Title

Longitudinal Fecal Calprotectin Profiles Characterize Disease Course Heterogeneity in Crohn's Disease

Short Title

Disease Course Heterogeneity & Longitudinal FCAL Profiles

Authors

Nathan Constantine-Cooke [1, 2], Karla Monterrubio-Gómez [1], Nikolas Plevris [2, 3], Lauranne A.A.P Derikx [4], Beatriz Gros [3], Gareth-Rhys Jones [3, 5], Riccardo E. Marioni [2], Charlie W. Lees[† 2, 3], and Catalina A. Vallejos[† 1, 6]

[†] Shared senior authorship

1. MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK
2. Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK
3. Edinburgh IBD Unit, Western General Hospital, Edinburgh, UK
4. Inflammatory Bowel Disease Center, Radboud University Medical Center, Nijmegen, The Netherlands
5. Centre for Inflammation Research, The Queen's Medical Research Institute, University of Edinburgh, Edinburgh, UK
6. The Alan Turing Institute, British Library, London, UK

Grant Support

This work was supported by the Medical Research Council & University of Edinburgh Precision Medicine PhD studentship (MR/N013166/1, to **NC-C**) and the UKRI Future Leaders Fellowship (MR/S034919/1, to **CWL**). **CAV** was funded by a Chancellor's Fellowship provided by the University of Edinburgh. **KM-G** was supported by an MRC University Unit grant to the MRC Human Genetics Unit. **GRJ** is supported by a Wellcome Trust Clinical Research Career Development Fellowship.

Abbreviations

5-ASAs: Aminosalicylates

AIC: Akaike information criterion

BIC: Bayesian information criterion

CD: Crohn's disease

FCAL: Fecal calprotectin

IQR: Interquartile range

LCMM: Latent class mixed model

Correspondence

Nathan Constantine-Cooke,

MRC Human Genetics Unit, Institute of Genetics and Cancer, The University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU

nathan.constantine-cooke@ed.ac.uk

+447503218115

Disclosures

NC-C: none declared; **KM-G**: none declared; **NP** has received consultancy fees from Takeda, speaker fees and/or travel support from Abbvie, Takeda, Norgine; **LAAPD** has received consultancy fees from Sandoz, speaking fees from Janssen; **BG** has received consultancy fees from Abbvie; **GRJ** has received speaker fees from Abbvie, Takeda, Pfizer, Ferring and Janssen; **REM**: none declared; **CWL** has received research support from Abbvie and Gilead, consultancy fees from Abbvie, Pfizer, Janssen, Gilead, Celltrion, Pharmacosmos, Takeda, Vifor, Iterative Scopes, Trellus Health, Galapagos, Vifor Pharma, Bristol Meyers Squibb, Boehringer Ingelheim, Sandoz, Novartis, Fresnius, and Kabi Tillotts; speaker fees and/or travel support from Janssen, Abbvie, Pfizer, Dr Falk, Ferring, Hospira, GSK, and Takeda; **CAV**: none declared.

Preprint Server

A preprint of this work is hosted on medRxiv; doi: 10.1101/2022.08.16.22278320.

Writing Assistance

None received.

Authorship

NC-C, **KM-G**, **NP**, **REM**, **CWL**, and **CAV** contributed to the conception and study design for the manuscript. **NC-C**, **NP**, **LAAPD**, **BG**, and **CWL** collected the data for this study. All authors except **REM** had access to the study data. **NC-C** performed all statistical analysis. **NC-C**, **BG**, and **KM-G** drafted the manuscript. All authors were involved with critical revision of the manuscript, and all authors reviewed and approved the final manuscript prior to submission.

Data Transparency Statement

The analytical reports generated for this study and corresponding code are hosted online (https://vallejosgroup.github.io/lcmm-site/).

The data used in this study is not publicly available, as it originates from patients who have not given consent for the data to be publicly shared. For access to the data, please contact **CWL**.

Abstract

Background and Aims

The progressive nature of Crohn's disease is highly variable and hard to predict. In addition, symptoms correlate poorly with mucosal inflammation. There is therefore an urgent need to better characterize the heterogeneity of disease trajectories in Crohn's disease by utilizing objective markers of inflammation. We aimed to better understand this heterogeneity by clustering Crohn's disease patients with similar longitudinal fecal calprotectin profiles.

Methods

We performed a retrospective cohort study at the Edinburgh IBD Unit, a tertiary referral center, and used latent class mixed models to cluster Crohn's disease subjects using fecal calprotectin observed within five years of diagnosis. Information criteria, alluvial plots, and cluster trajectories were used to decide the optimal number of clusters. Chi-squared, Fisher's exact test, and ANOVA were used to test for associations with variables commonly assessed at diagnosis.

