

Feature engineering of digital health data for machine learning prediction of COPD exacerbations in a real-world cohort

Henry M G Glyde

henry.glyde@bristol.ac.uk

University of Bristol

Alison M Blythin

my mhealth Limited

Tom M A Wilkinson

University of Southampton

Ian T Nabney

University of Bristol

James W Dodd

University of Bristol

Article

Keywords:

Posted Date: July 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9844535/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. The authors declare the following financial interests/personal relationships which may be considered as potential competing interests. Tom Wilkinson reports a relationship with my mhealth ltd that includes: board membership. Alison Blythin reports a relationship with my mhealth ltd that includes: employment. The other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Feature engineering of digital health data for machine learning prediction of COPD exacerbations in a real-world cohort

Henry M.G. Glyde^{1, 2}, Alison M. Blythin^{3, +}, Tom M.A. Wilkinson^{3, 4, +}, Ian T. Nanbney^{5, +}, and James W. Dodd^{1, 2, +}

¹Academic Respiratory Unit, Translational Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

²NIHR Bristol Biomedical Research Centre, Bristol, UK

³my mhealth, London, UK

⁴Clinical and Experimental Science, University of Southampton, Southampton, UK

⁵School of Engineering Mathematics and Technology, University of Bristol, Bristol, UK

*henry.glyde@bristol.ac.uk

+these authors contributed equally to this work

ABSTRACT

Chronic obstructive pulmonary disease (COPD) exacerbations are a major cause of morbidity, mortality and hospitalisation. Reliable early prediction could enable pre-emptive intervention, but effective short-term models remain limited. We used real-world longitudinal data from the myCOPD digital self-management platform, with and without linked wrist-worn step-count data, to develop and validate machine-learning models for exacerbation prediction. A retrospective cohort of 5,865 users with at least one exacerbation event was identified; 4,223 contributed at least one labelled app entry to the final analytic dataset. Of 625 users with linked wearable data, 459 contributed labelled app entries to the wearable analytic cohort. XGBoost models were trained to predict whether an exacerbation would occur by the patient's next 1–3 app entries and were evaluated using grouped nested cross-validation. In the main cohort, the primary model achieved mean precision 29.8%, recall 41.2%, PR-AUC 0.284 and ROC-AUC 0.834, improving precision from approximately 7% in earlier myCOPD modelling despite stricter patient-level validation. In the wearable cohort, app-only and app-plus-steps models performed similarly, with no material added value from step-count features. Positive alerts were followed by an exacerbation in nearly 30% of cases, with a median time to the first exacerbation entry of 5 days.

Introduction

Early identification of impending exacerbations in chronic obstructive pulmonary disease (COPD) has the potential to transform clinical management by enabling timely intervention, reducing hospital admissions, improving quality of life and lowering healthcare costs. This is particularly important given that COPD is the third leading cause of death worldwide, responsible for 3.23 million deaths in 2019¹. In the UK, COPD is one of the most common long-term conditions and a major contributor to preventable hospital admissions. In England alone, approximately 1.17 million people are diagnosed with COPD, representing nearly 2% of the population², although true prevalence is likely to be higher because of under-diagnosis. The burden is disproportionately borne by people living in more deprived areas, where smoking rates and occupational exposures are higher.

The consequences for the National Health Service (NHS) are substantial and increasing. In the year ending 2024, emergency hospital admissions for COPD rose by 9% compared with the previous year, reaching 208 admissions per 100,000 population³. Exacerbations account for the majority of these admissions, many of which are considered avoidable with earlier detection and treatment^{4–6}. The economic impact is also considerable: average annual management costs are estimated at approximately £1,523, £2,405 and £3,396 for patients experiencing no, one, or two or more moderate-to-severe exacerbations, respectively⁷. This stepwise increase in costs illustrates how preventing exacerbations could yield substantial savings.

Exacerbations are defined as an acute and sustained worsening of symptoms such as breathlessness, sputum volume and sputum purulence beyond usual daily variation, and they are pivotal events in COPD progression. They accelerate lung function decline, reduce quality of life and increase mortality risk. Patients and carers in the UK have identified finding better ways to prevent exacerbations as their top research priority⁸, underscoring the unmet need for strategies that enable earlier recognition and intervention. Evidence from UK cohorts further shows that delayed treatment following symptom onset prolongs recovery and worsens health-related quality of life⁴, highlighting the clinical importance of tools capable of anticipating exacerbations before they become clinically overt.

Digital technologies offer a potential route to scalable, low-burden monitoring. Several telemonitoring and remote monitoring programmes in the UK have shown reductions in hospitalisation and emergency department attendance, but their wider implementation has been hampered by cost, patient burden and the requirement for sustained clinician input⁹. In contrast, digital self-management platforms that are already embedded within NHS care pathways provide an opportunity to leverage routinely collected data at scale. The myCOPD app, for example, allows patients to record symptoms and medication use, follow a personalised self-management plan, access educational resources, including pulmonary rehabilitation, and set reminders. Such platforms generate longitudinal, symptom-driven datasets that could support near-term exacerbation prediction using machine-learning methods, potentially enabling patients to act sooner, reducing pressure on NHS services and improving outcomes.

We and others have previously developed prediction models using limited sets of self-reported features from digital health applications, with only modest precision for short-term exacerbation prediction despite acceptable discrimination^{10,11}. More broadly, a substantial body of work has investigated exacerbation prediction using remote patient monitoring and related approaches; however, multiple evidence syntheses indicate that many existing systems fail to demonstrate clinically meaningful improvements in outcomes or sufficient predictive performance for implementation^{9,12,13}. Reported positive predictive values (PPVs) for exacerbation prediction models vary widely, with 38.9% representing the highest performance reported for systems evaluated within randomised controlled trials; however, even at this level, no reduction in hospitalisations has been demonstrated, suggesting that clinically useful prediction tools will likely need to exceed this threshold while maintaining methodological robustness^{14,15}.

It remains unclear whether richer, systematically engineered representations of app-based symptom and self-management data can deliver clinically meaningful gains in precision at scale. In parallel, wrist-worn wearables are increasingly used by patients with COPD, raising the question of whether routinely captured physical activity data, such as step counts, provide additional predictive value beyond what is already contained in symptom data and exacerbation history.

In this study, we use real-world data from the myCOPD app to address these questions. We focus on users with at least one in-app reported exacerbation and develop models to predict, at each app entry, whether an exacerbation will occur by the patient's next 1-3 app entries. We substantially expand the set of digital features available from the app, moving from a small number of raw variables to a large, feature-engineered representation of symptom reports, COPD Assessment Test (CAT) scores, medication use and exacerbation history. We then train gradient boosting models to predict COPD exacerbations and assess whether this richer feature-engineering approach yields substantially improved precision compared with previous myCOPD modelling approaches, under a more stringent patient-level validation framework. Finally, in a subgroup of users with linked wristworn wearable data, we conduct an ablation study to quantify the specific contribution of step-count-derived features. Here, ablation refers to removing the steps feature set while holding the model architecture, optimisation strategy and evaluation pipeline constant. This enables us to distinguish performance gains driven by self-reported digital therapeutic data and model optimisation from any additional signal contributed by wearable-derived activity measures.

