Title: PROFILE Trial Statistical Analysis Plan

Date: 12 May 2023

TiBDCCTU/TPL007V2 Approved ??/??/????

Statistical Analysis Plan

TRIAL FULL TITLE	**PR**edicting **O**utcomes **F**or Crohn's d**I**sease using a mo**L**ecular biomark**E**r (PROFILE) trial
EUDRACT NUMBER	
SAP VERSION	2.0
ISRCTN NUMBER	ISRCTN11808228
SAP VERSION DATE	12 MAY 2023
TRIAL STATISTICIAN	Simon Bond
TRIAL CHIEF INVESTIGATOR	Miles Parkes
SAP AUTHOR	Simon Bond

1 SAP Signatures

I give my approval for the attached SAP entitled Profile Final Analysis dated Version 2.0 12 MAY 2023

Chief Investigator

Name: Miles Parkes

Signature: ____________________

Date: ____________________

Statistician

Name: Simon Bond

Signature: ____________________

Date: ____________________

2 Table of Contents

3 Abbreviations and Definitions

AE	Adverse Event
CRF	Case Report Form
CRP	C-reactive Protein
DMC	Data Monitoring Committee
DoB	Date of Birth
EMA	European Medicine Agency
EQ-5D	Five dimensional Euro Quality of Life score
HBI	Harvey Bradshaw Index
IBDQ	Irritable Bowel Disease Questionnaire
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
IMP	Investigational Medical Product
MRE	Magnetic Resonance Elastography
PI	Principle Investigator
PT	Preferred Term
QC	Quality Control
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SE	Standard Error
SES-CD	Simple Endoscopic Score in Crohn's Disease
SOC	System Organ Class
TSC	Trial Steering Committee

4 Introduction

4.1 Preface

The hypothesis is that biomarker-driven stratification will facilitate personalised therapy in Crohn’s disease, and will improve clinical care. It will do this through identification of a group of participants destined to develop a more severe, relapsing course and who will benefit from “Top-Down” therapy while protecting those participants destined to experience more indolent disease from the risks and side-effects of unnecessary immunosuppression.

This is a randomised, multi-centre, biomarker-stratified, open-label trial in participants newly diagnosed with Crohn’s disease (enrolled within 6 months of diagnosis). Participants are randomised to either “Top-Down” or “Accelerated Step-Up” therapy.

4.2 Purpose of the analyses

These will be the final analyses that will be performed, based on the full and complete data set of the trial. They will assess the efficacy endpoints and safety, and estimate the interaction effect between biomarker and treatment.

Of note, the primary analysis study-wide will be focusing on the biomarker-treatment interaction and not the main effects. However, a key and complementary analysis will be examining the relative treatment and safety effects between “Accelerated Step-Up” and “Top-Down” therapy arms.

5 Trial Objectives and Endpoints

5.1 Trial Objectives

(ICH E3; 8.)

To demonstrate that a whole-blood prognostic biomarker can improve clinical outcomes, quality of life, and health resource allocation by facilitating the delivery of personalised therapy from diagnosis in Crohn’s disease.

5.2 Endpoints

(ICH E9; 2.2.2)

5.2.1 Primary Endpoint

Incidence of sustained surgery and steroid free remission from completion of a protocolised (maximum 8-week regimen) steroid induction treatment through to week 48.

Remission at each visit is a composite of two conditions

- HBI score <=4
- Absence of objective evidence of inflammation: both CRP<=ULN and calprotectin<200 μg/g. If both values are missing then the condition is deemed missing. If just one value is missing then it is assumed to be below the threshold.

If either or both conditions hold then the participant is in remission at the visit.

This is equivalent to HBI<=4 at all assessments, or if symptoms are present (HBI>=5) there is no objective evidence of inflammation (CRP<=ULN and calprotectin<200). Requirement for an extended induction or additional course of systemic steroids or surgery for active Crohn's disease would result in failure to meet primary outcome measure.

The outcome of each possible permutation of HBI + CRP + calprotectin results is given below

HBI	CRP	calpro	Outcome
<5	<=ULN	<=200	remission
<5	<=ULN	missing	remission
<5	missing	<=200	remission
<5	missing	missing	remission
<5	<=ULN	>200	remission
<5	>ULN	<=200	remission
<5	>ULN	>200	remission
<5	>ULN	missing	remission
<5	missing	>200	remission
>=5	<=ULN	<=200	remission
>=5	<=ULN	missing	remission
>=5	missing	<=200	remission
>=5	missing	missing	missing
>=5	<=ULN	>200	flare
>=5	>ULN	<=200	flare
>=5	>ULN	>200	flare
>=5	>ULN	missing	flare
>=5	missing	>200	flare

5.2.2 Secondary Endpoints

1. Endoscopic remission at week 48 (defined by absence of ulceration i.e. SES-CD ulcer subscore = 0). Centrally-read endoscopic scores will be used where available. Where these are not available locally-read scores will be used.
2. Quality of life averaged across weeks 16, 32 and 48 (using disease specific IBD-Q score).
3.i Number of flares requiring treatment escalation by week 48 primary follow-up period.* Determined from adjudication of CRF data.
3.ii Cumulative steroid exposure by week 48 primary follow-up period (defined by number of courses of steroids for active Crohn's disease). Determined from adjudication of CRF data.
3.iii Number of hospital admissions and surgeries related to Crohn's disease by week 48 primary follow-up period. The number of hospital admissions will be derived from the AE information; the number of surgeries is directly captured in the CRF. The two counts will simply be added together without attempting to identify coincident hospitalisation-surgery events.

*flares will be considered as protocol definition of flare indicating need for treatment escalation

5.2.3 Tertiary endpoints

- Incidence of sustained surgery and steroid free remission from completion of a standard (maximum 8-week regimen) steroid induction treatment through to week 48 (when remission defined using clinical parameters alone, HBI < 5).
- Clinical remission (defined as HBI < 5) at weeks 4,16, 32 and 48.
- Average clinical disease activity (comparison of mean HBI scores in each group) at weeks 4,16,32 and 48.
- Biochemical remission (defined as CRP $\leq$ ULN and calprotectin <200) at weeks 4,16,32 and 48.
- CRP response (comparison of mean CRP scores in each group) at weeks 4,16,32 and 48.
- Calprotectin response (comparison of mean calprotectin scores in each group) at weeks 4,16,32 and 48.
- Incidence of 2 or more treatment escalations for flares of Crohn's disease.
- Time to event, time from baseline to first flare or need for surgery for Crohn's disease, which may occur during the protocolised induction course of steroid medication.
- Time to event, time from baseline to second flare or need for surgery for Crohn's disease.
- Time to event, time from baseline to starting on anti-TNF therapy for Crohn's disease.
- Patient reported clinical remission (using score generated from abdominal pain and stool frequency components of HBI score – abdominal pain $\leq$1 and stool frequency $\leq$3) at weeks 4,16,32,48.
- Steroid free clinical remission (defined as HBI < 5 and no current use of or plan to prescribe steroids) at weeks 4,16,32,48.
- Steroid-free biochemical remission (defined as CRP <ULN and calprotectin <200 and no current use of or plan to prescribe steroids) at weeks 4,16,32,48.
- Steroid-free endoscopic remission (defined as absence of ulceration i.e. ulcer subscore=0 and no current use of or plan to prescribe steroids) at week 48.
- Endoscopic remission at week 48 using video and images from end of trial. Defined by ulcer subscore=0 using central-reads from videos and images where available, in combination with local-reads (whenever video or imaging central reads not available).
- Endoscopic remission at week 48 incorporating total SES-CD score. Defined by ulcer subscore=0 + SES-CD score <4, using central-reads from videos where available, in combination with local-reads (whenever video central reads not available).

