate in DSSs or legacy systems used by clinicians, hence fully autonomous in its functioning, we do not currently support some fine-tuning operations that require manual intervention, such as feature engineering. However, we are aware that such interventions may improve predictive performance, and plan to further investigate automatisation of more pre-processing steps in our future works. Finally, a limitation of our study concerns the input dataset, that lacks imaging features that are notably useful in predicting various aspects of asthma and COPD conditions [8]. Nevertheless, our study is the first addressing both asthma and COPD predictions exclusively from primary care data, and the results achieved encourage further research along this line.

Before concluding the paper with some final remarks and an outlook to our planned future works, we take the chance to share with the reader some lessons we learnt from our experience in building and evaluating the ML pipelines described, so as to possibly deliver recommendations to those walking along the same path, or willing to.

Getting data takes time. Even if the data is *technically* readily available in a database, actually getting the hands on it may take a long time depending on the *organisational setting* of the provider: do not underestimate this. Is there the need for approval by an ethical committee? Add months to the expected delivery time. Is there administrative paperwork to carry out, such as for signing non-disclosure agreements? Add weeks. Will the data be handed over from a separate database / server / file than the one where it is used by the provider for daily practice? Again, add weeks. In our case, for example, we needed all of this, hence from the day we agreed to get the data to the day we actually got the data almost 6 months had passed. **Takeaway:** *plan ahead.*

Real-world data is a mess. This may be obvious to state, but a notable amount of research on ML happens through either synthetic datasets or carefully curated datasets storing only relevant data in a neat and clean format. Real-world data is rarely the same: redundant fields storing the same information in different ways, missing or inconsistent information, wrong data formats, high class imbalance, and other technical issues are omni-present. Furthermore, also *non-technical issues* complicate the picture, as they need a domain expert to be resolved: relevant data mixed in with useless data, wrong data, correlations to be confirmed, and so on. In our case, for instance, without the collaboration of the clinical co-authors, making sense of some of the data would have been almost impossible. **Takeaway:** *making data neat requires much more effort than training a ML model.*

Exploratory analysis is crucial. Domain experts may know the *meaning* of data, e.g. what a certain feature means from a clinical perspective, but may ignore the *hidden relationships* between data, e.g. correlation of clinical variables, predictive power, etc. Also, there are technical issues with data that only an exploration stage may bring to light, such as proxy predictors, inconsistencies between related features, imbalanced representation of classes, etc. In our case, the different classes to predict are not uniformly represented. Moreover, domain experts' knowledge is valuable to get insights about data, but may also (unintentionally) bias the exploratory analysis towards confirmation of already known facts. **Takeaway:** *explore data with and without domain experts.*

Involve domain experts. Also this point may appear obvious, but involving domain experts in the process of *designing* the ML pipeline adds value. Way too much research efforts include domain experts only in the validation stage, to assess performance of the models. However, domain experts bring added value at all stages of a ML pipeline, from initial conception to deployment in a production environment, passing through design and evaluation of the pipeline. In our case, for instance, clinicians participated from the very first stage of requirement elicitation, to define what goal to pursue with the ML pipeline, to the later stage of defining which prediction models to compare (mostly chosen for explainability), up to the latest stage of performance assessment. **Takeaway:** *domain experts add value at each stage of the pipeline.*

Overfitting is tempting. While in the iterative process of training and validating models, *hyper-specialising* the model to reach the best performance possible is alluring, but may be misleading. We are not only talking about the well known problem of model overfitting, but of the subtle habit of manually manipulating both the dataset and the model parameters to squeeze out a negligible performance increase, at the cost of sacrificing generalisation power, portability across populations, opportunities for automation of the ML pipeline. In our case, for example, while designing the ML pipeline we deliberately wanted to stick with pre-processing tasks and model building tasks easy to automate, as our concern was not only to find the best models, but also to produce an automated ML pipeline easy to adapt to different populations. **Takeaway:** *privilege automation over specialisation.*

Automation is good. Although increasing attention is recently being devoted to the *deployment* stage of a ML pipeline [40] (rather than to the training and evaluation stage only) and to *automatisation* of various stages of the pipeline (e.g. the AutoML movement [41]), way too many research works still present ML pipelines in which each step is performed manually or with little automation, and on a *ad-hoc* basis requiring costant human intervention. Although this may improve performance of the pipeline in the specific domain and use case it is being built, such practice limits its re-usability across populations and deployability in legacy environments. That's why we tried to keep the ML pipeline as automated as possible, at the cost of (perhaps) sacrificing performance. **Takeaway:** *automation adds value.*

