PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

Title (Provisional)

PRODIGY-ILD: A protocol for a prospective cohort study to predict outcomes in patients with interstitial lung disease using digital health technologies.

Authors

Gunne, Emer; Holden, Sinead; Franciosi, Alessandro N; Keane, Michael; McCarthy, Cormac; Doran, Peter

---

VERSION 1 - REVIEW

Reviewer	**1**
Name	**Dowman, Leona**
Affiliation	**Monash University, Allergy, Immunology and Respiratory Medicine**
Date	**15-Oct-2024**
COI	**None**

The protocol is very clearly written and in depth.

Minor comments

- Questionnaires - less content, more succint decription. Too nuch detail included

- Table 2 and Table 4 provide similar information. Don't both is necessary. The detail of table 2 is not required. Again more succint description

In the discussion, there is a alot of 'we will' statements. It is unknown if you will achieve what you set out to do. We 'will' should be replacement with we 'aim'

Currently you do not know what the predictions will be nor whether your data can produce usuable predictions therefore tone down language 'These predictions will be of significant value in clinical management...'

dpnt knpw if will develop predictions

What do you mean by ? 'Critically these models will be utilized as synthetic controls for clinical trials in these rare diseases, thereby enabling additional investigation of new therapeutic strategies'

You state a strength is passive collectively data, thereofre less burden however the participants need to answer a questionnaire weekly or monthly for the nextg 3 years. I would say that is onerous and more of a limitation

you have not completed this study therefore you don't know if it will be sucessful not translatable to other disease groups

50 is a small sample, particularly for prediction modelling. there may not be enough data for adequate prediction modelling. This should be listed as a limitation.

Reviewer	**2**
Name	**Harrison, Megan**
Affiliation	**The University of Sydney**
Date	**24-Oct-2024**
COI	**None**

Excellent and well-set out protocol. No concerns.

Reviewer	**3**
Name	**Albrecht, Steffen**
Affiliation	**The University of Auckland, School of Computer Science**
Date	**31-Dec-2024**
COI	**None**

This protocol gives a clear description of the goals of the study and provides a comprehensive description of the variables/measures to be collected during the study over three years and for which purpose they are collected, creating predictive models for the outcome of individual patients diagnosed with IPF or PPF.

The overall structure and aims of the study are well-described throughout the protocol. However, there are some details missing that would be valuable for the broad readership or scientists who might be interested in applying a similar approach of using wearables to obtain patient data. In my opinion, the protocol would benefit from discussing the following aspects, which I try to summarize in the following paragraphs.

Will patients be informed about the intermediate predictions from the specialist and/or models based on the first or second annual visit? How much impact could potential lifestyle changes, motivated by the intermediate predictions, have on the whole study? Can you speculate on this and assess if this should be addressed in the statistical analyses applied?

Would it be possible to provide the participation agreement with the supplementary material?

The description of the models should be more detailed, outlining a clear plan of what strategies will be applied. Regarding machine learning, will this be approached by unsupervised or supervised algorithms? If supervised algorithms are going to be used, will the labels be set by the specialists based on the annual visits? If labels will be used, are those categorical or continuous? For continuous labels, 50 samples might be too low. If labels are categorical, how many categories, considering especially the models for disease progression and acute exacerbation?

What exactly are the dynamic models? Please provide some names for the concepts planned to be used.

Why 50 patients? I assume and understand that more participants require more financial resources and staff to conduct the study. However, it appears to me that it takes time and effort to set up this study, and especially when it comes to machine learning analyses, they would strongly benefit from another 50 or better 100 samples. Increasing the number of participants is probably not possible; however, please discuss the limiting factors in more detail; what exactly is the challenge in scaling up? For scientists who are interested in pursuing a similar approach for other diseases it would be precious to share your experience.

Minor comments and questions:

The abbreviation PROM should be introduced earlier in the paragraph on page 8, line 3.

Table 2, Distance Walking Running (km): is the average taken per day for the last 7 days or the average over all 7 days? Is it possible to differentiate between walking and running? If it was possible to differentiate, would it be sensible to keep these as two different variables?

Table 2, Exercise minutes: Is it possible to differentiate between cardio training and strength training? Or will the cardio part be covered by “Distance Walking Running”? Please provide more details here.