Results

Our study cohort comprised of 365 patients with newly diagnosed Crohn's disease and 2856 fecal calprotectin measurements taken within five years of diagnosis (median 7 per subject). Four distinct clusters were identified by characteristic calprotectin profiles: a cluster with consistently high fecal calprotectin and three clusters characterized by different downward longitudinal trends. Cluster membership was significantly associated with smoking ($p = 0.015$), upper gastrointestinal involvement ($p < 0.001$), and early biologic therapy ($p < 0.001$).

Conclusions

Our analysis demonstrates a novel approach to characterizing the heterogeneity of Crohn's disease by using fecal calprotectin. The group profiles do not simply reflect different treatment regimens and do not mirror classical disease progression endpoints.

Keywords

Biomarker; Epidemiology; Monitoring.

1 Introduction

Crohn's disease (CD) affects around 1 in 350 people in the UK[1,2] with substantial variation in phenotypes and disease outcomes. Historically, 30% follow a quiescent disease course[3], whilst many will require surgery due to strictures, fistulas, or lack of response to medical therapy. Despite this heterogeneity, our ability to characterize disease variability remains poor and, in the case of Montreal location and behavior, involves invasive examinations which limits the suitability of frequent longitudinal measurements.

Endoscopy remains the gold standard for monitoring IBD, however it is costly, invasive and not without risk. As such, non-invasive stool markers like fecal calprotectin (FCAL) are increasingly used to objectively monitor inflammation in IBD[4,5]. FCAL is one of the most well characterized non-invasive biomarkers in IBD[6]. Multiple studies have shown that it accurately correlates with mucosal inflammation, in both UC and CD. Furthermore, the CALM study[7] demonstrated that a treat to target approach based on FCAL results in superior outcomes when compared to a treatment strategy based on symptoms alone. It is therefore sensible to consider using FCAL to characterize heterogeneity found in intestinal inflammation. By incorporating all FCAL data, instead of only FCAL measurements which can be dichotomized into specific time points, FCAL can be modelled as a continuous longitudinal process. Whilst FCAL has previously been modelled in this way, no published research has attempted to cluster CD patients by longitudinal FCAL profiles: instead capturing heterogeneity across patients through *a priori* selected covariates (such subjects in endoscopic or clinical remission and those who have relapsed)[8,9].

Disease heterogeneity in CD has previously been described longitudinally by the IBSEN study[3]. In the IBSEN study, subjects with CD chose which profile they believed best described their

disease activity out of four profiles specified *a priori*. We aimed to perform a modernized iteration of this work by instead using FCAL profiles to characterize patient heterogeneity. We hypothesize that an unsupervised analysis to uncover latent patient subgroups with distinct longitudinal FCAL patterns can lead to better disease characterization.

2 Materials and Methods

2.1 Study Design

We performed a retrospective cohort study at the Edinburgh IBD Unit to determine if there were subgroups within the CD patient population identifiable from FCAL measurements which had been collected within five years of diagnosis. We modelled longitudinal FCAL profiles using latent class mixed models (LCMMs)[10], an extension of linear mixed effects models, which enables the identification of distinct subgroups with shared longitudinal patterns. LCMMs have been used to model biomarker trajectories in many contexts [11, 12]).

The data were obtained from a cohort study by Plevris et al. which identified all incident CD cases between 2005 and 2017 at The Edinburgh IBD unit which fulfilled set inclusion criteria[13]. For all patients, electronic health records (TrakCare; InterSystems, Cambridge, MA) were used to extract demographic as well as outcomes and FCAL values (both up to June 2019). Data for drug treatments and disease location were also extracted.

2.2 Criteria & Definitions

First, the inclusion criteria from Plevris et al. were applied: (1) CD diagnosis between 2005 and 2017; (2) an initial FCAL measurement at diagnosis (or within 2 months) and prior to treatment; (3) initial diagnostic FCAL result ≥ $250 \mu g/g$; (4) an accurate date of diagnosis; (5) at least one additional FCAL measurement within 12 months of diagnosis; (6) at least 12 months of followup; (7) neither having surgery nor a Montreal disease progression/new perianal disease within 12 months of diagnosis. Second, the following additional criterion was applied in this study: (8) at least 3 FCAL measurements within 5 years of diagnosis.

The following information was available at diagnosis: sex, age, smoking status, FCAL, and Montreal location and behavior. Treatments prescribed within one year of diagnosis were also recorded: 5-ASAs (aminosalicylates), corticosteroids, thiopurines, methotrexate, exclusive enteral nutrition, and biologic therapies.