- Endoscopic remission at week 48 defined by ulcer subscore=0 + SES-CD score <4, using only locally-read endoscopic scores.
- Endoscopic response (defined by SES-CD drop of ≥50% from baseline SES-CD score) at week 48 using only locally-read scores from all participants.
- Deep endoscopic remission (defined by total SES-CD score of 0) at week 48, using centrally-read videos where available, in combination with local-reads when video central reads not available.
- Deep endoscopic remission (defined by total SES-CD score of 0) at week 48, using only locally-read endoscopic scores.

- Endoscopic remission at week 48 in only those who had ulcers at the index colonoscopy (i.e. ulcer subscore of ≥1 on the index colonoscopy). Endoscopic remission defined as absence of ulceration i.e. ulcer subscore= 0). Centrally-read endoscopic scores will be used where available, and locally-read scores will be used only if central scores are not available.
- IBD-specific quality of life response (comparison of mean IBD-Q scores in each group) at weeks 16, 32 & 48 individually.
- IBD-specific quality of life remission (defined by IBD-Q score of ≥170) at weeks 16,32 and 48.
- IBD-specific quality of life improvement/response (defined as IBD-Q increase of ≥16 from screening visit IBD-Q score) at weeks 16,32 and 48.
- Generic quality of life response (comparison of mean EQ-5D scores in each group) at each of weeks 16,32,48. The mapping used to calculate a utility value is taken from reference [Devlin et al. 2016.]
- Generic quality of life improvement/response (defined as EQ-5D increase of ≥10 from the screening visit EQ-5D score) at weeks 16,32,48.
- IBD-Q bowel symptom improvement/response (≥8 increase in IBDQ bowel symptom domain from the screening visit) at weeks 16,32,48.
- IBD-Q fatigue improvement/response (≥1 increase in IBDQ fatigue symptom domain from the screening visit) at weeks 16,32,48.
- Weight.
- Blood Cell Counts (Haemoglobin, White cell count, platelet count) at weeks 4,16,32,48.
- Biochemical levels (CRP, albumin) at weeks 4,16,32,48.
- Metabolite levels (6TGN & 6MMP) at weeks 16,32,48.
- Perianal disease (yes/no, and then 4 non-exclusive classifications of: anal tag, anal fissure, anal fistula, perianal abscess) at weeks 4,16,32,48.
- Development of peri-anal abscess or fistula (development of peri-anal abscess / fistula vs no development of peri-anal abscess / fistula) at weeks 4,16,32,48.
- Development of endoscopic stricture by week 48 (development of stricture vs no development of stricture).
- Anti-TNF therapy at last observation: no anti-TNF therapy; monotherapy anti-TNF; combination therapy anti-TNF. Determined from adjudication of CRF data.
- Thiopurine at last observation within each participant: no thiopurine; optimised metabolite levels (6-TGN ≥235); non-optimised levels (6-TGN <235).

5.2.4 Exploratory Endpoints

Exploratory analyses (of data obtained alongside and to be linked to trial database) may be produced if the data are available in a timely fashion.

- *HLA-DQA1*05* variants at baseline.
- *NUDT15* variants at baseline.
- Infliximab drug levels at week 48 or at the timepoint closest prior to stopping infliximab.

- Imaging bowel inflammation (by comparing mean simplified MaRIA scores in each group) at week 48.
- Imaging remission (defined as <6 using simplified MaRIA score) at week 48.*
- Imaging response (defined as drop of ≥50% from baseline simplified MaRIA score) at week 48.*
- Imaging bowel damage (by comparing mean Lemann Index scores in each group) at week 48.*

*MRI simplified MaRIA and Lemann Index scores will be obtained only from centrally-read MRI scans.

5.2.5 Other samples collected during PROFILE

The following are captured as samples, but will not be processed during the primary trial follow-up period nor included in the trial database. Hence they will not be considered within the scope of this SAP but may be analysed later.

- RNA transcriptomic data.
- Serum proteomic data.
- Faecal microbiome data.
- Faecal metabolomic data.
- Histopathological data.

6 Trial Methods/

6.1 General Trial Design and Plan

(ICH E3;9)

The trial is a parallel randomised control trial with two arms. The treatments are two different treatment strategies (Accelerated Step-up; Top-down) of corticosteroids, immunomodulators and Infliximab, with additional protocolised criteria and regimens for rescue therapy; see the protocol for precise descriptions.

The randomisation is open-label, however the biomarker stratification factor measured at baseline is kept blinded to participants and clinicians.

Participants are screened and if recruited are then put on an initial standard course of treatment. Randomisation occurs approximately 2 weeks later, if eligibility is confirmed, to one of two dosing regimens. Observations are subsequently taken at weeks 4, 16, 32, and 48. Ad hoc visits may occur if the condition deteriorates and modifications to the dosing, within the guidelines given in the protocol, need to be made.

6.2 Inclusion-Exclusion Criteria and General Trial Population

(ICH E3;9.3. ICH E9;2.2.1)

The trial population is adult participants diagnosed with Crohn’s disease within 6 months. See the protocol for a precise list of detailed criteria.

6.3 Randomisation and Blinding

(ICH E3; 9.4.3, 9.4.6. ICH E9; 2.3.1, 2.3.2)

Following biomarker assessment, participants in each biomarker subgroup were randomly assigned (1:1) to either "Top-Down" or "Accelerated Step-Up" therapy, using a computer-generated algorithm. At the outset of the trial this occurred within 14 days (plus/minus 5 days). However, given some delays for sites to get results of screening investigations returned in time for a baseline visit, an amendment was made so that this would occur 14 days after screening (plus 10 days). Stratified block randomisation was used, stratifying on biomarker subgroup, (IBDhi/IBDlo), mucosal inflammation (mild / moderate / severe) and disease location (colon-only/other) with a randomly generated block size.

6.4 Trial Variables

The following table is taken from the protocol

Procedure	**Screenin g Wk -2**	**Baseline Wk 0**	**Wk 4**	**Wk 16**	**Wk 32**	**Wk 48**	**Ad hoc**
Written consent	✓						
Demographics (DOB, gender, initials, ethnicity)	✓						
Medical History	✓	✓					
Disease assessment - Harvey Bradshaw Index (HBI)	✓	✓	✓	✓	✓	✓	✓
Perianal disease review	✓	✓	✓	✓	✓	✓	✓
Concomitant medication	✓	✓	✓	✓	✓	✓	✓
AE reporting		✓	✓	✓	✓	✓	✓
Weight in Kg	✓	✓	✓	✓	✓	✓	✓
Height in cm	✓						
Physical examination	✓	✓	✓	✓	✓	✓	✓
Eligibility confirmed		✓					
Randomisation		✓					
PAXgene RNA tube for biomarker assessment	✓						
PAXgene RNA tube for research sample	✓	✓		✓	✓	✓	✓

Procedure	Screening Wk -2	Baseline Wk 0	Wk 4	Wk 16	Wk 32	Wk 48	Ad hoc
Serum tube	✓			✓	✓	✓	
EDTA (Ethylenediaminetetraacetic acid) tube	✓						
Bloods (Full Blood Count, CRP, Urea & Electrolytes, Creatinine, Liver Function TEst)	✓[†]	✓	✓	✓	✓	✓	✓
Blood results review	✓	✓	✓	✓	✓	✓	✓
Screening (Hepatitis B & C, Varicella Zoster Virus, Thiopurine Methyltransferase)	✓						
Assessment of TB	✓						
Pregnancy test	✓						
Bloods (thiopurine metabolites)			✓[α]	✓[α]	✓[α]	✓[α]	✓[α]
Faecal Calprotectin	✓[†]	[†]		✓	✓	✓	✓[β]
Buffered stool sample	✓					✓	
Stool sample for Microscopy, culture and sensitivity							✓
Primary outcomes review (steroids and surgery)			✓	✓	✓	✓	✓
IBDQ & EQ-5D Qol	✓	✓		✓	✓	✓	
Resource usage questionnaire		✓		✓	✓	✓	
Endoscopy	✓*					✓	
MR enterography (MRE)	✓¥					✓	

Baseline visit should be 2 weeks +10 days from screening. Week 4 visit should be within -10 days to +4 weeks of the schedule. Week 16 to 48 visits should be within ± 4 weeks of the schedule.