Results

Cohort characteristics

A total of 5,865 myCOPD users had at least one self-reported exacerbation event in the available dataset. Of these, 4,223 contributed at least one labelled app entry under the final outcome definition and were included in the primary analytic cohort. Baseline demographic and clinical characteristics of this cohort, together with the wearable subgroup, are shown in Table 1. In the primary analytic cohort, mean age was 69.5 years (SD 10.8), 39.4% were female, and most users were ex-smokers (61.4%). Disease burden was substantial, with a median CAT score of 18.0 [IQR 12.0–24.0], 59.4% of users having mMRC dyspnoea scores of 2 or higher, and 67.3% reporting prior antibiotics and/or steroid treatment in the past year.

A total of 625 users had linked wearable step-count data available. Restricting this subgroup to users contributing at least one labelled app entry yielded a wearable analytic cohort of 459 users. Compared with the main analytic cohort, the wearable subgroup had broadly similar age and smoking history, but lower apparent disease burden, with lower CAT scores, fewer prior hospitalisations, and lower prior hospitalisation frequency (Table 1). Median days with non-missing step-count data in the wearable subgroup was 12.0 [IQR 2.0–80.5], although coverage was highly variable.

Calendar interpretation of the app entry prediction horizon

Because the primary outcome was the next 1-3 app entries rather than fixed calendar days, we examined the elapsed calendar time represented by this horizon (Table 2). Across all rows in the broader exacerbator cohort, the mean time to the next recorded observation was 4.71 days, increasing to 11.42 days for the third-next observation. Within the final analytic model rows, the corresponding mean times were shorter: 3.14 days to the next observation, 5.97 days to the second-next observation, and 8.58 days to the third-next observation. Thus, in the modelled dataset, the practical horizon of “next 1-3 app entries” corresponded on average to approximately 3-9 calendar days, although substantial variability was observed.

Calendar intervals differed markedly by class label. For rows labelled negative ($y = 0$), the mean time to the next observation was 2.64 days and the mean time to the third-next observation was 7.02 days. For rows labelled positive ($y = 1$), the corresponding means were 13.87 and 45.89 days, respectively (Table 2), indicating that positive windows were associated with substantially sparser subsequent app engagement.

To provide a more clinically interpretable summary of this prediction horizon, we additionally examined the elapsed calendar time from each positive row to the first future exacerbation app entry (Table 3). Among positive rows, the first exacerbation entry occurred after a median of 5 days [IQR 2–21], with 39.5% occurring within 3 days, 56.0% within 7 days, and 65.4% within 14 days. Although the overall horizon was variable, these findings indicate that a substantial proportion of true-positive alerts were followed by a self-reported exacerbation over a clinically relevant period.

Performance of the primary model

The final modelling dataset comprised 547,465 labelled app entries from 4,223 users, with a positive class prevalence of 4.8%. In grouped nested cross-validation, the XGBoost model achieved a mean PR-AUC of 0.284 (SD 0.016) and a mean ROC-AUC of 0.834 (SD 0.006) across outer folds (Table 4). At fold-specific thresholds selected within the training data by maximising the F1 score, mean precision was 0.298 (SD 0.014) and mean recall was 0.412 (SD 0.011).

Pooled out-of-fold predictions showed consistent discrimination (Figures 1 and 2). Given the low positive class prevalence, this represents substantial enrichment over baseline risk. In practical terms, approximately 30% of positive alerts corresponded to true forthcoming self-reported exacerbation events, compared with a baseline event rate of 4.8%.

Relative to our previous myCOPD-based modelling approach using a limited feature set, the present model therefore achieved markedly improved precision under a more stringent patient-level validation framework. This indicates that the performance gain also reflected improved generalisable predictive performance in unseen patients.

Dominant predictors in the primary model

Gain-based feature importance analysis showed that the strongest predictors were symptom-derived variables and recent exacerbation history, with additional contributions from CAT-derived, engagement-related, and selected clinical features (Figure 3). The highest-ranked feature was recent symptom variability, followed by recent exacerbation burden and the number of recorded CAT scores. Current symptom state (`symptom_curr`), lagged symptom values, observations since the last exacerbation, longer-window symptom variability, and short-term symptom change (`symptom_delta0`) were also among the most influential predictors. Several CAT-derived features, including lagged CAT values, CAT deltas, and CAT trends, contributed meaningfully but were generally less dominant than the leading symptom and exacerbation-history variables.

Static clinical features such as prior antibiotics/steroid treatment, GOLD group, prior hospitalisation, mMRC score, age, and smoking history were retained in the model and contributed modestly, but were generally less influential than within-person dynamic features. Wearable-derived activity features ranked substantially lower than the leading app-derived predictors, and we therefore conducted a formal ablation analysis within the wearable subgroup to assess their incremental predictive contribution under comparable cohort conditions.

Overall, the feature-importance profile suggested that the model was driven primarily by current symptom state, recent symptom dynamics, exacerbation history, and patterns of symptom and CAT reporting rather than static baseline risk alone.

Wearable subgroup and ablation analysis

The wearable analytic subgroup comprised 459 users contributing 148,408 rows, of which 54,153 had non-missing step-count values (Table 6). Mean daily step count in this subgroup was 5,566 steps/day, with a median of 4,649 steps/day and substantial variability (SD 4,025). A previous meta-analysis reported a pooled mean of 4,579 steps/day (95% CI 4,310–5,208) in people with COPD¹⁶. The higher mean step count observed in the present wearable analytic subgroup therefore suggests that myCOPD users with linked wearable data may have been somewhat more active than COPD populations reported in the wider literature. A total of 1,708 symptom-defined exacerbation onset events were identified in this wearable analytic cohort using the 14-day onset window definition.

Within this subgroup, we compared two otherwise identical models: an app-only model excluding step-derived features and an app-plus-steps model including them. Mean performance across outer folds was very similar between the two models (Table 5). Without steps, mean PR-AUC was 0.162 (SD 0.039) and mean ROC-AUC was 0.785 (SD 0.022), with mean precision 0.214 (SD 0.059) and recall 0.264 (SD 0.011). With steps, mean PR-AUC was 0.163 (SD 0.043) and mean ROC-AUC was 0.787 (SD 0.026), with mean precision 0.214 (SD 0.049) and recall 0.261 (SD 0.045).

Pooled out-of-fold curves showed the same overall pattern (Figures 4 and 5). These results indicate that adding wristworn step count features did not materially improve predictive performance in the wearable analytic subgroup.

Trajectories of CAT and activity before exacerbation

To further explore the limited incremental value of wearable activity data, we examined mean trajectories of CAT deviation from baseline and daily step count deviation from baseline in the 14 days preceding self-reported exacerbation events, compared with stable periods in the wearable cohort (Figure 6). CAT deviation showed a clearer separation between pre-exacerbation and stable periods, with worsening symptoms emerging prior to event onset. In contrast, step-count trajectories exhibited only modest and inconsistent differences, with substantial overlap between exacerbation and stable periods.

These descriptive patterns are consistent with the ablation analysis. Symptom-driven and CAT-derived app features contained clearer early warning signal than wrist-worn activity measures, and step count features did not provide strong incremental predictive value over the feature-engineered app data alone.

