DOI: 10.1561/9781638282372.ch11

Chapter 11

Monitoring Wellness in Chronic Obstructive Pulmonary Disease Using the myCOPD App

By Brian Pickering, Chris Duckworth, Michael Boniface, Alison Blythin and Tom Wilkinson

11.1 Introduction

According to the World Health Organization, Chronic Obstructive Pulmonary Disease (COPD) is the third leading cause of death worldwide. Risks include environmental factors, especially pollution, and behaviours such as smoking. It is persistent and progressive, with an early diagnosis and treatment benefit in controlling the progression of the disease and reducing flare-ups, known as exacerbations. Self-reporting digital apps provide a methodology for remote monitoring and management of patients with chronic conditions in the community. However, the accuracy of what is being reported has been questioned [1]. Both under- and over-estimations have been reported [2]. Underestimation may depend on the mode of collection, that is, whether remotely reported or in the presence of a clinician [3]. Other factors include the health condition itself or the perceived sensitivity (e.g., personal invasiveness) of the data being requested [4, 5]. Nevertheless, although self-reported data may not reflect the status of an individual, such data, when aggregated, may be useful to characterize a whole cohort [6].

The myCOPD app used in this study is a digital therapeutic web application designed by respiratory specialists and patients and is available on multiple Internet-connected devices. Using user-friendly multi-channel input to support a broad range of educational and literacy levels, it provides digital support for people with Chronic Obstructive Pulmonary Disease (COPD) and helps them to understand their condition and to effectively self-manage by recognizing their symptoms, thus supporting medication adherence and techniques. Education on healthy behaviours and self-management skills, including exacerbation management, is provided via the app through a 6-week Pulmonary Rehabilitation (PR) programme that has been approved by the National Health Service (NHS). In addition to the patient-facing app, myCOPD supports a connected clinician dashboard, which allows NHS services to monitor COPD patients' health status and interact with them remotely to support medication optimization, PR, and exacerbation management.

For system-generated information (such as time on the app, and how often instructional videos are accessed) and demographic data (age of the user and their (smoking) history), there are two self-reports that we focus on in particular:

1. the *COPD* Assessment Test or CAT Score: an eight-item checklist intended to be an objective indication of the app users' COPD health status [7].
2. the Symptom Score: a four-item, subjective indicator of health status equivalent to the perceived normal state or a mild, moderate, or severe deterioration [8].

Leveraging such data into prognostic models could provide increased personalization of care and reduce the burden of care for people who live with chronic conditions. This study evaluated the predictive ability of prognostic models to predict acute exacerbation events in people with chronic obstructive pulmonary disease based on data self-reported to a digital health app.

The assessments of the two scoring systems will be discussed one after the other.

11.2 Challenges Substudy 1: Prediction and Integration

The main focuses for this study are as follows:

- Can we predict COPD exacerbations from the data provided by the my*COPD* app?
- Can we integrate environmental data sources to inform those predictions?
- Can we identify patterns of behaviour associated with self-reported well-being?

It is also important to remember that the definition of exacerbation can cause problems and is subjective in nature. It is therefore dependent on the patient's own

perceptions of their normal state and on what they perceive as an acceptable deviation. But it may also rely on the judgement of a third party, such as a caregiver [9]. COPD sufferers reported using a divergent number of terms, such as a rapid deterioration in breathing (worsening breathlessness), coughing, greater production of sputum, and a change in the colour of sputum.[i]

We should also consider the ethics of using (special category) personal data from equipment that has not been provided by the relevant healthcare provider for study purposes, in this case the NHS in the United Kingdom (UK).

This work received ethics approval from the University of Southampton's Faculty of Engineering and Physical Science Research Ethics Committee (ERGO/FEPS/52137) and was reviewed by the University of Southampton Data Protection Impact Assessment (DPIA 0045) panel, with the decision that the research protocol was of low risk.

11.2.1 Data in Substudy 1 (Exacerbations)

A total dataset from 5,170 app users, providing 94,882 reports, was available over the period from 1 January 2017 through 31 December 2019. These were self-reported, including the objective COPD Assessment Test (CAT) score and a subjective, four-choice *Symptom Score* of general well-being: i.e., I feel normal, I feel worse than normal, and so forth. The reports were cleaned, as has been summarized in [8].

The data controller for these data is My mHealth Ltd., which is responsible for the myCOPD app. Their privacy notice makes it clear that the data provided to the app may be used for ethically approved research, unless the app user opts out. The data subject (e.g., the app user) is free to choose whether they are happy for their data to be used for research or not.

No data held by the NHS were used at this stage. Thus, these are personal data and in most cases special-category (i.e., hyphenate) personal data, since they relate to health. However, since the data are presented to the app by the data subject with the knowledge that they may be used for research if they do not opt out, there was no requirement to seek ethics review or approval from the relevant NHS Research Ethics Committee. Although the app users for this study are receiving treatment under the NHS for their diagnosed COPD, the data they input to the myCOPD app is not curated or governed by the NHS. This would also hold for special-category personal data – health data – from other non-healthcare provider apps.

i. See https://www.nice.org.uk/guidance/ng114/chapter/terms-used-in-the-guideline.