2.3 FCAL Assay

The Edinburgh IBD Unit has been using FCAL for diagnostic and monitoring purposes since 2005. Stool samples have been routinely collected at all healthcare interactions[13]. Samples are stored at -20ºC and FCAL is measured using a standard enzyme linked immunosorbent assay technique (Calpro AS, Lysaker, Norway). All FCAL measurements in this study were performed using the same protocol and assay.

2.4 Statistical Analysis

Descriptive statistics are presented as median and interquartile range (IQR) for continuous variables. Frequencies with percentages are provided for categorical variables.

FCAL measurements greater than 2500$\mu g/g$ were set to 2500$\mu g/g$, the upper range for the assay. Likewise, measurements reported as less than the lower range for the assay, 20$\mu g/g$, were set to 20$\mu g/g$. FCAL values were log-transformed before the models were fitted. To model the FCAL trajectories and find clusters, we used LCMMs with longitudinal patterns captured using natural cubic splines[10]. Natural cubic splines provide a flexible framework to model FCAL trajectories whilst remaining stable at either end of the study followup period[14]. Using natural cubic splines results in fewer parameters needing to be estimated compared to polynomial regression which requires a high-degree polynomial to achieve the same level of flexibility[15]. Between two and five knots were considered for the splines and their performance was compared using Akaike information criterion (AIC). Three knots were found to produce the optimal AIC within this range which were placed at the quartiles of the FCAL measurement times. A full model description is provided as an Appendix.

LCMMs assuming two to six clusters were fitted. For each number of clusters, the optimal model was found via a grid search approach (50 runs with 10 maximum iterations) following the vignette provided as part of the lcmm R package. Models converged based on parameter and likelihood stability, and on the negativity of the second derivatives. After each optimal model was found, the log-likelihood, AIC, and Bayesian information criterion (BIC) were calculated. An alluvial plot was produced to provide intuition of how additional clusters are formed as the number of assumed clusters increases. These findings were used to decide on the appropriate number of clusters in our study population. As suggested in the lcmm package vignette, goodness-of-fit for the selected model was assessed by exploring whether model residuals were normally distributed. Uncertainty in cluster assignments was quantified using posterior classification probabilities. To visualize overall trajectories within

each cluster, point estimates for each of the model parameters were used, and statistical uncertainty was visualized using 95% confidence intervals.

Marginal associations between cluster membership and information available at the time of diagnosis were explored. Chi-square tests and Fisher's exact tests, dependent on suitability, were used for categorical variables. ANOVA was used for continuous variables. Upper gastrointestinal inflammation (L4) and perianal disease (P) were tested separately to Montreal location (L1-L3) and Montreal behavior (B1-B3) respectively.

Potential evidence of treatment effects was garnered by testing for associations between cluster membership and whether each treatment was prescribed within one year of diagnosis using Fisher's exact test. Biologic prescriptions within three months of diagnosis were also considered to study potential earlier treatment effects.

A 5% significance level was used for all statistical tests. Bonferroni adjustments have also been used to provide adjusted p-values (p_{adj}).

As an exploratory analysis, a multinomial logistic regression model[18] and a random forest classifier[19] were used to predict cluster allocations using information available at the time of diagnosis and biologic prescriptions. For this purpose, a 75:25 train:test split with 4-fold cross validation was used[20]. Classification performance was assessed via area under the curve (AUC) extended to multiple classes[21].

R[22] (v.4.2.1) was used for all statistical analyses using the lcmm[23] (v.1.9.5), nnet[24] (v.7.3-17), ranger[25] (v.0.13.1), datefixR[26] (v.0.1.4), tidyverse[27] (v.1.3.1), tidymodels[28] (v.0.2.0), vip[29] (v0.3.2)

and ggalluvial[30] (v.0.12.3) R libraries. The analytical reports generated for this study and corresponding source code are hosted online*.

2.5 Ethics

As this study was considered a retrospective audit due to all data having been collected as part of routine clinical care, no ethical approval or consent was required as per UK Health Research Authority guidance. Caldicott guardian approval (NHS Lothian) was granted (Project ID: 18002).

3 Results

3.1 FCAL Measurements

The study by Plevris et al.[13] found 1390 incident CD cases. After removing individuals without an accurate diagnostic date or diagnostic FCAL (+/- 60 days of diagnosis), 50 patients had diagnostic FCAL < 250µg/g. Once the additional inclusion criterion of at least three FCAL measurements was also applied, 356 subjects met the inclusion criteria for this study (Figure 1, Table 1). Across these patients, 2856 FCAL measurements were recorded within five years of diagnosis. The median frequency of FCAL measurements for a subject within this period was 7 (IQR 5-10). The distributions of all FCAL measurements and the number of measurements per subject are presented as supplemental material (Figure S1, Figure S2).