Discussion

In this study, we developed and evaluated an XGBoost model for prediction of self-reported COPD exacerbations using routinely collected data from the myCOPD digital self-management platform, and then examined whether wearable-derived step counts provided additional predictive value in a linked wearable subgroup. Within the broader exacerbator population of 5,865 myCOPD users, 4,223 contributed at least one labelled app entry to the final analytic cohort. This cohort represented an older, symptomatic population with substantial functional limitation. Using grouped nested cross-validation at the patient level, the model achieved stable performance, with mean PR-AUC 0.284, mean ROC-AUC 0.834, mean precision 29.8%, and mean recall 41.2%. Compared with our earlier myCOPD modelling approach, this represents a substantial improvement in precision under a more stringent patient-level validation framework. Across analyses, predictive signal was driven primarily by current symptom state, recent symptom dynamics, exacerbation history, and CAT related features, whereas static baseline characteristics contributed relatively little. Within the wearable analytic subgroup of 459 users, step count features did not materially improve performance, and descriptive trajectories of step count change before exacerbation showed only small and inconsistent deviations. Together, these findings suggest that self-reported symptom signals captured through digital self-management can provide useful early warning of impending exacerbation, whereas wristworn step counts, as captured here, add limited incremental value.

An important feature of this study is that the primary prediction task was set to app entries rather than calendar days. Specifically, the model predicted whether an exacerbation would occur by the patient's next 1-3 app entries. In the final analytic cohort, this horizon corresponded on average to approximately 3.1 days to the next observation, 6.0 days to the second-next observation, and 8.6 days to the third-next observation, although variability was substantial. The model therefore reflects a pragmatic real-world use case in which risk is updated when patients engage with the app, rather than at fixed daily intervals. This framing is clinically relevant for digital self-management, but it also means that the effective calendar horizon varies between users and between labelled windows. Future work should compare app-entry and strictly calendar-indexed approaches directly to determine which framing offers the best balance between realism and clinical actionability.

In this low-prevalence prediction task, discrimination metrics alone do not fully capture clinical usefulness. We therefore prioritised precision-recall evaluation and optimised hyperparameters against average precision, which is more informative than accuracy or ROC-based measures under marked class imbalance. To define a clinically usable alert threshold, we selected the operating point within each outer training fold by maximising the F1 score, thereby favouring a balance between positive predictive value and sensitivity rather than applying an arbitrary 0.5 threshold. In practical terms, this approach aimed to improve the proportion of alerts that were followed by true exacerbation while retaining sufficient sensitivity to detect a meaningful proportion of events. This is important because an alerting system with too many false positives would be unlikely to be acceptable in routine care. Future work should evaluate alternative operating thresholds depending on the downstream action triggered by an alert, for example whether alerts are intended to prompt self-management advice, increased monitoring, or clinician review.

The dominance of symptom-derived variables, together with recent exacerbation history and CAT-related features, is consistent with the idea that exacerbations are preceded by measurable short-term changes in patient-reported burden rather than by isolated single observations. Features capturing current status, lagged values, rolling summaries, deviations from an individual baseline, and short-window changes provided a richer representation of within-person change than static variables alone. The strong contribution of recent exacerbation history, including recency and recent counts, also supports the concept of temporal clustering and heightened vulnerability following an event, likely reflecting incomplete recovery, disease instability, or both. The prominence of variables such as the number of recorded CAT scores further suggests that patterns of symptom and CAT reporting may themselves carry predictive information, potentially reflecting engagement, perceived deterioration, or both.

These findings reinforce the value of combining current state, recent trajectory, and personalised baseline information, rather than relying solely on static risk markers or raw symptom values alone. They also suggest that future models may benefit from explicit representation of recovery and instability phases, for example through post-event features or interaction terms

linking symptom trajectories to recent event burden. Importantly, feature importance should not be interpreted causally; rather, it indicates which signals most strongly supported short-horizon prediction within this dataset and modelling framework.

The limited incremental contribution of step-derived features may reflect several factors. First, the wearable analytic subgroup was substantially smaller than the main analytic cohort, reducing statistical power for detecting modest incremental gains. Second, wearable step count data showed highly variable coverage, with a median of 12 days with step-count data per participant despite a much higher mean, indicating strong skew and heterogeneous engagement. Third, step counts are susceptible to behavioural and measurement noise, including non-wear, inconsistent syncing, routine variation, and symptom-driven changes in device use itself. Even with feature engineering, activity may therefore function more as a downstream correlate of symptom worsening than as an earlier independent warning signal. This interpretation is consistent with the descriptive trajectory analysis, in which CAT deviation from baseline showed clearer pre-exacerbation separation than did step count deviation. In addition, the higher mean step count observed in this wearable subgroup than that reported in some COPD cohorts in the wider literature may suggest a relatively more active and engaged subgroup, which could also influence the generalisability of our findings.

These findings do not imply that wearable data are clinically irrelevant. Rather, they suggest that step counts alone, as captured within this group, offered limited additional information beyond app-derived symptom and CAT data for this specific prediction task. Wearable-derived value may be greater under different conditions, including higher adherence, explicit modelling of missingness and engagement, alternative prediction horizons, or inclusion of richer physiological channels such as oxygen saturation, heart rate, respiration rate, or sleep quality.

This study has several limitations. First, the primary outcome was based on symptom scores entered within myCOPD rather than externally validated clinical exacerbation events, so outcome misclassification is possible. Second, the app-entry modelling design means that the effective calendar prediction horizon varied according to app engagement. Third, although patient-level grouped cross-validation provides a more realistic estimate of generalisation to unseen users, this remains a retrospective modelling study and cannot establish clinical utility in practice. Fourth, the wearable subgroup was smaller than the main cohort and step count completeness was variable, limiting our ability to exclude smaller incremental benefits with high confidence. Finally, users of a digital self-management platform may not be representative of the wider COPD population, particularly with respect to digital engagement, health literacy, and disease-management behaviours.

Overall, this study shows that prediction of self-reported COPD exacerbations is feasible at scale using routinely collected myCOPD app data. In a large analytic cohort evaluated using patient-level grouped nested cross-validation, an XGBoost model using feature-engineered digital self-management data achieved stable discrimination and substantially improved precision compared with our earlier myCOPD modelling approach, with predictive performance driven primarily by current symptom state, recent symptom dynamics, exacerbation history, and CAT-related features. In contrast, in the wearable analytic subgroup, adding wrist-worn step-count features did not materially improve performance, consistent with the weak and inconsistent pre-exacerbation activity changes. These findings reinforce the central importance of symptom and health-status trajectories for early warning and suggest that step counts, as captured here, offer limited incremental value beyond app-derived data alone. Clinically, the model functioned as a near-term early warning system rather than a fixed-day predictor: among positive rows, the first exacerbation-defining entry occurred after a median of 5 days [IQR 2–21], with 39.5% occurring within 3 days and 56.0% within 7 days. Future work should focus on improving outcome definitions, handling missingness and engagement patterns, directly comparing calendar-day and app-entry prediction, and integrating richer physiological signals and clinically calibrated decision thresholds to better translate predictive models into actionable exacerbation support.

Methods

Study design and data source

We conducted a retrospective machine learning model development study using pseudonymised real-world data from the myCOPD digital self-management platform (my mhealth Ltd, UK). myCOPD is used within NHS services and enables people living with COPD to record symptoms and medication use, complete the COPD Assessment Test (CAT), access personalised self-management content, and engage with educational resources including pulmonary rehabilitation. For some users, linked wrist-worn wearable step-count data were also available.