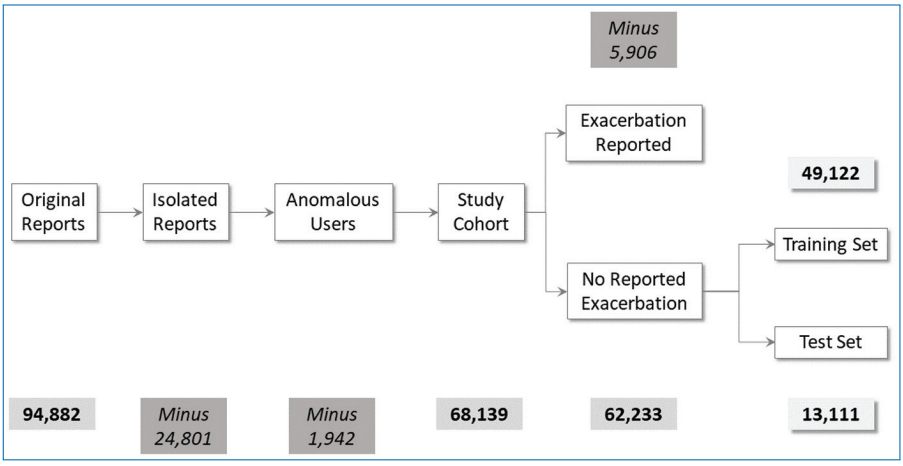

Figure 11.1. Selection of self-reported data in our study cohort, finally containing 2,374 patients. Number of retained self-reports after each filtering criteria given in box. In this study, some of the reports were cleaned and retrospectively analysed to train prognostic models. To this end, only those reports in which the patient did not report an exacerbation event were used (n = 49,122).[ii] We then randomly assigned reports from 19.2% (13,111/68,139) of the patients to a holdout test set.

From the dataset provided by the myCOPD app, 24,801 isolated reports were removed, namely those that were not part of a set with at least three consecutive reports (Figure 11.1). Then, all reports submitted by 1,942 so-called anomalous users were removed. These were internal test users or those submitting reports without being registered. This led to a study cohort of 68,139 reports. Of these, 5,906 reported a severe flare-up (exacerbation). The remaining 62,233, which therefore contained no self-reports of exacerbation events, were divided into a test set (13,111 reports) and a training set (49,122); the test set represented approximately 21% of the circa 62,000 reports (not reporting an exacerbation) registered with the app in the final dataset.

11.2.2 Predicting Exacerbations

A goal of the machine learning in this study was to investigate whether exacerbation events in the near future could be predicted from the data input to the my*COPD* app. If so, this would allow an alert to be raised as part of the care regime for the patient. Furthermore, demonstrating reliable prediction of this sort could then

ii. Note: At this stage, we do not have access to other data, such as hospitalizations.

open up the discussion about who should receive such alerts (the patient, the clinician, or both) and therefore what regulatory approval is needed.[iii]

Briefly, predictions sent to the clinician would allow them, seen as a decision-support prompt, to make an informed decision about how to alter the treatment of the patient. This respects the human-in-the-loop recommendations for AI-based technologies.[iv] Furthermore, this allows the clinician to bring to bear their own experience and knowledge of the patient in making the decision to intervene.

11.3 Methods of Substudy 1

Clinicians with the relevant experience and training recommended a 3-day analysis window. This would be close enough to a future exacerbation event to affect the course of the potential exacerbation and, thus, what goes into the reports. That is, the patient would be expected to be aware of changes (or returning stability) to their well-being. But at the same time, 3 days would be sufficiently far from the projected event to allow time for a range of pre-emptive actions for the patient and clinician. The clinician, for instance, might consult with the patient to develop an appropriate plan of action, which could include changes to medication or (brief, pre-emptive) hospitalization. This could lead to significant savings in resources. For the patient, this would demonstrate that their data are being used for their benefit (making their use of the app more transparent – see also Chapter 26 (Technology acceptance in healthcare)) and perhaps leading to increased awareness of their own health status.

To understand the added value of machine learning, we created a baseline heuristic model based only on a user's most recently reported subjective Symptom Score. The model assigns users to two risk groups:

- Users reporting a Symptom Score of 1 are predicted to be at low risk (1.7% risk) of exacerbation within 3 days.
- Users reporting a Symptom Score of 2 are predicted to be at heightened risk (7.2% risk) of exacerbation within 3 days.

Percentages in brackets correspond to the mean 3-day exacerbation rate for all reports in the training set with Symptom Scores of 1 or 2, respectively. A score of "1" indicates no exacerbation prediction, and "2" indicates the prediction of an exacerbation in the near future. This heuristic model is equivalent to a decision

iii. For a more detailed description of the procedure and results presented in this paragraph, see [8].

iv. See https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Table 11.1. Variables used for modelling.

Variable	Description
Age	User age at time of registration
Gender	User gender
Symptom Score	User-reported Symptom Score
CAT Score	User-reported CAT Score
Smoking Status	One of: smoker, ex-smoker, non-smoker
Smoking Years	How many years smoked
Time from Last Report	Time (in days) since the user last reported
Last Symptom Score	Last user-reported Symptom Score
Last CAT Score	Last user-reported CAT Score
Mean 7-Day Symptom Score	Mean Symptom Score for user over last 7 days
Mean 7-Day CAT Score	Mean CAT Score for user over last 7 days
7-Day Exacerbation Count	Number of days on which user reported an exacerbation event over the last 7 days
Mean 14-Day Symptom Score	Mean Symptom Score for user over last 14 days
Mean 14-Day CAT Score	Mean CAT Score for user over last 14 days
14-Day Exacerbation Count	Number of days on which user reported an exacerbation event over the last 14 days

tree with a depth of 1. Supervised machine learning models make use of patient demographics, lifestyle information, self-reported information, and aggregate features that summarize a patient's (recent) self-reporting history. The variables we used to generate our models are presented in Table 11.1. We built both logistic regression and random forest classifiers to predict exacerbations. Each model was trained using 5-fold cross-validation (grouped by user), which means that reports from individual users appear exclusively in either the training or the test set.

