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

Statistical Analysis Plan

Version No	1.0
Date	18/12/2020
Authors	PREdiCCt Analytical Committee
CI Name	Professor Charlie Lees
CI Email address	charlie.lees@ed.ac.uk
Study Statistician	Professor Christopher Weir

Signatures	
Trial Statistician: Christopher J. Weir	**Date: 21 April 2021**
Chief Investigator:	**Date: 19 April 2021**

Table of Contents

Document Control		
Version No	**Date**	**Summary of Revisions**
1.0 (draft)	19/03/2018	Initial Creation
1.0 (draft)	18/05/2018	Restructuring of sections 4.3 onwards
1.0 (draft)	08/08/2018	Incorporation of microbiota analysis section 6
1.0 (draft)	09/04/2019	Further edits to sections 5 and 6
1.0 (draft)	21/02/2020	Edits following TSC feedback
1.0 (draft)	06/04/2020	Final draft following analytical committee Feb 2020
1.0 (draft)	11/08/2020	Addition of descriptive summaries pre- and post- Covid-19 lockdown
1.0	18/12/2020	Finalised version 1.0 to incorporate protocol v5.0 amendments

List of Abbreviations

Abbreviation	Full name
CD	Crohn's disease
CI	Confidence Interval
ECTU	Edinburgh Clinical Trials Unit
FFQ	Food Frequency Questionnaire
HLA	Human leukocyte antigen
IBD	Inflammatory bowel disease
IBDU	Inflammatory bowel disease unclassified
IQR	Interquartile-range
ITT	Intention-to-treat
PUFAs	Polyunsaturated fatty acids
SAP	Statistical Analysis Plan
SCFA	Short chain fatty acids
SD	Standard Deviation
SOP	Standard operating procedure
UC	Ulcerative colitis

1. Introduction

This document details the statistical methodology for analysis of the PREdiCCT study, a prospective multi-centre cohort study to investigate which environmental and microbial factors are associated with disease flare in patients with established Crohn's disease, ulcerative colitis and inflammatory bowel disease unclassified (IBDU) who are currently in clinical remission.

This document has been compiled with reference to the Edinburgh Clinical trials Unit (ECTU) standard operating procedure (SOP) ECTU-ST-04 "Statistical Analysis Plans", and version 5.0 of the PREdiCCT protocol dated 17 Dec 2020.

The aim is to recruit 3100 participants, consisting of 1550 patients with Crohn's disease and 1550 with ulcerative colitis and /or inflammatory bowel disease unclassified (IBDU). Recruitment will continue until there are at least 1550 in each patient group.

2. Statistical Methods section from the protocol

Clinical Statistics

The following analyses will be conducted for the pooled sample of Crohn's and UC participants as the primary analysis. Secondary analysis, separately for the UC and CD patients, will be performed where evidence from a likelihood ratio test indicates significant heterogeneity of associations between UC and Crohn's. We will initially produce descriptive analyses of the measured exposures in the study cohort, split by clinical flare / no flare and overall. These will be presented as mean (SD) for normally distributed variables, and median (interquartile range) for those not normally distributed. Binary variables will be presented as number (percentage). We will assess the differences between these populations using a t-test for normally distributed variables, the Mann–Whitney test for non-normally distributed continuous variables, and the Fisher's Exact test for differences in proportions. We will then divide the primary exposures variables (intakes of PUFA, SCFA, and dietary fibre and bacterial diversity) into quartiles and will examine the relationship between these variables and time to first clinical flare using Cox frailty regression models which take into account variation in detectable flare rates across hospital site. For each exposure variable of interest, we will firstly fit a model adjusting for hospital site only (as a random effect); and then in a second model we will additionally adjust for potential confounders measured in the baseline questionnaire, which will be included in the model as fixed effects. Results will be expressed as hazard ratios with 95% confidence intervals. The proportional hazards assumption made when fitting each Cox regression model will be assessed. The above analysis will then be repeated for the secondary outcome of hard clinical flare.