Reviewer	**4**
Name	**Liao, Weiqi**
Affiliation	**University of Leicester**
Date	**03-Jan-2025**
COI	**None**

Comments for the article titled “PRODIGY-ILD: Data driven predicted outcomes in interstitial lung disease” submitted by Emer Gunne et al to the BMJ Open.

There are significant methodological weaknesses in this submitted article, to name a few:

1. “Interstitial lung disease (ILD) is a heterogeneous group (over 200 disorders)” – a clear clinical definition is needed, i.e. including the ICD-10 codes and other relevant clinical codes (e.g. SNOMED CT codes) in the appendix is necessary, which will be helpful for clinical colleagues, researchers, and readers.

2. Sample size – why 50? It is necessary to explain how the sample size was calculated, and how the authors got 50.

3. What kinds of measures will the authors take to minimise and mitigate selection bias of participants?

4. “Participants will be followed for 3 years” – How would the authors plan to deal with loss to follow-up and missing data? The authors need to discuss this and mention in the statistical analysis plan.

5. “Patients to wear a smart watch for the duration of the study” – What is the compliance of wearing the device? How would the authors plan to mitigate the impact of different wear durations among patients for the prediction? Discussion of methodological and practical considerations will be helpful.

6. “Interim analysis will be conducted at 6 monthly intervals.” – What is the rationale of interim analysis? What do the authors expect to get from the interim analysis for prediction models?

7. The three outcomes: what would the authors consider positive events for “disease progression” and “acute exacerbation”? As to “mortality”, is it all-cause mortality or cause-specific? Clearer definitions for the three outcomes are needed. ICD-10 codes and other relevant clinical codes in the appendix are necessary.

8. What kinds of methods will the authors use for prediction? How would the authors deal with missing data when developing prediction models? It is vague just saying “traditional regression analysis along with dynamic prediction methods”. A more detailed description and relevant references are needed.

9. How would the authors plan to evaluate the developed model? A detailed description of validation and evaluation of the prediction model is needed.

10. Are authors confident to have sufficient positive events with a sample size of only 50?

The current section on statistical methods is not clear. Above are some major limitations. I strongly encourage the authors to consult with a statistician and get advice for the methodology. There are lots of advanced methodological issues needed to be considered.

VERSION 1 - AUTHOR RESPONSE

Reviewer: 1

Dr. Leona Dowman, Monash University, Austin Health

Comments to the Author:

The protocol is very clearly written and in depth.

Response: thank you for the positive comments on the protocol.

Minor comments

1. Questionnaires - less content, more succint decription. Too much detail included

Response: This text has now been reduced as follows.

1.*Breathlessness Questionnaire – modified Medical Research Council dyspnoea scale (mMRC)*[32,33], a self-rating tool used to assess the degree of baseline functional disability due to breathlessness on a scale from 0 to 4.

2.*Cough Severity Questionnaire – VAS*[34], A Visual analogue scale for cough severity where the patient indicates the severity of their cough over the last two weeks on a scale of 0 to 100. A ≥ 30-mm reduction in cough is considered a clinically meaningful change threshold for clinical trials in chronic cough[34].

3.*Cough Quality of Life Questionnaire – The Leicester Cough questionnaire (LCQ)*[35], is a19-item cough questionnaire comprising three health domains: physical, psychological, and social to assess the impact of cough in the previous two weeks which takes less than 5 minutes to complete.

4.*Fatigue Questionnaire – Fatigue Assessment Scale (FAS)*[36,37], FAS is a 10-item self-report scale (1-5) evaluating symptoms of physical and mental fatigue. A total FAS score < 22 indicates no fatigue, a score ≥ 22 indicates fatigue.

5.*King's Brief Interstitial Lung Disease – KBILD*[38,39], is a self-completed 15-item validated ILD-specific measure of health-related quality of life consists of three domains Physiological (3,5,6,8,12,14), Breathlessness and activities (1,4,11,13) and chest symptoms (2,7,9). The KBILD domain and total score ranges are 0-100; 100 represents best health status.