Missing CAT Scores were imputed through forward-filling at the user level where possible. All other missing values were filled using mean imputation within that fold. Either target or ordinal encoding was used for all categorical variables (Table 11.1).

Model hyperparameters were optimized on the out-of-fold validation samples by Bayesian optimization via the Tree Parzen Estimator algorithm as implemented in the HyperOpt Python library [10, 11]. Model performance was evaluated on the holdout test set, and 95% Confidence Intervals (CIs) were estimated by bootstrapping. To create a binary decision about exacerbation risk, model predictions were dichotomized with thresholds chosen to yield either a fixed specificity or the maximum Youden's J statistic on the test set [12].

11.4 Results of Substudy 1

On the holdout test set, the baseline heuristic model obtained an Area Under the Receiver Operating Characteristic (AUROC) of 0.655 (95% CI 0.676–0.689). The logistic regression model yielded an AUROC of 0.697 (95% CI 0.689–0.711) and the random forest model 0.727 (95% CI 0.720–0.735) on the holdout test (Table 11.2).[v] The significantly higher performance of the random forest model ($p < 0.001$) suggests either interactions between variables are important in discriminating between reports associated with exacerbation within 3 days or nonlinear relations are present.

In Table 11.2, we show the sensitivity and specificity of the baseline model and the machine learning models evaluated on the holdout test set. Although the baseline model is already dichotomized, a threshold must be chosen to binarize the continuous exacerbation risks it produces for the machine learning models. The baseline model obtained a sensitivity of 0.551 (95% CI 0.508–0.596) with specificity of 0.759 (95% CI 0.752–0.767), as shown in the table. Although neither machine learning model significantly outperforms the baseline model at the same specificity (e.g., compare models A and E in Table 11.2), the tuning of the threshold used to dichotomize the machine learning model predictions can lead to a range of sensitivities and specificities (compare models C, D, and E in Table 11.2) on the holdout test, which could be tuned to match different escalation policies and interventional strategies. For example, the random forest model can be tuned to yield a sensitivity of 0.921 (95% CI 0.907–0.935) or 0.576 (95% CI 0.553–0.594) with respective specificities of 0.250 (95% CI 0.246–0.254) or 0.750 (95% CI 0.749–0.751).

Table 11.2. Model performances evaluated on the holdout test set.

Name	Model	Area under the receiver operating characteristic curve	Threshold	Sensitivity	Specificity
A	Baseline model	0.655	N/A	0.551	0.759
B	Logistic regression	0.697	Youden's J statistic	0.708	0.644
C	Random forest	0.727	Youden's J statistic	0.755	0.629
D	Random forest	0.727	Specificity = 0.25	0.921	0.250
E	Random forest	0.727	Specificity = 0.75	0.562	0.750

v. The 95% CIs are given in [11].

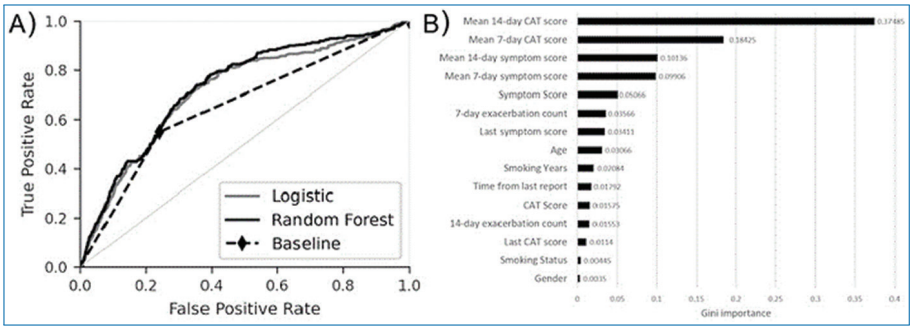

Figure 11.2. Model performance evaluated on the patient holdout test set.

Figure 11.2 shows the Receiver Operating Characteristic (ROC) curve comparing a set of prognostic models to predict exacerbation in the left-hand panel. This includes a baseline model alongside the two machine learning models – a logistic regression model (the continuous grey line) that does not consider variable interactions and a random forest classifier that does (the continuous black line). The baseline prognostic model captures the key features about exacerbation events observed in our data: people reporting a deterioration of symptoms are significantly more likely to experience an exacerbation event in the next 3 days compared to those reporting normal symptoms ($p < 0.001$), with a relative risk of 4.16 (95% CI 3.8–4.5).

In the right panel of Figure 11.2, we present the Gini importance of the features used in our random forest model.[vi] The most important features include the patients' recent CAT Scores (mean 14-day CAT Score and mean 7-day CAT Score). The importance of the CAT Score per se is to be expected since the 8-item instrument has been separately validated [7]. It is also consistent with research identifying CAT Scores as an effective way to quantify the severity of a patient's COPD, which is linked in turn to their exacerbation risk [13]. The next most important features are those quantifying recently reported Symptom Scores. Symptom Scores reflect the symptoms a patient is (or was recently) experiencing, and Figure 11.3(e) shows that people reporting higher Symptom Scores are more likely to report an exacerbation event within 3 days after having reported the step-up in score compared to those reporting lower Symptom Scores. It follows that this information would be helpful in developing a machine learning model.

vi. It is worth noting at this stage that CAT scores are typically provided monthly, whilst Symptom Scores are provided each time the app user accesses the app.

