BMJ Digital Health & AI

Machine-learning approach to dissect the clinical heterogeneity of IBD-associated fatigue

Cher S Chuah,[1] Rebecca Hall,[2] Robert J Whelan,[1] Peter D Cartlidge,[2] Beatriz Gros,[3,4] Eva Iglesias-Flores,[3] Nikita Parkash,[5] Ray K Boyapati,[6] Clara Ramos-Belinchon,[7] Solomon Ong,[2] Emily F Brownson,[8] Iona A M Campbell,[8] Craig Mowat,[9] John P Seenan,[8] Jonathan C MacDonald,[8] MUSIC Patient-Public Involvement Group, Gwo-Tzer Ho [1]

To cite: Chuah CS, Hall R, Whelan RJ, *et al*. Machine-learning approach to dissect the clinical heterogeneity of IBD-associated fatigue. *BMJ Digit Health* 2026;**2**:e000037. doi:10.1136/bmjdh-2026-000037

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjdh-2026-000037).

Received 27 January 2026
Accepted 27 January 2026

For numbered affiliations see end of article.

Correspondence to
Professor Gwo-Tzer Ho;
gwo-tzer.ho@glasgow.ac.uk

ABSTRACT

Objective Extreme and persistent fatigue affects >50% of individuals with inflammatory bowel disease (IBD), with similar prevalence across many common immune-mediated inflammatory diseases (IMIDs). Despite its ubiquity, human scientific studies have yet to fully explain the mechanistic basis of this complex symptom. One fundamental reason is our inability to account for the clinical heterogeneity and multifactorial nature of fatigue.

Methods and analysis We present the conceptual machine-learning (ML) framework to dissect fatigue using one of the largest prospectively captured, real-world patient-reported outcome (PROs) on well-being from three contemporaneous cohorts (2020–present), totalling 2970 responses from 2290 participants across the UK and internationally, including non-IBD controls with 100 lines of clinical metadata. In parallel, our patient public involvement group performed thematic analysis of this PRO dataset, which identified fatigue as a key research priority (www.musicstudy.uk).

Results We systematically defined the (1) threshold of fatigue as our primary outcome (≥10/14 fatigue days in 1604 patients (1151 responses in active disease and 1061 responses in remission; some patients measured longitudinally; median fatigue days 14 vs 7, respectively; p<0.001) to build our ML approach, (2) used routinely available clinical data that can be used at a population-level analysis, (3) employed seven different ML methods with external validation in three different cohorts in the UK, Spain and Australia (n=252), (4) employed Shapley Additive Explanations (SHAP) analysis to break down clinical heterogeneity and allow the examination of clinical predictive factors at an individual level; and finally, (5) investigated whether there are distinct clusters of fatigue patients. We found that ML models performed comparably (area under the curve/C-index ~0.7) on external validation with SHAP analysis showing interpretable, individualised fatigue drivers and five distinct fatigue cluster groups, including a subgroup with lower fatigue burden.

Conclusions Our data provide the ML 'roadmap' to predict and deconstruct fatigue in IBD and potentially more widely in IMIDs, enabling patient-level dissection beyond symptom-based classification with the ability to integrate deep molecular data. This is a step towards future clinical-scientific artificial intelligence models with immediate clinical application to stratify patients for human experimental studies to better identify patient-level patterns associated with fatigue.

Trial registration number NCT04760964.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Fatigue in inflammatory bowel disease (IBD) is complex, heterogeneous, has significant negative impact on quality of life and is identified as a top research priority by patients.

⇒ Clinical trials aiming to treat fatigue in IBD have been limited due to inability to account for patient-level heterogeneity.

WHAT THIS STUDY ADDS

⇒ We developed an end-to-end machine-learning framework using patient-reported outcomes to dissect the heterogeneity of fatigue, enabling individual-level prediction compared with traditional population-level statistics.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Provides a scalable approach to stratify patients and guide mechanistic or interventional studies, laying the foundation for future individualised treatment strategies for fatigue in IBD and other diseases with similar complex, multifactorial symptoms.

INTRODUCTION

Many patients with inflammatory bowel diseases (IBD) suffer from extreme fatigue, even when in remission.[1–3] From a patient's perspective, fatigue is a major priority area for further research.[4] Extreme fatigue is also a common symptom across many immune-mediated inflammatory conditions (IMIDs) such as rheumatoid arthritis (RA), systemic lupus erythematosus and sarcoidosis.[5–7] This

suggests there may be unrecognised cross-disease mechanisms that are potentially independent of organ-specific inflammation[8] and hence also poorly treated by conventional immune suppression. There are increasingly powerful scientific tools from functional neuroimaging to metabolic assays to study the mechanistic basis of this pervasive symptom in IMIDs.[9 10] However, human experimental and interventional studies have been stymied by patient heterogeneity. Pertinently, in IBD, few interventional studies have specifically targeted fatigue, and those that have often show poor results,[11–13] are small and open-labelled in design[14] and include ill-defined patient groups with fatigue.

IBD is a challenging condition with considerable heterogeneity in its presentation. Many patients have different disease courses, from mild to severe activity, and with variable IBD-related complications from anaemia to nutritional deficiencies. In addition, they may experience and respond variably to clinical and gastrointestinal symptoms from pain, urgency to diarrhoea that can affect their well-being and sleep on an individual basis. Medical therapies such as steroids and immunosuppressants and lifestyle factors (eg, diet and smoking) that can contribute to fatigue are also different within the IBD patient populations. IBD individuals can be at different stages of this lifelong condition; in remission or at an advanced disease phase; and undergoing IBD flares during different developmental stages—adolescence to advanced age groups with different physiological demands. While in some IBD patients, fatigue can be explained by IBD-inflammation related effects and improve with medical therapy, many continue to experience fatigue when in clinical remission and without obvious cause. How we can account for such heterogeneity in addressing the complex construct of extreme fatigue in IBD presents a formidable challenge to understanding the contributors to this debilitating symptom.

Our aim is to develop a conceptual machine-learning (ML) approach that can characterise this heterogeneity vis-à-vis the complex symptom construct of fatigue (figure 1). Here, we present an end-to-end ML roadmap in IBD using as first basis, patient-reported outcomes (PROs) and combining routinely available clinical and laboratory data to do this at a population level to predict significant fatigue in IBD. We investigate the potential for our ML approach to define the clinical characteristics on the individual level who are fatigued, and thus a working transparent ML framework that can be further developed and enhanced, working towards the objective of deeper patient stratification is 'to find the right patients for the right studies' in the future.