2. Table 2 and Table 4 provide similar information. Don't both is necessary. The detail of table 2 is not required. Again more succint description

Response: we have removed Table 2 and provided as supplementary material

3. In the discussion, there is a alot of 'we will' statements. It is unknown if you will achieve what you set out to do. We 'will' should be replacement with we 'aim' Currently you do not know what the predictions will be nor whether your data can produce usuable predictions therefore tone down language 'These predictions will be of significant value in clinical management...'

dpnt knpw if will develop predictions

Response: Thank you for highlighting this. We have now toned down the language with “we aim” and “predictions are expect” as suggested;

Discussion

The primary objective of this study is to develop useful prediction models for clinical outcomes in ILD using a combination of clinical, physiological, activity and patient reported data fields. Through the comprehensive collection and analysis of these data sets, we aim to identify better predictors of disease progression, acute exacerbations, and mortality. Predictions will be compared to actual outcomes to validate prediction models. These predictions are expected to offer meaningful value for clinical management by providing clinicians with an improved tool for prognostication.

4. What do you mean by ? 'Critically these models will be utilized as synthetic controls for clinical trials in these rare diseases, thereby enabling additional investigation of new therapeutic strategies'

Response: sentence has been replaced to provide clarity around individual patients acting as their own control as follows:

By developing a clinical prediction model of ILD outcomes, we anticipate that such predictors will be useful as patient level controls for clinical trials, where predicted outcomes for individual patients can be compared to actual recorded outcomes following an intervention. In this manner individual patients will act as their own control, enabling additional investigation of new therapeutic strategies.

5. You state a strength is passive collectively data, thereofre less burden however the participants need to answer a questionnaire weekly or monthly for the nextg 3 years. I would say that is onerous and more of a limitation

Response: wearable biometric data is collected passively. Questionnaires have been piloted to ensure they take a minimum amount of time. Weekly questionnaires take less than 30 seconds to complete, monthly and three-monthly questionnaires take approximately 5 mins each to complete. Please see modified text as follows:

A strength of this study lies in the ability of the wearable to passively collect a large quantity of biometric data without burdening the patient to record active measures. However, the patient reported outcome measures may represent a burden to patients.

6. you have not completed this study therefore you don't know if it will be sucessful not translatable to other disease groups

Response: This has now been removed.

~~On successful completion of this study, it is envisaged that this study method could be applied~~

~~to other disease groups. This could be particularly useful for diseases in which clinical course varies by individual and rare disease populations where it is difficult to recruit large numbers for clinical trials of new treatments.~~

7. 50 is a small sample, particularly for prediction modelling. there may not be enough data for adequate prediction modelling. This should be listed as a limitation.

Existing strengths and limitations section sentence included in main text:

Strengths and limitations of this study

- Interstitial lung disease is a rare disease therefore our sample size is small at 50 participants; however, we will collect a large amount of data points.

Suggested strengths and limitations section sentence to include in main text:

Strengths and limitations of this study

- Interstitial lung disease is a rare disease; therefore, our sample size is small at 50 participants, which may limit the robustness of prediction modeling. However, we will collect a large volume of data points and anticipate a high event rate in this population over the three-year follow-up period.
- By using the rule of thumb of 10 events per variable, we expect to have sufficient information for the development of the model; though we acknowledge a larger data set and patient population will likely be required for the further validation of the predictor

Reviewer: 2
Dr. Megan Harrison, The University of Sydney
Comments to the Author:
Excellent and well-set out protocol. No concerns.
Response: Thank you for taking the time to review our protocol, and for the positive comments.

Reviewer: 3
Dr. Steffen Albrecht, The University of Auckland
Comments to the Author:
This protocol gives a clear description of the goals of the study and provides a comprehensive description of the variables/measures to be collected during the study over three years and for which purpose they are collected, creating predictive models for the outcome of individual patients diagnosed with IPF or PPF.
The overall structure and aims of the study are well-described throughout the protocol. However, there are some details missing that would be valuable for the broad readership or scientists who might be interested in applying a similar approach of using wearables to obtain patient data. In my opinion, the protocol would benefit from discussing the following aspects, which I try to summarize in the following paragraphs.
1. Will patients be informed about the intermediate predictions from the specialist and/or models based on the first or second annual visit? How much impact could potential lifestyle changes, motivated by the intermediate predictions, have on the whole study? Can you speculate on this and assess if this should be addressed in the statistical analyses applied?

Response: We have not stated that participants will be informed about the intermediate predictions in the patient information leaflet. Our ethical approval for this study does not provide for the return of individual results.