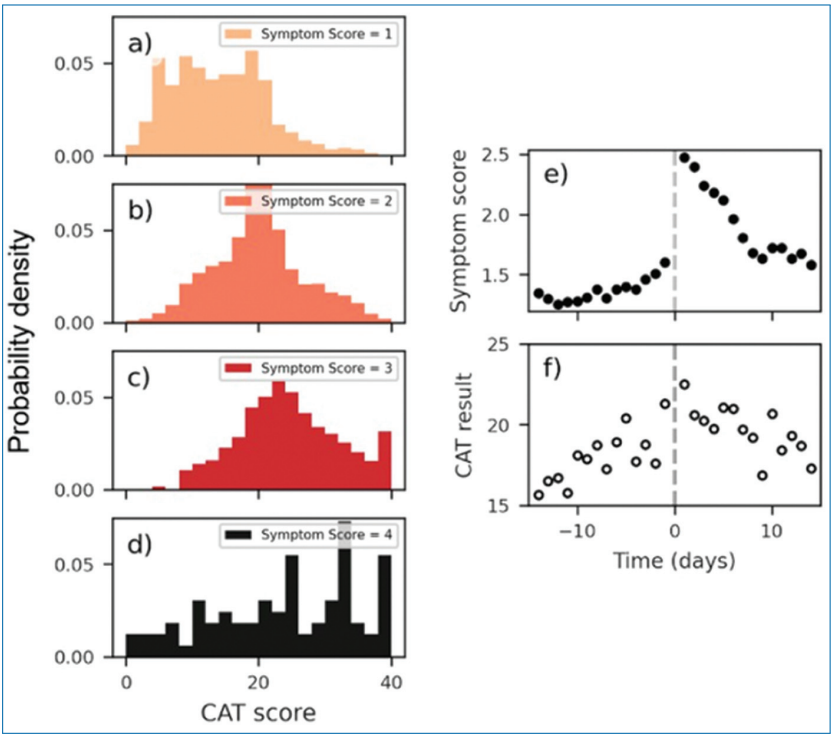

Figure 11.3. Self-reported Symptom Scores and results of Chronic Obstructive Pulmonary Disease Assessment Test (CAT) for reports in our overall 2,374 patient cohort.

11.5 Learnings from Substudy 1

There are three main areas to highlight here. First, our results suggest that self-reported data submitted to a digital health app, designed for the management of people with COPD, can be used to identify users at risk of exacerbation within 3 days after a step-up in scores, with moderate discriminative ability (AUROC 0.727, 95% CI 0.720–0.735). Further research utilizing additional linked data (particularly from medical devices such as smart inhalers, physiological monitoring sensors, and environmental sensors) is expected to increase the accuracy of these models.

We also investigated the potential to include environmental factors in building our models. We found that linking environmental data to improve prediction might be expected to be non-trivial. Referring to the traditional 'V's for big data:

- Velocity: pollution data, for instance, is delivered at regular intervals that are difficult to synchronise with the self-reported frequencies;
- Volume: pollution and weather data may not be available at the level of granularity relevant to many app users; for instance, sensor stations are typically

available across multiple streets, not close enough for an individual to track their movements between locations;
- Variety: pollutants are different in the home versus outside. Without knowing where an app user is at any given time makes it difficult to match data sources and the specific environment they would be exposed to; and
- Veracity: it is still not clear which pollutants, for instance, are more or less relevant (see [14, 15]).

The status of the data should be carefully considered. If the data to be used for modelling are clearly curated and owned by a healthcare provider, then there should be a requirement to seek ethical review and approval from an external body to avoid double jeopardy for the patient.

However, there needs to be a broader debate about the status of personal data submitted to a self-reporting app. In this study, we did not use NHS data, even though the data we accessed were special-category personal data relating to health. Although the ethical responsibilities remain the same when exploiting such data, institutional ethics review is sufficient.

11.6 Substudy 2: Behaviours Around App Usage

Although our primary focus was to examine the predictability of exacerbation events from the data reported in the my*COPD* app, the self-reported data might also show changes in behavioural patterns over time. For example, Symptom and CAT Scores may both deteriorate (i.e., become higher) moving from autumn into winter, as a function of worsening weather. Subsequently, they may improve (i.e., become lower) as the weather gets better again in the spring. In addition, the scores reported by the app users are based on their own perceptions of their 'normal state' and therefore on how they evaluate any deviation from that state. Furthermore, app users may be prone to anxiety about changing weather conditions, which may intrinsically affect their subjective evaluation of their COPD status. It is important, therefore, to consider what we might infer from reporting behaviours, in particular as the seasons change.

In addition, though, and as highlighted in the right panel of Figure 11.2, permutations of the CAT and the Symptom Scores contributed significantly to the model reported in the first part of this chapter. Indeed, the Gini importance in the figure suggests greater significance for these scores than smoking status and gender. Although the CAT Score has been separately validated, [7] the Symptom Scores are low-dimensional (there are only four options). Furthermore, COPD status in general has been associated, in the most severe of cases, with a subjective response

from the patient but is also potentially influenced by their carer [9]. Therefore, it is important to consider the reliability of the scores or, in big data terms, their veracity.

11.7 Methods of Substudy 2: Behaviours Around App Usage

An extract was created of the mean CAT and Symptom Scores for the periods autumn 2018 (September through November), winter (December through February), and spring 2019 (March through May). We used the data from those who reported three or more CAT Scores during a period, which resulted in 128 app users. From the 128 app users, the mean CAT Score and corresponding mean Symptom Score were calculated for each of the three periods, resulting in six scores per app user.

This study was approved by the Faculty of Engineering and Physical Sciences research ethics committee at the University of Southampton, reference ERGO/FEPS/56580.