Handle scoring metrics with care. It is already known that some of the most common scoring metrics used in ML are often *abused* [42], either by applying them disregarding the underlying assumptions supporting their validity, or simply blindly using them "following the masses" without questioning their appropriateness (e.g. F-measure [43] and ROC [39]). Even more so in the case of imbalanced datasets [44]. In our case, for instance, we commented how ROC AUC may be misleading in the case of imbalanced classes representation, and how precision-recall curves AUC may complement it to give a more comprehensive picture of models performance (see Subsection 3.2). Researchers must keep in mind at all times that each performance metric *(i)* is usually meant to assess *one* facet of model performance, and *(ii)* comes with its own *applicability* requirements (or assumptions) dictating when the metric has meaning. **Takeaway:** *scoring has goals and assumptions, don't ignore them.*

6 | CONCLUSIONS

In this paper, we tackled the problem of building a ML pipeline for clinicians treating asthma and COPD patients. On the clinical side, we developed an automated ML pipeline and compared performance of a few prediction models to predict the number of exacerbations, the CCQ category, and the ACQ category of patients, to deliver advices about usage of LAMA, ICS and LABA, and β_2 medications, and to automatically diagnose asthma, COPD, or ACOS. On the technical side, we implemented in Python the automated pipeline behind each compared model, from pre-processing to models scoring.

We found at least one good prediction model for each task. We emphasise that all the prediction models evaluated in our study have been trained and tested on real data coming from a Dutch primary care service between 2007 and 2017, that has been pre-processed by carefully avoiding to introduce distortion, such as due to over/under-sampling techniques. As the results achieved are satisfactory, we plan to advance the current work in two directions. First, further improve performance of the predictions by trying to cluster patients based on similarities amongst clinical variables first, and then apply classification separately within clusters. Also, we could consider using neural networks in combination with techniques for explainable AI, as interpretability of the models is of primary importance in the healthcare domain. Then, embed our automated and configurable Python pipeline into a web service able to serve

models as REST resources across different platforms.

Finally, we hope that our "lessons learnt" may serve well researchers and clinicians willing to start similar investigations as reference guidelines to avoid common pitfalls in machine learning pipelines design and development.

acknowledgements

This work has been supported by the CONNECARE (Personalised Connected Care for Complex Chronic Patients) project (EU H2020-RIA) under contract no. 689802.

conflict of interest

The authors declare no conflicts of interest whatsoever.

references

[1] World Health Organization (WHO), Chronic respiratory diseases: Burden of COPD;. https://www.who.int/respiratory/copd/burden/en/. online.

[2] World Health Organization (WHO), Chronic respiratory diseases: Asthma;. https://www.who.int/news-room/q-a-detail/asthma. online.

[3] Johns DP, Walters JAE, Walters EH. Diagnosis and early detection of COPD using spirometry. Journal of thoracic disease 2014 11;6(11):1557–1569. https://pubmed.ncbi.nlm.nih.gov/25478197.

[4] Moreira MWL, Rodrigues JJPC, Korotaev V, Al-Muhtadi J, Kumar N. A Comprehensive Review on Smart Decision Support Systems for Health Care. IEEE Systems Journal 2019;13(3):3536–3545.

[5] Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013;309(13):1351–1352.

[6] Cano I, Lluch-Ariet M, Gomez-Cabrero D, Maier D, Kalko S, Cascante M, et al. Biomedical research in a Digital Health Framework. Journal of Translational Medicine 2014 Nov;12(2):S10.

[7] Andreu-Perez J, Poon CCY, Merrifield RD, Wong STC, Yang GZ. Big Data for Health. IEEE Journal of Biomedical and Health Informatics 2015 July;19(4):1193–1208.

[8] Angelini E, Dahan S, Shah A. Unravelling machine learning: insights in respiratory medicine. European Respiratory Journal 2019;54(6). https://erj.ersjournals.com/content/54/6/1901216.