2. Would it be possible to provide the participation agreement with the supplementary material?

Response: Yes, we have now included the Informed Consent Form/Patient Information Leaflet as part of the supplementary material

3. The description of the models should be more detailed, outlining a clear plan of what strategies will be applied. Regarding machine learning, will this be approached by unsupervised or supervised algorithms? If supervised algorithms are going to be used, will the labels be set by the specialists based on the annual visits? If labels will be used, are those categorical or continuous? For continuous labels, 50 samples might be too low. If labels are categorical, how many categories, considering especially the models for disease progression and acute exacerbation?

What exactly are the dynamic models? Please provide some names for the concepts planned to be used.

Response: we have now given more detail outlining the models and the different approaches used under the Data analysis section as follows:

Data analysis

The overall objective of this study is to leverage the data collected to generate predictive models of outcome. Specifically, we will seek to develop individual patient level models, of the following categorical outcomes of

1. Acute exacerbation
2. Hospitalisation
3. Mortality

Models will be developed using a number of different approaches, with all models evaluated to determine best performance. These may include:

1) Logistic regression for mortality, acute exacerbation and hospitalization modelling the relationship between independent variables (for example biometric or PROM data) and each of the model outcomes of mortality, acute exacerbation and hospitalization.
2) Proportional Hazards to predict time to event (mortality) estimate the hazard ratio for input variables.
3) Linear regression, for continuous outcomes (exacerbations) where a linear relationship between variables (for example cough questionnaire) is anticipated
4) Survival Analysis: For time to event outcomes, where outcome (for example mortality) will be explored in the context of data collected from patients

To complement these approaches, we will also utilize machine learning approaches to explore

more complex models, including

1) Random Forest, to classify outcomes based on biometric or PROM variable data, as both single and multiple predictors
2) Support Vector Machines- classifying patients as progressors or not based on multiple variables including PROM and biometric data, as well as clinical phenotype
3) K-Nearest neighbours - to make predictions about individual patients based on similarities in the dataset to other patients with known outcomes
4) Pattern discovery: Unsupervised machine learning methods will be utilized to uncover patterns in the data applying methods such as clustering to identify for example high symptom burden patients from PROMs data

All models will be evaluated and compared to determine performance. Key evaluation measures will include

- Accuracy: based on assessment of predicted vs observed outcomes, as participants continue in the cohort
- Sensitivity and Specificity: analysis of true predictors, based on comparison with actual participants outcome
- C-Statistics: to measure performance of models in discriminating between outcomes

Six-month patient data will be collected and analysed. Outcomes predicted will be compared with actual outcomes recorded at month twelve, allowing refinement, enhancement, and validation of the developed models. Data analysis will proceed as follows:

Step 1: Data quality control and validation:

Individual patient level data will be reviewed for completeness. Given the lack of previous studies integrating clinical, physiological, activity and patient reported data in this population, we will employ a conservative approach to data completeness with a requirement for at least 70% completeness for key variables. Missing data will be summarised and dealt with through Case deletion, where the 70% threshold is not reached. Rational substitution will be employed where possible. For missing data at random, arithmetic imputation methods will be employed including worst case imputation for dropouts and interpolation/extrapolation where prior and after data is available.

Step 2: Descriptive analysis: Data from activity and physiological measurements will be summarised as per Table. 4. Data summarisation will involve providing a concise overview of key characteristics of the dataset including central tendency values, variance to give insights into the spread of the data and frequency distributions, assessing for normality, while highlighting any notable patterns or peaks in the dataset.

Step 3: Feature selection, potential predictors of progression, exacerbation and mortality will

be reviewed by specialist respiratory clinicians. The rationale behind the feature selection process will be clearly documented and the chosen predictors will be rigorously validated to ensure clinical relevance to the target outcomes. In parallel, activity and physiological measurements (individual and composite, single point and time-rolling average trends, raw values and individual-normalised) will be analysed to identify potential features, predictive of clinical outcome.

Step 4: Predicted outcomes, the prediction model will be simplified by the elimination of unnecessary variables. Methods such as correlation analysis will be performed to identify potential redundancies. Statistical tests will be used to rank features and their association with the target outcomes.