MATERIALS AND METHODS

Patient data

We used PRO fatigue data from two prospective mechanistic biomarker studies, namely (1) Investigation into Gastrointestinal Damage Associated Molecular Patterns (DAMPs) that is a cross-sectional IBD study and (2) Mitochondrial DAMPs as mechanistic biomarkers of mucosal inflammation in Crohn's Disease (CD) (MUSIC study; www.musicstudy.uk; ClinicalTrials.gov NCT04760964), a prospective IBD cohort study. Both are cohort studies carried out in Scotland (Glasgow, Edinburgh and Dundee; 2020–present) that recruit participants with IBD along with ~100 lines of clinical metadata including IBD activity, treatment, comorbidities and laboratory parameters. In addition to this, we conducted an online survey to further collect PRO data on fatigue (2023–2024; n=1643 within the UK, n=112 internationally).

In all three studies (online supplemental table 1), we used the validated Crohn's and Ulcerative Colitis Questionnaire-32 (CUCQ32) questionnaire that is applicable to CD and ulcerative colitis (UC), the two subtypes of IBD.[15] CUCQ32 contains 32 questions that measure 4 domains of well-being (gastrointestinal, social, psychological and general well-being) with 5 main questions pertaining to fatigue, generating a score ranging from 0 to 272 (online supplemental table 2). Collectively, these studies involved 2290 total participants (including 336 non-IBD participants who responded online, and 27 non-IBD symptomatic controls) and provide the scale of data to establish a clinical threshold of patient-reported fatigue as a baseline for our ML approach.

Model development and evaluation of ML pipeline

Input features include clinical (eg, diagnosis group, Montreal classification), demographics, exposome (smoking, alcohol, seasonality), laboratory (eg, C reactive protein (CRP), faecal calprotectin and IBD drug exposure data (full list in online supplemental table 4). Seven supervised ML models were employed: XGBoost, Random Forest, AdaBoost, multilayer perceptron (MLP), support vector machine, logistic regression (scikit-learn) and a custom feedforward deep neural network (DNN) implemented in TensorFlow (online supplemental figure 4c). Conventional logistic regression (statsmodels), which uses maximum likelihood estimation (traditional method in premachine learning era medical literature) compared with gradient descent with regularisation from scikit-learn, was included as a reference comparator on the final external dataset comparison. Analyses were conducted using Python (V.3.11.9), with the following libraries: scikit-learn (V.1.5.2), scipy (V.1.14.1), xgboost (V.2.1.2), pandas (V.2.2.3), numpy (V.2.0.2), shap (V.0.46.0), tensorflow (V.2.18.0), statsmodels (V.0.14.4), seaborn (V.0.13.2) and matplotlib (V.3.10.0).

Missing data were imputed according to variable type. Continuous variables with non-normal distributions (eg, CRP, albumin, faecal calprotectin) were imputed using the median to minimise bias (imputation sensitivity analyses in online supplemental table 10). Categorical variables were transformed using one-hot encoding. Numerical variables were standardised (zero mean, unit variance) using StandardScaler (scikit-learn). Derived features included seasonality (calculated from

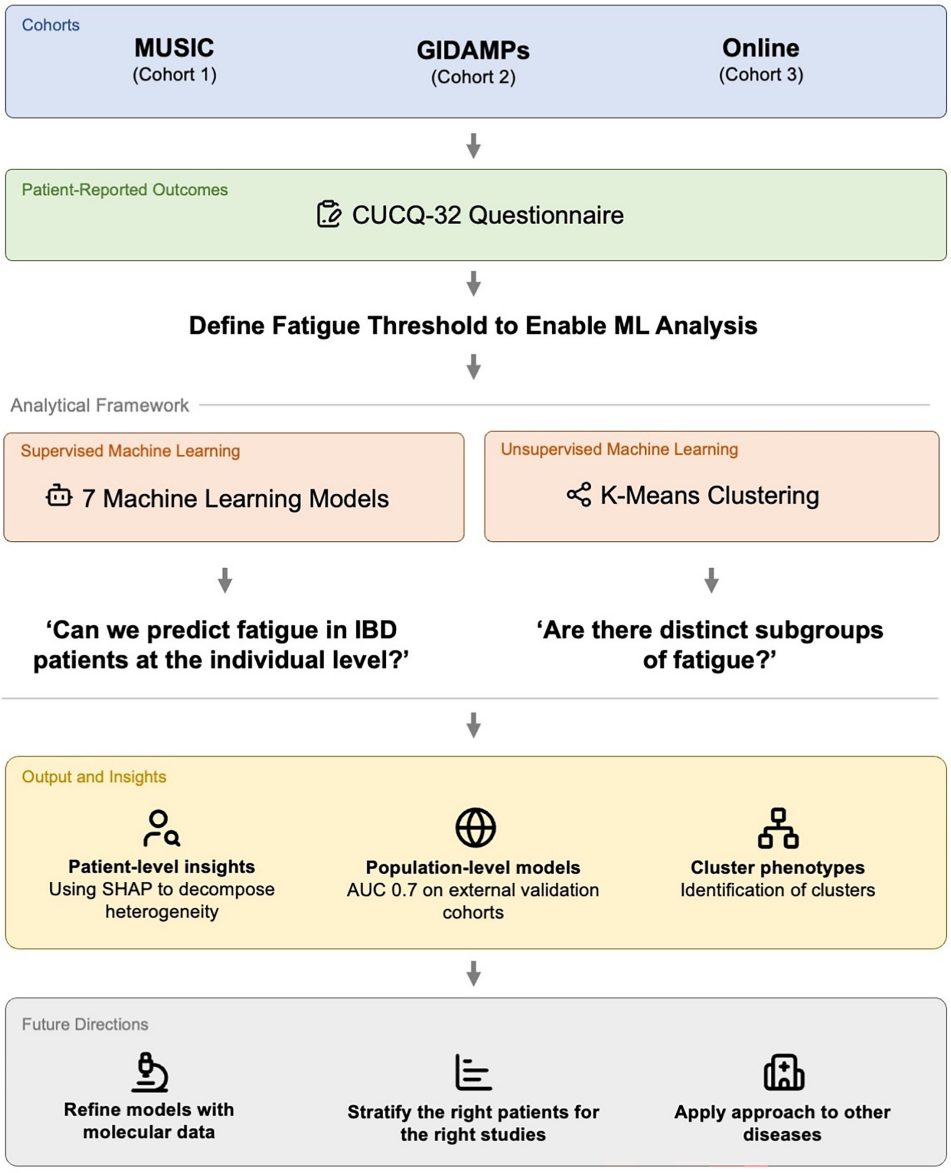

Figure 1 Summary and overview of machine-learning approach. AUC, area under the curve; CUCQ-32, Crohn's and Ulcerative Colitis Questionnaire-32; IBD, inflammatory bowel disease; ML, machine learning; SHAP, Shapley Additive Explanations.