For much of the time during the COVID pandemic from March 2020 onwards, many clinic visits were carried out virtually. For most of the observations this had little impact, the exceptions being the colonoscopy, which could not take place, and two components of the HBI score, that would normally require a trained clinician to perform an assessment but was instead done by the participant themselves under virtual guidance. The final week 48 visit window for colonoscopy and MRE was extended up to three months following the end of trial visit. This timeline was selected based on expert clinical opinion, agreed with by the TSC, that results within this window would still be valid and accurate.

Variables collected outside of the time windows will not be used in efficacy analyses, but safety events will be included.

HBI values are a composite of several individual questions, with ordered categorical responses that are mapped to integers (0,1,2,..) and then added together. One of the components is a physical examination, which is normally undertaken by a trained clinician to assess for any abdominal masses. However, during the pandemic from March 2020, this in-person physical examination was not possible in many instances and a proxy assessment was made over the telephone by the participant guided by the local investigators. Collection of this data as reported by participants, was considered appropriate given the growing literature demonstrating reliability of participant-reported information and notable in this instance high association between HBI abdominal mass sub-score reported by participants compared to physician-assessed HBI abdominal sub-score (Echarri et al. Telemed J E Health, 2020).

A scaled version of the HBI will be considered that removes this abdominal mass component, as a complimentary endpoint for summary statistics and figures. The full HBI score, using either the clinician’s or participant’s assessment, will be used in the primary analysis, but sensitivity analyses will treat these as missing, and potentially impute the missing component. The primary endpoint is an aggregate over all visits, that assesses if HBI <5 and/or absence of objective evidence of inflammation, and if surgery occurred or an additional or extended induction course of steroids have been taken; so the value of this binary endpoint can still be determined, even if the HBI at one or more visits has the component missing.

7 Sample Size

(ICH E3; 9.7.2. ICH E9; 3.5)

Remission rates were estimated using data pertaining to the clinical phenotype of IBDhi and IBDlo participants and data from the literature regarding response to early anti-TNFα (including the original “Step Up-Top Down” trial, D’Haens et al. Lancet, 2008; the SONIC trial, Colombel et al. NEJM, 2010; and subgroup analyses of large anti-TNFα trials, D’Haens. Nat Rev Gastro Hepatol 2010).

Estimated remission rates were: IBDhi: “Accelerated Step-Up” 0.3, “Top-Down” 0.7; IBDlo: “Accelerated Step-Up” 0.8, “Top-Down” 0.9. The prevalence of IBDhi/ IBDlo is 0.5/0.5 based on all of the cohorts in whom the classifier has been assessed. The primary comparison is powered as an interaction analysis, where the interaction refers to the difference between the relative benefits of "Top-Down" over "Accelerated Step-Up" in each subgroup. Based on the remission rates and subgroup prevalence rates above, an interaction of 0.3 can be detected with a power of 92% at a 2-sided 5% significance level with a total sample size of 333.

To allow for a ~17% drop out rate, from the outset of the trial the goal was to recruit 400 participants across approximately 50 sites around the United Kingdom.

8 General Considerations

8.1 Timing of Analyses

The trial closed to new participant screening in December 2021, with the final randomisations for the trial in January 2022. The recruitment rate had initially slowed substantially during the COVID pandemic but subsequently increased, so that the trial almost recruited to target (n=390 participants randomised). In addition, the dropout rate was less than planned for, as monitored during ongoing DMC and TSC meetings, hence the final recruitment number of n=390 was felt sufficient to maintain adequate power according to the original trial design.

8.2 Analysis Populations

(ICH E3; 9.7.1, 11.4.2.5. ICH E9; 5.2)

- **Full population**
    - All participants who were randomised and met eligibility criteria (all inclusion and no exclusion criteria) for the PROFILE trial.
- **Safety population**
    - All participants who were randomised and received any trial treatment (including the initial induction course of steroid medication).
- **Modified per-protocol treatment population** - All participants who did not substantially deviate from the protocol for treatment to be determined on a per-participant basis. All participant cases will be reviewed by an expert adjudication committee (Prof Miles Parkes, Dr James Lee, Prof James Lindsay).
    - Treatment escalation without meeting criteria for flare.
        - Treatment escalation in absence of HBI score ≥7, AND CRP >ULN OR calprotectin ≥200ug/ml.
    - Treatment not escalated despite meeting criteria for flare.
        - HBI score ≥7, AND raised CRP >ULN OR calprotectin ≥200ug/ml.
    - Treatment escalated but not in accordance with the trial protocol.
        - E.g. immunomodulator step skipped and started on infliximab; infliximab skipped and started on adalimumab; infliximab or immunomodulator not prescribed but given additional course of steroid medication. This list is not exhaustive.
- **Modified per-protocol treatment, schedule and procedures population** - All participants who did not substantially deviate from the protocol including all aspects of treatment, trial visits and performance of trial procedures, to be determined on a per-participant basis. All participant cases will be reviewed by an expert adjudication committee (Prof Miles Parkes, Dr James Lee, Prof James Lindsay).
    - Observations made outside the visit window, using the amended time window from Baseline for all participants.
- **Pre-COVID population**
    - All participants completing the study prior to March 2020.
- **Peri-COVID population**
    - All participants randomised or in follow-up after March 2020.

8.3 Covariates and Subgroups

(ICH E3; 9.7.1, 11.4.2.1. ICH E9; 5.7)

The analyses will estimate for the main effect of the biomarker and the interaction between the biomarker and treatment, adjusting for baseline variables:

- Mucosal Inflammation (mild / moderate / severe)
- Disease location (ileal / colonic / ileocolonic) as recorded locally at baseline on the Medical history form
- Disease behaviour (inflammatory / other) as recorded locally at baseline on the Medical history form
- Smoking status (never / former / current)
- Age (16-39,40-64,65+)
- BMI (0-19.9, 20-24.9,25-29.9, 30+)
- CRP (continuous)
- Calprotectin (continuous)
- Course of glucocorticoids prior to trial enrolment (yes/no)
- Time from date of endoscopy to date of screening
- HBI score

Subgroup analyses of potential interest. Subgroups with potential biological plausibility for differential effects on biomarker-treatment interaction effect were selected for assessment.