3.2 Modelling FCAL Trajectories

LCMMs fitted with two to six assumed clusters all converged as per default convergence criteria. As seen in Figure 2, cluster assignments were largely stable across differing assumed clusters,

* https://vallejosgroup.github.io/lcmm-site/

particularly when comparing the 3-cluster, 4-cluster, and 5-cluster models. Performance metrics for each model considered are provided in Table S1. BIC suggested the 2-cluster model was most appropriate, but this model was discarded as visual inspection of the inferred trajectories suggested a larger number of distinct clusters (Figure S3-S4). AIC favored the 5-cluster model. However, this model was found to overfit the data as some of the inferred trajectories were similar (Figure S5). As a parsimonious choice, we selected the 4-cluster model.

The distribution of the model residuals for the 4-cluster model is bell-shaped, but the quantile-quantile plot suggests some deviations from normality in the tails of the distribution within and across clusters (Figure S6-S7). Figure 3 presents the log mean profiles for the 4-cluster model alongside subject-specific observed FCAL trajectories. The model identified three main groups of patients: clusters 1, 2 and 3 (92, 191, and 58 subjects, respectively) and a small cluster 4 with 15 subjects. Clusters 1 and 3 display similar profiles — both showing a sharp decrease in FCAL which then remains low. However, cluster 1 is differentiated by the decrease occurring immediately after diagnosis, whilst this decrease does not occur until around a year after diagnosis for cluster 3. In contrast, cluster 2 is characterized by a mean profile which remains consistently high: never dropping below the $250 \mu g/g$ clinical threshold for disease activity. Finally, the mean profile for cluster 4 exhibits an initial decrease, but this is not sustained during the first 3 years.

3.3 Association with Variables Available at Diagnosis

Out of the eight variables typically available at diagnosis we tested for association with class membership, two variables were found to be significant at the 5% significance level before applying a Bonferroni adjustment: smoking status ($p = 0.01$; $p_{adj} = 0.08$) and the presence of

upper gastrointestinal inflammation ($p < 0.001$; $p_{adj} = 0.002$). 24% and 23% of cluster 1 and cluster 2 respectively were smokers when they were diagnosed, whereas only 7% of cluster 3 and cluster 4 smoked during this period. Only 9% of cluster 1 had upper gastrointestinal involvement at diagnosis in comparison to the 27%, 34%, and 33% in cluster 2, cluster 3, and cluster 4 respectively.

3.4 Association with Treatments

A difference in the percentage of subjects prescribed a biologic therapy within one year of diagnosis was observed across classes (Table 1): 46% of class 1 were prescribed one of these treatments, compared to 18% and 21% for class 2 and class 3 respectively.

Out of the prescriptions considered, being prescribed a thiopurine within one year of diagnosis ($p = 0.023$; $p_{adj} = 0.16$) and being prescribed a biologic either within three months ($p < 0.001$; $p_{adj} = 0.004$) or one year of diagnosis ($p < 0.001$; $p_{adj} < 0.001$) were found to be significant before Bonferroni adjustment. However, class membership could not be predicted from demographic data and biologic prescriptions (AUC of 0.68 for the multinomial logistic regression model and 0.66 for the random forest classifier).

4 Discussion

In this study, four patient clusters in the CD population with distinct FCAL trajectories have been identified and described (Figure 3). To the best of our knowledge, we are the first to apply LCMMs to characterize latent patient heterogeneity using FCAL data, although others have applied linear mixed models to FCAL data[8,9] or have applied LCMMs in other disease contexts[31,32].

We have demonstrated cluster membership is associated with smoking and upper gastrointestinal inflammation. A comparatively high number of subjects who smoked at diagnosis were found in both cluster 1 and cluster 2 despite cluster 1 being characterized by an overall decrease in FCAL and cluster 2 being characterized by a consistently high profile. The interpretation of this finding is not clear from our data. Previous research has found smoking to be associated with low drug concentrations for infliximab and adalimumab, mediating low remission rates in CD patients[33] in addition to being associated with undergoing surgery and disease progression[34]. Upper gastrointestinal involvement is likely a proxy for a more severe CD sub-phenotype. We also observed cluster membership to be associated with early biologic treatment. This is reasonable given the often-reported association between FCAL and endoscopic activity and an association between biologic treatments and endoscopic healing for CD patients[35, 36].

Whilst examining the residuals for the 4-cluster model, we found evidence against normality in the tails of the distribution, where some outliers can be observed. In most cases, this is driven by FCAL observations being truncated to be within the limits of detection of the assay (20 μg/g to 2500 μg/g). As subjects were required to have a FCAL above 250 μg/g at diagnosis to be included in the study, only the upper truncation applies for diagnostic FCAL. We believe the impact of any violation of the assumption of normally distributed residuals on our findings is minimal. If this assumption is violated, then an inappropriate number of classes can be found when solely relying on model selection metrics such as AIC and BIC[16]. Instead, we also considered visual inspection (alluvial plots, mean cluster profiles versus subject-specific trajectories) in addition to AIC and BIC as a more rigorous approach to determining the number of latent classes, avoiding the identification of spurious clusters.