The secondary outcomes of (i) total number of clinical flares and (ii) total number of hard flares in the 24 months follow-up period will be analysed using negative binomial mixed effects regression models, where hospital site is the random effect and all others are fixed effects. Results will be expressed as rate ratios with 95% confidence intervals.

Our primary hypothesis is that the following factors are significantly related to time to first disease flare:

1. Total animal protein intake (red meat, dairy, poultry, fish)
2.. Dietary fibre (non-starch polysaccharides)
3. N-6 polyunsaturated fatty acids

4. Dietary emulsifiers (lecithin)
5. Total bacterial gene count in stool
The results from the secondary hypotheses (listed in Table 1, Section 2.1) and secondary outcomes will be suitably cautious to reflect the high number of variables considered. Exposure variables that are significant in isolation will be interpreted more cautiously than exposure variables which are consistently significant across all secondary endpoint analyses.

Analysis of FFQs and food diaries
Food frequency responses are converted into nutrient intakes using an in-house calculation package linked to the UK food databank and pre-defined portion size or measure. The dietary intakes will be divided into quartiles across the distribution of the whole cohort, with the lowest assigned as the reference value. Hazard ratios will be calculated using Cox frailty regression, for risk of relapse according to quartiles of nutrients. Analyses will be adjusted for smoking and total energy intake, the latter allowing consideration of: body size, physical activity and metabolic rate, as well as correction for measurement error.

In the nutrient analysis, we will assess the effects of total fibre and that from cereals, fruit and vegetables. The study by Roberts in 2010 reported that the translocation of adhesive invasive E. coli was potentiated by the emulsifier polysorbate-80, which is present in ice creams and yoghurts. Correspondingly in the food-frequency questionnaire we will measure the intake of the later. Analyses will be adjusted for fibre and foods containing polysorbate-80 and lecithin.

Analysis of baseline microbiota data
Microbial DNA will be extracted from stool collected in buffer using standardised, locally validated protocols in Edinburgh (SOPs available on request). We will undertake 16S profiles and full meta-genomic sequencing of stool samples from each individual in the study. The 16S profiles are the most widely used (and inexpensive) means for quantifying microbial diversity and will allow our data to be easily integrated with published data. The additional metagenomic profiles will produce high-resolution species classification, assay variation within individual bacterial genes and survey non-bacterial organisms (e.g. viruses and fungi). This will enable us to ask a variety of additional questions beyond primary analyses of diversity, such as whether a particular strain (classified by genetic variants) within a species has the potential to produce flares. Metagenomic sequencing will be performed using the Illumina platform through an existing collaboration with the Sanger Institute.

Analysis of baseline genetic data
Genomic DNA will be extracted from salivary samples at the Edinburgh Clinical Research Facility. Aliquots of DNA will be sent to the Wellcome Sanger Institute. Whole-exome sequencing (50x coverage) will be performed as part of the Sanger Institute's 5-year plan to sequence 25,000 IBD genomes or exomes. The UK IBD Bioreource is the major recruiting mechanism for this effort. Patients will be allowed to consent to enter PREdiCCt and / or the Bioresource. All patients will be genotyped using the Affymetrix biobank array, either as part of the IBD bioresource initiative or, for those patients enrolled in the PREdiCCT study only, via a bespoke genotyping initiative. All PREdiCCT patients will have given explicit consent to cover whole-genome sequencing. The sequencing data will be made available to the PREdiCCt analytical team. The primary analysis will be of genetic associations with time to first flare. The data will in addition allow us to incorporate genetic risk profiles into our analyses, which our previous work has shown to be useful in sub-classifying disease. While testing for genetic signatures predisposing to flare will be underpowered in the current collection, the data will be valuable to contribute to international collaborative efforts.

3. General Analysis Principles

Unless otherwise stated in the sections below, the following general analysis principles will be adopted:

- The primary analyses will be conducted for the pooled sample of Crohn's, UC and IBDU participants. Secondary analyses, separately for the UC and CD patients, will be performed where evidence from a likelihood ratio test indicates significant heterogeneity between UC and Crohn's in the association of a given exposure with flare, or where there is prior knowledge, such as for smoking status, of different associations for UC and Crohn's..