- Age in years at baseline (<60 vs ≥60)
- Gender (male vs female)
- Ethnicity (European origin vs other)
- Smoking status at baseline (current smoker vs other)
- Obesity (BMI <30 vs ≥30)
- Time from diagnosis to inclusion in days (<30 vs ≥30)
- Disease location (ileal / colonic / ileocolonic) as recorded locally at baseline on the medical history form
- Stricturing disease at baseline (passable vs non-passable stricture based on SES-CD sub-score)
- Severe endoscopic severity at baseline using SES-CD score (≥6 vs <6 for ileocolonic and colonic disease, and ≥4 vs <4 for ileal disease)
- Presence of ulcers at baseline using SES-CD ulcer sub-score (0 vs ≥1 for ulcer sub-score in total SES-CD)
- Steroid use at baseline (prednisolone vs budesonide)
- Previous abdominal surgery at baseline (unrelated to Crohn's disease) vs no previous abdominal surgery
- Any other immune-mediated inflammatory disease at baseline (IMID) vs no other IMID

8.3.1 Post-Baseline Subgroups

The following groupings are only observable post-baseline. As such they cannot be used to determine treatment choice, but may be indicative of a mediating effect of the treatment mechanism. They are included as tertiary endpoints in 5.2.3.

The primary and secondary endpoints will have summary tables provided using these as subgroups, but not formal regression analysis. Any comparison between such post-baseline subgroups is dependent on there not being any confounding between the subgroup incidence and the endpoint, which is untestable and thus such comparisons will be considered with caution.

- Development of peri-anal fistula by week 48 (development of peri-anal fistula vs no development of peri-anal fistula)
- Development of endoscopic stricture by week 48 (development of stricture vs no development of stricture)
- Three levels of therapy: no anti-TNF therapy; monotherapy; combination therapy.
- Thiopurine at last observation within each participant: no thiopurine; optimised metabolite levels (6-TGN $\geq$235); non-optimised levels (6-TGN <235)

8.3.2 Potential Data

The following data may be available at the time of final analysis, and if so, will be used to define subgroups.

- Optimised infliximab trough drug levels ($\geq$5ug/g) vs non-optimised infliximab trough drug levels ($\leq$5ug/g). This will be treated as a post-baseline subgroup.
- HLA-DQA1*05 variants vs no HLA-DQA1*05 variants

8.3.3 Coding Details

The majority of variables to be used as predictor variables in the regression models are binary. The parameterisation used to represent the effect of a binary variable will be of the form +β/2, and -β/2 for each of the two possible values. Thus β directly estimates the difference in the expected values between the two values {+β/2 – (-β/2)}= β . This is achieved in the coding by setting the *contrast* to be (+1/2, -1/2). Furthermore when interactions with a second variable are included, this interpretation of the *main* effect is preserved, as the equal-weighted average of the predicted effect of the first variable within each of the levels of the second variable. To verify this, one can form the design matrix for each possible combination of predictor variables, from which the predicted value is obtained by calculating the matrix product with the vector of coefficients; inverting the design matrix gives the interpretation of the coefficients—differences of weighted averages of the group-level expected values—which is given below for the treatment-biomarker interaction model

Treatment	Top-up		Step-down	
Biomarker	Hi	Lo	Hi	Lo
Intercept	1/4	1/4	1/4	1/4
Treatment Main Effect	1/2	1/2	-1/2	-1/2
Biomarker Main Effect	1/2	-1/2	1/2	-1/2
Interaction	1	-1	-1	1
Stratified Treatment Effects				

Treatment in Hi Group = Main Treatment+1/2.Interation	1	0	-1	0
Treatment in Lo Group= Main Treatment - 1/2 .Interaction	0	1	0	-1

If 3 such variables are combined, as per a subgroup analysis, then the equivalent set of contrasts is below, using smoking as an example subgroup. The marginal distribution of the subgroups will be used to weight the subgroups when calculating main effects, and 2nd-order interactions. So if the marginal distribution has weights p & q, the contrasts will be (q , -p) . For the smoking illustration below we take p, q as ¼, ¾ respectively.

Treatment	Top-up				Step-down			
Biomarker	Hi		Lo		Hi		Lo	
Smoking	Current	Other	Current	Other	Current	Other	Current	Other
Intercept	0.0625	0.1875	0.0625	0.1875	0.0625	0.1875	0.0625	0.1875
Main Effects								
Treatment	1/8	3/8	1/8	3/8	-1/8	-3/8	-1/8	-3/8
Biomarker	1/8	3/8	-1/8	-3/8	1/8	3/8	-1/8	-3/8
Smoking	1/4	-1/4	1/4	-1/4	1/4	-1/4	1/4	-1/4
2-way interactions								
Treatment-Biomarker	1/4	3/4	-1/4	-3/4	-1/4	-3/4	1/4	3/4
Treatment-Smoking	1/2	-1/2	1/2	-1/2	-1/2	1/2	-1/2	1/2
Smoking-Biomarker	1/2	-1/2	-1/2	1/2	1/2	-1/2	-1/2	1/2
3-way Interaction	1	-1	-1	1	-1	1	1	-1

8.4 Missing Data

(ICH E3; 9.7.1, 11.4.2.2. ICH E9;5.3. EMA Guideline on Missing Data in Confirmatory Clinical Trials)

Generally, total population size will be reported in summary tables (and will highlight any missing values). Summary statistics will use complete case analysis, which assumes Missing Completely At Random. Regression analysis that adjust for covariates, or account for within-participant correlation in repeat-measures analysis, will be performed which assumes Missing At Random.

See section 5.2.1 for a specific discussion and sensitivity analyses to consider the partially missing component of the HBI score.

If the proportion of missing values falls below 5% for an endpoint then no further sensitivity analyses will be performed as the scope to influence the conclusion is too small.

To deal with any missing baseline covariates Multivariate Imputations using Chain Equations (Buuren & Groothuis-Oudshoorn 2011) is used to provide 5 imputed complete data sets. This is repeated separately for each endpoint/analysis as the endpoint is used as predictor of the missing baseline values; for any longitudinal analysis of repeated observations, the per-patient average of the repeated observations is used, along with an average of the visit numbers with observations. Rubin’s rules are used to combine the analysis from the multiple imputed data sets.