Step 5: Validation: we will evaluate the predictive model's effectiveness by comparing its performance to actual events per predictor both within individual and across individuals. By doing this we will gain insights into the practical relevance of individual predictors, and we will refine the model accordingly.

4. Why 50 patients? I assume and understand that more participants require more financial resources and staff to conduct the study. However, it appears to me that it takes time and effort to set up this study, and especially when it comes to machine learning analyses, they would strongly benefit from another 50 or better 100 samples. Increasing the number of participants is probably not possible; however, please discuss the limiting factors in more detail; what exactly is the challenge in scaling up? For scientists who are interested in pursuing a similar approach for other diseases it would be precious to share your experience.

Response: Further justification has now been added to the sample size section of the article as follows

The primary objective of this study is to develop prediction models for clinical outcomes (disease progression, acute exacerbations and mortality) by looking at patterns in data types and the significance of those patterns for individual participants. The development of a prediction model requires firstly a developmental dataset containing the likely predictor values, which are used to develop the prediction model. The sample size for the development data set must be sufficiently large to enable a model to be developed which can subsequently be tested. In this study we are utilizing the 10 events per variable rule of thumb to estimate the sample size for the developmental dataset (Peduzzi et al[1,2]). Using this approach, we estimate that 50 participants will provide sufficient events to enable the development of the prediction modeling. It should be noted that given the progressive nature of this disease, and the proposed 3 year follow up, we anticipate that all patients will

progress with exacerbations, hospitalisations and death occurring within the study time frame, in most participants. Depending on the predictor variable ultimately uitilised, a larger sample size for the validation set will be enrolled. The sample size was determined based on the need to achieve reliable and generalizable results for this objective. Sampling for rare diseases is inherently challenging due to limited patient populations. Given the prevalence of ILD (rare disease) a sample size of 50 is feasible within the available time frame and resources.

References:

1.Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol1995;48:1503-10.

2.Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol1996;49:1373-9

Minor comments and questions:

5. The abbreviation PROM should be introduced earlier in the paragraph on page 8, line 3.

Response: the abbreviation PROM is introduced in the paragraph below

Digital health technology

Study data will be collected and managed using REDCap electronic data capture hosted at UCD Clinical Research Centre[30,31]. REDCap (Research Electronic Data Capture) is a secure, web-based software platform designed to support data capture for research studies. As part of enrolment patients will be assigned a unique study identification number, in the REDCap study database. The unique study identification number will be used to name the participants wearable device and to anonymously track their data throughout the study. The study team will provide each participant with a wearable device (Apple watch series 6 or above) and an iPhone series 8 or above if required. Training and on-boarding will take place at the baseline visit. In addition, MyCap, a companion App to RedCap research software, will be used to collected Patient reported outcome measure (PROM) data electronically, to enable comprehensive remote PROM collection over time[32].

6. Table 2, Distance Walking Running (km): is the average taken per day for the last 7 days or the average over all 7 days? Is it possible to differentiate between walking and running? If it was possible to differentiate, would it be sensible to keep these as two different variables?

Distance Walking Running (km)	Calculates an average distance you have walked and run over the last 7 days

Walking and running distance is one of the variables calculated by Apple watch, the iPhone health app displays this as a daily average metric by day, week, month, 6 months and year. The export from the Apple watch contains the detail of each distance taken during the day but does not separate out which distance was walked or ran. We cannot make any changes to this variable but can use the raw data to calculate daily distance and a 7 day rolling average.

7. Table 2, Exercise minutes: Is it possible to differentiate between cardio training and strength training? Or will the cardio part be covered by "Distance Walking Running"? Please provide more details here.

Exercise minutes	Measure of how many minutes of brisk activity you do

Exercise minutes is one of the variables calculated by Apple watch, the iPhone health app displays daily average metric by day, week, month, 6 months and year. Every full minute of movement equal to or exceeding the intensity of a brisk walk counts toward daily Exercise minutes. The export from the Apple watch contains the detail of each exercise minute during the day but does not separate out cardio training and strength training. We cannot make any changes to this variable but can use the raw data to calculate daily exercise minutes and a 7 day rolling average.

Reviewer: 4

Dr. Weiqi Liao, University of Leicester

Comments to the Author:

Comments for the article titled "PRODIGY-ILD: Data driven predicted outcomes in interstitial lung disease" submitted by Emer Gunne et al to the BMJ Open.

