PREdiCCt

The PRognostic effect of Environmental factors
in Crohn's and Colitis

Statistical Analysis Plan (Baseline Data)

Version No	Final 2.0
Date Finalised	28/112022
Author(s)	Dr Linda Williams
CI Name	Dr Charlie Lees
CI Email address	charlie.lees@ed.ac.uk

Signatures	
Trial Statistician: Professor Chris Weir Christopher J. Weir	**Date:** 7 December 2022
Chief Investigator: Dr Charlie Lees	**Date:** 7th December 2022

Document Control		
Version No	**Date**	**Summary of Revisions**
1.0	13/10/2022	Finalised version
2.0	28/11/2022	Addition of disease location to predictors

Contents

Introduction

The PREdiCCt study is a prospective observational cohort study looking to assess the effect of habitual diet, genetic variation, gut microbiota and environmental factors on disease flare in inflammatory bowel disease (IBD). The study aimed to recruit 3100 patients in self-reported clinical remission and follow them up to flare of disease. To date, 2629 IBD patients have been recruited, with the study closing early due to the global pandemic.

Baseline data was collected for DNA, gut microbiome, dietary, environmental exposure and Quality of Life (QoL) analysis.

The primary objectives of the main longitudinal study are to determine which aspects of a) baseline habitual diet, b) the environment, c) genetic variation and d) the gut microbiota, predict disease flare in Crohn's disease and /or ulcerative colitis. The main longitudinal analysis of PREdiCCt is covered by the main analytical plan (PREdiCCt SAP_v1.0_18Dec2020_signed.pdf).

Faecal calprotectin is an established biomarker of gut inflammation in the diagnosis and monitoring of patients with inflammatory bowel disease. In the diagnostic phase faecal calprotectin of <50mcg/g has a negative predictive value of approx. 99% in excluding IBD. A level of >250mcg/g would typically trigger further investigation (endoscopy and imaging); intermediate levels (50-250mcg/g) are often repeated and taken in the context of other symptoms / results.

During routine monitoring of patients the threshold of <250mcg/g is widely established as a target for remission. We have established this with the faecal calprotectin ELISA assay used in the Biochemistry Department since 2005 and used for all faecal samples in the PREdiCCt Study (REF: Plevris and Lees – Gastroenterology 2022).

There is now a substantial body of literature on the role of calprotectin in the monitoring of IBD. An elevated calprotectin is known to increase the risk of subsequent disease flare and disease progression in patients presently in clinical remission. However to date, no study has detailed the effect of residual gut inflammation – as determined by faecal calprotectin – on symptoms beyond standard clinical activity indices.

The aim of the present study is characterise the effect of residual gut inflammation on symptoms, lifestyle and other psychosocial variables on patients with IBD in clinical remission.

Baseline Analytical Plan for the PREdiCCt study

The baseline analysis plan only concerns data that were collected at the participants' first recruitment visit, their first baseline questionnaires and their baseline stool and saliva samples.

Primary hypothesis Residual gut inflammation - as determined by faecal calprotectin measurement - in patients in self-reported clinical remission, affects patient symptoms, lifestyle and psycho-social well-being.

For the purposes of this baseline analysis, participants are defined as having no residual gut inflammation where their faecal calprotectin <50mcg/g, low residual gut inflammation where their faecal calprotectin is 50-250mcg/g, and high residual gut inflammation where their faecal calprotectin > 250mcg/g.

Primary Objectives

The primary objective of the baseline analysis are as follows:

1) Describe the patients in PREdiCCt with sufficiently complete baseline data – demographics, self-described disease phenotype, medical history, phenotype data (including detailed disease location), by levels of residual gut inflammation. A note of the number of missing datapoints per variable will be made.

2) Characterize the main disease associated symptoms (frequency of bowel movements, pain, urgency, blood in stool, number of flares etc) by levels of residual gut inflammation, by disease, disease location (colonic [L2 CD and any UC/IBD-U] vs ileal [L1 CD] vs ileocolonic [L3 CD]), age group and gender, smoking, medication, prior resectional surgery (CD), QoL, anxiety and depression, diet quality

3) Characterize the key QoL variables (IBD control, fatigue, depression, anxiety, sleep) by levels of residual gut inflammation, by disease, disease location (colonic [L2 CD and any UC/IBD-U] vs ileal [L1 CD] vs ileocolonic [L3 CD]), age group and gender, medication and prior resectional surgery (CD), diet quality, physical exercise, anxiety and depression

4) Characterize the habitual diet of those patients by levels of residual gut inflammation, also comparing to that of healthy controls, by disease, disease location (colonic [L2 CD and any UC/IBD-U] vs ileal [L1 CD] vs ileocolonic [L3 CD]), age group and gender and prior resectional surgery (CD)

- Do the participants meet (or fail to meet) the current UK dietary and nutrient intake recommendations impact on self-reported disease status (IBD-Control-8) and inflammation?
- Does the addition of life-style factors (smoking, sleep, anxiety etc) change this relationship?
- Do the individual dietary variables impact on self-reported disease status, inflammation and lifestyle factors (smoking, sleep, stress)?