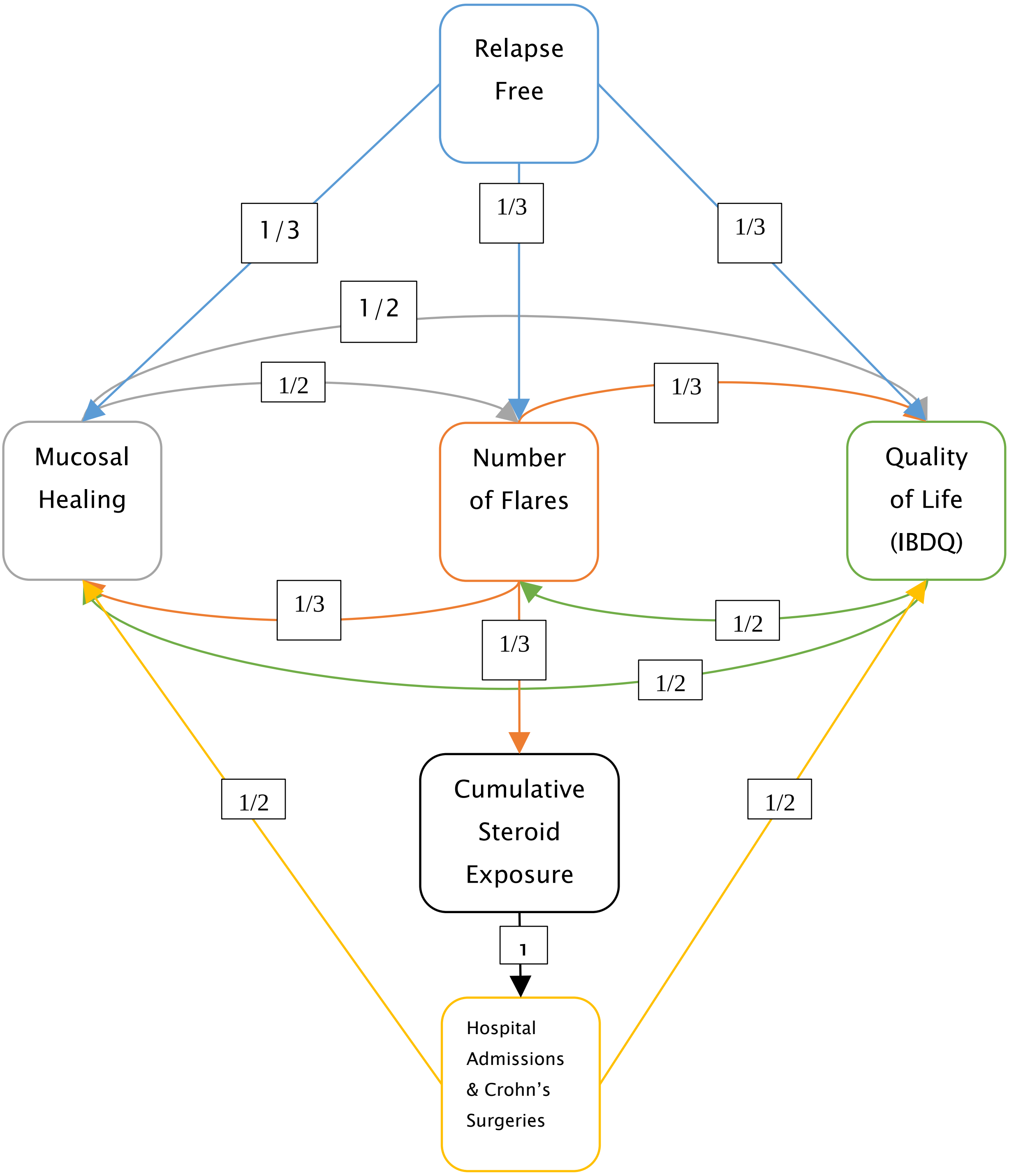
Where bootstrapping is used to provide confidence intervals around standardised effects on the additive scale, then bootstrapping is performed as an outer loop around repeated multiple imputations (Schomaker & Heumann 2018)

8.5 Multiple Testing

(ICH E3; 9.7.1, 11.4.2.5. ICH E9; 2.2.5)

To formally control for multiple testing, we will perform a closed testing procedure over the primary and five secondary endpoints (see section 10.1 and 10.2) testing the biomarker-treatment interaction and restricting the family-wise type 1 error rate to an overall 5% significance level.

The methodology to combine together gate-keeping and Holm-Bonferroni methods in formal hypothesis testing will be used (Bretz et al. Stat Med 2009), with the diagram below defining how the significance levels will be transitioned assuming an initial configuration of 5% at the primary endpoint (relapse-free remission) and 0% on all other tests.



The diagram shows the flow of significance level spending in the sequential multiple testing procedure.

9 Summary of Trial Data

All continuous variables will be summarised using the following descriptive statistics: n (non-missing sample size), mean, standard deviation, median, maximum and minimum. The frequency and percentages (based on the non-missing sample size) of observed levels will be reported for all categorical measures. In general, all data will be listed, sorted by treatment, biomarker and participant, and when appropriate by visit number within participant. All summary tables will be structured with a column for each treatment-biomarker combination, and marginally for each treatment. The tables will be annotated with the total population size relevant to that table/treatment, including any missing observations.

9.1 Participant Disposition

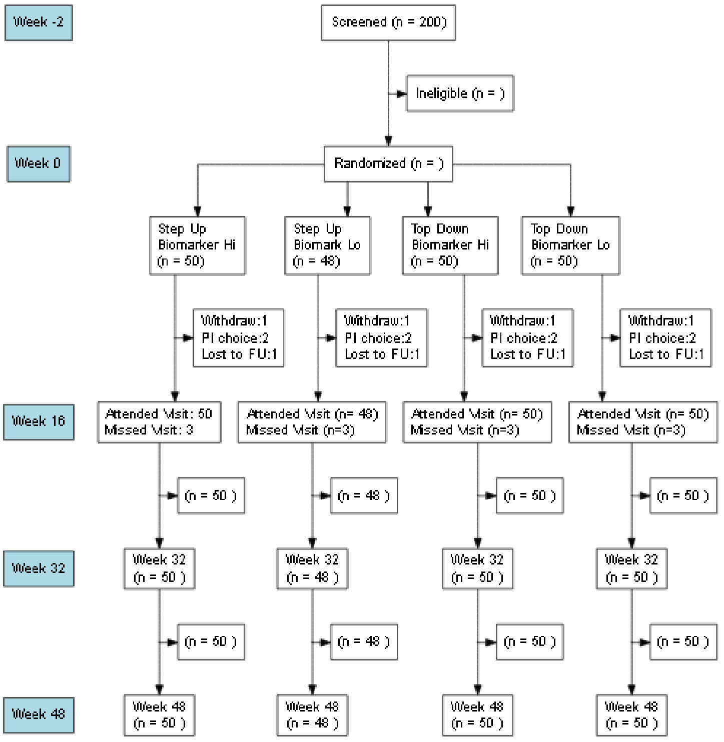
The Visit Completion form and Withdrawal form will be used to determine how far each participant reached within the trial, and provide reasons for withdrawal from data collection, and from treatment.

A figure of cumulative recruitment over time will be provided.

- Full population
    - All participants who were randomised and meet eligibility criteria (all inclusion and no exclusion criteria) for the PROFILE trial.
- Safety population
    - All participants who received any trial treatment
- Per-protocol population
    - All participants who did not substantially deviate from the protocol as to be determined on a per-participant basis. For cases where it is not clear, those participant cases will be referred for review by the central trial team to an expert adjudication committee (Prof Miles Parkes, Dr James Lee, Prof James Lindsay).

Pandemic sensitivity analyses will compare pre-1 March 2020 with 1 March 2020 and beyond. This month was selected as this was the date of the first COVID-19 wave building in the UK, resulting in a national period of lockdown. Thereafter, for the duration of recruitment and follow-up, subsequent waves of incident COVID-19 followed in the UK.

A skeleton CONSORT diagram should be provided in this section that provides an explicit statement of what statistics are to be provided. http://www.consort-statement.org/.