There are significant methodological weaknesses in this submitted article, to name a few:

1. "Interstitial lung disease (ILD) is a heterogeneous group (over 200 disorders)" – a clear clinical definition is needed, i.e. including the ICD-10 codes and other relevant clinical codes (e.g. SNOMED CT codes) in the appendix is necessary, which will be helpful for clinical colleagues, researchers, and readers.

Response: this study is targeted at Interstitial Lung Disease patients with Idiopathic Pulmonary Fibrosis (ICD-10 code J84.112) or with Progressive Pulmonary Fibrosis (ICD-10 code J84.170)

Existing article references 3 and 6 clinical guidelines

3. Raghu G et al. Idiopathic Pulmonary Fibrosis (an Update) and Progressive Pulmonary Fibrosis in Adults: An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. Am J Respir Crit Care Med. 2022 May 1; 205(9): e18-e47. Published online 2022 May1. Doi:10.1164/rccm.202202-0399ST

6. Raghu G, Collard HR, Egan JJ, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. Am J Respir Crit Care Med 2011; 183: 788–824.

These two articles provide a comprehensive background to Idiopathic Pulmonary Fibrosis and Progressive Pulmonary Fibrosis, disease progression, acute exacerbations and mortality as interpreted by our study.
Reference 3: This American Thoracic Society, European Respiratory Society, Japanese Respiratory Society, and Asociación Latinoamericana de Tórax guideline updates prior idiopathic pulmonary fibrosis (IPF) guidelines and addresses the progression of pulmonary fibrosis in patients with interstitial lung diseases (ILDs) other than IPF.

2. Sample size – why 50? It is necessary to explain how the sample size was calculated, and how the authors got 50.

Response: Further justification has now been added to the sample size section of the article as follows

The primary objective of this study is to develop prediction models for clinical outcomes (disease progression, acute exacerbations and mortality) by looking at patterns in data types and the significance of those patterns for individual participants. The development of a prediction model requires firstly a developmental dataset containing the likely predictor values, which are used to develop the prediction model. The sample size for the development data set must be sufficiently large to enable a model to be developed which can subsequently be tested. In this study we are utilizing the 10 events per variable rule of thumb to estimate the sample size for the developmental dataset (Peduzzi et al[1,2]). Using this approach, we estimate that 50 participants will provide sufficient events to enable the development of the prediction modeling. It should be noted that given the progressive nature of this disease, and the proposed 3 year follow up, we anticipate that all patients will progress with exacerbations, hospitalisations and death occurring within the study time frame, in most participants. Depending on the predictor variable ultimately uitilised, a larger sample size for the validation set will be enrolled. The sample size was determined based on the need to achieve reliable and generalizable results for this objective. Sampling for rare diseases is inherently challenging due to limited patient populations. Given the prevalence of

ILD (rare disease) a sample size of 50 is feasible within the available time frame and resources.

References:
1.Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol1995;48:1503-10.
2.Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol1996;49:1373-9

3. What kinds of measures will the authors take to minimise and mitigate selection bias of participants?

Response: Participants will be recruited from a tertiary referral ILD clinic based on diagnosis of interstitial lung disease. All eligible patients attending the clinical will be invited to participate However we are limited to clinic attendees willing to take part in the study during the recruitment timeframe window. Every effort will be made to recruit a balanced mix of gender, ethnicity, age and disease category. Details of participants characteristics will be included in the manuscript.

4. "Participants will be followed for 3 years" – How would the authors plan to deal with loss to follow-up and missing data? The authors need to discuss this and mention in the statistical analysis plan.

Response: We anticipate a proportion of patients will be lost to follow up and there will be missing PROMs, watch and clinical data. We intend to minimise missingness by ensuring robust follow-up methods using PROM reminders and engaging with patients to collect watch data on a regular basis to ensure smartwatch wear-time compliance. The specific detail of missing data imputation methods will be outlined in our results article.

5. "Patients to wear a smart watch for the duration of the study" – What is the compliance of wearing the device? How would the authors plan to mitigate the impact of different wear durations among patients for the prediction? Discussion of methodological and practical considerations will be helpful.

