



PREdiCt: The PRognostic effect of Environmental factors
in Crohn's and Colitis
Updated Statistical Analysis Plan

Version No	1.9.99
Date	16/10/2025
Author	Dr Nathan Constantine-Cooke
CI Name	Professor Charlie Lees
CI Email address	charlie.lees@ed.ac.uk

Signatures

Statistician:	Date:
Chief Investigator:	Date:

Contents

1	Introduction	3
2	Core dietary set	3
3	Additional analyses	4
3.1	Imputation of smoking status	4
3.2	Adjustment for multiple testing	4
3.3	Sensitivity analysis excluding participants with high baseline faecal calprotectin . . .	4
3.4	Report cumulative incidence for primary variables	5
3.5	Modelling continuous variables	5
3.6	Control for additional covariates	5

1 Introduction

The statistical analysis plan (SAP) for the PREdiCCt study was first devised over seven years ago, with conception of the study occurring over three years prior. In that time, the world of inflammatory bowel disease (IBD) research has changed substantially and, as a result, some of the analyses outlined in the original SAP have become outdated, with modification required to reach modern expectations.

This document acts as a time stamped amendment to the SAP. Whilst the outcomes (time-to-flare) are unchanged, this amendment clarifies the core dietary dataset and outlines additional sensitivity analyses.

2 Core dietary set

Some of the pre-specified diet variables of interest in the original SAP were not available for analysis. It was not possible to separate N-6 polyunsaturated fatty acids (PUFAs) from overall PUFA intake using food frequency questionnaire (FFQ) data. Therefore, instead of N-6 PUFA intake, we will consider overall PUFA intake recorded using FFQs. The SAP also specified dietary emulsifiers (lecithin) as a primary exposure, which was also unavailable from FFQs and instead we use calorie intake attributable to ultra processed (NOVA score 4) food, which should be closely related to emulsifiers.

The remaining pre-specified variables can successfully be calculated from FFQs. For clarity, the following variables are considered to be the core dietary dataset as part of the updated statistical analysis:

- Total animal protein intake
- Dietary fibre
- Total polyunsaturated fatty acid (PUFA) intake
- Calorie intake attributable to ultra processed (NOVA Score 4) food

Calorie intake attributable to ultra processed (NOVA Score 4) food has been obtained using the University of Aberdeen's in-house software package. All other dietary variables are considered to be exploratory in nature.

3 Additional analyses

The following analyses are planned, with the majority being sensitivity analyses.

3.1 Imputation of smoking status

Due to the widely-reported association between poor outcomes and smoking status in Crohn’s disease (CD)^{1,2} and the possible protective effect in UC,³ the SAP previously mandated that we control for smoking status in time-to-flare models. However, of the 2629 people living with IBD recruited to PREdiCCt, 715 (27.2%) did not report smoking status. In the primary analysis, smoking status was not controlled for in order to not lose 27.2% of the cohort without making assumptions regarding the true status of smoking status for participants. However, as a sensitivity analysis, we propose using multiple imputation with chained equations (MICE) to impute smoking status and then control for smoking status across models. The imputation model used in the MICE algorithm will include the model covariates, the event status, and the cumulative hazard calculated at the event or right-censoring time, estimated using a Nelson-Aalan estimator⁴.

3.2 Adjustment for multiple testing

Due to the large amount of data collected by PREdiCCt, a high number of statistical tests have been performed, increasing the probability of Type I (false positive) errors occurring. As such, for variables which were not pre-specified (see Section 2), Bonferroni multiple testing correction will be applied to provide adjusted p-values in a supplementary table. Adjusted p-values this will take the form of

$$p_{adj} = \min \{np, 1\}, \tag{1}$$

where n indicates the total number of tests and p is the unadjusted p-value.

3.3 Sensitivity analysis excluding participants with high baseline faecal calprotectin

For inclusion into the PREdiCCt study, participants were required to have felt that their disease was under control, as well as having not undergone a recent change in therapy, or be receiving steroids. However, we did not require participants to have faecal calprotectin (FC) levels, a proxy for endoscopic disease activity, below a set threshold. For some participants with elevated FC, it may be possible they were in a ‘pre-flare state’. As such, we will conduct a sensitivity analysis where models are rerun excluding subjects with FC above a set threshold. This threshold will initially be

250 μ g/g, but alternative thresholds (e.g. 125 μ g/g) may also be considered. Results are expected to be presented as a supplementary table.

3.4 Report cumulative incidence for primary variables

Whilst reporting hazard ratios is standard practice for Cox models, there remains an inherent issue regarding interpretation.⁵ Therefore, in addition to hazard ratios, we will report 1- and 2-year cumulative incidence of flares for primary variables, which are more clinically relevant statistics.

3.5 Modelling continuous variables

Previously, continuous variables such as FC and dietary data were discretised. FC was categorised into $FC < 50$, $50 \leq FC \leq 250$, and $FC > 250$, and dietary data was categorised into quantiles. Further analyses will be conducted without discretisation, with continuous variables modelled using splines to account for any non-linear relationships. FC will be log-transformed due to positive skewness. If there are extreme outliers for dietary data, these may be excluded.

3.6 Control for additional covariates

Previously, only sex, social deprivation, and FC have been controlled for across all time-to-flare analyses, with body mass index and diet quality index also controlled for in dietary analyses. Additional common risk factors will be included to ascertain if findings are also independent of these factors. We will control for age, previous surgery, disease duration, and body mass index across all analyses and aim to also control for prescribing data if this is feasible.

References

- [1] Tiago Nunes, Maria Josefina Etchevers, Olga Merino, Sonia Gallego, Valle García-Sánchez, Ignacio Marín-Jiménez, Luis Menchén, Manuel Barreiro-de Acosta, Guillermo Bastida, Sara García, Elena Gento, Daniel Ginard, Fernando Gomollón, Maite Arroyo, David Monfort, Esther García-Planella, Benito Gonzalez, Carme Loras, Carles Agustí, Carolina Figueroa, and Miquel Sans. Does smoking influence Crohn’s disease in the biologic era? The TABACROHN Study. *Inflammatory Bowel Diseases*, 19(1):23–29, January 2013. doi: 10.1002/ibd.22959. Publisher: Oxford University Press (OUP).
- [2] H. S. Odes, A. Fich, S. Reif, A. Halak, A. Lavy, D. Keter, R. Eliakim, J. Paz, E. Broide, Y. Niv, Y. Ron, Y. Villa, N. Arber, and T. Gilat. Effects of current cigarette smoking on clinical course

- of Crohn's disease and ulcerative colitis. *Digestive Diseases and Sciences*, 46(8):1717–1721, 2001. doi: 10.1023/a:1010609722315. Publisher: Springer Nature.
- [3] P. C. Lunney and R. W. L. Leong. Review article: Ulcerative colitis, smoking and nicotine therapy. *Alimentary Pharmacology & Therapeutics*, 36(11-12):997–1008, 2012. ISSN 1365-2036. doi: 10.1111/apt.12086. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/apt.12086>.
- [4] Mai Zhou. *Empirical Likelihood Method in Survival Analysis*. Chapman and Hall/CRC, 0 edition, June 2015. ISBN 978-0-429-09694-5. doi: 10.1201/b18598. URL <https://www.taylorfrancis.com/books/9781466554931>.
- [5] Elise Dumas and Mats J. Stensrud. How hazard ratios can mislead and why it matters in practice. *European Journal of Epidemiology*, 40(6):603–609, June 2025. ISSN 0393-2990, 1573-7284. doi: 10.1007/s10654-025-01250-9. URL <https://link.springer.com/10.1007/s10654-025-01250-9>.