Week -2
Screened (n = 200)
Ineligible (n =)
Week 0
Randomized (n =)
Step Up Biomarker Hi (n = 50)
Step Up Biomark Lo (n = 48)
Top Down Biomarker Hi (n = 50)
Top Down Biomarker Lo (n = 50)
Withdraw:1 PI choice:2 Lost to FU:1
Withdraw:1 PI choice:2 Lost to FU:1
Withdraw:1 PI choice:2 Lost to FU:1
Withdraw:1 PI choice:2 Lost to FU:1
Week 16
Attended Visit: 50 Missed Visit: 3
Attended Visit (n= 48) Missed Visit (n=3)
Attended Visit (n= 50) Missed Visit (n=3)
Attended Visit (n= 50) Missed Visit (n=3)
(n = 50)
(n = 48)
(n = 50)
(n = 50)
Week 32
Week 32 (n = 50)
Week 32 (n = 48)
Week 32 (n = 50)
Week 32 (n = 50)
(n = 50)
(n = 48)
(n = 50)
(n = 50)
Week 48
Week 48 (n = 50)
Week 48 (n = 48)
Week 48 (n = 50)
Week 48 (n = 50)

9.2 Derived variables

- Primary endpoint, see section 6.4 and 5.2.1 for detailed description of how the endpoint is derived.
- Several endpoints are the cumulative incidence or total over all visits
    - Number of Flares
    - Cumulative Steroid Use
    - Hospitalisation
    - Surgery

9.3 Protocol Deviations

Modifications to participants' dosing or medication had a set of guidelines described in the protocol, termed as “escalation”. Deviations did occur in both directions of escalating or not escalating, in breach of the protocol guidelines. But this is a very broad description, and the large variety of medications and doses used make it impractical to give a definition based on trial data to identify non-compliance. A 3-member committee will be assembled (Prof Miles Parkes, Dr James Lee, Prof James Lindsay) to take expert judgement based on the totality of a participant's notes as to whether, and at which visit, they were non-compliant, and thus would leave the visit-specific Per Protocol treatment population and a more stringent Per Protocol treatment, schedules and procedures population.

The COVID pandemic meant that face to face clinic visits had to be paused across many sites, and so the data collection process and visit windows were modified. See section 6.4 for a full description.

9.4 Demographic and Baseline Variables

Each participant undergoes a screening visit at which data is observed, and if eligible and consenting they commence an initial treatment, common to both arms. Hence the data from the screening visit is representative of the participants underlying state of health and will be used to present “Baseline” data.

The following variables will be summarised as per section 0

- Gender
- Age, Age group (16-39,40-64, 65+)
- Ethnicity
- Smoking Status
- BMI, BMI group (0-19, 20-24,25-29, 30+)
- HBI score
- Modified HBI score (removing the abdominal examination component)
- IBDQ score
- CRP
- Albumin
- Haemoglobin
- Faecal Calprotectin
- SES-CD score
- Disease Location
- Disease Behaviour
- Time from diagnosis to consent
- Biomarker

9.5 Concurrent Illnesses and Medical Conditions

A fixed list of specific medical conditions is assessed systematically at baseline visit, which will be tabulated as to the count and proportion of participants with each condition as per section 0.

9.6 Prior and Concurrent Medications

A fixed list of prior steroid medications is assessed systematically at screening, which will be tabulated as to the count and proportion of participants for each medication as per section 0.

Concurrent medications are captured during the trial and will be surmised with line listings.

9.7 Treatment Compliance

Treatment compliance is assessed as a binary variable at each visit, and participants being included or excluded from the per-protocol population is where this information is captured. See section 9.3

Any individual participants and specific visits, where not clear if treatment compliance has occurred, will be discussed and reviewed by the adjudication committee (Prof Miles Parkes, Dr James Lee, Prof James Lindsay). Majority consensus from this committee will be used to determine whether treatment compliance has occurred in accordance with the protocol or not.

10 Efficacy Analyses

The main focus will be on estimating the interaction between treatment and biomarker for each of the endpoints, defined as a difference in differences. To compliment this, the treatment effect within each biomarker subgroup will also be estimated, along with a marginal effect averaging equally across the subgroups. The marginal effect corresponds to the treatment effect when knowledge of the biomarker is not available for individual participants, for a population where the distribution of the biomarker is equally split 1:1 between Hi and Lo.

Each analysis will have 4 comparisons:

- Treatment effect for Hi.
- Treatment effect for Lo.
- Biomarker-Treatment interaction: the difference between the previous two comparisons.
- Main treatment effect: average of the Hi- and Lo- treatment effects.

For endpoints that are analysed on an absolute difference scale these contrasts are simple linear transformations of the model parameters. For endpoints/models that use odds ratios, relative risks or any non-linear function of the parameter values, the model parameters will be used to obtain predicted values on the scale of the endpoint. To calculate the main treatment effect, averages across biomarkers will be calculated, and the contrasts then potentially transformed back onto the non-linear scale of choice, using the standardisation, a.k.a. parametric g-formula, method (Hernán & Robins, Chapter 13). If adjustment for covariates leads to non-convergence on the absolute difference scale for binary variables, then an alternative will be to use a logistic scale, and the same standardisation approach taken to obtain predicted values on the endpoint scale, averages taken for combinations of treatment, biomarker and subgroup, and contrasts taken on the absolute difference scale as desired. Inference to provide SE, 95% CIs, and p-values will be done analytically for the linear scale, or using non-parametric bootstrapping for the non-linear scale.

The primary endpoint, and mucosal healing (using SES-CD score) are both binary variables, and will be analysed using a generalised linear model with an additive link function to estimate absolute differences in incidence. Number of Flares, number of Steroid Courses, and count of Hospitalisation or Surgery will be analysed using a linear model. Quality of life is observed over repeated visits and will be analysed using a mixed effect repeat measure analysis with a clustered participant-level residual error with unstructured covariance over visits, fixed effects for visit, and all other covariates assumed to have a constant fixed effect over time.

The randomisation stratum will be adjusted for as baseline covariates: biomarker subgroup, (IBD^{hi}/IBD^{lo}), mucosal inflammation (mild / moderate / severe) and disease location (colon-only/other) plus the main effect of the covariates listed in section 8.3. Treatment and the interaction between biomarker and treatment will also be included in the models. Quality of life and Mucosal Healing (endoscopic remission) will also adjust for the values observed at screening.

The Full Analysis population will be used for all presentations of efficacy variables, complimented by using the Per Protocol population in some select cases. The p-values from the full analysis population will be used in the process described in section 8.5 to control for multiple testing, using just the treatment-biomarker parameter.

The Safety Population will be used to summarise the safety endpoints.

Pandemic sensitivity analyses will compare pre-1 March 2020 with 1 March 2020 and beyond. This month was selected as this was the date of the first COVID-19 wave building in the UK, resulting in a national period of lockdown. Thereafter, for the duration of recruitment and follow-up, subsequent waves of incident COVID-19 followed in the UK.

In formal statistical terms the null hypothesis is that the interaction between treatment and biomarker is zero. This is asking if the treatment effect in each of the biomarker subgroups is the same. As such it does depend directly on the choice of scale used to measure the (non-zero) treatment effect, and the modelling assumptions. Statistical significance will be assessed in a two-sided manner using a 5% level. However, the focus is on estimation of the magnitude of the interaction and providing a measure of uncertainty around the estimate using 95% confidence intervals. All assumptions for regression models will be assessed by viewing plots of the residual values

10.1 Primary Efficacy Analysis

The primary endpoint is defined as:

Incidence of sustained surgery and steroid free remission from completion of a protocolised (maximum 8-week regimen) steroid induction treatment through to week 48.

*remission = HBI≤4 and/or absence of objective evidence of inflammation.

This is equivalent to HBI≤4 at all assessments, or if symptoms are present (HBI≥5) there is no objective evidence of inflammation (CRP<ULN and calprotectin<200). Requirement for an additional or extended induction course of systemic steroids or surgery for active Crohn's disease would result in failure to meet primary outcome measure.