- We will include all patients consenting to the Predicct study regardless of compliance with the Predicct protocol or subsequent withdrawal. Only patients withdrawing consent to use their data will be excluded.

- Numbers of participants with missing phenotypic data will be reported. Where a substantial proportion of data is missing, and appropriate data to support imputation is available, then missing data will be imputed, and sensitivity analyses performed to test the robustness of the assumptions in the imputation.

- Categorical data will be presented using counts and percentages, whilst continuous variables will be presented using the mean and standard deviation (SD) (or median and interquartile range if not normally distributed).

- There will be no formal interim analysis of the trial data and the final analyses will be performed after all follow-up data have been collected.

- All statistical tests and confidence intervals will be two-sided. In analysis plan Sections 4, 5 and 8, 95% confidence intervals will be presented with the significance of p-values assessed based on a 5% significance level. Appropriate multiplicity adjustments and/or false discovery rates will be implemented in the microbiota and genetics analyses; see Sections 6 and 7 for details.

- Except in the microbiota and genetics analyses (Sections 6 and 7), no adjustment for multiplicity will be made. Where no adjustment is made the interpretation of results will be suitably cautious to reflect the high number of variables considered. Exposure variables that are significant in isolation will be interpreted more cautiously than exposure variables which are consistently significant across all endpoint analyses. This is especially true of the secondary hypotheses and secondary outcomes.

- Centre will be adjusted for as a random effect in statistical models unless otherwise stated below.

- All analyses will include outliers unless otherwise stated.

- The study will be reported in a manner consistent with the STROBE checklist for cohort studies (von Elm, 2008).

4. Clinical Statistics Analysis Plan

4.1 Recruitment of patients, retention, and questionnaire completion

We will report number of patients confirmed eligible, included in the study, remaining in the study for two years follow-up (i.e. not withdrawing), and analysed. Where available, reasons will be given for participants formally withdrawing from the study.

The number and percentage of participants with baseline and monthly questionnaire data will be reported, at each time point and overall. Responses will be categorised as complete questionnaire; questionnaire submitted but <10% of questions unanswered; questionnaire submitted but 10-20% of questions unanswered; questionnaire submitted but 20% or more questions unanswered; or questionnaire missing. Number of patients with at most 10% of questions unanswered in at least 75% of the monthly follow-up questionnaires will be given.

4.2 Descriptive analysis

The total and average follow-up time will be computed across all patients, where follow-up time is measured as the length of time before the patient's final questionnaire is completed or patient withdrawal (whichever is later).

Frequency tables for the total number of hard and clinical (soft) flares reported will be constructed. A further descriptive summary will assess any association between the Covid-19 pandemic and flare event reporting. For the subset of participants whose follow-up spans the periods before and after commencement of lockdown in the UK, the hard and clinical (soft) flare tabulations will be further stratified into flares occurring up to and including 23 March 2020 and those occurring after 23 March 2020.

A Kaplan-Meier survival plot will be constructed showing time until first clinical flare and time to first hard flare on the same plot.

Descriptive analyses of demographic, clinical and social characteristics of study participants and their measured baseline exposures (see Table 1, section 4.4) will be calculated, split by clinical flare / no clinical flare and overall. These will be presented as mean (SD) for normally distributed variables, and median (interquartile range) for those not normally distributed. Binary, multinomial nominal and multinomial ordinal variables will be presented as number (percentage) in each category.

We will assess the differences between the clinical flare and no clinical flare populations using a t-test for normally distributed variables, the Mann–Whitney test for non-normally distributed continuous variables, and the Fisher's Exact test for differences in proportions.

4.3 Analysis of the Primary Outcome

The primary outcome is time to first clinical flare as assessed by the monthly questionnaires. A clinical flare is defined as the patient answering "no" to the question "Do you think your disease has been well controlled in the past 1 month?".