the date of CUCQ32 completion) and disease duration (expressed in weeks since diagnosis). To avoid information leakage from repeated measures, group-aware data splitting was performed using GroupKFold and GroupShuffleSplit (scikit-learn), with participant identifiers used as grouping variables. For conventional ML models, data were partitioned into training-validation (80%) and independent testing (20%) sets, with hyperparameter tuning conducted via fivefold group cross-validation (hyperparameter details in online supplemental table 9). For DNNs, data were split into training (65%), validation (15%) and testing (20%) sets using the same group-aware strategy. For the DNN, a fixed validation set was required for hyperparameter tuning and early stopping, which precluded the use of cross-validation; however, the train-validation-test split ratios were matched as closely as possible to the conventional ML workflow, and all models were evaluated on the same testing and external validation datasets.

Model interpretability was evaluated using Shapley Additive Explanations (SHAP).[16] Model generalisability was evaluated using fully anonymised datasets from three prospectively collected independent validation cohorts based in Scotland, Spain and Australia (online supplemental table 6, predominantly recruited in the outpatient IBD clinic setting). These cohorts were selected because they were recruited in different health systems, geographic regions and clinical workflows, which provided intentionally heterogeneous casemix. We also deliberately focused on the outpatient IBD clinic setting, where such a model would most likely be applied in practice. These datasets were entirely independent of model development, and

no subjects overlapped with the training population. Prior to prediction, clinical units were harmonised to match training data, and feature standardisation was applied using parameters derived exclusively from the training dataset to avoid data leakage. This approach ensured that the evaluation reflected performance on truly out-of-distribution data. Model performance was assessed using receiver operator curves, area under the curve (AUC), sensitivity, specificity, positive and negative predictive values, using a standard probability threshold of 0.5 (ie, model output >0.5 is classified as $Fatigue_{High}$). Model reporting adheres to Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis-Artificial Intelligence guidelines (online supplemental file 3).

Unsupervised clustering was performed using K-means. Following exclusion of a single outlier datapoint, the optimal number of clusters was empirically determined by evaluating solutions for k=2 to k=10, informed by prior clinical knowledge that the cohort contained a subgroup with active IBD. A solution with k=5 was selected, balancing cohort heterogeneity while avoiding over-fragmentation or generation of spurious small clusters (online supplemental figure 5).

All anonymised datasets, dependencies and code to reproduce this study are publicly available at https://github.com/1-gut/machine_learning_for_ibd_fatigue and Zenodo.[17]

Patient and public involvement

Our patient and public involvement group has been embedded throughout the duration of the study, with patient-led thematic analysis of our PRO dataset (available open access via www.musicstudy.uk/patients-taking-the-lead) which identified fatigue as a major patient and research priority.[18] Our patient group provided direct input into the collection of data via an online survey and overview on the analysis and presentation of the MUSIC/GIDAMPs dataset.

RESULTS

Defining the threshold of fatigue

We modelled our ML algorithms to predict IBD patients with PRO fatigue of ≥10 days over the past 14 days at time of data entry, as the primary outcome variable (Fatigue PRO). This is based on CUCQ32 question: 'On how many days over the last 2 weeks did you feel tired?' (0–14 days). Across the combined dataset, IBD patients with active disease reported significantly more fatigued days than those in remission (medians 14 vs 7 days; p<0.001). In IBD patients in remission, fatigue days remained significantly elevated compared with non-IBD controls (7 vs 4 days; p<0.001; online supplemental figure 1a). These findings were consistent across individual cohorts. Fatigue days were strongly correlated with total CUCQ32 scores (r=0.73, p<0.0001; online supplemental figure 1b). The distribution of fatigue days was skewed (online supplemental figure 1c). Therefore, we used the median value of ≥10 days of fatigue over 14 days and defined it as the threshold for clinically significant fatigue ($Fatigue_{High}$). Using this definition, 53.9% (1602/2970) of observations met criteria for $Fatigue_{High}$, comprising 75.2% of active IBD patients, 43.6% of those in remission, and 14.2% of non-IBD controls. Patients in the $Fatigue_{High}$ group had significantly higher overall CUCQ32 scores than those below this threshold (median 130 vs 38; p<0.0001; online supplemental figure 1d, full detailed breakdown in online supplemental table 8).

Fatigue PRO is highly correlated to all measures of CUCQ32 domains (all p<0.001, online supplemental table 3). This shows that a simple question is relevant in the complex construct of fatigue. PROs ('what patients tell clinicians') are strongly correlated with clinician-based assessment ('what clinicians think of patients') (online supplemental figure 2a,b) and are associated with clinical parameters such as CRP and faecal calprotectin (online supplemental figure 2c,d, both p<0.0001). Among 147 MUSIC patients with longitudinal follow-up over 12 months, serial CUCQ32 measurements demonstrated heterogeneity between individuals. Patients achieving clinical remission (Harvey Bradshaw Index, HBI≤5 for CD and Simple Clinical Colitis Activity Index (SCCAI≤2) for UC) at 12 months tended to show improvements in both overall CUCQ32 scores and fatigue-specific domains (online supplemental figure 3). However, no significant difference in CUCQ32 was observed between those with or without mucosal healing at 12 months (median 68 vs 62; online supplemental figure 2e), indicating that endoscopic remission does not fully capture patient-reported well-being. Collectively, this demonstrates the clinical heterogeneity of fatigue when assessed in accordance with conventional measures of inflammation in IBD.

Development of ML algorithms and external validation to predict fatigue

Having defined our threshold for fatigue, we incorporated self-reported fatigue assessments and all available clinical parameters such as disease activity indices (HBI and SCCAI), laboratory parameters (CRP, faecal calprotectin, full blood count), body mass index (BMI), smoking status, drug therapy and seasonality. Here, we generated an initial training dataset of 1215 observations from cohorts 1 and 2 for seven ML algorithms: XGBoost, random forest, AdaBoost, MLP classifier, support vector classifier, logistic regression and a custom-built DNN (figure 2A). The purpose of comparing these models was to benchmark different methodological families and assess whether more complex models provide meaningful gains over traditional logistic regression. GroupKFold cross-validation ensured that repeated measures from individual patients remained in the same dataset split.