The analysis dataset was assembled from multiple source files containing longitudinal symptom, CAT, pulmonary rehabilitation (PR), medication-use, step count, and app-engagement data, together with a separate patient-level clinical file. Time-series datasets were merged by anonymous user identifier (`patient_key`) and calendar date using outer joins. Static clinical variables were merged at the patient level after deduplication to one row per patient. All data were de-identified before analysis.

All methods were carried out in accordance with relevant guidelines and regulations, including applicable data protection requirements. This study was a retrospective secondary analysis of anonymised or pseudo-anonymised service data derived from users of the myCOPD digital health platform. No prospective recruitment, intervention, product testing, experimental

procedures, or collection of new data from human participants was undertaken by the study team. Permission to access and analyse the dataset for the current study was granted by the data custodians, my mhealth Limited.

Users of myCOPD are required to accept the my mhealth Terms and Conditions and acknowledge the my mhealth Privacy Policy before accessing the service. The my mhealth Privacy Policy states that health information and digital health/device information may be processed for scientific research purposes and that anonymised or pseudo-anonymised information may be shared with trusted research partners for scientific research. The study team received only anonymised or pseudo-anonymised data and did not have access to information that would allow individual users to be identified.

All methods were carried out in accordance with relevant guidelines and regulations, including the Declaration of Helsinki and applicable data protection requirements. Ethical approval for this study was granted by the University of Bristol WS1 Research Ethics Review Committee (Ref: 21806). Permission to access and analyse the pseudonymised/de-identified myCOPD service data was granted by the data custodians, my mhealth Limited. Users of myCOPD are required to accept the my mhealth Terms and Conditions and acknowledge the my mhealth Privacy Policy before accessing the service. The my mhealth Privacy Policy permits anonymised or pseudo-anonymised information to be shared with trusted research partners for scientific research. The requirement for obtaining additional individual informed consent for this retrospective secondary analysis was waived by the University of Bristol WS1 Research Ethics Review Committee.

Data preprocessing

Dates were parsed and harmonised to calendar date using a mixed-format parsing procedure. Rows with missing `patient_key` or unparseable dates were removed during preprocessing. Duplicate time-series records within the same patient-date were aggregated before merging. For PR data, records with `duration == 1` or `rpe == 1` were treated as artefactual and excluded. The remaining PR records were then aggregated within patient-date by summing duration and taking the maximum daily rating of perceived exertion (RPE); PR load was calculated as duration multiplied by RPE.

For medication and app-engagement data, duplicate numeric values within patient-date were summed. For step-count data, common source column variants were harmonised to a single `steps` variable where necessary. Step-count records dated before 1 January 2016 and values greater than 20,000 steps/day were excluded.

As part of upstream dataset construction, a forecasting label based on symptom-defined exacerbation onset was also generated using a 14-day post-onset exclusion window. Rows occurring within this post-onset window and other unlabelled rows were removed before model development. These variables were retained in the exported dataset for reference but were explicitly excluded from the final model feature set.

Participants and analytic cohorts

For the main modelling analysis, we first identified users with at least one symptom-defined exacerbation event in the available app data, defined as any symptom score of 3 or 4 at any time. This created the broader exacerbator cohort. The final analytic cohort for the primary model consisted of exacerbators who contributed at least one labelled app entry under the final app-entry outcome definition described below. Because this outcome required three subsequent recorded app entries, users who never contributed a valid labelled entry were not included in the final analytic modelling cohort.

For the wearable ablation analysis, we identified the subset of exacerbators with any non-missing step count data from linked wrist-worn devices. The final wearable analytic cohort consisted of those users who also contributed at least one labelled app entry under the same outcome definition. Within this wearable cohort, we trained matched models with and without step-derived predictors so that the incremental contribution of wearable activity data could be evaluated directly.

Outcome definition

Symptom scores in myCOPD were recorded on a four-point ordinal scale, with higher scores indicating worse symptom burden. Consistent with previous work, symptom scores of 3 or 4 were treated as exacerbation symptom events and scores of 1 or 2 as stable symptom states¹⁰.

The primary prediction task was app-entry rather than calendar-day. For each retained app entry within a patient, we identified the next one, two, and three recorded app entries in chronological order. The binary outcome variable (`exac_next3obs`) was assigned a value of 1 if any of those next three app entries had a symptom score of 3 or 4, and 0 otherwise. Rows were labelled only when all three subsequent app entries existed; rows without three future observations were excluded from modelling.

Because this outcome was set to app usage rather than fixed calendar time, the effective calendar horizon varied according to reporting frequency. To characterise this, we summarised the elapsed calendar time between an app entry and the next, second-next, and third-next recorded app entries within patient. To aid clinical interpretation, we also summarised, among positive rows, the elapsed calendar time from the app entry to the first future exacerbation app entry within the next 1-3 app entries.

Predictor variables and feature engineering

Time-varying predictors were engineered to avoid future information leakage. For any row, the model could use information available at that app entry, including current-row values, while lagged and rolling features were constructed so that future app entries did not contribute to predictors.

Symptom and CAT features

Raw symptom scores and CAT scores were coerced to numeric values and sorted by patient and date. Missing CAT values were forward-filled within patient to preserve continuity between questionnaire completions. The model was allowed to use the current app-entry values, reflecting the intended use case in which prediction is made immediately after a patient submits an app entry. In addition, we derived lagged symptom and CAT features at the previous 1, 2, 3 and 7 app entries.

To capture deviation from each patient's usual state, we estimated a personalised CAT baseline from historically stable app entries, defined as rows with symptom score equal to 1, using an expanding mean restricted to those entries and shifted by one entry. Deviations from this baseline were then calculated for both current and lagged CAT values. We also derived rolling summaries, including 7-entry rolling means and standard deviations, together with short-term change features based on differences between current and recent prior values and between successive lagged values. These variables were intended to capture current symptom burden, recent instability, and deviation from each user's own typical state.

Exacerbation history and longer-window features

To represent longer-term instability and prior event burden, we derived several exacerbation-history variables at the patient level. These included the number of app entries since the most recent prior exacerbation event, counts of prior exacerbation entries over the preceding 28 and 90 app entries, and 28-entry rolling means and standard deviations of symptom scores together with a 28-entry rolling mean of CAT. These features were intended to summarise recent exacerbation burden and longer-horizon symptom variability.

App-derived side-channel and static clinical features

Where available, we included additional app-derived variables reflecting self-management behaviour and treatment-related activity. These included pulmonary rehabilitation duration, pulmonary rehabilitation RPE, derived pulmonary rehabilitation load, counts of regular and reliever inhaler use, and app engagement measured using tile-count variables. For these time-varying variables, the model was allowed to use current-row values together with lagged features at 1, 2 and 7 prior app entries and rolling summaries over the preceding 7, 14 and 28 app entries, using sums, means, or maxima as appropriate to the variable type. We also included calendar covariates for day of week and month.

Static numeric clinical features available within the dataset were included where present, including age, pack-year history, modified Medical Research Council (mMRC) dyspnoea score, prior antibiotic/steroid treatment variables, and prior hospitalisation variables. Static categorical variables, including gender, smoking status, and GOLD group, were one-hot encoded. To avoid information leakage, variables representing labels generated during preprocessing were explicitly excluded from the model feature set.