The primary exposure variables are:

1. Total animal protein intake (red meat, dairy, poultry, fish)
2. Dietary fibre (non-starch polysaccharides)
3. N-6 polyunsaturated fatty acids (PUFA)

4. Dietary emulsifiers (lecithin)
5. Total bacterial gene count in stool

Further details of the analysis of the food frequency questionnaire (FFQ) primary exposure variables 1-4 are provided in Section 5. Additional details of primary exposure variable 5 are provided in section 6. The remainder of Section 4.3 summarises the general statistical modelling strategy for time to event on the primary outcome, while Section 4.4 indicates the specific clinical exposures which will be investigated in this way.

Exposure variables will be split into quartiles and we will examine the relationship between each of these variables and time to first clinical flare using separate Cox regression models which include a gamma-distributed frailty term for site which takes into account variation in detectable flare rates across hospital site.

For each exposure variable of interest, we will firstly fit a model adjusting for hospital site only (as a random effect); and then in a second model we will additionally adjust for potential confounders measured in the baseline questionnaire, which will be included in the model as fixed effects. The confounders we will adjust for are: socioeconomic status (decile of the national index of multiple deprivation for the participant), current cigarette smoking, gender.

Results will be expressed as hazard ratios with 95% confidence intervals. To provide an indication of absolute as well as relative effects as recommended in the STROBE guidance (von Elm, 2008), the difference in event rates at 2 years will be presented. For categorical exposures, this will be calculated for each level versus a reference category. For continuous exposures, differences in event rate between quartiles will be reported, with the lowest quartile being used as a reference category.

The proportional hazards assumption made when fitting each Cox regression model will be assessed using the significance of continuous time-dependent covariates (incorporating a cubic B-spline or fractional polynomial shape if necessary) corresponding to each exposure of interest. The Akaike Information Criterion (AIC) will be used to assess whether including a time-dependent effect for an exposure improves the model fit and therefore whether hazards are not proportional. Log-log hazard plots will also be used to describe the proportionality of hazards. Due to the nature of the study design we would not expect hazards to be non-proportional, but due to the large number of exposure variables considered it would be unsurprising if some were associated with non-PH. If hazards are non-proportional for an exposure, then we will allow effect of exposure to vary by year (years 1 and years 2) by fitting an appropriate interaction term. If still non-proportional, we will allow the effect of exposure to vary by six-month periods, again via a suitable interaction term.

The following intercurrent events will all be censored in the analysis and assumed unrelated to flare outcome: loss to follow-up (i.e. no longer completing monthly questionnaires); formal withdrawals; deaths which are of unknown cause or which are not due to condition/flare. Sensitivity analyses will be performed to evaluate the robustness of the conclusions to these assumptions. Deaths due to IBD or possible flare will be treated as a flare event in the analysis.

4.4 Analyses of Clinical Exposures and the Primary Outcome

We will run the analysis described in section 4.3 for secondary exposure variables listed in the Environmental Factors section of Table 1.

Table 1. In patients in clinical remission, flares of IBD are hypothesized to be associated with (see protocol for references):

DIETARY FACTORS (Baseline habitual diet)
low levels of **dietary fibre** intake, with emphasis on examining the source of fibre from food groups (fruit, vegetables , cereals , legumes)
high levels of **sugar** and **starch** intake
high intake of **n-6 PUFAs**, with emphasis on examining source from animal souces (meat & eggs) and vegetable based oils
low levels of **n-3 PUFA**, with emphasis on examining source from fish & marine sources, eggs & dairy sources
high levels of **dietary protein** (g), with emphasis on the role of animal, dairy and vegetables sources
High levels of **emulsifiers**
ENVIRONMENTAL FACTORS
A low socioeconomic status* *controlled for in primary analysis*
People who live in urban areas
Being in a stable relationship with a single partner
People who consider themselves disabled
People working anti-social hours
A recent history of gastrointestinal infection
A recent history of non-gastrointestinal infection
A recent history of antibiotic use
A recent history of NSAID use
A history of persistent use of paracetamol-containing drugs
A recent history of oral or depot contraceptive use
Poor medication adherence
Current cigarette smoking* *controlled for in primary analysis*
A sedentary lifestyle
Exposure to significant amounts of air pollution
Recent and sustained exposure to altitude, including air travel
Poor sleep quality
A recent significant life event or persistent stress
A poor quality of life
MICROBIAL FACTORS
Lower microbial alpha diversity (as measured by e.g. inverse Simpson index)
Lower abundance of *F. prausnitzii*
Lower / Higher abundance of *Bacteroides*
Higher abundance of G*ammaproteobacteria*
Higher abundance of *Enterococcaceae*
Higher abundance of *Veillonellaceae*
On metagenomic sequencing, lower abundance of genes associated with butyrate production
GENETIC FACTORS
Male/female sex* *controlled for in primary analysis*
NOD2 mutations