11.8 Results of Substudy 2: Behaviours Around App Usage

For the periods from 2018 going into 2019, the mean temperature was 9.77°C, 5.17°C, and 8.40°C for autumn, winter, and spring, respectively.[vii] There was little difference in the mean CAT or Symptom Scores across the same period: the average CAT Scores for autumn, winter, and spring were 16.75, 16.89, and 17.53, respectively. The average Symptom Scores for autumn, winter, and spring were 1.44, 1.46, and 1.48, respectively. The differences were not significant (Wilcoxon test).

This suggests that neither objective CAT nor subjective Symptom Score changes over the three seasons; there does not appear to be any seasonal effect with these changes in terms of temperature. In addition, Figure 11.4 shows the fraction of self-reported exacerbations (i.e., >2) of all registered Symptom Scores on a given day. Days with a higher fraction of reported exacerbations have a darker shade. Visually, we see little discernible systematic seasonality in the daily exacerbation rate, although there are some indications that winter may be worse. At any rate, it is difficult to identify consistent trends at this time.

To investigate this further, we analysed the scores to establish if responses could be clustered in any meaningful way. For this, the six scores for each of the 128

vii. https://www.metoffice.gov.uk/research/climate/maps-and-data/summaries/index.

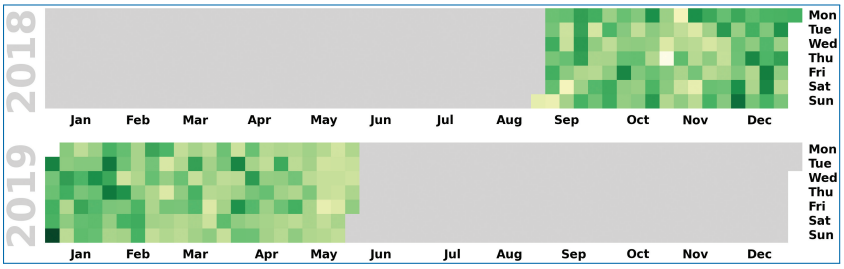

Figure 11.4. Calendar plot of exacerbation rate for study period. Each individual day shows the fraction of users reporting an exacerbation (3 or 4) out of all users reporting Symptom Scores on that day.

app users were converted to normalized (z) scores and clustered as follows: first, the amount of variance accounted for by the number of clusters was plotted for each season.[viii] For all four seasons, a four-cluster solution consistently showed good separation between the clusters. Figure 11.5 shows all 128 combinations of CAT and Symptom Scores for winter (left panel), along with the mean value for the assumed four clusters for each of the three time periods (right panel). The four clusters may be summarized as follows:

- Cluster 1: both CAT and Symptom Scores are low.
- Cluster 2: CAT Score is high; Symptom Score is low. The app user seems to be underestimating their symptoms.
- Cluster 3: both CAT and Symptom Scores are high.
- Cluster 4: CAT Scores are low, Symptom Scores are high. The app user seems to be overestimating their symptoms.

We would expect CAT and Symptom Scores to correlate: the objective CAT would be reflected by a corresponding subjective Symptom Score. Clusters 1 and *3* represent the ideal situation.

Table 11.3 shows how many individuals were assigned to a given cluster across the three seasons of autumn, winter, and spring.[ix] The average temperature (Deg C in the table) came from the UK Met Office as referenced. The final column shows the number of app users who were clustered into Cluster 1 or Cluster 3, where CAT and Symptom Scores correspond, namely those whose objective and subjective self-reports are in line.

The left panel of Figure 11.6 shows a schematic representation of the clusters from Figure 11.5. Clusters 1 and 3 are shown in blue in the figure. These two clusters are expected because CAT and Symptom Scores correspond, as stated, to

viii. That is, the difference between the sum of squares and the total sum of squares (using R-Studio).

ix. K-means clustering using IBM SPSS Ver 28.

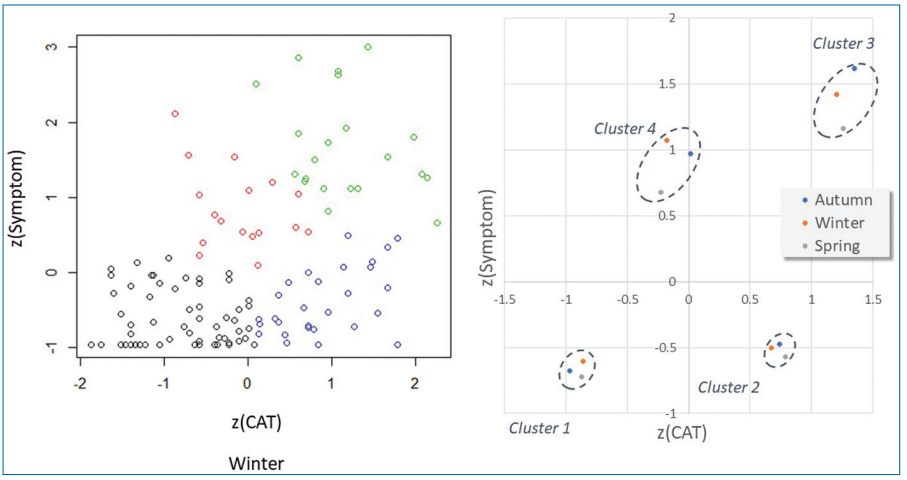

Figure 11.5. Clusters of CAT and Symptom Scores over the seasons (with k = 4).

Table 11.3. Cluster membership by season (total N = 128).