5) Characterise the association between smoking and alcohol intake, painkiller usage and herbal remedies by QoL and anxiety / depression, by levels of residual gut inflammation and diet quality/dietary pattern

6) Characterise female reproductive health (menstruation / menopause etc) by levels of residual gut inflammation, gut symptoms, QoL variables and diet quality/dietary pattern

Statistical methods

The majority of the objectives reported above will be simple tabulations by groups and subgroups. Continuous and pseudo-continuous data will be reported as means and standard deviations (or medians and IQRs, as appropriate), categorical data will be reported as counts and percentages.

Modeling of continuous outcomes by continuous explanatory variables will be by linear regression. Binary outcomes by continuous explanatory variables will be modelled using logistic regression. Time to flare data will be explored with survival type analyses.

ROC analysis will be used in the exploratory analysis of the current thresholds for inflammation in the FC and CRP measurements.

All analyses are dependent on the availability of data. Effect sizes will be reported as point estimates and 95% Confidence Intervals, p-values (where reported) will be 2-sided.

Missing data
Data missing due to inapplicability (eg female only questions) will be excluded from descriptions of missingness. Missingness within the primary outcomes will be explored to ensure data is missing at random. Variables containing data not missing at random will be flagged, described and excluded from multivariable modelling.

Datasets
The following datasets will be used for the baseline analyses:

- Detailed baseline phenotyping data provided by the local clinician and collected in RedCap. Due to the varying levels of completion of the redcap form, only participants with a diagnosis will be included in the first instance. However, for variables where at least one option should be selected, any participants with none selected will be excluded, and this excluded number recorded. Where this is less clear, the category ‘unknown’ will be reported.
- Baseline questionnaire where completed.

The following explorations of missingness and outliers will be conducted by Nathan Constantine-Cooke, and included as a non-ECTU section of the report.

Missingness

Data which are missing due to not being applicable to the subject, for example data relating to oral contraceptive pills for male subjects, will be coded as a unique factor. This will allow the true missingness for these data to be determined. Missingness will be explored via heatmap visualisations, which present missingness patterns across subjects.

Tables of quantities of missing observations, stratified by variables suspected to influence the missingness mechanism, will be produced to determine the category of missingness (which will likely either be missing at random or missing not at random).

Outliers

For all laboratory variables, histograms will be produced to visualise the observation distributions and to identify outliers. Each observation will be standardised to a consistent unit of measurement for that variable. Where variables are typically reported as a logarithm, logarithms will be applied if a logarithm has clearly not already been applied. A team of clinicians, consisting of three team members, will determine the appropriate cut off for lab results by considering the histograms and physiological validity.

Primary Exposures and Outcomes

Variables will be collected from the patient-completed baseline questionnaire unless otherwise stated.

1) Demographics

Tabulation of the following variables by levels of residual gut inflammation:

- Primary diagnosis at study entry (CD vs UC/IBDU)
- Age
- Gender
- Ethnicity
- Socioeconomic status *(if practicable)*
- Smoking status (Current, Ex, never); e-cigarette use
- Alcohol intake (from FFQ)
- Surgical history (CD only, resectional)
- Disease location (L1/L2/L3/L4 and E1/E2/E3) & behavior (B1/B2/B3 +/-p) (from Redcap)
- Duration of disease (from date of diagnosis to date of study recruitment) (from Redcap)
- A history recent/current NSAID (ibuprofen, diclofenac, aspirin) or COX-2 (celecoxib) – daily vs once or twice /week vs once or twice /month
- A history of antibiotic exposure within the previous 6 months
- Co-morbidity
- Biologics (current / prior / never) = infliximab, adalimumab, vedo, uste, tofa (from Redcap if the baseline questionnaire has not been completed)

- Immunosuppressants (current / prior / never) = azathioprine, mercaptopurine, methotrexate (from Redcap where the baseline questionnaire has not been completed)
- Blood test results (haemoglobin, WCC, platelets, albumin, Ferritin, CRP, creatinine)
- Employment status
    - Employed
    - Unemployed
    - Retired
    - Students
    - Long term sick
    - Other (but we might consider this as unemployed. It included patients who reported it as looking after someone else or long term sick)
- Patient reported causes of flares
- Diet quality score

2) Symptoms

Stratification and comparison by levels of residual gut inflammation, disease, age group, gender, diet quality, smoking status, QoL (SF12), anxiety and depression, medication (biologic or not), disease location, prior resectional surgery (CD only)

- Stool frequency (HBI/Mayo)
- Blood in stool (HBI/Mayo)
- Number of flares in last year

3) Quality of Life

Stratification and comparison by levels of residual gut inflammation, disease, age group, gender, QoL (SF12), anxiety and depression, and disease control (IBD_Control-8), diet quality, medication, disease location, prior resectional surgery (CD only), exercise (GPAQ)