So if the HBI is 5 or greater at any scheduled or unscheduled visit due to active Crohn’s disease as assessed by local PI and corroborated by either a raised CRP>ULN or raised calprotectin >200 or both, or surgery occurs, or steroids are prescribed (above the protocolised maximum 8 week induction treatment), then the primary endpoint will not have been met.

The primary analysis will proceed as per section 0 using a generalised linear model, and additionally adjusting for the HBI total score observed at screening, with an additive link function, or logistic link if convergence problems arise.

A sensitivity analysis will be performed using the intersection across all visits of the visit-specific per-protocol populations.

An exploratory figure will present the proportion of participants meeting the primary endpoint at or before each time point using a Kaplan-Meier estimate with 95% confidence intervals.

10.2 Secondary Efficacy Analyses

10.2.1 Mucosal Healing

This is the SES-CD score observed at week 48 and will focus on endoscopic remission (defined as absence of ulceration including aphthous ulcerations (ie ulcer subscore=0).

The analysis will use a generalised linear model with an additive link function, as per section 0, and in addition to the standard adjustment covariates, will add the baseline SES-CD score.

The Week 48 Per Protocol population will be used for a sensitivity analysis.

10.2.2 Number of Flares

This endpoint is defined cumulatively across all visits, including ad hoc visits. The analysis will use linear regression as per section 0.

A sensitivity analysis will be performed using the intersection across all visits of the visit-specific per-protocol populations.

10.2.3 Quality of Life IBDQ (Inflammatory Bowel Disease Questionnaire)

There are 32 questions, each of which is recorded on a seven-point Likert scale, where 1 is the worst score and 7 the best (Guyatt et al. Gastroenterology 1989). The total score is derived by adding these numerical values up over the 32 questions for each participant-visit.

Any missing values to individual questions will be ignored and the score scaled to reflect the number of “non-missing” values.

The endpoint is observed over repeated visits and the values at the scheduled visits (but not ad hoc) will be analysed using a mixed effect repeat measure analysis with a clustered participant-level residual error with unstructured covariance over visits, fixed effects for visit, and all other covariates assumed to have a constant fixed effect over time. The standard set of covariates will be used for adjustment with the addition of the baseline value for IBDQ. A treatment effect estimate assumed to be constant over time will be used for the formal closed multiple testing procedure. But individual visit-specific treatment effects at all scheduled visits will also be presented.

Observations taken at participant-visits belonging to the visit-specific per-protocol population will be used for a sensitivity analysis.

10.2.4 Cumulative Steroid Exposure

This endpoint is defined as the total number of courses of steroids taken, including those that were increased in line with the protocol, and any non-compliant increases to treat active Crohn’s disease. The analysis will use linear regression as per section 0.

A sensitivity analysis will be performed using the intersection across all visits of the visit-specific per-protocol populations.

10.2.5 Number of Hospital Admissions and Surgeries for Crohn’s disease

The incidence of surgery for Crohn’s disease is captured directly within the CRF. An SAE that requires hospitalisation will capture hospital admissions. Hence the count is taken across all visits, including ad hoc visits, and aggregated into a single endpoint per participant.

The analysis will proceed as per section 0 using a generalised linear model with an additive link function.

A sensitivity analysis will be performed using the intersection across all visits of the visit-specific per-protocol populations.

10.2.6 HBI

HBI score will be presented using spaghetti plots, and mean values with CIs at each visit and biomarker-treatment presented. A mixed effect repeat measure analysis will be performed with a clustered participant-level residual error with unstructured covariance over visits, fixed effects for visit, and all other covariates assumed to have a constant fixed effect over time. The standard set of covariates will be used for adjustment with the addition of the baseline value.

The analysis will performed twice on two variations of the endpoint:

- a scaled version of the score that excludes the component based on a physical examination
- using the participant reported physical examination.

10.2.7 Incidence of Surgery

The individual component of the primary endpoint will be considered in isolation.

The analysis will proceed as per section 0 using a generalised linear model with an additive link function.

10.2.8 Incidence of Steroid Use

The individual component of the primary endpoint will be considered in isolation. This is not the cumulative count of steroids, but rather simply if steroids were used at any point in time above the maximum 8 week initial induction treatment.

The analysis will proceed as per section 0 using a generalised linear model with an additive link function.

10.3 COVID sensitivity analyses

For each of the primary and secondary endpoints, the analyses will be repeated for the pre- and peri-covid populations.

10.3.1 Seemingly Unrelated Regression

The components of the primary and secondary endpoints, in particular those that relied on colonoscopy visits, were not able to be performed during the lockdown restrictions, and thus have higher incidence of missing values. Hence a joint model across several key endpoints, for the week 48 value, may increase precision to the comparisons by accounting for the within-participant correlation and thus make better use of participants with partial observations for a subset of the endpoints. A joint linear model that allows an unstructured residual error covariance matrix will be fitted to the following endpoints on their original, untransformed, scales.

- HBI score
- Centrally-read SES-CD
- Locally-read SES-CD

- Number of Flares
- IBDQ
- Cumulative Steroid exposure
- Number of Hospital admissions and surgeries
- Faecal Calprotectin – log transformed

Fixed effects will be stratified by the endpoint, and adjust for treatment-biomarker interaction plus the baseline covariates detailed in 8.3. The baseline or screening value of the endpoint will be used as an adjustor where a meaningful baseline value exists, or set to a dummy value of 0. Values observed at visits before week 48 will not be considered.

The same set of comparisons as per the primary analysis will be presented with estimates, SE, confidence intervals and p-values.

The R code will use the gls function taken from the nlme package and will be of this nature, as an illustration:

```
gls(value~variable/(treatment*biomarker + covariate +
baseline_value),correlation = corSymm(form=~var_num|subjid),
weights=varIdent(form=~1|variable))
```

where var_num is the endpoint, variable, mapped to a sequence of consecutive integers, as demanded by the software.

Censored observations for Faecal Calprotectin, where the observation is above the limit of detection will be handled using the EM algorithm [Dempster, Laird & Rubin, 1977], and a simulation exercise will be performed to establish that the implementation with bespoke code provides an unbiased estimate with correct coverage of confidence intervals.

10.4 Subgroup Analyses

For each of the primary and secondary analysis, the subgroup analyses will be performed using the full analysis population.

Summary tables broken down by treatment-biomarker-subgroup will be produced

Sequentially, and considering in turn each subgroup from section 8.3, the main effects, 2nd- and 3rd-order interactions will be presented, with the estimates and 95% confidence intervals, but not the p-values. Interpretation of the estimates' magnitude is more important that consideration of statistical significance, including whether the subgroup could modify the optimal choice of treatment. Should a subgroup interaction prompt reporting in academic journals, then a stratified version of the model may be reported, as distinct from the difference-of-difference-of-difference that a 3-way interaction represents, and would be calculated outside the main statistical report.

10.5 Tertiary and Exploratory Efficacy Analyses

Summary statistics as described in section 0 will be provided for the tertiary and exploratory endpoints as detailed in sections 5.2.3 and 5.2.4, stratified by visit, treatment and biomarker. For continuous variables, the change from baseline (or screening value where appropriate) will be plotted over time, with lines connecting the observations within an individual participant (a spaghetti plot), with colour coding used to distinguish between the treatment-biomarkers. Box and Whisker plots will also be provided to represent their distribution in a cross-sectional manner. Kaplan-Meier plots will be presented for time-to-event endpoints.

11 Safety Analyses

11.1 Adverse Events

Safety Analyses will use the Safety Population and be broken down by biomarker-treatment, and treatment (pooling across biomarker).