Wearable-derived step-count features

In the wearable subgroup analysis, daily step-count data from linked wrist-worn devices were processed as numeric variables and merged by patient and date. The model was allowed to use the current app-entry step count value, where available, together with lagged and rolling step-derived features. Step count features were engineered using the same general temporal design as the app-derived time-series predictors so that behavioural and symptom-derived signals were represented on comparable scales.

We generated lagged step counts at the previous 1, 2, 3 and 7 app entries, together with 7-, 14-, and 28-entry rolling means and standard deviations. Short-term change features were also derived, including step-count deltas and a 7-entry trend. To represent individual activity norms, a personal step baseline was calculated from a 60-entry rolling median with an expanding-median fallback. Absolute and relative deviations from this baseline were derived for both the current app entry and the most recent prior app entry. Additional features captured activity variability and low-activity patterns, including coefficients of variation over 14 and 28 app entries, binary indicators for reduced activity relative to baseline or recent averages, and a 14-entry low-activity streak feature. Lagged and rolling step-derived predictors were constructed so that they used only prior information, while current-row step count values were retained as candidate predictors. In ablation models without wearable input, all step-derived variables were removed while the remainder of the modelling pipeline was left unchanged.

Handling of missing data and numeric representation

Symptom scores and CAT scores were coerced to numeric values. Missing CAT values were forward-filled within patient before feature generation. Remaining missing predictor values were handled using median imputation within the modelling pipeline, with imputation estimated separately within each training fold during cross-validation. Predictors were stored as 32-bit floating-point values where possible. Features were not standardised, as tree-based gradient boosting models are not sensitive to monotonic scaling.

Model development

We trained gradient boosting classifiers using XGBoost to predict the binary app-entry-indexed outcome. To prevent information leakage between training and testing data, all splitting was performed at the patient level.

For the primary analysis, we used grouped nested cross-validation. The outer loop used 5-fold `GroupKFold` splitting, with `patient_key` as the grouping variable, so that each outer test fold contained only unseen users. Within each outer training set, 3-fold grouped cross-validation was used for hyperparameter optimisation via `RandomizedSearchCV`.

The modelling pipeline consisted of median imputation of missing predictor values, conversion to 32-bit floating-point arrays, and fitting of an `XGBClassifier` using histogram-based tree construction and average precision as the tuning objective. To address class imbalance, we calculated the ratio of negative to positive observations within each outer training fold and passed this value to XGBoost as `scale_pos_weight`. Hyperparameter tuning explored the number of trees, learning rate, maximum depth, minimum child weight, subsample fraction, column subsampling fraction, gamma, and L2 regularisation strength. Twelve candidate hyperparameter configurations were evaluated per outer fold.

The full feature-engineered app-derived model was trained in the main analytic cohort. In the wearable analytic cohort, we trained two otherwise identical models: an app-only model excluding all step-derived features and an app-plus-steps model including the full step-derived feature set. In the wearable analysis, the same labelled row set and the same outer patient-level cross-validation splits were used for both models to isolate the incremental contribution of wearable step-count features.

Model evaluation and feature importance

For each outer fold, predicted probabilities for the positive class were generated in the held-out test set. A decision threshold was selected within the corresponding outer training set by maximising the F1 score and was then applied unchanged to the outer test set.

Test-set performance was summarised using area under the precision-recall curve (PR-AUC; average precision), area under the receiver operating characteristic curve (ROC-AUC), precision and recall at the selected threshold, and confusion-matrix counts. Results were summarised as means and standard deviations across outer folds. In addition, pooled out-of-fold predictions were retained for generation of combined precision-recall and ROC curves.

Feature importance was assessed using XGBoost gain-based importance. Gain values were extracted from each outer-fold model, normalised within fold, and averaged across folds to produce mean normalised gain values for ranked feature-importance summaries.

Software

All analyses were conducted in Python using `pandas` for data handling, `scikit-learn` for preprocessing, cross-validation and evaluation, and `xgboost` for gradient boosting model development.

References

1. Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease: 2024 Report (2024).
2. NHS England Digital. Quality and Outcomes Framework, 2020-21 (2021).
3. Office for Health Improvement & Disparities. Respiratory disease profile: statistical commentary, June 2025. <https://www.gov.uk/government/statistics/update-of-indicators-in-the-respiratory-disease-profile-june-2025/respiratory-disease-profile-statistical-commentary-june-2025> (2025). Accessed: 16 September 2025.
4. Wilkinson, T. M., Donaldson, G. C., Hurst, J. R., Seemungal, T. A. & Wedzicha, J. A. Early therapy improves outcomes of exacerbations of chronic obstructive pulmonary disease. *Am. journal respiratory critical care medicine* **169**, 1298–1303 (2004).
5. Chandra, D., Tsai, C.-L. & Camargo Jr, C. A. Acute exacerbations of copd: delay in presentation and the risk of hospitalization. *COPD: J. Chronic Obstr. Pulm. Dis.* **6**, 95–103 (2009).
6. Lenferink, A. *et al.* Self-management interventions including action plans for exacerbations versus usual care in patients with chronic obstructive pulmonary disease. *Cochrane Database Syst. Rev.* (2017).
7. Punekar, Y. S., Shukla, A. & Müllerova, H. Copd management costs according to the frequency of copd exacerbations in uk primary care. *Int. journal chronic obstructive pulmonary disease* 65–73 (2014).
8. James Lind Alliance. COPD Exacerbation Priority Setting Partnership (2024).
9. Glyde, H. M. G., Morgan, C., Wilkinson, T. M., Nabney, I. T. & Dodd, J. W. Remote patient monitoring and machine learning in acute exacerbations of chronic obstructive pulmonary disease: Dual systematic literature review and narrative synthesis. *J. Med. Internet Res.* **26**, e52143 (2024).

10. Glyde, H. M., Blythin, A. M., Wilkinson, T. M., Nabney, I. T. & Dodd, J. W. Exacerbation predictive modelling using real-world data from the mycopd app. *Helixyon* (2024).
11. Chmiel, F. P. *et al.* 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* **10**, e26499 (2022).
12. Sanchez-Morillo, D., Fernandez-Granero, M. A. & Leon-Jimenez, A. Use of predictive algorithms in-home monitoring of chronic obstructive pulmonary disease and asthma: a systematic review. *Chronic respiratory disease* **13**, 264–283 (2016).
13. Al Rajeh, A. M. & Hurst, J. R. Monitoring of physiological parameters to predict exacerbations of chronic obstructive pulmonary disease (copd): a systematic review. *J. Clin. Medicine* **5**, 108 (2016).
14. Patel, N., Kinmond, K., Jones, P., Birks, P. & Spiteri, M. A. Validation of copdpredict™: unique combination of remote monitoring and exacerbation prediction to support preventative management of copd exacerbations. *Int. journal chronic obstructive pulmonary disease* 1887–1899 (2021).
15. Gkini, E. *et al.* Use of a personalised early warning decision support system for acute exacerbations of chronic obstructive pulmonary disease: Results of the “predict & prevent” phase iii trial. *COPD: J. Chronic Obstr. Pulm. Dis.* **22**, 2544719 (2025).
16. Saunders, T. *et al.* Objectively measured steps/day in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis. *J. Phys. Activity Heal.* **13**, 1275–1283 (2016).
17. Hao, Z., AghaKouchak, A., Nakhjiri, N. & Farahmand, A. Global integrated drought monitoring and prediction system (GIDMaPS) data sets. *figshare* <http://dx.doi.org/10.6084/m9.figshare.853801> (2014).

