Methods for Staging disease activity in ulcerative colitis; A comparison of endoscopy and histology (MESIAH)

Statistical Analysis Plan
Bilal Mateen, Mehul Patel and Bu’Hussain Hayee
Department of Gastroenterology, King’s College Hospital, London, UK

For all hypotheses, data will be summarized as counts and proportions for categorical data, and the significance of any differences from the expected distribution using a test. For continuous data, summaries will be presented as medians and ranges for non-normally distributed data, and means with standard deviations for data that conforms to the assumptions of the normal distribution. Significance testing will then be carried out using a parametric or non-parametric test as appropriate. In circumstances where multiple statistical hypothesis tests are carried out simultaneously, a Benjaimi-Hochberg method will be applied to control the false discovery rate. The analysis plan for each separate hypothesis are detail below:

A) Determining agreement between endoscopic and a novel histological definition of remission

Agreement statistics (incl. Fleiss’ and Cohen’s Kappa) will be calculated based on different definitions of remission using the Mayo endoscopic score and the presence or absence of neutrophils (Personal Communication; B. Hayee, 2020 - forthcoming peer reviewed consensus statement). The sample of interest is the histology sub-study in the PURSUIT trial, with n = 94. Agreement will be tested at the 0 week time point, and the re-randomisation (post-induction) time-point, based on the availability of data.

B) Identification of a diagnostic (upper) threshold for relapse/flare based on faecal calprotectin results

The primary aim is the identification of the upper threshold cut-off at which faecal calprotectin is diagnostic of active disease on endoscopy (the reference standard). Endoscopic remission/mucosal healing will be defined as either a mayo score 0 or 1. A secondary analysis will also be carried out with just the mayo score 0 population. The basis of the analysis will be decision curve modelling of faecal calprotectin, against the index (reference) test results from endoscopy. Modelling will be undertaken in the VARSITY study (Vedolizumab versus Adalimumab), based on data available at the final time-point (i.e. 46 weeks). Validation will then be undertaken using the GEMINI and PURSUIT study final follow-up time points (i.e. 52 weeks). Analysis will conform to the recommendations of the Standards for Reporting Diagnostic accuracy studies (STARD) guidelines for diagnostic accuracy studies.

C) Evaluating the increase in diagnostic accuracy from knowledge of an index calprotectin measurement

The primary aim is to assess the potential role, if any, of knowing the index calprotectin measurement when interpreting a result. The methodological approach will be
as follows; using the GEMINI study for derivation the diagnostic accuracy of the faecal calprotectin results (i.e. sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under curve (AUC); visual representation with receiver operating characteristic curve (ROC)), will be determined for the 6 weeks (using 0 weeks as the index), and 52 week (using 0 weeks, and 6 weeks as the index) time points. The reference standard will be active disease on endoscopy. Endoscopic remission/mucosal healing will be defined as either a mayo score 0 or 1. A secondary analysis will also be carried out with just the mayo score 0 population. At each time point of the above time points (i.e. 6 weeks and 52 weeks), 3 values will be compared: 1) the diagnostic accuracy of the calprotectin; 2) the absolute change in calprotectin from the index value; 3) the proportion change in calprotectin from the index ( = Δ[index-latest]/index). The 9 diagnostic thresholds will then be applied to the PURSUIT study as a form of external validation. Sub-group results for each treatment pathway will be presented to demonstrate that these do not affect the results (our hypothesis given that the double blind randomised controlled trial (RCT) nature should mean that the observation and selection bias is attenuated to non-significance).

D) Identification of covariates that are associated with variation in the faecal calprotectin results for individuals with endoscopically confirmed remission.

Exploratory analysis will be based on visualizing, and quantifying the proportion of variation explained by each covariate in the univariate setting. Quantification will be carried out using the appropriate correlation coefficient, and hypothesis tested for significance with a threshold of p < 0.05. The above exploratory analysis will be complemented by a series of predictive modelling experiments to facilitate triangulation of the contribution of each of the measured features. The benchmark experiment will entail a k-fold cross validation set-up using all the covariates to predict the faecal calprotectin result for each individual. All models will be compared to the statistical (regression) baseline (i.e. predicting the sample median for all cases) to demonstrate that they are superior to an informed guess. The measures of predictive power utilized will include: root mean squared error, and mean absolute error. Confidence intervals for the measures will be generated using an algorithm based on the jackknife estimator of variance, or derived analytically. A series of different models will be utilized, including regularized linear regression, and tree-based methods. Specifically, a gradient boosted tree-based method will be utilized to allow for the generation of Shapley-additive values, to allow for in-depth exploration of the relative contribution of each feature to the outcome, both at the global and individual levels. Comparison of model performance will be undertaken using a Wilcoxon signed rank test to compare the paired residual errors. All analysis will be carried out using R, and the package MLR.