Across classical ML models, predictive performance was broadly comparable with AUC values ranging from 0.69 to 0.73 for $Fatigue_{High}$ prediction (figure 2B, online

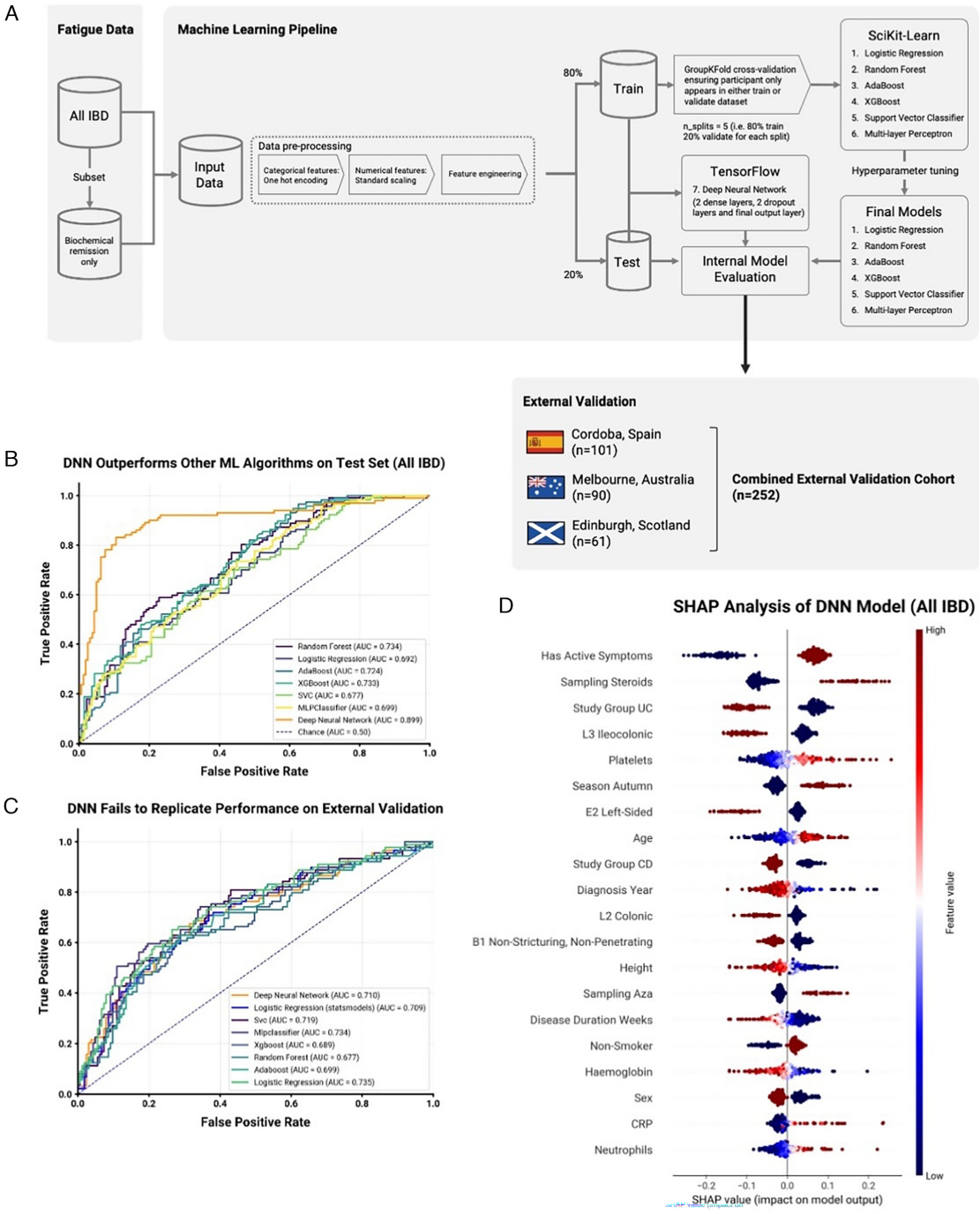

Figure 2 Machine-learning workflow, performance and interpretability of models predicting fatigue in IBD patients. (a) Overview of the machine-learning workflow. Two pipelines were implemented: one including all IBD patients (cohorts 1–2), and another restricted to patients in biochemical remission (defined as calprotectin <250 µg/g and CRP<5 mg/dL). Model performance was subsequently evaluated on an external validation cohort comprising patients with IBD from Spain, Australia and Scotland. (b) Model performance (AUC) in the full IBD cohort. The deep neural network (DNN) achieved the highest test set performance compared with other models, motivating external validation. (c) Model performance (AUC) in the external validation cohort. All models performed similarly, though the DNN showed a notable decrease in AUC, suggesting possible overfitting in the initial test dataset. (d) SHAP summary plot for the DNN, illustrating the impact of individual features on model predictions. Each dot represents a patient's data point for a specific feature. The horizontal position of the dot shows how much that feature pushes the prediction towards fatigue (positive values) or away from fatigue (negative values). Dot colour reflects the actual feature value (eg, red for high platelet counts, blue for low). Features are listed vertically in order of their overall impact on the model, with the most influential features at the top. This visualisation reveals which features most strongly influence fatigue predictions and highlights that the impact of features can vary between patients—for example, very high platelet counts increase predicted fatigue in some individuals but not in others. Assessments conducted in autumn also increased the likelihood of fatigue prediction. AUC, area under the curve; CD, Crohn's disease; CRP, C reactive protein; IBD, inflammatory bowel disease; ML, machine learning; SHAP, Shapley Additive Explanations; UC, ulcerative colitis.

supplemental table 5). To explore whether complex nonlinear patterns might enhance predictive accuracy, we developed a custom DNN implemented in TensorFlow. The network architecture comprised four layers (two dense layers interspersed with two dropout layers; online supplemental figure 4c). On internal test data, the DNN outperformed classical ML models, achieving an AUC of 0.89 for all IBD cases and 0.94 in patients in biochemical remission (figure 2B). However, external validation using an independent cohort (n=252; Córdoba, Spain n=101; Melbourne, Australia n=90; Edinburgh, Scotland n=61) revealed that the DNN's performance did not generalise, with AUCs falling back to the 0.69–0.73 range observed across all models (figure 2C, online supplemental table 5). This suggests that the superior internal performance likely reflected model overfitting rather than true generalisability. No single ML model demonstrated consistent superiority across all datasets (online supplemental table 11).