LaTeX formats citations and references automatically using the bibliography records in your .bib file, which you can edit via the project menu. Use the cite command for an inline citation, e.g.¹⁷.

For data citations of datasets uploaded to e.g. *figshare*, please use the `howpublished` option in the bib entry to specify the platform and the link, as in the `Hao:gidmaps:2014` example in the sample bibliography file.

Acknowledgements

This study was supported by the National Institute for Health and Care Research Bristol Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Author contributions statement

H.M.G.G. and J.W.D. co-conceived the study. H.M.G.G. led the study, designed the analytical approach, performed the analysis, interpreted the findings, prepared the figures, and wrote the original manuscript draft. A.M.B. provided my mhealth platform and data support, contributed to data access and interpretation, and revised the manuscript. T.M.A.W. supported the study design, contributed clinical expertise, supported interpretation of the findings, and revised the manuscript. I.T.N. contributed machine learning expertise, supported the analytical approach and interpretation of findings, and revised the manuscript. J.W.D. provided supervision and mentorship, supported the study design, contributed clinical expertise, supported interpretation of the findings, and revised the manuscript. All authors reviewed and approved the final manuscript.

Data availability

The data that support the findings of this study are available from my mhealth Limited, but restrictions apply to the availability of these data, which were used under licence for the current study and are therefore not publicly available. Data may be made available from the corresponding author upon reasonable request and with permission from my mhealth Limited, subject to appropriate approvals and data-sharing agreements.

Additional information

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests. Tom Wilkinson reports a relationship with my mhealth Ltd that includes: board membership. Alison Blythin reports a relationship with my mhealth Ltd that includes: employment. The other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

H.M.G.G was funded by the AI for Collective Intelligence Hub (AI4CI).

Table 1. Baseline demographics and clinical characteristics of the myCOPD cohort and wearable analytic subgroup. Values are mean $\pm$ SD, median [IQR], or n (%).

Characteristic	myCOPD cohort (N=4,223)	Wearable cohort (N=459)
Age, years, mean $\pm$ SD	69.5 $\pm$ 10.8	68.3 $\pm$ 9.8
Age, years, median [IQR]	70.0 [63.0, 77.0]	69.0 [62.0, 75.0]
Gender, n (%)		
Female	1,664 (39.4)	239 (52.1)
Male	1,664 (39.4)	209 (45.5)
Not recorded	895 (21.2)	11 (2.4)
Smoking status, n (%)		
Ex-smoker	2,591 (61.4)	266 (58.0)
Non-smoker	319 (7.6)	25 (5.4)
Smoker	583 (13.8)	46 (10.0)
Not recorded	730 (17.3)	122 (26.6)
Pack-years, mean $\pm$ SD	34.6 $\pm$ 25.0	31.1 $\pm$ 22.6
Pack-years, median [IQR]	30.0 [18.0, 45.0]	30.0 [18.0, 40.0]
CAT score, mean $\pm$ SD	18.4 $\pm$ 8.3	16.7 $\pm$ 8.2
CAT score, median [IQR]	18.0 [12.0, 24.0]	16.0 [11.0, 23.0]
mMRC score, n (%)		
0	304 (7.2)	39 (8.5)
1	1,040 (24.6)	166 (36.2)
2	1,179 (27.9)	122 (26.6)
3	780 (18.5)	80 (17.4)
4	550 (13.0)	44 (9.6)
Not recorded	370 (8.8)	8 (1.7)
GOLD group, n (%)		
A	336 (8.0)	40 (8.7)
B	1,645 (39.0)	125 (27.2)
C	136 (3.2)	17 (3.7)
D	1,349 (31.9)	104 (22.7)
Not recorded	757 (17.9)	173 (37.7)
Antibiotics and/or steroids in past year, n (%)		
False	1,017 (24.1)	130 (28.3)
True	2,844 (67.3)	324 (70.6)
Not recorded	362 (8.6)	5 (1.1)
Antibiotics/steroids courses in past year, mean $\pm$ SD	2.3 $\pm$ 2.5	2.1 $\pm$ 2.3
Antibiotics/steroids courses in past year, median [IQR]	2.0 [0.0, 3.0]	1.0 [0.0, 3.0]
Hospitalisation in past year, n (%)		
False	3,106 (73.5)	391 (85.2)
True	755 (17.9)	63 (13.7)
Not recorded	362 (8.6)	5 (1.1)
Hospitalisations in past year, mean $\pm$ SD	0.4 $\pm$ 1.0	0.2 $\pm$ 0.8
Days with step-count data, mean $\pm$ SD	–	118.0 $\pm$ 237.4
Days with step-count data, median [IQR]	–	12.0 [2.0, 80.5]

Table 2. Calendar time between observations in the modelling framework.

Group	Quantity	Mean (days)	Median (days)	IQR (days)	90th percentile (days)
All cohort rows	To next observation	4.71	1	1–2	4
All cohort rows	To 2nd next observation	8.35	2	2–3	9
All cohort rows	To 3rd next observation	11.42	3	3–5	15
Model rows	To next observation	3.14	1	1–1	3
Model rows	To 2nd next observation	5.97	2	2–3	7
Model rows	To 3rd next observation	8.58	3	3–5	12
Negative rows ($y = 0$)	To next observation	2.64	1	1–1	3
Negative rows ($y = 0$)	To 2nd next observation	4.89	2	2–3	6
Negative rows ($y = 0$)	To 3rd next observation	7.02	3	3–4	9
Positive rows ($y = 1$)	To next observation	13.87	2	1–10	27
Positive rows ($y = 1$)	To 2nd next observation	30.55	15	3–25	61
Positive rows ($y = 1$)	To 3rd next observation	45.89	18	16–37	95

Table 3. Calendar timing of the first future exacerbation-defining app entry among positive rows in the modelling framework.

Metric	Value
Positive rows, n	26,491
Mean time to first future exacerbation, days	35.38
Median time to first future exacerbation, days	5
IQR, days	2–21
90th percentile, days	84
Within 3 days, %	39.5
Within 7 days, %	56.0
Within 14 days, %	65.4
First exacerbation at next app entry, %	36.0
First exacerbation at second-next app entry, %	32.0
First exacerbation at third-next app entry, %	32.0

Table 4. Performance of the XGBoost model across outer cross-validation folds. Values are reported as mean and standard deviation (SD).

Metric	Mean	SD
PR-AUC	0.284	0.016
ROC-AUC	0.834	0.006
Precision	0.298	0.014
Recall	0.412	0.011

Table 5. Performance comparison for the wearable-cohort ablation study. Values are reported as mean (SD) across outer cross-validation folds.