[9] Kaplan A, Cao H, FitzGerald JM, Iannotti N, Yang E, Kocks JWH, et al. Artificial Intelligence/Machine Learning in Respiratory Medicine and Potential Role in Asthma and COPD Diagnosis. The Journal of Allergy and Clinical Immunology: In Practice 2021;https://www.sciencedirect.com/science/article/pii/S221321982100194X.

[10] Axelrod R, Vogel D. Predictive Modeling in Health Plans. Disease Management and Health Outcomes 2003 12;11:779–787.

[11] Wagstaff K. Machine Learning that Matters. CoRR 2012;abs/1206.4656. http://arxiv.org/abs/1206.4656.

[12] Baker K, Dunwoodie E, Jones RG, Newsham A, Johnson O, Price CP, et al. Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy. International Journal of Medical Informatics 2017;103:32 – 41.

[13] Ye Y, Jiang Z, Diao X, Yang D, Du G. An Ontology-Based Hierarchical Semantic Modeling Approach to Clinical Pathway Workflows. Comput Biol Med 2009 Aug;39(8):722–732.

[14] Alexandrou D, Xenikoudakis F, Mentzas G. SEMPATH: Semantic Adaptive and Personalized Clinical Pathways. In: 2009 International Conference on eHealth, Telemedicine, and Social Medicine; 2009. p. 36–41.

[15] Yao W, Kumar A. CONFlexFlow: Integrating Flexible clinical pathways into clinical decision support systems using context and rules. Decision Support Systems 2013;55(2):499 – 515.

[16] Huang Z, Lu X, Duan H. Using Recommendation to Support Adaptive Clinical Pathways. Journal of Medical Systems 2012;36(3):1849–1860.

[17] Bins JE, Metting EI, Muilwijk-Kroes JB, Kocks JWH, in 't Veen JCCM. The use of a direct bronchial challenge test in primary care to diagnose asthma. npj Primary Care Respiratory Medicine 2020;30(1):45. https://doi.org/10.1038/s41533-020-00202-y.

[18] Messinger AI, Luo G, Deterding RR. The doctor will see you now: How machine learning and artificial intelligence can extend our understanding and treatment of asthma. The Journal of allergy and clinical immunology 2020 02;145(2):476–478. https://pubmed.ncbi.nlm.nih.gov/31883444.

[19] Exarchos KP, Beltsiou M, Votti CA, Kostikas K. Artificial intelligence techniques in asthma: a systematic review and critical appraisal of the existing literature. European Respiratory Journal 2020;56(3). https://erj.ersjournals.com/content/56/3/2000521.

[20] Mekov E, Miravitlles M, Petkov R. Artificial intelligence and machine learning in respiratory medicine. Expert Review of Respiratory Medicine 2020;14(6):559–564. https://doi.org/10.1080/17476348.2020.1743181, pMID: 32166988.

[21] Fathima M, Peiris D, Naik-Panvelkar P, Saini B, Armour CL. Effectiveness of computerized clinical decision support systems for asthma and chronic obstructive pulmonary disease in primary care: a systematic review. BMC pulmonary medicine 2014;14(1):189.

[22] Roshanov PS, Misra S, Gerstein HC, Garg AX, Sebaldt RJ, Mackay JA, et al. Computerized clinical decision support systems for chronic disease management: a decision-maker-researcher partnership systematic review. Implementation Science 2011;6(1):92.

[23] Velickovski F, Ceccaroni L, Roca J, Burgos F, Galdiz JB, Marina N, et al. Clinical Decision Support Systems (CDSS) for preventive management of COPD patients. Journal of translational medicine 2014;12(S2):S9.

[24] Mariani S, Zambonelli F, Tényi Á, Cano I, Roca J. Risk Prediction as a Service: a DSS Architecture Promoting Interoperability and Collaboration. In: 32nd IEEE International Symposium on Computer-Based Medical Systems, CBMS 2019, Cordoba, Spain, June 5-7, 2019 IEEE; 2019. p. 300–305. https://doi.org/10.1109/CBMS.2019.00069.

[25] Franssen FM, Alter P, Bar N, Benedikter BJ, Iurato S, Maier D, et al. Personalized medicine for patients with COPD: where are we? International journal of chronic obstructive pulmonary disease 2019 07;14:1465–1484. https://pubmed.ncbi.nlm.nih.gov/31371934.