The approach demonstrated here has notable advantages over the methodology used by the IBSEN study which required participants to choose which diagram they believed best described their disease activity out of four possible options[3]. Using FCAL profiles allows the quantification of inflammation in an objective manner rather than using patient reported symptom activity which may be influenced by recency bias and the tendency for patient-reported data to exhibit extreme responses[37]. Furthermore, using FCAL allows longitudinal profiles to be generated in a data-driven manner. Instead of profiles needing to be generated based on prior beliefs and opinion, we can allow these profiles to be formed naturally. Finally, FCAL profiles can be readily generated for many CD patients from electronic healthcare records without requiring active involvement from patients.

Some similarities can be observed between the clinically derived profiles in the IBSEN cohort patterns and the cluster-specific mean profiles uncovered in this study. Both studies identified a large group of patients that exhibit a decline in severity of symptoms (cluster 1 and cluster 3 in our study) and a group with chronic continuous symptoms (cluster 2 in our study). However, the IBSEN study identified a group with increasing intensity of symptoms which was not found by our analysis. Such differences may be due to the disconnect between symptoms and inflammation which is commonly seen when using endoscopic activity scores[38]. Moreover, the IBSEN study findings were gathered before the widespread emergence of biologic therapies for CD and may not represent more modern trends which may not be well known a priori: demonstrating the advantage of being able to infer subgroup profiles in a data-driven manner.

In this study, eight potential associations with variables typically available at diagnosis, and seven potential associations with treatments have been explored. As such, we potentially invite criticism due to multiple testing. Indeed, some associations reported here (e.g., between cluster

membership and smoking) fail to be significant after applying Bonferroni corrections. However, we believe our findings here are biologically plausible and in line with other published literature.

The retrospective design of this study remains a limitation, and the results reported may be due to observational biases and should not be assigned a causal interpretation. Quantifying causal treatment effects from such observational data is an active area of research and beyond the scope of this study[39,40]. The data gathering process is observational and whilst FCAL is collected routinely at all clinical interactions, subjects with more complicated disease are likely to have more measurements available. The retrospective study design also means all subjects did not have the same treatment options at the same stage in their disease trajectories, as subjects may have been diagnosed at any time between 2005 and 2017. However, the date of diagnosis, converted to the number of days the subject was diagnosed after 01/01/2001, was considered for potential association with cluster membership, and no significant association was found (p = 0.12). We also acknowledge the potential for inclusion bias in this study. The study by Plevris et al. required subjects to have an FCAL of at least 250$\mu g/g$ at diagnosis and excluded subjects which met one of the endpoints within a year of diagnosis. The former potentially excludes subjects with milder disease, whilst the latter potentially excludes subjects with more aggressive disease. Extreme trajectories may therefore have been underrepresented in our analysis. Whilst we are confident in the classes described here, additional classes may be found if the inclusion criteria were relaxed.

The clusters reported here are intended purely for exploring heterogeneity in CD and are not intended for use as predictors in a risk score. Indeed, some FCAL measurements were taken after typical outcomes of interest (e.g., surgery), hence cluster membership information is not a suitable risk factor. However, our approach provides an objective way to characterize disease

trajectory heterogeneity using a routinely collected inflammation marker, providing a proof of concept for novel longitudinal patient stratification in the context CD. It should also be noted that only a single cohort was examined in this study and generalizing these results to other populations requires caution.

5 Conclusion

We have demonstrated the suitability and utility of latent class mixed modelling for identifying clusters within the CD population based on FCAL profiles. After we found and described four clusters, we reported cluster membership to be significantly associated with smoking and upper gastrointestinal involvement. We believe our findings are an important first step towards embracing longitudinal FCAL measurements to explain disease heterogeneity in CD.

6 Figure Legends

Figure 1: Flowchart demonstrating data processing steps. FCAL: fecal calprotectin.

Figure 2: Alluvial plot demonstrating how cluster membership obtained from the fecal calprotectin profiles of Crohn's disease patients changes as the assumed number of clusters increases. The height of each band indicates the size of each cluster.

Figure 3: Log-transformed subject-specific five-year fecal calprotectin profiles for the study cohort for **A**, cluster 1; **B**, cluster 2; **C**, cluster 3; **D**, cluster 4. The red solid line represents the predicted mean trajectory for each cluster, whilst the red dotted lines represent 95% confidence intervals. The grey lines indicate the trajectory of each subject. The blue dotted line indicates an FCAL of log (250 $\mu g/g$): the common threshold for biochemical remission in Crohn's disease. See Figure S8 for the fits in the original measurement scale.