When restricting the analysis to patients in biochemical remission (CRP <5 mg/L and calprotectin <250 µg/g), model performance declined (AUCs 0.61–0.66 (online supplemental figure 4a). In this subgroup, inflammatory markers contributed less to fatigue prediction, while features such as anaemia, elevated lymphocytes, reduced urea (potentially reflecting sarcopenia), age and seasonality emerged as more influential (online supplemental figure 4b). This reduction in performance indicates that the model is partly driven by inflammatory activity and that routine clinical metadata alone is insufficient to capture the complexity of fatigue once overt inflammation is absent.

Personalised fatigue profiling using SHAP-driven model interpretation

To better understand how individual factors contributed to model predictions, we applied SHAP analysis.[16] SHAP values quantify the relative contribution of each input variable to individual model predictions, offering a transparent, patient-specific decomposition of the model's decision-making process. SHAP interpretation revealed distinct patient-level patterns, underscoring the clinical (and possibly biological) heterogeneity of fatigue in IBD. For some individuals, fatigue was closely linked to inflammatory markers, while for others, factors such as weight, medication use or seasonal variation were more influential. In the DNN model, key predictive features for all IBD patients included clinical symptom activity, steroid use, IBD diagnosis (CD vs UC), platelet count and autumn season (figure 2D). By enabling granular, patient-specific insights into the drivers of fatigue, SHAP analysis provides a proof-of-concept pathway towards personalised characterisation of factors associated with fatigue in IBD (figure 3).

Identification of distinct fatigue phenotypes using K-means clustering

To further characterise potential fatigue subtypes, we applied unsupervised k-means clustering (figure 4). This analysis is an exploratory, data-driven view of fatigue heterogeneity and highlights patient subgroups that may differ in their fatigue profiles despite similar routine clinical features. Although SHAP groups related features by their contribution patterns, this reflects feature behaviour rather than patient similarity. The k-means analysis therefore identifies groups of individuals with shared multidimensional profiles and provides a starting point for defining subpopulations for future mechanistic or interventional studies.

Five distinct patient clusters emerged (table 1), namely those with (1) active IBD, (2) young tall males, (3) CD patients with long disease duration, (4) more elderly UC patients and (5) young female IBD patients. Cluster 0, representing active IBD, exhibited the highest prevalence of $Fatigue_{High}$ (81.4%, n=113). In contrast, cluster 1 (young tall males) demonstrated a significantly lower fatigue prevalence (34.9%) compared with all other groups (analysis of variance F=22.6, p<0.0001; Tukey HSD post hoc comparisons; figure 4C). The remaining clusters exhibited intermediate fatigue prevalence ranging from 50% to 56% (online supplemental table 7). The low-fatigue male cluster (cluster 1) was distinct not only in sex (96% male), height (median 1.80 m) and weight (median 84 kg), but also exhibited higher haemoglobin levels (145 g/L vs 120–137 g/L in other clusters). The reasons for their relative fatigue resistance remain unclear but may reflect biological resilience, under-reporting or unidentified protective mechanisms. Notably, deviations from low baseline fatigue within this group may represent meaningful signals warranting focused mechanistic investigation. Together, these findings demonstrate that IBD-associated fatigue remains prevalent despite disease control, exhibits marked individual variability and can be modelled using supervised and unsupervised ML approaches that integrate patient-reported data and current level multidimensional clinical data that can be further optimised in future work.

DISCUSSION

Artificial intelligence and powerful ML are now firmly established and are increasingly used in medicine.[19] A key consideration is how we construct the practical steps and apply these tools to a complex and relevant medical patient-centric problem, such as fatigue. Here, we present a transparent conceptual ML framework 'roadmap' to characterise fatigue-related patterns towards a practical outcome of better patient stratification to aid mechanistic studies; using one of the largest prospectively captured, real-world PROs on well-being from three contemporaneous cohorts, totalling 2970 responses from 2290 participants across the UK and internationally, including non-IBD controls. We systematically defined the relevance of the (1) primary outcome of fatigue to base our ML approach, (2) used routinely available clinical data that can be used at a population-level analysis, (3) employed seven different ML methods, (4) employed SHAP analysis to break down the clinical heterogeneity at