4.5 Secondary outcome analyses

The same analysis as described in sections 4.3 and 4.4 will be repeated for the secondary outcome of hard clinical flare. This is defined as a clinical flare plus commencement of any new medication; altered dosing of existing medication for the treatment of IBD flare, with an increase in CRP (>5mg/L) and / or faecal calprotectin (>200mcg/g).

We will investigate the relationship of the exposures of interest with the secondary outcomes of (i) total number of clinical flares and (ii) total number of hard flares in the 24 months follow-up period using a negative binomial mixed effects regression analysis, where hospital site is the random effect and all others are fixed effects. Results will be expressed as rate ratios with 95% confidence intervals.

4.6 Validation and QC

A second statistician will separately program and check the primary analysis on the primary outcome (section 4.3), and all statistically significant effects on the primary outcome among secondary exposure variables. If the number of statistically significant effects is more than 20 then a random selection of 20 out of the total will be validated.

5. FFQ analysis plan

5.1 Descriptive analysis

Descriptive analysis of the dietary factors listed in Table 1 will be performed as described in section 4.2.

5.2 Primary analysis of the primary outcome

The primary dietary exposures of interest will be analysed with respect to the primary outcome as outlined in section 4.3. The exposures are:

1. Total animal protein intake (red meat, dairy, poultry, fish)
2. Dietary fibre (non-starch polysaccharides)
3. N-6 polyunsaturated fatty acids (PUFA)
4. Dietary emulsifiers (lecithin)

In addition to the covariates summarised in section 4.3, a categorical factor for season (as per the Met Office definition of spring [March/April/May], summer [June/July/August], autumn [September/ October/November] and winter [December/January/February]) will be included as an adjustment variable in the analysis. Possible interactions between season and the associations between the primary dietary exposures and flare rate will also be investigated in exploratory analyses.

5.3 Secondary analysis of the primary outcome

The secondary exposures of interest N-3 polyunsaturated fatty acids (PUFA), sugar intake levels and starch intake levels will be analysed as in section 5.2.

5.4 Secondary outcome analyses

The secondary outcome analyses of section 4.5 will be applied to the FFQ exposures outlined in sections 5.2 and 5.3.

6. Microbiota data analysis plan

6.1 Quality control, mapping and quantification

Sequence data will be generated and undergo initial sequencing quality control. Quality control will include inspection of base quality scores, base composition and sequence duplication levels, and will identify poor quality sequencing lanes, runs or libraries to be removed. We will align the reads to the human reference genome, plus currently unassigned human contigs, using BWA-MEM to remove human contamination. We will then remove low quality reads using SolexaQA++ (Cox, 2010) on the default settings.

We will use MetaPhlan2 (Truong, 2015) to quantify the relative abundance of different bacterial clades. We will use StrainPhlAn (Truong, 2017) and RAxML (Stamatakis, 2014) to construct phylogenetic trees of specific strains of interest across samples (see section 6.4). We will use HUMAnN2 (Abubucker, 2012) to calculate pathway abundances.