		Cluster				
Season	**Deg C**	**1**	**2**	**3**	**4**	**1+3**
Autumn	9.77	50	36	16	26	66
Winter	5.17	57	32	25	14	82
Spring	8.40	52	28	23	25	75

52%, 64%, and 59% of the 128 app users in autumn, winter, and spring, respectively. 48% for the autumn, 36% for winter, and 41% for spring therefore did not report their expected status. These individuals cluster either in Cluster 2, where they underestimate on the Symptom Score, or Cluster 4, where the CAT Score is lower than expected.

Looking at the clusters over the three seasons, 85 app users (66.4% of 128) stayed within the same cluster, either the expected ones (57% or 67% of the 85 in Clusters 1 and 3, respectively) or not (28% or 33% in Clusters 2 and 4, respectively). Staying in the same cluster suggests they noticed no seasonal effects: the differences in temperature reported by the UK Met Office did not provoke any changes in reported COPD status. Note that seasonal changes may involve dampness and other factors as well as temperature.

Around one-third (128 − 85 = 43, or 33.6%) did change clusters. The arrows on the right panel in Figure 11.6 show possible changes: the blue arrow between Clusters 1 and 3 represents situations where reports remain consistent in that both CAT and Symptom Scores vary together. An app user's condition might have been

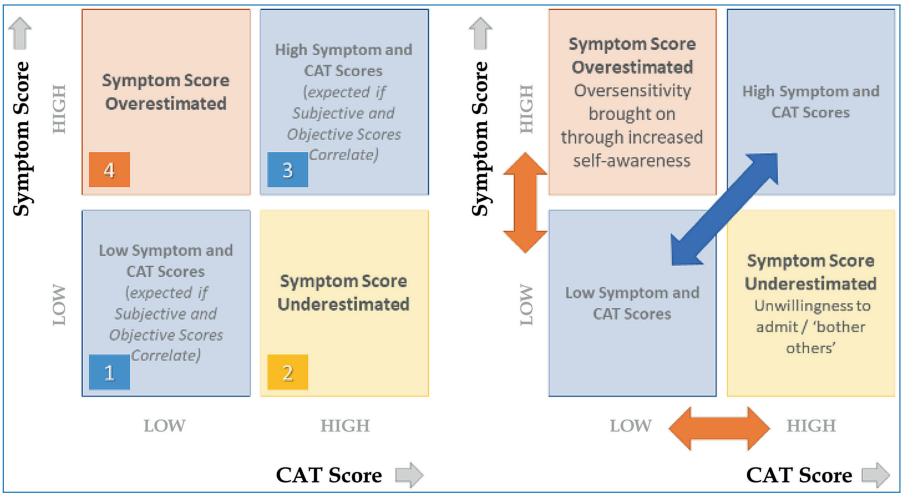

Figure 11.6. Schematic of the clustering of CAT and Symptom Scores.

expected to deteriorate (move from 1 to 3), for example, from one season to the next, and then either return to what it was or remain unchanged.

The central aspect was: which clusters did individuals move to, and can we account for any such moves? In the right panel of Figure 11.6, the arrows suggest expected changes in cluster. For instance, the blue arrow (labelled a) would represent the following scenario: an app user reporting low CAT and Symptom Scores (Cluster 1) might start to feel worse as the weather deteriorates. This would most obviously involve a move to Cluster 3. As the weather improves, this would either result in a return to Cluster 1: the app user feels better, and hence the arrow is double-ended. Alternatively, the app user may stay in Cluster 3, because their condition has genuinely worsened, and thus the assessment persists despite the better weather.

Table 11.4 summarizes which clusters those who move clusters (43 of the total 128 app users) are in for which season. There appears to be an order effect, if not a seasonal one. In autumn, there are less users reporting in Clusters 1 and 3 – the expected clusters where CAT and Symptom Scores correlate highly and follow the blue areas in Figure 11.6. The majority of reports fall in Clusters *2* and *4*, with 17 reports in each. By winter, they are at least more consistent across the clusters, with 13, 13, and 12 in Clusters 1, 2, and 3, respectively. By spring, Symptom Scores seem to be exaggerated (Cluster 4), perhaps by an oversensitivity to changeable conditions in a British spring.

Table 11.5 summarizes which clusters app users start in and where they move to in the following season. Note that the movement expressed in this table could be from autumn to winter or from winter to spring. As can be seen from the principal diagonal (numbers in **bold** in the table), 23 (or 30%) of the 79 entries in this table

Table 11.4. Cluster membership by season for the population changing cluster (total N = 43).

	Cluster			
Season	**1**	**2**	**3**	**4**
Autumn	6	17	3	17
Winter	13	13	12	5
Spring	8	9	10	16

Table 11.5. Cluster membership changes by season (total N = 43).

	To Cluster			
From Cluster	**1**	**2**	**3**	**4**
1	**8**	3	0	8
2	6	**6**	11	5
3	0	8	**4**	3
4	5	0	7	**5**

show no season-to-season movements; they remain for at least two seasons in the same cluster. Furthermore, as mentioned previously and shown by the blue areas in Figure 11.6, there are no cases from one season to the next where app users move from Cluster 1, where both CAT and Symptom Scores are low to Cluster 3 where both are high. This would be expected as an indication of a worsening condition. The same progression in reverse (from Cluster 3 back to 1), where the condition settles down again, is also not found. What is shown, however, is that 56 out of the 79 movements (70% of the movements) shift from one cluster to a different one.