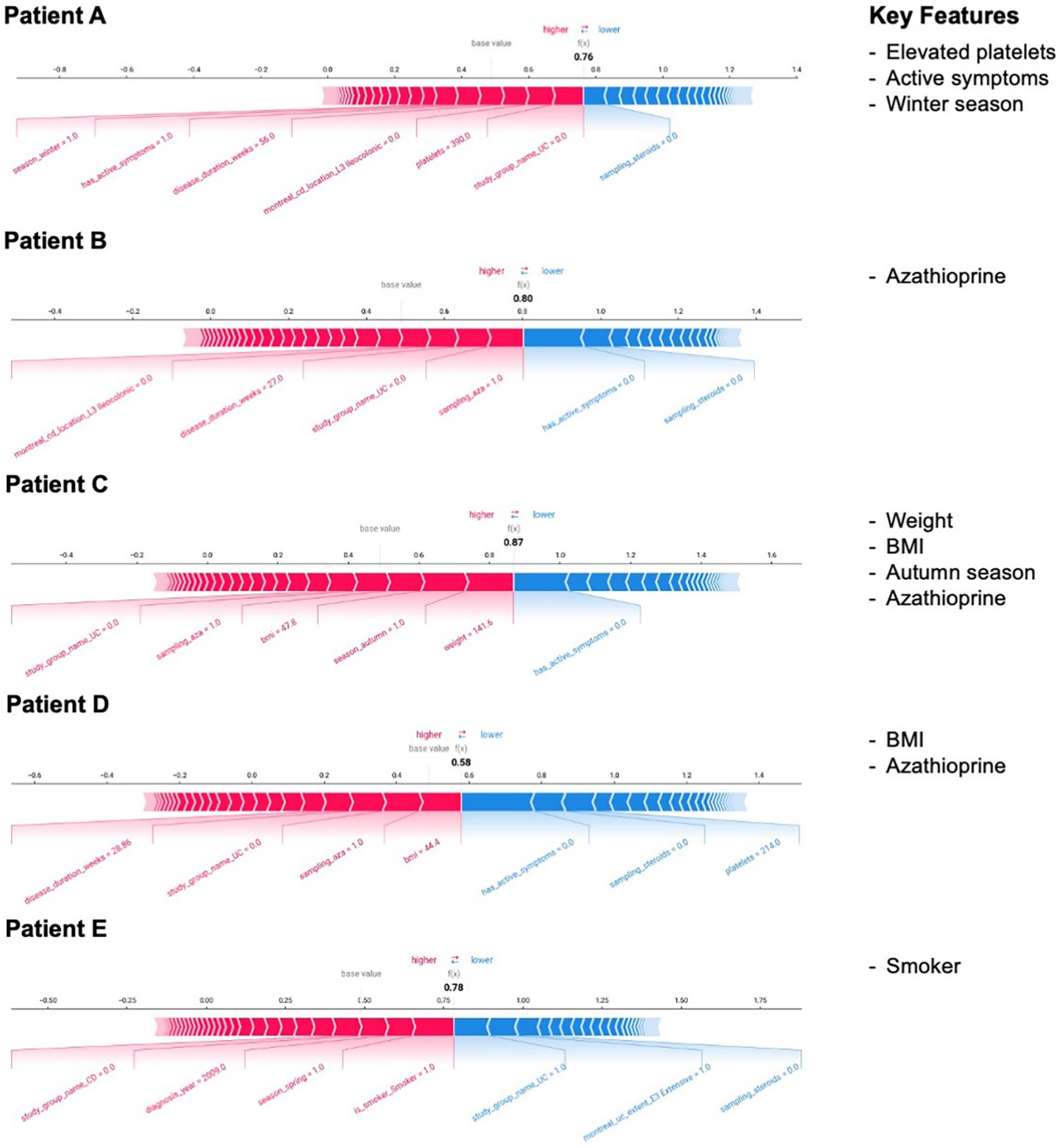

Figure 3 Individualised decomposition of IBD-related fatigue using SHAP. SHAP force plots for true positive cases (ie, instances where the patient reports fatigue and the model predicts fatigue). f(x) represents the model-predicted probability of $Fatigue_{High}$; for example, f(x) = 0.76 corresponds to a 76% probability of $Fatigue_{High}$. Values closer to 1 indicate higher model confidence in predicting fatigue, while values closer to 0 indicate confidence in predicting low fatigue. This proof of concept illustrates the model's ability to decompose feature contributions at the individual level. For example, in patient A, active IBD is a major contributor to fatigue; in patient B, azathioprine is a more influential factor. Patients C and D exhibit multifactorial contributions, with an important weight-related component, while patient E's smoking status emerges as a key driver of fatigue. BMI, body mass index; CD, Crohn's disease; IBD, inflammatory bowel disease; SHAP, Shapley Additive Explanations; UC, ulcerative colitis.

an individual level; and finally, (5) investigated whether there are distinct clusters of fatigue patients based simply on routine clinical observations.

Notwithstanding the complexities of IBD, we found that our ML approach can predict patients with significant fatigue (AUC 0.70) with external validation across three independent cohorts in different geographical locations, but the performance drops within the IBD patient group that is in remission (AUC 0.60). We systematically compared seven ML models against traditional logistic regression to evaluate their ability to incorporate deep clinical data. DNNs demonstrated superior internal performance but similar external validation performance across all models. However, substantial unexplained

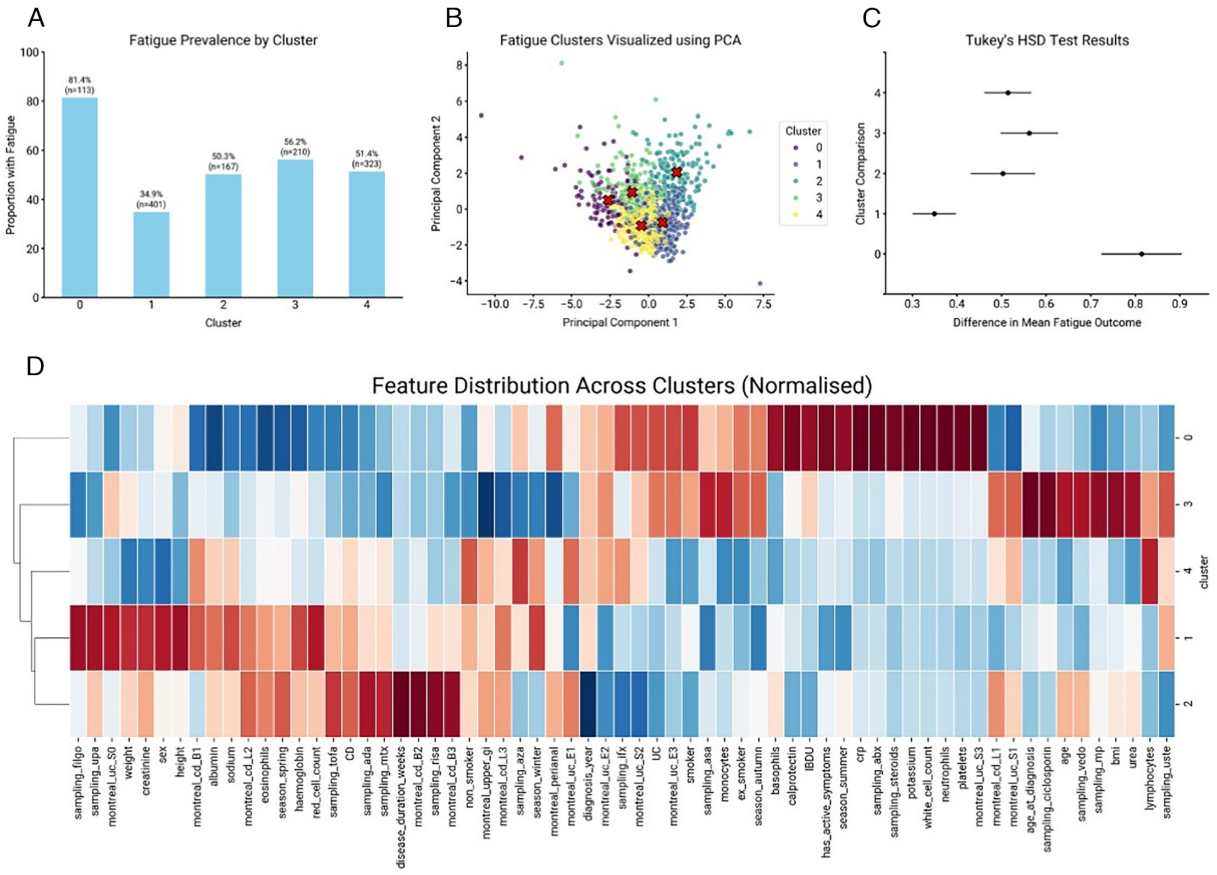

Figure 4 K-means clustering identifies five distinct fatigue clusters. (a) Fatigue prevalence by cluster: cluster 0 shows high fatigue prevalence (81%), cluster 1 shows low fatigue prevalence (35%), and clusters 2–4 show intermediate fatigue prevalence (50%–56%). (b) Visualisation of fatigue clusters using principal components analysis (PCA). (c) Tukey's HSD test indicates that cluster 0 has significantly higher fatigue compared with the other clusters, while cluster 1 has significantly lower fatigue. No significant differences are observed among clusters 2–4. (d) Heatmap visualisation of the five identified subgroups. CD, Crohn's disease; HSD, honestly significant difference; IBU, inflammatory bowel disease unclassified; UC, ulcerative colitis.