Metric	Without steps	With steps
PR-AUC	0.162 (0.039)	0.163 (0.043)
ROC-AUC	0.785 (0.022)	0.787 (0.026)
Precision	0.214 (0.059)	0.214 (0.049)
Recall	0.264 (0.011)	0.261 (0.045)

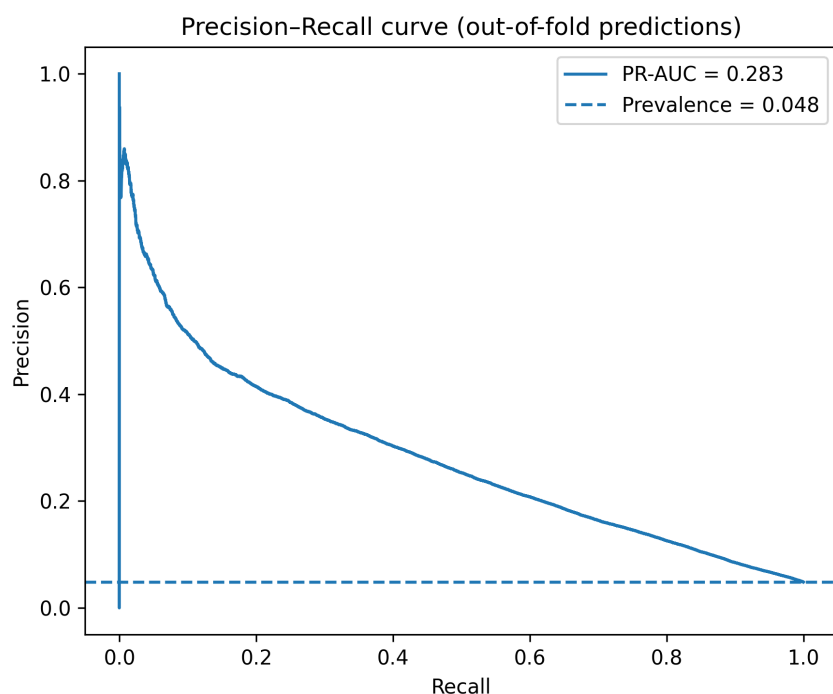


Figure 1. Precision–recall performance of the XGBoost model. Precision–recall curve derived from pooled out-of-fold predictions across grouped nested cross-validation folds for the task of predicting whether an exacerbation (symptom score 3 or 4) would occur by the patient’s next 1–3 app entries. The dashed horizontal line indicates the positive class prevalence in the analytic dataset (4.8%). The pooled out-of-fold area under the precision–recall curve was 0.283.

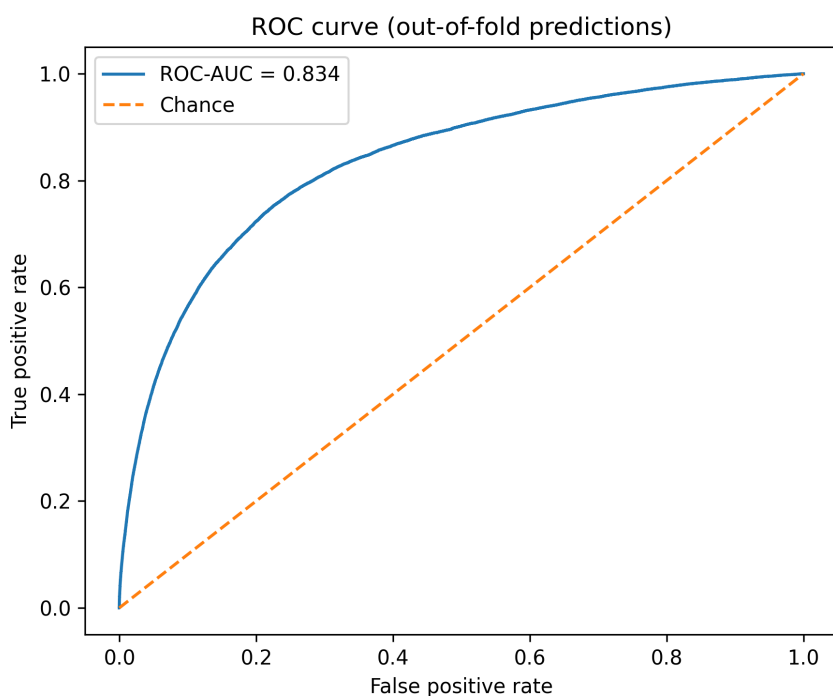


Figure 2. Receiver operating characteristic performance of the XGBoost model. Receiver operating characteristic curve derived from pooled out-of-fold predictions across grouped nested cross-validation folds for the task of predicting whether an exacerbation (symptom score 3 or 4) would occur by the patient’s next 1–3 app entries. The dashed diagonal line indicates chance performance. The pooled out-of-fold area under the receiver operating characteristic curve was 0.834.

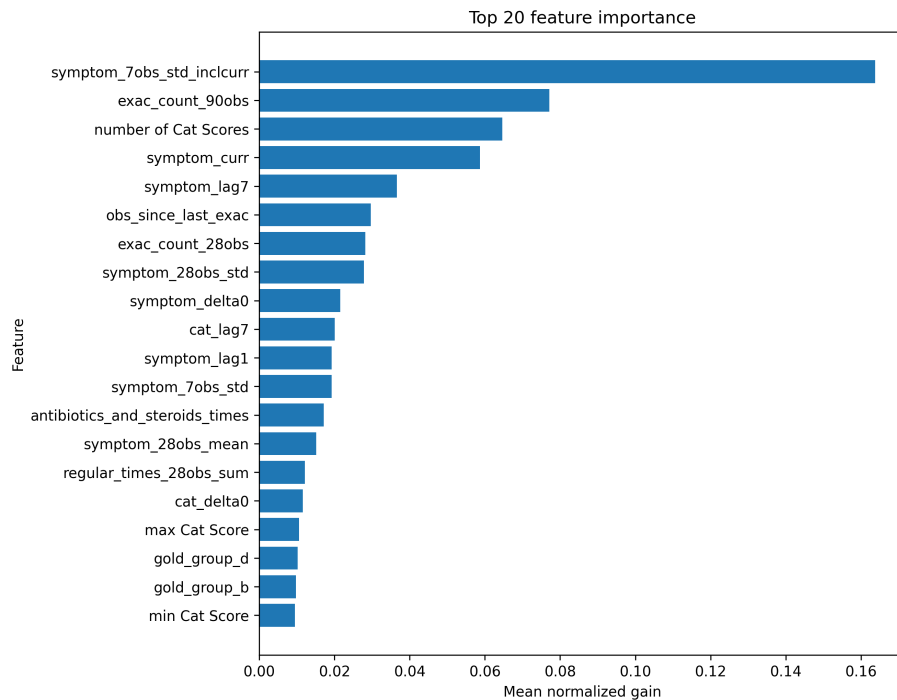


Figure 3. Bars show the mean normalised gain for each predictor, where gain reflects the average improvement in the model’s split objective attributable to that feature. Larger values indicate greater contribution to model performance. Features include symptom dynamics, recent exacerbation history, CAT-derived summaries, and demographic/clinical measures.

Table 6. Cohort size and overall step-count summary statistics (non-missing) for the wearable analytic cohort.