7 References

[1] Jones GR, Lyons M, Plevris N, et al. IBD prevalence in Lothian, Scotland, derived by capture–recapture methodology. *Gut* 2019; 68(11): 1953–1960. doi: 10.1136/gutjnl-2019-318936

[2] Hamilton B, Green H, Heerasing N, et al. Incidence and prevalence of inflammatory bowel disease in Devon, UK. *Frontline Gastroenterol.* 2021; 12(6): 461–470. doi: 10.1136/flgastro-2019-101369

[3] Henriksen M, Jahnsen J, Lygren I, et al. Clinical course in Crohn's disease: Results of a five-year population-based follow-up study (the IBSEN study). *Scand J Gastroenterol* 2007; 42(5): 602-610. PMID: 17454881 doi: 10.1080/00365520601076124

[4] D'Amico F, Nancey S, Danese S, Peyrin-Biroulet L. A practical guide for faecal calprotectin measurement: myths and realities. *J Crohns Colitis* 2020; 15(1): 152–161. doi: 10.1093/ecco-jcc/jjaa093

[5] Kennedy NA, Jones GR, Plevris N, Patenden R, Arnott ID, Lees CW. Association between level of fecal calprotectin and progression of Crohn's disease. *Clin Gastroenterol Hepatol* 2019; 17(11): 2269-2276.e4. doi: 10.1016/j.cgh.2019.02.017

[6] Plevris N, Lees CW. Disease monitoring in inflammatory bowel disease: Evolving principles and possibilities. *Gastroenterology* 2022; 162(5): 1456-1475.e1. doi: 10.1053/j.gastro.2022.01.024

[7] Colombel JF, Panaccione R, Bossuyt P, et al. Effect of tight control management on Crohn's disease (CALM): a multicentre, randomised, controlled phase 3 trial. *The Lancet*. 2017;390(10114):2779-2789. doi:10.1016/S0140-6736(17)32641-7

[8] De Vos M, Dewit O, D'Haens G, et al. Fast and sharp decrease in calprotectin predicts remission by infliximab in anti-TNF naïve patients with ulcerative colitis. *J Crohns Colitis* 2012; 6(5): 557–562. doi: 10.1016/j.crohns.2011.11.002

[9] Zhulina Y, Cao Y, Amcoff K, Carlson M, Tysk C, Halfvarson J. The prognostic significance of faecal calprotectin in patients with inactive inflammatory bowel disease. *Aliment Pharmacol Ther* 2016; 44(5): 495–504. doi: 10.1111/apt.13731

[10] Proust-Lima C, Philipps V, Liquet B. Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *J Stat Softw* 2017; 78(2): 1–56. doi: 10.18637/jss.v078.i02

[11] Courvoisier D, Alpizar-Rodriguez D, Gottenberg J, et al. Rheumatoid arthritis patients after initiation of a new biologic agent: trajectories of disease activity in a large multinational cohort study. *EBioMedicine* 2016; 11: 302–306. doi: 10.1016/j.ebiom.2016.08.024

[12] Jiang G, Luk AOY, Tam CHT, et al. Progression of diabetic kidney disease and trajectory of kidney function decline in Chinese patients with Type 2 diabetes. *Kidney Int* 2019; 95(1): 178–187. doi: 10.1016/j.kint.2018.08.026

[13] Plevris N, Fulforth J, Lyons M, et al. Normalization of fecal calprotectin within 12 months of diagnosis is associated with reduced risk of disease progression in patients with Crohn's disease. *Clin Gastroenterol Hepatol* 2021; 19(9): 1835–1844.e6. doi: 10.1016/j.cgh.2020.08.022

[14] Elhakeem A, Hughes RA, Tilling K, et al. Using linear and natural cubic splines, SITAR, and latent trajectory models to characterise nonlinear longitudinal growth trajectories in cohort studies. *BMC Med Res Methodol* 2022; 22(1): 68. doi: 10.1186/s12874-022-01542-8

[15] James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning* ch. 7: 297-317; Springer Texts in Statistics. Springer US. 2nd ed. 2021

[16] Bauer DJ, Curran PJ. Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods* 2003; 8(3): 338–363. doi: 10.1037/1082-989X.8.3.338

[17] Bauer DJ. Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research* 2007; 42(4): 757–786. doi: 10.1080/00273170701710338

[18] Kwak C, Clayton-Matthews A. Multinomial logistic regression. *Nurs Res* 2002; 51(6).