[26] Kerkhof M, Freeman D, Jones R, Chisholm A, Price DB, Group RE. Predicting frequent COPD exacerbations using primary care data. International journal of chronic obstructive pulmonary disease 2015 11;10:2439–2450. https://pubmed.ncbi.nlm.nih.gov/26609229.

[27] Loymans RJB, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Assendelft WJJ, Schermer TRJ, et al. Identifying patients at risk for severe exacerbations of asthma: development and external validation of a multivariable prediction model. Thorax 2016;71(9):838–846. https://thorax.bmj.com/content/71/9/838.

[28] Al-ani S, Spigt M, Hofset P, Melbye H. Predictors of exacerbations of asthma and COPD during one year in primary care. Family Practice 2013 10;30(6):621–628. https://doi.org/10.1093/fampra/cmt055.

[29] Tibble H, Tsanas A, Horne E, Horne R, Mizani M, Simpson CR, et al. Predicting asthma attacks in primary care: protocol for developing a machine learning-based prediction model. BMJ Open 2019;9(7). https://bmjopen.bmj.com/content/9/7/e028375.

[30] Min X, Yu B, Wang F. Predictive Modeling of the Hospital Readmission Risk from Patients'Claims Data Using Machine Learning: A Case Study on COPD. Scientific Reports 2019;9(1):2362. https://doi.org/10.1038/s41598-019-39071-y.

[31] Zarrin PS, Wenger C. Implementation of Siamese-Based Few-Shot Learning Algorithms for the Distinction of COPD and Asthma Subjects. In: Farkaš I, Masulli P, Wermter S, editors. Artificial Neural Networks and Machine Learning – ICANN 2020 Cham: Springer International Publishing; 2020. p. 431–440.

[32] Bose S, Kenyon CC, Masino AJ. Personalized prediction of early childhood asthma persistence: A machine learning approach. PLOS ONE 2021 03;16(3):1–17. https://doi.org/10.1371/journal.pone.0247784.

[33] Cortes C, Vapnik V. Support-Vector Networks. Machine Learning 1995 Sep;20(3):273–297.

[34] Prajapati GL, Patle A. On Performing Classification Using SVM with Radial Basis and Polynomial Kernel Functions. In: 2010 3rd International Conference on Emerging Trends in Engineering and Technology; 2010. p. 512–515.

[35] Dudani SA. The Distance-Weighted k-Nearest-Neighbor Rule. IEEE Transactions on Systems, Man, and Cybernetics 1976;SMC-6(4):325–327.

[36] Breiman L. Random forests. Machine learning 2001;45(1):5–32.

[37] Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Technical Report SIE-07-001 from the School of Informatics and Engineering of the Flinders University, Adelaide, Australia 2011;https://csem.flinders.edu.au/research/techreps/SIE07001.pdf.

[38] Ware JH. The Limitations of Risk Factors as Prognostic Tools. New England Journal of Medicine 2006;355(25):2615–2617.

[39] Davis J, Goadrich M. The Relationship between Precision-Recall and ROC Curves. In: Proceedings of the 23rd International Conference on Machine Learning ICML '06, New York, NY, USA: Association for Computing Machinery; 2006. p. 233–240.

[40] Baier L, Jöhren F, Seebacher S. Challenges in the Deployment and Operation of Machine Learning in Practice. In: vom Brocke J, Gregor S, Müller O, editors. 27th European Conference on Information Systems - Information Systems for a Sharing Society, ECIS 2019, Stockholm and Uppsala, Sweden, June 8-14, 2019; 2019. https://aisel.aisnet.org/ecis2019_rp/163.

[41] Guyon I, Bennett K, Cawley G, Escalante HJ, Escalera S, Tin Kam Ho, et al. Design of the 2015 ChaLearn AutoML challenge. In: 2015 International Joint Conference on Neural Networks (IJCNN); 2015. p. 1–8.

[42] Japkowicz N. Why question machine learning evaluation methods. In: AAAI workshop on evaluation methods for machine learning; 2006. p. 6–11.

[43] Powers DMW. What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes. CoRR 2015;abs/1503.06410. http://arxiv.org/abs/1503.06410.

[44] Jeni LA, Cohn JF, De La Torre F. Facing Imbalanced Data–Recommendations for the Use of Performance Metrics. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction; 2013. p. 245–251.