Metric	Value
Patients kept	459
Rows	148,408
Non-missing step-count observations (<i>n</i>)	54,153
Mean steps	5,566
Median steps	4,649
Standard deviation	4,025
5th percentile (P05)	530
25th percentile (P25)	2,543
75th percentile (P75)	7,804
95th percentile (P95)	13,573
Exacerbation onset events found (window=14 days)	1,708

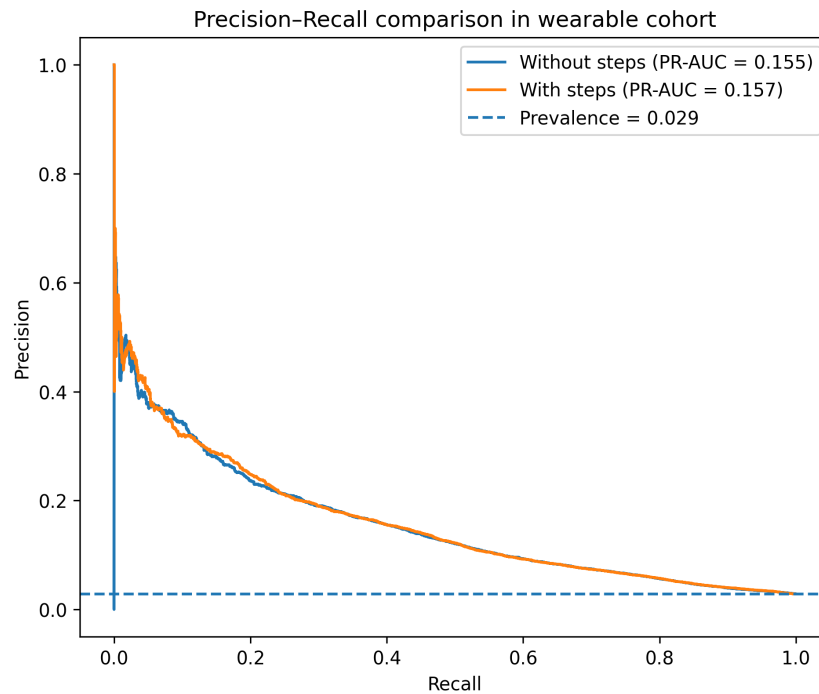


Figure 4. Precision–recall comparison for the wearable-cohort ablation analysis. Precision–recall curves derived from pooled out-of-fold predictions across grouped nested cross-validation folds for the task of predicting whether an exacerbation (symptom score 3 or 4) would occur by the patient’s next 1–3 app entries. Both models were trained and evaluated in the same wearable analytic cohort, with one model excluding step-derived predictors and the other including them. The dashed horizontal line indicates the positive class prevalence in the wearable analytic dataset (2.9%). Inclusion of step features did not materially improve discrimination, with very similar pooled out-of-fold PR-AUC values (0.157 with steps vs 0.155 without steps).

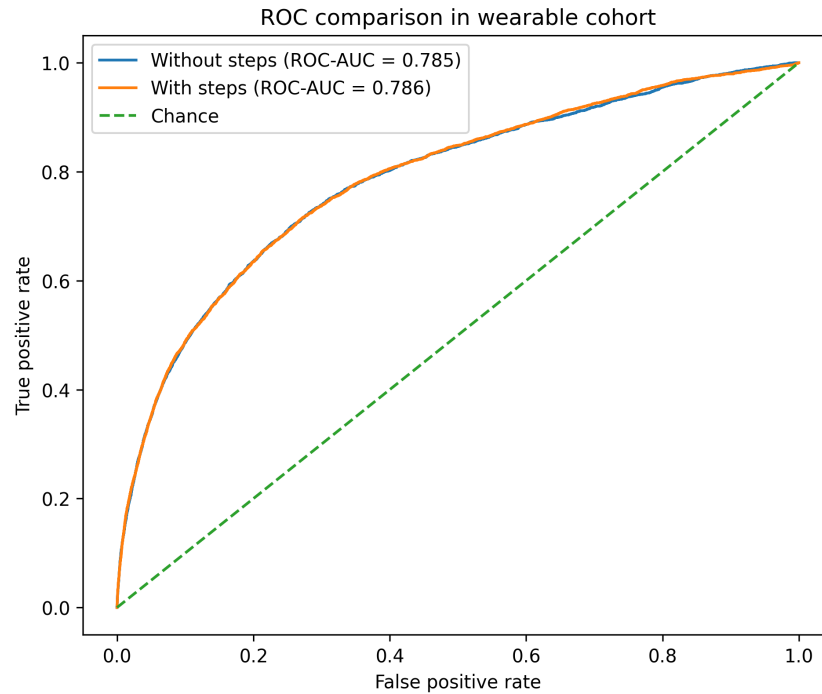


Figure 5. Receiver operating characteristic comparison for the wearable-cohort ablation analysis. Receiver operating characteristic curves derived from pooled out-of-fold predictions across grouped nested cross-validation folds for the task of predicting whether an exacerbation (symptom score 3 or 4) would occur by the patient's next 1-3 app entries. Both models were trained and evaluated in the same wearable analytic cohort, with one model excluding step-derived predictors and the other including them. The dashed diagonal line indicates chance performance. Inclusion of step features did not materially improve discrimination, with very similar pooled out-of-fold ROC-AUC values (0.786 with steps vs 0.785 without steps).

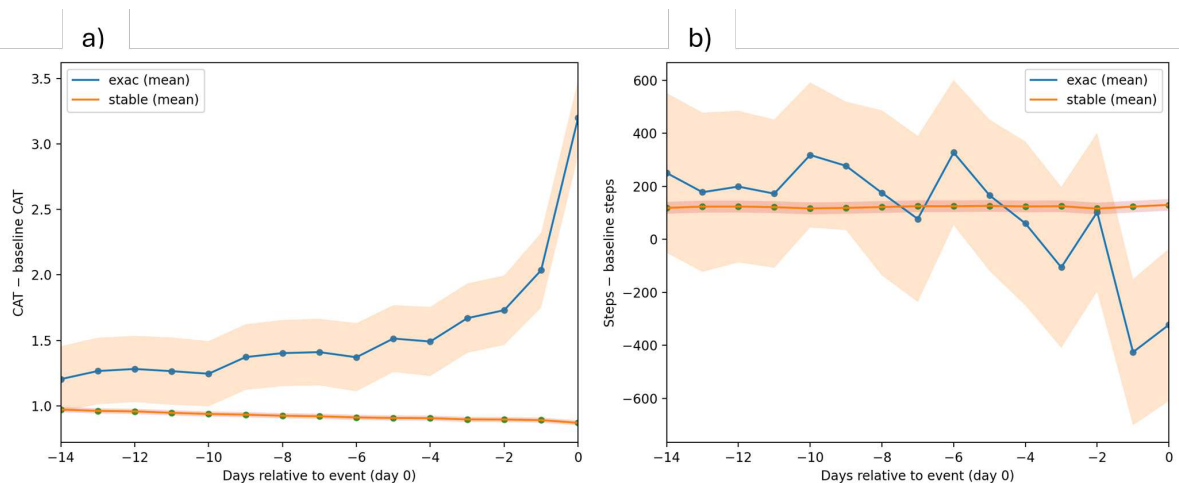


Figure 6. Average trajectories of symptoms and physical activity change from a stable baseline in the 14 days preceding symptom-defined exacerbation events and stable health in the wearable cohort. (a) Mean change in COPD Assessment Test (CAT) score relative to individual baseline (CAT - baseline CAT). (b) Mean change in daily step count relative to individual baseline steps. Blue lines represent trajectories prior to exacerbation and orange lines represent trajectories during periods of stable health; points denote daily mean values. Shaded regions indicate 95% confidence intervals across participants. Day 0 corresponds to the exacerbation event.