[19] Breiman L. Random forests. *Mach Learn* 2001; 45(1): 5–32. doi: 10.1023/A:1010933404324

[20] Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*: 241-247; Springer Series in Statistics. Springer New York. 2nd ed. 2009

[21] Hand DJ, Till RJ. A Simple Generalisation of the area under the ROC curve for multiple class classification problems. *Mach Learn* 2001; 45(2): 171–186. doi: 10.1023/a:1010920819831

[22] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2022.

[23] Proust-Lima C, Philipps V, Diakite A, Liquet B. lcmm: extended mixed models using latent classes and latent processes. https://cran.r-project.org/package=lcmm; 2021. R package version: 1.9.3.

[24] Venables WN, Ripley BD. *Modern applied statistics with S*. New York: Springer. fourth ed. 2002. ISBN 0-38795457-0.

414 [25] Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional
415 data in C++ and R. *J Stat Softw* 2017; 77(1): 1–17. doi: 10.18637/jss.v077.i01

416 [26] Constantine-Cooke N. datefixR: fix really messy dates in R. https://CRAN.R-
417 project.org/package= datefixR; 2022. R package version 0.1.4

418 [27] Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. *J Open Source Softw*
419 2019; 4(43): 1686. doi: 10.21105/joss.01686

420 [28] Kuhn M, Wickham H. Tidymodels: a collection of packages for modeling and machine
421 learning using tidyverse principles. https://CRAN.R-project.org/package=tidymodels; .

422 [29] Greenwell BM, Boehmke BC. Variable importance plots—An introduction to the vip
423 package. *The R Journal* 2020; 12(1): 343–366. doi: 10.32614/RJ-2020-013

424 [30] Brunson JC. ggalluvial: layered grammar for alluvial plots. *J Open Source Softw* 2020;
425 5(49): 2017. doi: 10.21105/joss.02017

426 [31] Chapuis N, Ibrahimi N, Belmondo T, et al. Dynamics of circulating calprotectin accurately
427 predict the outcome of moderate COVID-19 patients. *EBioMedicine* 2022; 80. doi:
428 10.1016/j.ebiom.2022.104077

429 [32] Vistisen D, Andersen GS, Hulman A, Persson F, Rossing P, Jørgensen ME. Progressive
430 decline in estimated glomerular filtration rate in patients with diabetes after moderate loss in
431 kidney function—Even without albuminuria. *Diabetes Care* 2019; 42(10): 1886–1894. doi:
432 10.2337/dc19-0349

433 [33] Kennedy NA, Heap GA, Green HD, et al. Predictors of anti-TNF treatment failure in anti-
434 TNF-naive patients with active luminal Crohn's disease: A prospective, multicentre, cohort
435 study. *Lancet Gastroenterol Hepatol* 2019; 4(5): 341–353. doi: 10.1016/s2468-
436 1253(19)30012-3

[34] Lawrance IC, Murray K, Batman B, et al. Crohn's disease and smoking: Is it ever too late to quit?. *J Crohns Colitis* 2013; 7(12): e665–e671. doi: 10.1016/j.crohns.2013.05.007

[35] Jusué V, Chaparro M, Gisbert JP. Accuracy of fecal calprotectin for the prediction of endoscopic activity in patients with inflammatory bowel disease. *Dig Liver Dis* 2018; 50(4): 353–359. doi: 10.1016/j.dld.2017.12.022

[36] Narula N, Wong EC, Dulai PS, Marshall JK, Jairath V, Reinisch W. Comparative effectiveness of biologics for endoscopic healing of the ileum and colon in Crohn’s disease. *Am J Gastroenterol* 2022; Publish Ahead of Print. doi: 10.14309/ajg.0000000000001795

[37] Vaerenbergh YV, Thomas TD. Response styles in survey research: A literature review of antecedents, consequences, and remedies. *Int J Public Opin Res* 2012; 25(2): 195–217. doi: 10.1093/ijpor/eds021

[38] Koutroumpakis E, Katsanos K. Implementation of the simple endoscopic activity score in Crohn's disease. *Saudi J Gastroenterol* 2016; 22(3): 183. doi: 10.4103/1319-3767.182455

[39] Nogueira AR, Pugnana A, Ruggieri S, Pedreschi D, Gama J. Methods and tools for causal discovery and causal inference. *Wiley Interdiscip Rev Data Min Knowl Discov* 2022; 12(2). doi: 10.1002/widm.1449

[40] Hammerton G, Munafò MR. Causal inference with observational data: the need for triangulation of evidence. *Psychol Med* 2021; 51(4): 563–578. doi: 10.1017/s0033291720005127