E) Developing a prediction model for endoscopic remission based on features collected routinely in clinical practice

N.B: This aim will be split into two different questions:

The first question will be with regards to predicting whether an individual has active disease or not. The dataset used for this question will be the information collected at the final
follow-up, and active disease will either be defined by the endoscopic scoring that is paired with final clinical assessment for each individual.

The second question will pertain to predicting likelihood of achieving remission/active disease when treated with a biologic drug. The data utilised here will be the data at the index visit, and the outcome defined at the final follow-up. Similar definitions of active disease as per question 1 will be utilised again.

The general methods are as follows. A series of predictive modelling experiments will be undertaken, however this time the framework will be classification instead of regression. The benchmark experiment will entail a k-fold cross validation set-up using all the covariates to predict disease state. Note that we will limit the number of covariates based on domain expert knowledge, via a literature search to identify previously informative/prognostic features, and based on an exploration of a randomly selected 10% (hold-out) development pool [with appropriate safeguards for information leak from the prior work]. Exploration for non-linear effects will also be undertaken in the development pool, and fractional polynomials will be used to transform the data in the presence of any identified useful non-linear relationships with the outcome. Missing data will be handled using multiple imputation by chained equations, as per best practice guidance; although we a priori acknowledge that this is a hotly contested issue, and that the issue of portability of the missingness mechanism is harder to justify in going from the highly controlled setting of an RCT to the real world – which may limit generalisation. We may consider automated feature selection methods, but we would prefer to avoid this given the greedy nature of these algorithms, and the difficulty with interpreting the outputs from a cross-validation set-up. All models will be compared to the statistical (probabilistic and deterministic) baseline (i.e. predicting the mean probability or the majority class for all cases) to demonstrate that they are superior to an informed guess. The measures of predictive power utilized will include: mean misclassification error, sensitivity, specificity, PPV, NPV, and F1 score (harmonic mean of sensitivity and PPV). Confidence intervals for the measures will be generated using an algorithm based on the jackknife estimator of variance, or derived analytically. A series of different models will be utilized, including regularized logistic regression, logistic regression, support vector machine, and tree-based methods. A nested out-of-sample tuning of the models will also be undertaken. Comparison of model performance will be undertaken using a Wilcoxon signed rank test to compare the paired residual errors. All analysis will be carried out using R, and the package MLR. Bespoke code written by the applicants for determining the confidence intervals around the errors and the significant testing can be manually entered or a dedicated file provided to the Vivli admins for transfer into the secure computational environment.

F) Rasch analysis [psychometric validation] of the Mayo Tool

To ascertain whether there is an underlying interval scale in the Mayo tool, rasch analysis (using the conditional maximum likelihood to fit the item parameters, and a maximum likelihood estimator for the person locations) will be utilised to reframe the ordinal scores into their interval equivalents. No anchoring method will be utilised, because although the index time-point is not reflective of the potential range of scores (i.e. only individudals with active disease are recruited), the final time-point should be, given the
relapsing-remitting nature of IBD. Rasch-Andrish thresholds will be determined using the partial credit model, and any post-hoc changes that are necessary to improve fit will be undertaken and described. The final rasch-modified variant of the mayo tool will be presented, alongside a description of the interval scale to illustrate whether it is reasonable to continue using the tool as has been previously proposed (i.e. a 4 item, 3 point ordinal scale).

**References**

The following references provide a brief overview of previous published applications of the above methods carried out by the quantitative lead for this proposed series of work, as well as links to international working group guidance document published by the BMJ on best practice in prediction modelling for health on which the applicants are named lead authors.


Denaxas S, Shah A, Mateen BA, et al.. A semi-supervised approach for rapidly creating clinical biomarker phenotypes in the UK Biobank using different primary care EHR and clinical terminology systems. JAMIA. 2020 [Accepted, pending proofs]


**Other relevant material**