Response: existing paragraph included in main text:

Step 1: Data quality control and validation:

Individual patient level data will be reviewed for completeness. Given the lack of previous studies integrating clinical, physiological, activity and patient reported data in this population, we will employ a conservative approach to data completeness with a requirement for at least 70% completeness for key variables. Missing data will be summarised and dealt with through Case deletion, where the 70% threshold is not reached. Rational substitution will be employed where possible. For missing data at random, arithmetic imputation methods will be employed including worst case imputation for dropouts and interpolation/extrapolation where prior and after data is available.

The patient information leaflet requests that participants wear the watch for 20 hours per day. Other studies have required a wear time of at least 10 hours per day. For example Days are excluded from analysis if wear time is less that 10 hours. Apple watch takes a heart rate reading every 5 minutes, we will use this to calculate wear time. All of the detail of watch wear time will be outlined in the results article per patient. Individual patient models do not require patients to have the same wear time.

6. "Interim analysis will be conducted at 6 monthly intervals." – What is the rationale of interim analysis? What do the authors expect to get from the interim analysis for prediction models?

Response: thank you for highlighting this we have changed the abstract to clarify it is not interim analysis but that the first 6 months data will be used to develop clinical prediction models.

Modified paragraph to include in abstract:

Participants will be followed for 3 years to assess rate of disease progression, occurrence of acute exacerbations and mortality. Initial data will be used to develop clinical prediction models. These models will be further evaluated for accuracy using regular follow up data.

(i.e. we will analyse data at 6 months, make predictions, then compare to actual outcomes at 12, 18, 24 months until the end of the study.)

7. The three outcomes: what would the authors consider positive events for "disease progression" and "acute exacerbation"? As to "mortality", is it all-cause mortality or cause-specific? Clearer definitions for the three outcomes are needed. ICD-10 codes and other relevant clinical codes in the appendix are necessary.

Response: Disease progression and acute exacerbation definitions are referenced below. Mortality is all cause mortality.

Reference 3: Raghu G et al (2022)

In a patient with ILD of known or unknown aetiology other than Idiopathic Pulmonary Fibrosis who has radiological evidence of pulmonary fibrosis, Progressive pulmonary Fibrosis is

defined as at least two of the following three criteria occurring within the past year with no alternative explanation:

Worsening respiratory symptoms

Physiological evidence of disease progression (either of the following):

a. Absolute decline in forced vital capacity ⩾5% predicted within 1 year of follow-up

b. Absolute decline in diffusion capacity of the lung for carbon monoxide (corrected for haemoglobin) ⩾10% predicted within 1 year of follow-up

Radiological evidence of disease progression (one or more of the following):

a. Increased extent or severity of traction bronchiectasis and bronchiolectasis

b. New ground-glass opacity with traction bronchiectasis

c. New fine reticulation

d. Increased extent or increased coarseness of reticular abnormality

e. New or increased honeycombing

f. Increased lobar volume loss

Reference 9 (Collard et al 2016):

Acute exacerbations typically develop in less than 1 month and are accompanied by new radiologic abnormalities on high resolution computed topography such as diffuse, bilateral ground-glass opacification with or without consolidation, and the absence of other obvious clinical causes like fluid overload, left heart failure, or pulmonary embolism.

8. What kinds of methods will the authors use for prediction? How would the authors deal with missing data when developing prediction models? It is vague just saying "traditional regression analysis along with dynamic prediction methods". A more detailed description and relevant references are needed.

Response: As per answer to query 4 & 5 and we have now given more detail outlining the models and the different approaches used under the Data analysis section as follows:

Data analysis

The overall objective of this study is to leverage the data collected to generate predictive models of outcome. Specifically, we will seek to develop individual patient level models, of the following categorical outcomes of

1. Acute exacerbation
2. Hospitalisation
3. Mortality

Models will be developed using a number of different approaches, with all models evaluated to determine best performance. These may include:

5) Logistic regression for mortality, acute exacerbation and hospitalization modelling the relationship between independent variables (for example biometric or PROM data) and each of the model outcomes of mortality, acute exacerbation and hospitalization.