8 Tables

		Clusters					
	Population	Cluster 1	Cluster 2	Cluster 3	Cluster 4	*p*	*p*adj
Sex						0.157	1
Male	183 (51%)	43 (47%)	107 (56%)	24 (41%)	9 (60%)		
Female	173 (49%)	49 (53%)	84 (44%)	34 (59%)	6 (40%)		
Age at diagnosis						0.851	1
First quartile (q_1)	16.0	20.4	15.3	15.1	14.5		
Median (q_2)	27.3	29.6	26.4	26.8	21.7		
Third quartile (q_3)	48.7	45.3	49.9	50.9	51.2		
Smoking status						0.015*	0.116
Smoker	70 (20%)	22 (24%)	43 (23%)	4 (7%)	1 (7%)		
Non-smoker	286 (80%)	70 (76%)	148 (77%)	54 (93%)	14 (93%)		
Diagnostic FCAL ($\mu g/g$)						0.131	1
First quartile (q_1)	590	500	610	630	592		
Median (q_2)	820	725	900	825	660		
Third quartile (q_3)	1140	986	1270	1180	1160		
Montreal behavior						0.494	1
Inflammatory (B1)	323 (91%)	80 (87%)	175 (92%)	55 (95%)	13 (87%)		
Stricturing (B2)	30 (8%)	10 (11%)	15 (8%)	3 (5%)	2 (13%)		
Penetrating (B3)	3 (1%)	2 (2%)	1 (1%)	0 (0%)	0 (0%)		
Perianal disease (P)†						0.776	1
Present	57 (16%)	17 (18%)	28 (15%)	9 (16%)	3 (20%)		
Not present	299 (84%)	75 (82%)	163 (85%)	49 (84%)	12 (80%)		
Montreal location						0.125	1
Ileal (L1)	95 (27%)	22 (24%)	58 (30%)	9 (16%)	6 (40%)		
Colonic (L2)	140 (39%)	39 (42%)	68 (36%)	30 (52%)	3 (20%)		
Ileocolonic (L3)	121 (34%)	31 (34%)	65 (34%)	19 (33%)	6 (40%)		
Upper gastrointestinal (L4)‡						< 0.001***	0.002**
Present	84 (24%)	8 (9%)	51 (27%)	20 (34%)	5 (33%)		
Not present	272 (76%)	84 (91%)	140 (73%)	38 (66%)	10 (67%)		
5-ASAs (aminosalicylates)						0.326	1
Yes	76 (21%)	15 (16%)	48 (25%)	11 (19%)	2 (13%)		
No	280 (79%)	77 (84%)	143 (75%)	47 (81%)	13 (87%)		
Thiopurine						0.023*	0.161

Yes	250 (70%)	62 (67%)	127 (66%)	50 (86%)	11 (73%)		
No	106 (30%)	30 (33%)	64 (34%)	8 (14%)	4 (27%)		
Corticosteroids						0.983	1
Yes	298 (84%)	78 (85%)	159 (83%)	48 (83%)	13 (87%)		
No	58 (16%)	14 (15%)	32 (17%)	10 (17%)	2 (13%)		
Methotrexate						0.139	0.975
Yes	15 (4%)	8 (9%)	6 (3%)	1 (2%)	0 (0%)		
No	341 (96%)	84 (91%)	185 (97%)	57 (98%)	15 (100%)		
Exclusive enteral nutrition						0.779	1
Yes	80 (22%)	22 (24%)	39 (20%)	14 (24%)	5 (33%)		
No	213 (60%)	54 (59%)	115 (60%)	35 (60%)	9 (60%)		
Not known	63 (18%)	16 (17%)	37 (19%)	9 (16%)	1 (7%)		
Biologic within 3 months						< 0.001 ***	0.004 **
Yes	44 (12%)	23 (25%)	14 (7%)	5 (9%)	2 (13%)		
No	312 (88%)	69 (75%)	177 (93%)	53 (91%)	13 (87%)		
Biologic						< 0.001 ***	<0.001 ***
Yes	94 (26%)	42 (46%)	35 (18%)	12 (21%)	5 (33%)		
No	262 (74%)	50 (54%)	156 (82%)	46 (79%)	10 (67%)		

Table 1: Cohort characteristics and treatments prescribed to the cohort. All prescriptions were prescribed within one year of diagnosis unless otherwise stated. Percentages when stratified across clusters are out of the total number of subjects in the cluster. Biologic is defined as either infliximab, adalimumab, ustekinumab or vedolizumab prescription. † Perianal disease may be present concomitantly to B1, B2 or B3 disease behavior or separately. ‡ Upper gastrointestinal inflammation may be present in addition to ileal, colonic, or ileocolonic inflammation. *p* unadjusted p-value. p_{adj} p-value after Bonferroni correction. * Significant at a 5% significance level. ** Significant at a 1% significance level. *** Significant at a 0.1% significance level.