Returning to the right panel of Figure 11.6, the changes in cluster across seasons suggest instability in reporting rather than specific effects of the changing seasons. In the figure, we suggest that movement to Cluster 4 (high Symptom Score, low CAT score), which represents 16 (29%) of the 56 movements in the table, reflects increased awareness by the app users of their condition: app-users are perhaps oversensitive to minor changes in their overall well-being and therefore exaggerate symptoms. Movements to Cluster 2, 11 (20%) of the 56 movements, perhaps reflect denial: their condition is deteriorating (as evidenced by the higher CAT score), but they are unwilling to admit it by underestimating the Symptom Score.

Although we have not looked specifically at three-stage reporting across autumn, winter, and spring here, movement from Cluster 4 or Cluster 2 into either Cluster 1 or Cluster 3 (29% or 52% of the 56 movements) could represent increasing awareness of health status and thereby increasing confidence about reporting in that the Symptom Scores would reflect the independently validated CAT Scores. Alternatively, their health status as measured by the CAT Score changes, and yet their subjective view as shown by the Symptom Score does not. The suggested explanations for the changing behaviours here, although finding some support in the literature, require further investigation in follow-on work.

The reliability of self-reporting has been examined in many different studies (for instance, [1]). In healthcare specifically, this may result in characteristics of the cohort reporting status that potentially influence their condition or the context in which reporting is done and an unwillingness to be seen as a burden on health professionals [16–18]. For COPD, Stelmach and colleagues suggested deliberate misreporting under some conditions, [19] whilst Sigurgeirsdottir and colleagues found evidence that the complex interaction between different factors, including anxiety and feelings of isolation, might influence how patients are willing to engage [20]. Similarly, in a series of studies with COPD patients in the Netherlands, Korpershoek and colleagues found that not everyone is suited for self-management [21]. Patient engagement with self-reporting may be influenced by whether they feel they can affect the outcome of their treatment or are dependent on (other) external factors [22]. In all, COPD patients are subject to denial, inexperience, their support system, underplaying their health status, and their trust in the healthcare ecosystem [23].

11.9 Learnings From Substudy 2: Behaviours Around App Usage

Examining the two data types – *CAT* Score and Symptom Score and which have been shown to contribute significantly to the machine-learning models (see the right panel of Figure 11.2) – does not show any clear seasonal effects: scores did not worsen as the weather deteriorated going into winter or improve going into spring.

This could, of course, be the result of app users staying indoors more during the winter or even that the temperature changes were not sufficient to affect their perceptions of their well-being. However, deviation from expected behaviours – Clusters 1 and 3 in the figures of this paragraph, where CAT and Symptom Scores correlate – may either reflect fluctuating sensitivity by the app users to their condition, a required period of adaptation to changing perceptions, or even an unwillingness to acknowledge the seriousness of their condition. Those using self-reporting apps that take subjective responses as input may need time to develop user trust in

how the healthcare app fits into their existing healthcare regime. At all events, the Veracity of the data needs to be seen in the context of reporting behaviours.

11.10 Discussion and Conclusion Both Substudies

In this chapter, we have reported on the initial findings of two approaches to the use of apps and data analysis for clinical effects. Both the substudies sought to investigate whether data coming from self-reporting apps could support the development of a machine-learning model that could then be able to predict exacerbation events. In the first sub-study, we show promising results based on a relatively small set of data. This was encouraging, not least because we did have to deal with the challenges involved in integrating environmental factors (i.e., COVID-19) directly into the models. In the second substudy, we examined behavioural aspects which could be inferred from the data. Using simple *k*-means clustering, we identified four intuitively plausible clusters. Looking at how these changed across reporting periods suggested that instability in self-reports (changes in reporting or movement between clusters) probably reflects app-user perceptions and adaptations to the app as part of their healthcare rather than external, seasonal changes.

Taken together, this study has demonstrated the potential and value of machine learning in exacerbation prediction whilst also highlighting the need to investigate behavioural aspects of app usage. In this respect, our study differs from some of the other studies where self-reporting is based on more objective measures (e.g., blood sugar levels in an app supporting gestational diabetes). It therefore contributes to big data healthcare research based on subjective self-reports.

Acknowledgements

We would like to acknowledge our former colleagues, Dr. Francis Chmiel, who finalized some of the machine learning results reported in this study, and Dr. Banafshe Arbab-Zavar, who contributed to earlier modelling work.

References

[1] Newell SA, Girgis A, Sanson-Fisher RW, Savolainen NJ. The accuracy of self-reported health behaviors and risk factors relating to cancer and cardiovascular disease in the general population: a critical review. American Journal of Preventive Medicine. 1999. 17(3), 211–229. https://doi.org/10.1016/S0749-3797(99)00069-0.

[2] Prince SA, Cardilli L, Reed JL, Saunders TJ, Kite C, Douillette K, Fournier K, Buckley JP. A comparison of self-reported and device measured sedentary behaviour in adults: a systematic review and meta-analysis. International Journal of Behavioral Nutrition and Physical Activity. 2020. 17(1), 31. https://doi.org/10.1186/s12966-020-00938-3.

[3] Scribani M, Shelton J, Chapel D, Krupa N, Wyckoff L, Jenkins P. Comparison of bias resulting from two methods of self-reporting height and weight: a validation study. JRSM Open. 2014. 5(6), 2042533313514048. https://doi.org/10.1177/2042533313514048.

[4] Wu SC, Li CY, Ke DS. The agreement between self-reporting and clinical diagnosis for selected medical conditions among the elderly in Taiwan. Public Health. 2000. 114(2), 137–142. https://doi.org/10.1016/s0033-3506(00)00323-1.

[5] Barrett EM, Maddox R, Thandrayen J, Banks E, Lovett R, Heris C, Thurber KA. Clearing the air: underestimation of youth smoking prevalence associated with proxy-reporting compared to youth self-report. BMC Medical Research Methodology. 2022. 22(1), 108. https://doi.org/10.1186/s12874-022-01594-w.