variance remains, particularly in remission. At the cohort level, the model reflects the strong association between fatigue and disease activity, and its lower performance in remission indicates that routine inflammatory and clinical markers have limited predictive value once overt inflammation is controlled. This suggests that current classifiers partially reflect disease activity, although they still capture some non-inflammatory components of fatigue (online supplemental figure 4b). This is pertinent as it points towards the need to build in more expansive molecular data in this difficult group of patients. Most pertinently, future studies will require a larger cohort of IBD patients in remission.

The use of our simplified outcome threshold of ≥10/14 fatigue days deserves further discussion. First, the prevalences of fatigue defined by this agree with clinical studies of IBD and IMID fatigue.[20 21] Second, this allows us to harmonise fatigue measurement across cohorts, scale up our dataset and facilitate predictive ML analyses that can be applicable to different IMID cohorts in the future. Thirdly, our data from the broader CUCQ32 shows that this simple PRO is highly correlated to other domains of well-being and therefore a good representation of the deeper construct of the symptom fatigue that may be linked to poor sleep, depression, anxiety and organic disease-related causes. Furthermore, regression modelling (available at code repository) of the continuous 0–14 fatigue outcome demonstrated minimal explanatory ability (R^2=0.17), indicating current clinical metadata lack the granularity required for finer-scale prediction. Hence, our goal is to build an ML approach that can capture these complexities from a simple starting point that is easily captured in the clinic.

A potential critique is that fatigue prediction models may merely recapitulate the fatigue question itself, i.e. reinforcing a circular argument. We address this by explicitly excluding the fatigue question and other CUCQ questions as input variables in our ML algorithms. Instead, fatigue was predicted from a broad range of independently collected clinical metadata, including laboratory markers (eg, CRP, calprotectin), medication exposure, BMI, smoking status and seasonality. These predictors are biologically and clinically orthogonal to self-reported fatigue. Furthermore, SHAP analysis revealed interpretable, participant-specific predictors such as inflammatory markers and smoking, providing

Table 1 Summary cluster characteristics and descriptive label

Cluster ID	Descriptive label	Number of observations, n	Fatigue prevalence	Top differentiating features
0	Active IBD	113	81.4%	Low haemoglobin, albumin Elevated WCC, platelets, CRP, calprotectin
1	Young tall males	401	34.9%	96% male, Median age 32 More Crohn's (64%) Less steroids, less vedolizumab use, more Upadacitinib use
2	Experienced Crohn's patients	167	50.7%	Median age 52 Higher adalimumab use (31%) Median disease duration 25 years Median diagnosis year 1997 More CD (69%)
3	Older UC patients	210	56.2%	Median age 61 More UC (58%) More steroid use
4	Young female IBD	323	51.4%	Median age 29 87% female More azathioprine use More CD (59%)

CD, Crohn's disease; CRP, C reactive protein; IBD, inflammatory bowel disease; UC, ulcerative colitis; WCC, white cell count.

face validity to the modelling. External validation across independent cohorts confirmed that fatigue can be predicted to an extent without direct fatigue self-report, demonstrating the capacity of ML to uncover latent patterns underlying fatigue.

This suggests that (1) ML is equivalent to classical methods for structured clinical data and (2) the consistent performance ceiling likely reflects a 'hidden' multifactorial pathobiological compartment of factors currently unmeasured. We, therefore, present the first comprehensive framework using ML algorithms to dissect the complex presentation of fatigue. This is initially based on using routinely available clinical data, but this will be expanded to include molecular data such as genetics, nutritional factors, microbiome and metabolomics in the future. We acknowledge that our dataset lacked key variables known to influence fatigue, including sleep quality (we excluded the use of CUCQ sleep question as it was highly correlated to the outcome), mental health, socioeconomic status and more detailed comorbidities. However, with the ML framework established, these datasets can be included with relative ease in future studies. We think that with further training of the models, there will be direct clinical application by identifying 'clusters of patients' or individualised stratification based on SHAP analysis to human experimental studies of fatigue. These insights may ultimately support patient stratification to targeted interventional studies (figure 1).

In conclusion, fatigue profoundly impacts well-being in IBD, even in remission. Our data provide credible support for the utility of PROs as endpoints for future translational scientific research. In addressing IBD-associated fatigue, we show the conceptual ability to reduce the dimensionality via modelling (thus flattening out the heterogeneity of IBD-associated fatigue) as the first-pass method to identify fatigue subgroups. This simplified approach will allow the future incorporation of multiple streams of complex scientific metadata (such as genetics and microbiome, for example) for larger scale analyses at a cohort level. There are many better scientific tools to study central and peripheral fatigue—from imaging, metabolism and mitochondrial function. We envisage that our work provides a step towards identifying subgroups suitable for future mechanistic studies and therapeutic development, shifting beyond symptom-based classification towards data-driven and personalised fatigue management in IBD.

Author affiliations
[1]School of Infection and Immunity, University of Glasgow, Glasgow, UK
[2]University of Edinburgh Institute of Regeneration and Repair, Edinburgh, UK
[3]IMIBIC, Reina Sofia University Hospital, Córdoba, Spain
[4]CIBEREHD, Madrid, Spain
[5]Department of Gastroenterology, Monash Health, Clayton, Victoria, Australia
[6]Monash University Faculty of Medicine Nursing and Health Sciences, Melbourne, Victoria, Australia
[7]Department of Gastroenterology, Western General Hospital, Edinburgh, UK
[8]Department of Gastroenterology, NHS Greater Glasgow and Clyde, Glasgow, UK
[9]University of Dundee School of Medicine, Dundee, UK

Acknowledgements This work is part of the MUSIC IBD study.