Counts for each microbiome measure (each species, genera, phyla and pathway) will be transformed using the centred log-ratio transformation, log(x/G(X)), where G(X) is the harmonic mean of read counts across all measures within that category (e.g. all species, for a species measure), as recommended by (Gloor, 2017). Zero counts will be imputed using the Zcompositions R package (Palarea-Albaladejo, 2015). Finally, we will remove microbiome measures that are inaccurately measured, due to a low abundance in a large number of samples (<2 reads observed in >50% of samples).

6.2 Descriptive analysis

We will generate a summary table of patient metagenomic characteristics, broken down by flaring and non-flaring patients. This will including total read count, % QC+, % human contamination, total number of observed species, total number of observed genes, and alpha diversity (measured by the inverse Simpson and Shannon indices).

We will test for associations between hospital site and microbiome measures, listing all phyla, genera, species or pathway that correlate significantly (Benjamini-Hochberg FDR < 0.05) with hospital site under a Kruskal-Wallis test.

We will generate two plots, each showing all samples on the first two principal coordinates of clades (species) and pathways (KO terms), coloured by hospital site. Principal coordinates will be calculated using the Bray–Curtis distance, using the R package vegan.

6.3 Primary analysis of the primary outcome

The primary microbiome exposure of interest will be gene richness, defined as the total number of bacterial genes observed in a given sample at a given sequencing depth.

We will calculate the total bacterial gene count using the method described by (Cotillard, 2013). We will uniformly down-sample the number of reads for each sample to be equal to a constant depth across samples, where that depth is chosen such that at least 90% of samples have at least this depth (removing samples below this depth). We will perform this down-sampling 30 times for each sample, and calculate the average gene richness across these samples. The bacterial gene count will be calculated as the total number of unique genes using HuMann2's gene alignments (using the *ChocoPhlAn* representative database).

The primary outcome will be defined as in Section 4.3. We will carry out two separate Cox regression analyses on time to flare. The first will control for hospital site (as a random variable), and the second will additionally control for confounders known to correlate with flare rate (smoking status, gender and cigarette smoking).

Representation of hazard ratios, confidence intervals and absolute effect sizes will be reported as in Section 4.3. We will test for non-proportionality in the survival model as described in Section 4.3. We will also test for robustness to the transformation method used, by calculating hazard ratios using log and logit transformations on normalized abundance.

6.4 Secondary analysis of the primary outcome

We will carry out secondary analyses of the abundance of candidate clades and functional pathways, as well as overall taxonomic diversity. We will test for association between time-to-flare and the transformed abundances of candidate bacterial clades, including the species *Faecalibacterium prausnitzii*, the genus *Bacteroides*, the family *Enterococcaceae* and the classes Gammaproteobacteria and Veillonellaceae, measured using MetaPhlan2. We will also test for association between time-to-flare and abundance of genes involved in butyrate production, measured as the total abundance of genes in the KEGG pathway "butanoate metabolism" measured via HUMAnN2.Finally, we will test for association between flare and the species, genus and pathway diversity of the stool samples, quantified using the Shannon and inverse Simpson indices. Time-to-flare analyses will be carried out using cox regression analyses as described in section 6.3.

6.5 Additional analyses of primary outcome

Microbiome-wide analysis: We will also carry out microbiome-wide analyses using cox regression for each transformed microbiome abundance measure (phyla, genera, species and pathway). We will calculate p-value thresholds using two different techniques. Firstly, we will use the Benjamini-Hochberg procedure (applied separately to phyla, genera, species and pathway analyses) to establish a p-value threshold corresponding to a false discovery rate of 0.05. Secondly, we will use label permutation (permuting phenotypic labels randomly to microbiome samples) to establish a "microbiome-wide significance value" for this experiment, i.e. a value that gives a family-wise p-value of 0.05 across all tests. We will report all associations that meet either of the significance criteria, but note that the latter is expected to provide a higher degree of certainty.

Principal co-ordinate analysis: Using the same survival model as in section 6.3, test for correlations between principal coordinates and time-to-flare. We will also carry out a more general PERMANOVA, to test whether time to flare correlates with Bray–Curtis distance.

Strain analysis: For species identified as significantly associated with time-to-flare in the above analyses we will test for a relationship between the specific strains present in the individuals and the time to flare, using the miLineage method (Tang 2017).