- Disease control as perceived by patient (IBD_Control-8) + VAS
- Pain (Q17.1, Q43.1, Q44.5)
- Fatigue (Q17.1)
- Depression and anxiety (HADS)
- Overall QoL (SF12 – if possible)?
- Sleep (PSQI)
- Diet (following specific diets for disease control, Q40.1)

4) Characterize the habitual diet of those patients

Stratification and comparison by levels of residual gut inflammation, disease, age group, gender, prior surgery, disease location, prior resectional surgery (CD only)

- Energy and nutrient intake
- Food groups (IBD dietary patterns)
- Diet quality score
- Ultra-processed foods intake (NOVA score)
- Total and individual sugars

- Plant-based foods intake
- Total animal protein
- Fibre (total fibre, starch, non-starch polysaccharides, soluble rich foods)
- Fatty acids

5) Lifestyle factors

Stratification and comparison by levels of residual gut inflammation, disease, age group, gender, QoL (SF12), anxiety and depression, disease control (IBD_Control-8), diet quality, disease location, prior resectional surgery (CD only)

- Smoking status
- Alcohol intake
- Painkiller usage (Q10 / Q11)
- Herbal remedies (Q12 / Q13)

6) Female reproductive health

Stratification and comparison by levels of residual gut inflammation, disease, age group, QoL (SF12), anxiety and depression, disease control (IBD_Control-8), diet quality, disease location, prior resectional surgery (CD only)

- Pregnancy / breastfeeding (Q46/ Q47)
- Contraception use (Q48)
- Period duration and symptoms and impact on IBD related symptoms (Q49)

Secondary Objectives

1) To further explore any factors univariately associated ($p<0.1$) with inflammation in a multivariable analysis. If sufficient numbers are available we will do an internal validation exercise. If results indicate further research, an external database will need to be sourced for validation.

2) To explore whether any of the lifestyle and dietary factors (sleep, exercise, smoking status, anxiety, depression, diet types) influence FC and/or CRP.

Appendix A – calculation of composite variables

SF12

Some questions within the SF12 section of the questionnaire were taken from version 1 and others from version 2. This means that the composite scores defined by the authors are incalculable, and cannot be used with reference populations. However, a within study summary can be used to compare between and within patients. This should NOT be used in any subsequent meta-analysis.

Montreal classification

For UC, IBDU and Other patients

Take variable Maximum macroscopic extent ever (max_extent):

- Rectum: E1
- Rectosigmoid + < splenic: E2
- < Hepatic + total: E3

For CD patients

- Ileum (no rectum/colon): L1 (mac_extent___4=checked)
- Rectum / colon (no ileum): L2 (mac_extent___4=unchecked, (mac_extent___5=checked AND/OR (mac_extent___6=checked))
- Rectum / colon + ileum: L3 (mac_extent___4=checked, (mac_extent___5=checked AND/OR (mac_extent___6=checked)

Partial Mayo score (UC and IBD)

The pMayo score will be calculated from the MovementsToNormal and BloodInStool variables from the baseline questionnaire. If there is sufficient data from Redcap (>50% completion), then the General Wellbeing will be included in the calculation for patients in whom it is available. If Endoscopy is also available, a complete Mayo score may be calculated.

Partial Harvey Bradshaw index (CD)

The pHBI will be calculated from the GeneralWellbeing, AnyAbdominalPain and LiquidStoolsPerDay variables in the baseline questionnaire. If there is sufficient data from Redcap (>50% completion), the abdominal mass scores may be used to calculate a full HBI for those patients.

	Baseline questionnaire	**RedCap (available by not well completed)**
HBI		
- **Abdominal pain**	V	V
- **Abdominal mass**	X	V
- **Number of liquid stools**	V	V
- **General well being**	V	V
- **EIM**	X	V
Mayo		
- **Number stools**	V	V
- **Blood in stool**	V	V
- **General well being**	X	V
- **Endoscopy**	X	V

Activity index score

Where the answer to “Are you currently in employment?” is Yes, then participant is defined as being employed.

Where the answer to “Are you currently in employment?” is No,

- Where the answer to the question “How would you describe your current status?” is “Looking after home or family”, or “Other”, then the worktype is given as “Other”

- Where the answer to the question “How would you describe your current status?” is either of the “Unemployed” options, then worktype is given as “Unemployed”.

- Where the answer to the question “How would you describe your current status?” is “Retired” then worktype is given as retired

- Where the answer to the question “How would you describe your current status?” is one of the three student options, then worktype is given as “Student”

Note, the following definitions have been provided by me, since there were 72 participants with “N/A” as the answer to “Are you currently in employment?”. If you don’t want them coded this way, just let me know.

Where the answer to “Are you currently in employment?” is “N/A” then.

- Where age is given to be less than 19, worktype is given as “Student”
- Where age is given to be 65 or older, worktype is given as “Retired”
- Where participants are of working age (19-65) and have given no further information, worktype is given as “N/A, no further info”

The two <16year olds in the Other category have details of “Young person at school” and “school pupil” and so really should be recoded as full time students in primary/secondary education.