Collaborators The MUSIC Patient Public Involvement Group comprises Molly Halligan, Aerin Thompson, Destiny Docherty, Patricia Kelly and Emma Pryde.

Contributors CSC was the first author and performed the machine-learning work and analysis. RH and PC analysed the CUCQ data and collected the online survey in collaboration with the PPI group. RJW contributed to the machine-learning analysis. RH, PC and RJW contributed to the writing of the manuscript. BG and EI-F validated the machine-learning model in Spain, NP and RKB validated the model in Melbourne, and CR-B and SO validated the model in Scotland. EFB, IAMC, CM, JPS and JCM established the MUSIC IBD study in Glasgow and Dundee, recruited patients and collected clinical and patient-reported outcome data. G-TH is chief investigator of the MUSIC IBD study, supervised the study and is the guarantor for this work. All authors reviewed and approved the final manuscript.

Funding This work was funded by The Leona M. and Harry B. Helmsley Charitable Trust (G-1911-03343) and the Anne Ferguson Memorial Fund to G-TH. RJW is funded by UKRI grant MR/N013166/1.

Competing interests None declared.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Ethics approval Ethical approval for Investigation into Gastrointestinal Damage Associated Molecular Patterns (GI-DAMPs) was obtained from East Scotland Ethics Committee (REC 18/ES/0090). Ethical approval for Mitochondrial DAMPs as mechanistic biomarkers of mucosal inflammation in Crohn's Disease (MUSIC study; www.musicstudy.uk; ClinicalTrials.gov NCT04760964) was obtained from East Scotland Ethics Committee (REC 19/ES/0087).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. All anonymised datasets, dependencies and code to reproduce this study are publicly available at: (1) GitHub: https://github.com/1-gut/machine_learning_for_ibd_fatigue. (2) Zenodo: https://doi.org/10.5281/zenodo.17857416.

ORCID iD

Gwo-Tzer Ho https://orcid.org/0000-0002-6014-372X

REFERENCES

1 Farrell D, McCarthy G, Savage E. Self-reported Symptom Burden in Individuals with Inflammatory Bowel Disease. *J Crohns Colitis* 2016;10:315–22.
2 D'Silva A, Fox DE, Nasser Y, *et al*. Prevalence and Risk Factors for Fatigue in Adults With Inflammatory Bowel Disease: A Systematic Review With Meta-Analysis. *Clin Gastroenterol Hepatol* 2022;20:995–1009.
3 Borren NZ, van der Woude CJ, Ananthakrishnan AN. Fatigue in IBD: epidemiology, pathophysiology and management. *Nat Rev Gastroenterol Hepatol* 2019;16:247–59.
4 Jagt JZ, van Rheenen PF, Thoma SMA, *et al*. The top 10 research priorities for inflammatory bowel disease in children and young adults: results of a James Lind Alliance Priority Setting Partnership. *Lancet Gastroenterol Hepatol* 2023;8:690–1.
5 Arnaud L, Gavand PE, Voll R, *et al*. Predictors of fatigue and severe fatigue in a large international cohort of patients with systemic lupus erythematosus and a systematic review of the literature. *Rheumatology (Oxford)* 2019;58:987–96.
6 Drent M, Lower EE, De Vries J. Sarcoidosis-associated fatigue. *Eur Respir J* 2012;40:255–63.
7 Nikolaus S, Bode C, Taal E, *et al*. Fatigue and factors related to fatigue in rheumatoid arthritis: a systematic review. *Arthritis Care Res (Hoboken)* 2013;65:1128–46.
8 Davies K, Dures E, Ng WF. Fatigue in inflammatory rheumatic diseases: current knowledge and areas for future research. *Nat Rev Rheumatol* 2021;17:651–64.
9 Stefanov K, Al-Wasity S, Parkinson JT, *et al*. Brain mapping inflammatory-arthritis-related fatigue in the pursuit of novel therapeutics. *Lancet Rheumatol* 2023;5:e99–109.
10 Naviaux RK, Naviaux JC, Li K, *et al*. Metabolic features of chronic fatigue syndrome. *Proc Natl Acad Sci U S A* 2016;113:E5472–80.
11 Bager P, Hvas CL, Rud CL, *et al*. Randomised clinical trial: high-dose oral thiamine versus placebo for chronic fatigue in patients with quiescent inflammatory bowel disease. *Aliment Pharmacol Ther* 2021;53:79–86.
12 Farrell D, Artom M, Czuber-Dochan W, *et al*. Interventions for fatigue in inflammatory bowel disease. *Cochrane Database Syst Rev* 2020;2020:CD012005.
13 Truyens M, Lobatón T, Ferrante M, *et al*. Effect of 5-Hydroxytryptophan on Fatigue in Quiescent Inflammatory Bowel Disease: A Randomized Controlled Trial. *Gastroenterology* 2022;163:1294–305.
14 Moulton CD, Young AH, Hart AL. Modafinil for Severe Fatigue in Inflammatory Bowel Disease: A Prospective Case Series. *Clin Gastroenterol Hepatol* 2024;22:1737–40.
15 Alrubaiy L, Cheung W-Y, Dodds P, *et al*. Development of a Short Questionnaire to Assess the Quality of Life in Crohn's Disease and Ulcerative Colitis. *J Crohns Colitis* 2015;9:66–76.
16 Ponce-Bobadilla AV, Schmitt V, Maier CS, *et al*. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clin Transl Sci* 2024;17:e70056.
17 Chuah CS, Hall R, Whelan RJ, *et al*. Data from: machine learning approach to dissect the clinical heterogeneity of ibd-associated fatigue. Zenodo; 2025. Available: https://doi.org/10.5281/zenodo.17857415
18 Halligan MJ, Thompson AE, Docherty D, *et al*. Patient-public involvement (PPI)-led qualitative analysis of >400 inflammatory bowel disease patient responses in United Kingdom: an independent patient-led review to provide insights to improve care and research. *medRxiv* [Preprint] 2025.
19 Maddox TM, Embí P, Gerhart J, *et al*. Generative AI in Medicine - Evaluating Progress and Challenges. *N Engl J Med* 2025;392:2479–83.
20 Villoria A, García V, Dosal A, *et al*. Fatigue in out-patients with inflammatory bowel disease: Prevalence and predictive factors. *PLoS One* 2017;12:e0181435.
21 Cohen BL, Zoëga H, Shah SA, *et al*. Fatigue is highly associated with poor health-related quality of life, disability and depression in newly-diagnosed patients with inflammatory bowel disease, independent of disease activity. *Aliment Pharmacol Ther* 2014;39:811–22.