The data have MedDRA coding provided for each event reported, along with an assessment of causality, seriousness, and expectedness. For each preferred term, there will be a count of the number of participants (any repetitions of an AE within the same participant will be ignored), proportion of the population, and number of events (to include any repetitions), broken down as described above, and also split into serious and non-serious AEs. The preferred terms will be grouped by the System Organ Class.

For more meaningful reporting, a grouping of the preferred terms will be defined (see appendix) and the same format of reporting provided.

A figure will be provided, a “Dot Plot” that, for each grouping of preferred terms, gives the absolute risk for each biomarker-treatment on a line representing the unit interval and also give the relative risk estimate and 95% CI on a logarithmic relative risk scale.

A line listing of each AE, split into serious and non-serious AEs, will be provided sorted by treatment-biomarker, SOC, PT, causality and expectedness, in that order of sorting.

11.2 Pregnancies

A line list of pregnancies that occurred for participants will be produced.

12 Reporting Conventions

P-values ≥0.001 will be reported to 3 decimal places; p-values less than 0.001 will be reported as “<0.001”. The mean, standard deviation, and any other statistics other than quantiles, will be reported to one decimal place greater than the original data. Quantiles, such as median, or minimum and maximum will use the same number of decimal places as the original data. Estimated parameters, not on the same scale as raw observations (e.g. regression coefficients) will be reported to 3 significant figures. Percentages will be rounded to whole numbers unless they are less than 1%.

13 Technical Details

R version 4.1 will be the software tool used, correct at the time of writing. Full documentation of all versions and add-on packages will be recorded as the report is generated.

Any outputs will have

- The date and time included

- The name of the code file that produced the analysis
- The author
- A log capturing the version of the software and any external add-on code used.

14 Summary of Changes

14.1 From Protocol

At the TSC held on 10th September 2018, at which 30 participants had been randomised, the biomarker thresholds had been further optimised, and a decision was taken to accept the modified thresholds used to define IBDhi and IBDlo using this biomarker blood-test. A few of the existing participants' biomarker value were changed but given this was very early on in the trial, subsequent stratified randomisation was able to be maintained and remained blinded to the trial management group, the TSC, participants and site principal investigators.

14.2 From SAP V1

- Modification of adjusting covariates to include baseline HBI.
- Added details on using multiple imputation in conjunction with bootstrapping, to deal with missing values in baseline covariates.
- Additional endpoints relating to endoscopy (central & local reading) and only using incidence of ulcers, rather than composite with the total SES-CD.
- Additional Kaplan-Meier curve to describe time to primary endpoint.
- Seemingly Unrelated regression uses faecal calprotectin on a log-transformed scale and uses EM algorithm to deal with censored observations above limit of detection.
- Clarification of primary outcome measure to include objective markers of active inflammation (CRP and calprotectin) in the definition of presence and absence of remission (reflecting their inclusion in definition of flare in the protocol). This was agreed by the TSC 26/04/2023:
- Incidence of sustained surgery and steroid free remission from completion of a protocolised (maximum 8-week regimen) steroid induction treatment through to week 48.
- Clarification of secondary endpoint re method to define endoscopic remission
  - Endoscopic remission at week 48 (defined by absence of ulceration i.e. SES-CD ulcer subscore = 0). Centrally-read endoscopic scores will be used where available. Where these are not available locally-read scores will be used
- Modification of secondary endpoint relating to timepoints for assessment of quality of life
  - Quality of life averaged across weeks 16, 32 and 48 (using disease specific IBD-Q score).
- Addition of the following tertiary end-points for analysis
  - Incidence of sustained surgery and steroid free remission from completion of a standard (maximum 8-week regimen) steroid induction treatment through to week 48 (when remission defined using clinical parameters alone, HBI < 5).
  - Clinical remission (defined as HBI < 5) at weeks 4,16, 32 and 48.
  - Average clinical disease activity (comparison of mean HBI scores in each group) at weeks 4,16,32 and 48.

- Time to event, time from baseline to first flare or need for surgery for Crohn's disease, which may occur during the protocolised induction course of steroid medication.
- Time to event, time from baseline to second flare or need for surgery for Crohn's disease.
- Time to event, time from baseline to starting on anti-TNF therapy for Crohn's disease
- Patient reported clinical remission (using score generated from abdominal pain and stool frequency components of HBI score – abdominal pain $\leq$1 and stool frequency $\leq$3) at weeks 4,16,32,48.
- Development of peri-anal abscess or fistula (development of peri-anal abscess / fistula vs no development of peri-anal abscess / fistula) at weeks 4,16,32,48.
- Anti-TNF therapy at last observation: no anti-TNF therapy; monotherapy anti-TNF; combination therapy anti-TNF.
- Thiopurine at last observation within each participant: no thiopurine; optimised metabolite levels (6-TGN $\geq$235); non-optimised levels (6-TGN <235).

15 Quality Control

The derivation of the primary endpoint is complex as it involves multiple visits and multiple questions from the CRF. Hence it will be derived by a separate QC statistician and compared, and also a random set of 10 participants will be checked by hand.

The primary analysis will be replicated by a separate QC statistician. Overall, the code and data manipulations will be reviewed by a separate statistician.

16 References

Bretz F, et al. (2009). A graphical approach to sequentially rejective multiple test procedures. *Stat Med* 28:586-604.

Colombel JF, et al. (2010). Infliximab, azathioprine or combination therapy for Crohn's disease. *NEJM* 362:1383-1395.

D'Haens G, et al. (2008). Early combined immunosuppression or conventional management in participants with newly diagnosed Crohn's disease: an open randomised trial. *Lancet* 371:660-667.

D'Haens G, et al. (2010). Top-down therapy for IBD: rationale and requisite evidence. *Nat Rev Gastroent Hepatol* 7:86-92.

Devlin, N., Shah, K., Feng, Y., Mulhern, B. and van Hout, B. (2016). Valuing Health-Related Quality of Life: An EQ-5D-5L Value Set for England. *Health Economics* 27, 7-22.

Echarri A, et al. (2020). The Harvey-Bradshaw Index adapted to a mobile application compared with in-clinic assessment: The MediCrohn Study. *Telemed J E Health* 26, 80-88.

Guyatt G, et al. (1989). A new measure of health status for clinical trials in inflammatory bowel disease. *Gastroenterology* 96, 804-810.

Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.

White IR and Thompson SG (2005). Adjusting for partially missing baseline measurements in randomised trials. *Stat Med* 24**:**993-1007.

Dempster, A. P., N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm." Journal of the Royal Statistical Society. Series B (Methodological) 39, no. 1 (1977): 1–38. http://www.jstor.org/stable/2984875.

Schomaker, M, Heumann, C. Bootstrap inference when using multiple imputation. *Statistics in Medicine*. 2018; 37: 2252– 2266. https://doi.org/10.1002/sim.7654

Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. DOI 10.18637/jss.v045.i03.

17 Listing of Tables, Listings and Figures

An appendix document gives precise details for each table, listing or figure to be produced. As a minimum it will be a tabulation of the following aspects unique to each table or listing.

- Title
- Footnotes
- Numbering
- Population
- Endpoint(s)
- Time Points or details of how to conglomerate multiple observations
- Covariates or subgroups used to break down summary statistics
- Which summary statistics will be calculated
- Or, what formal analysis will be used

Further details to aid the writing of code or improve the report, may be included as well, for example a list of variable names from the database; section numbers or titles, subtitles; comments;