[6] Sandmo SB, Gooijers J, Seer C, Kaufmann D, Bahr R, Pasternak O, Lipton ML, Tripodis Y, Koerte IK. Evaluating the validity of self-report as a method for quantifying heading exposure in male youth soccer. Research in Sports Medicine. 2021. 29(5), 427–439. https://doi.org/10.1080/15438627.2020.1853541.

[7] Jones P, Harding G, Berry P, Wiklund I, Chen W, Leidy NK. Development and first validation of the COPD assessment test. European Respiratory Journal. 2009. 34(3):648–654. https://doi.org/10.1183/09031936.00102509.

[8] Chmiel FP, Burns DK, Pickering JB, Blythin A, Wilkinson TM, Boniface MJ. Prediction of chronic obstructive pulmonary disease exacerbation events by using patient self-reported data in a digital health app: statistical evaluation and machine learning approach. JMIR Medical Informatics. 2022. 10(3):e26499. https://doi.org/10.2196/26499.

[9] Rodriguez-Roisin R. Toward a consensus definition for COPD exacerbations. Chest. 2000. 117(5):398S–401S. https://doi.org/10.1378/chest.117.5_suppl_2.398S.

[10] Bergstra J, Bardenet R, Bengio Y, Kegl B. Algorithms for hyper-parameter optimization. Advances in Neural Information Processing Systems. 2011. 24.

[11] Bergstra J, Yamins D, Cox D. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In International Conference on Machine Learning. 2013. Pages 115–123. PMLR.

[12] Youden WJ. Index for rating diagnostic tests. Cancer. 1950. 3(1):32–35. https://doi.org/10.1002/1097-0142(1950)3%3A1<32%3A%3AAID-CNCR2820030106>3.0.CO%3B2-3.

[13] Mackay AJ, Donaldson GC, Patel AR, Jones PW, Hurst JR, Wedzicha JA. Usefulness of the chronic obstructive pulmonary disease assessment test to evaluate severity of COPD exacerbations. American Journal of Respiratory and Critical Care Medicine. 2012. 185(11):1218–1224. https://doi.org/10.1164/rccm.201110-1843OC.

[14] Hansel NN, McCormack MC, Kim V. The effects of air pollution and temperature on COPD. COPD: Journal of Chronic Obstructive Pulmonary Disease. 2016. 13(3):372–379. https://doi.org/10.3109/15412555.2015.1089846.

[15] Schikowski T, Mills IC, Anderson HR, Cohen A, Hansell A, Kauffmann F, Kramer U, Marcon A, Perez L, Sunyer J, *et al.* Ambient air pollution: a cause of COPD? European Respiratory Journal. 2014. 43(1):250–263. https://doi.org/10.1183/09031936.00100112.

[16] Kriegsman DMW, Penninx BWJH, van Eijk JTM, Boeke AJP, Deeg DJH. Self-reports and general practitioner information on the presence of chronic diseases in community dwelling elderly: a study on the accuracy of patients' self-reports and on determinants of inaccuracy. Journal of Clinical Epidemiology. 1996. 49(12), 1407–1417. https://doi.org/10.1016/S0895-4356(96)00274-0.

[17] Sallis JF, Saelens BE. Assessment of physical activity by self-report: status, limitations, and future directions. Research Quarterly for Exercise and Sport. 2000. 71(sup2), 1–14. https://doi.org/10.1080/02701367.2000.11082780.

[18] Johnsen H, Clausen JA, Hvidtjørn D, Juhl M, Hegaard HK. Women's experiences of self-reporting health online prior to their first midwifery visit: a qualitative study. Women and Birth. 2018. 31(2), e105–e114. https://doi.org/10.1016/j.wombi.2017.07.013.

[19] Stelmach R, Fernandes FLA, Carvalho-Pinto RM, Athanazio RA, Rached SZ, Prado GF, Cukier A. Comparison between objective measures of smoking and self-reported smoking status in patients with asthma or COPD: are our patients telling us the truth? Jornal Brasileiro de Pneumologia. 2015. 41, 124–132. https://doi.org/10.1590/S1806-37132015000004526.

[20] Sigurgeirsdottir J, Halldorsdottir S, Arnardottir RH, Gudmundsson G, Bjornsson EH. COPD patients' experiences, self-reported needs, and needs-driven strategies to cope with self-management. International Journal of Chronic Obstructive Pulmonary Disease. 2019. 14, 1033. https://doi.org/10.2147/COPD.S201068.

[21] Korpershoek YJ, Bos-Touwen I, de Man-van Ginkel J, Lammers J-W, Schuurmans MJ, Trappenburg J. Determinants of activation for self-management

in patients with COPD. International Journal of Chronic Obstructive Pulmonary Disease. 2016. 11, 1757. https://doi.org/10.2147/COPD.S109016.

[22] Korpershoek Y, Vervoort SC, Nijssen LI, Trappenburg JC, Schuurmans MJ. Factors influencing exacerbation-related self-management in patients with COPD: a qualitative study. International Journal of Chronic Obstructive Pulmonary Disease. 2016. 11, 2977. https://doi.org/10.2147/COPD.S116196.

[23] Korpershoek Y, Vervoort SC, Trappenburg JC, Schuurmans M. Perceptions of patients with chronic obstructive pulmonary disease and their health care providers towards using mHealth for self-management of exacerbations: a qualitative study. BMC Health Services Research. 2018. 18(1), 1–13. https://doi.org/10.1186/s12913-018-3545-4.