6) Proportional Hazards to predict time to event (mortality) estimate the hazard ratio for input variables.
7) Linear regression, for continuous outcomes (exacerbations) where a linear relationship between variables (for example cough questionnaire) is anticipated
8) Survival Analysis: For time to event outcomes, where outcome (for example mortality) will be explored in the context of data collected from patients

To complement these approaches, we will also utilize machine learning approaches to explore more complex models, including

4. Random Forest, to classify outcomes based on biometric or PROM variable data, as both single and multiple predictors
5. Support Vector Machines- classifying patients as progressors or not based on multiple variables including PROM and biometric data, as well as clinical phenotype
6. K-Nearest neighbours - to make predictions about individual patients based on similarities in the dataset to other patients with known outcomes
7. Pattern discovery: Unsupervised machine learning methods will be utilized to uncover patterns in the data applying methods such as clustering to identify for example high symptom burden patients from PROMs data

All models will be evaluated and compared to determine performance. Key evaluation measures will include

- Accuracy: based on assessment of predicted vs observed outcomes, as participants continue in the cohort
- Sensitivity and Specificity: analysis of true predictors, based on comparison with actual participants outcome
- C-Statistics: to measure performance of models in discriminating between outcomes

Six-month patient data will be collected and analysed. Outcomes predicted will be compared with actual outcomes recorded at month twelve, allowing refinement, enhancement, and validation of the developed models. Data analysis will proceed as follows:

Step 1: Data quality control and validation:

Individual patient level data will be reviewed for completeness. Given the lack of previous studies integrating clinical, physiological, activity and patient reported data in this population, we will employ a conservative approach to data completeness with a requirement for at least 70% completeness for key variables. Missing data will be summarised and dealt with through Case deletion, where the 70% threshold is not reached. Rational substitution will be employed where possible. For missing data at random, arithmetic imputation methods will be employed including worst case imputation for dropouts and interpolation/extrapolation where prior and

after data is available.

Step 2: Descriptive analysis: Data from activity and physiological measurements will be summarised as per Table. 4. Data summarisation will involve providing a concise overview of key characteristics of the dataset including central tendency values, variance to give insights into the spread of the data and frequency distributions, assessing for normality, while highlighting any notable patterns or peaks in the dataset.

Step 3: Feature selection, potential predictors of progression, exacerbation and mortality will be reviewed by specialist respiratory clinicians. The rationale behind the feature selection process will be clearly documented and the chosen predictors will be rigorously validated to ensure clinical relevance to the target outcomes. In parallel, activity and physiological measurements (individual and composite, single point and time-rolling average trends, raw values and individual-normalised) will be analysed to identify potential features, predictive of clinical outcome.

Step 4: Predicted outcomes, the prediction model will be simplified by the elimination of unnecessary variables. Methods such as correlation analysis will be performed to identify potential redundancies. Statistical tests will be used to rank features and their association with the target outcomes.

Step 5: Validation: we will evaluate the predictive model's effectiveness by comparing its performance to actual events per predictor both within individual and across individuals. By doing this we will gain insights into the practical relevance of individual predictors, and we will refine the model accordingly.

9. How would the authors plan to evaluate the developed model? A detailed description of validation and evaluation of the prediction model is needed.

Response: We will have the actual outcome at 12 months and will be able to compare it to the predicted outcome. Detailed description now included in the data analysis section as follows:

All models will be evaluated and compared to determine performance. Key evaluation measures will include

- Accuracy: based on assessment of predicted vs observed outcomes, as participants continue in the cohort
- Sensitivity and Specificity: analysis of true predictors, based on comparison with actual participants outcome
- C-Statistics: to measure performance of models in discriminating between outcomes

10. Are authors confident to have sufficient positive events with a sample size of only 50?

Response: As per answer to question 2, due to the severity and progressive nature of the disease we anticipate sufficient positive events with a sample size of 50.

The current section on statistical methods is not clear. Above are some major limitations. I strongly encourage the authors to consult with a statistician and get advice for the methodology. There are lots of advanced methodological issues needed to be considered.
Response: we and have in-house expertise within UCD Clinical Research Centre

---

VERSION 2 - REVIEW

Reviewer	**3**
Name	**Albrecht, Steffen**
Affiliation	**The University of Auckland, School of Computer Science**
Date	**11-Mar-2025**
COI	

Thanks to the authors for comprehensively addressing all the questions. The protocol improved considerably and the key aspects are much clearer now. I am looking forward to reading about the results of this study.