7. Genetic data analysis plan

7.1 Genotype calling and quality control

Genome-wide microarray-based genotyping
Patient DNA will undergo genotyping using the Affymetrix biobank array, either as part of the IBD Bioresource initiative or, for those patients enrolled in the PREdiCCT study only, via a bespoke genotyping initiative. Genotypes from each batch will be called separately at the Wellcome Sanger Institute, using a standard genotype calling algorithm with default settings (e.g. optiCall). Variants will be aligned to the positive strand and any duplicated variants removed. In order to identify potential sample swaps, will use the X data chromosome to impute the biologic sex via the F-statistic and compare it to the patient's self-reported sex. Data from sample swaps that can't be resolved will be removed from further analysis. Samples with a genotype call rate <80% will be removed. Next, variants with a call rate < 80% will be excluded. Following this initial QC, we will remove samples with a call rate < 95% and markers with call rate < 95% (or <98% if the marker has a minor allele frequency less than 1% in Europeans). We will exclude duplicated or closely-related samples (closer than third-degree relatives). We will use the 1000 Genome Project's principal component projections to identify the genetic ancestry of the individuals in the PREdiCCt cohort. Individuals that are within well-defined genetic clusters will be analysed separately, in order to avoid population stratification during the association testing. Variants significantly deviated from Hardy-Weinberg equilibrium within a single ancestry group will be identified and removed. Monomorphic variants will also be excluded.

Following completion of this quality control, we will undertake genotype imputation using the optimal haplotype reference dataset for European samples (currently the Haplotype Reference Consortium plus 1000genomes data). Variants with a post-imputation info score <0.4 will be excluded. To maximise resolution of the HLA , we will undertake HLA imputation using the HiBag software.

Whole-exome sequencing
Patient DNA will undergo whole-exome sequencing at the Sanger Institute using Illumina sequencing technology. Exomes will be sequenced at a target depth of 50x to ensure that >80% of target bases will have coverage of 20x or higher. Agilent V5 exome capture will be used. Libraries will be merged by the Sanger Institute's New Pipeline Group and stored in CRAM format to facilitate future data sharing on the European Genome-Phenome Archive. Variant calling will be handled by the Sanger Institute's Human Genome Informatics group using the latest version of the GATK variant caller (currently version 4). In order to improve the calling, samples will be jointly called with 15,000 other IBD exomes and large collections of population controls, including 50,000 individuals from UK Biobank.

Sample and variant quality control will be performed by the Anderson team at the Sanger Institute. Poorly genotyped samples with low mean depth, low call rate, high chimeric read rate, and a high rate of contamination will be identified and remove. Data from samples that have an outlying transition to transversion ratio, number of variants or heterozygosity rate (as defined by the median plus/minus four time the median absolute deviation) will also be removed. In order to identify potential sample swaps, will use the X data chromosome to impute the biologic sex via the F-statistic and compare it to the patient's self-reported sex. Data from sample swaps that can't be resolved will be removed from further analysis. We will exclude duplicated or closely-related samples (closer than third-degree relatives). We will use the 1000 Genome Project's principal component projections to identify the genetic ancestry of the individuals in the PREdiCCt cohort. Individuals that are within well-defined genetic clusters will be analysed separately, in order to avoid population stratification during the association testing. Finally, we will calculate the principal components for samples within the cohort, in order to account for potential batch effects. Overall, the sample quality procedures will be similar to those used by Karczewski for gnomAD.

Once sample QC has been completed we begin quality controlling variant sites and genotype calling. Initially, we will apply machine learning methods (variant quality score recalibration (VQSR), based on Gaussian mixed models) to identify variants that have similar quality metric values to known 'high-quality' variants in external cohorts. Quality metrics like depth, mapping quality, quality by depth will be used. We will evaluate the tradeoff between sensitivity and error rate to select the most appropriate VQSR score cutoff. Further manual filtering is likely to be required, but the specific parameters and metrics will have to be empirically derived for the specific cohort. We will validate that the majority of identified post-filtering variants are found in large scale genomics databases, such as gnomAD.

7.2 Analysis of genetic associations with the primary outcome

Time to flare will be used as the primary phenotype for genetic association analysis. Statistically significant clinical, dietary and lifestyle covariates will be identified for inclusion in subsequent association analyses. We will incorporate time-to-event information in order to increase the power to detect associations. Cox proportional hazard models will be fit to perform the genome- and exome-wide association analyses. An extension of the Cox model – time-varying proportional hazard models – will be used to account for the situations where one or several of the covariates change between the start of the study and the flare (e.g. treatment regime change).

In addition, we plan to analyse secondary outcomes described in 4.5 (e.g. the number of flares). The majority of these are quantitative or binary traits and will be analysed using standard GWAS techniques.

We will perform quality control of the association analyses summary statistics via quantile-quantile plots and chi-squared tests to ensure that we sufficiently control for cryptic population structure. Individual significant associations will be further examined to ensure that they are not false-positive associations (e.g. by checking that the association is not driven by a particular sequencing batch or that its frequency is consistent with the reported population-scale frequencies). To correct for testing of many millions of genetic markers, we will use the standard genome-wide significant threshold of $5x10^{-8}$ for declaring significance in the GWAS. A standard exome-wide significance threshold of $1x10^{-5}$ will be used for declaring variants in the exome-sequencing to be significantly associated with outcomes (both primary and secondary).

8. Predictive model analysis

Finally, we will address the objective to build predictive models of IBD prognosis and natural history utilising clinical, environmental, microbial and genetic data. Given the likely number of primary outcome flare events observed, this analysis will be exploratory and will focus on model development and internal validation. As with any predictive model an external validation step will be required subsequently.

A multiple Cox proportional hazards regression model for time to first flare will be developed, in which study centre will be included as a random effect. If several centres have recruited a small number of participants, grouping of centres will be considered to make it more feasible to fit such a random effect. Candidate predictors will be those clinical, environmental, microbial and genetic variables for which a significant univariate association with the primary outcome was identified.

Where dimension reduction is required due to an excessive number of candidate predictors, clinician knowledge and literature external to PREdiCCt will be used to decide which variables to omit. If the number of variables is still too large relative to the number of flare events following this process, statistical approaches such as principal components analysis will be used to move to a smaller number of readily interpretable variables.

The adjusted association between each variable and time to first flare will be reported as a hazard ratio and 95% confidence interval. Model discrimination will be quantified using the c-index. The discrimination and calibration performance of the model will be determined using optimism-corrected bootstrap internal validation. If feasible, conditional independence and mediation analyses will be investigated to identify potential causal pathways to flare outcomes.

9. References

Abubucker S, Segata N, Goll J, Schubert AM, Izard J, et al. Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. *PLoS Comput Biol* 2012;8(6):e1002358. doi:10.1371/journal.pcbi.1002358

Cotillard A, Kennedy SP, Kong LC, Prifti E, Pons N, Le Chatelier E, Almeida M, Quinquis B, Levenez F, Galleron N, Gougis S, Rizkalla S, Batto JM, Renault P; ANR MicroObes consortium, Doré J, Zucker JD, Clément K, Ehrlich SD. Dietary intervention impact on gut microbial gene richness. *Nature* 2013;500(7464):585-8.

Cox et al.: SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 2010;11:485.

Gloor GG, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 2017;8:2224.|doi:10.3389/fmicb.2017.02224

Karczewski KJ et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. https://doi.org/10.1101/531210

Palarea-Albaladejo J, Martín-Fernández JA. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems* 2015;143:85-96.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30(9):1312–1313.

Tang Z-Z, Chen G, Alekseyenko AV, Li H. A general framework for association analysis of microbial communities on a taxonomic tree. *Bioinformatics* 2017;33(9):1278-1285.

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* 2015;12:902-903.

Truong DT, Tett A, Pasolli E, et al. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* 2017;27:626-638.

von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol* 2008;61(4):344-9.