Plasma protein-based and polygenic risk scores serve complementary roles in predicting inflammatory bowel disease

Jakob Woerner[1†], Thomas Westbrook[1†], Seokho Jeong[2], Manu Shivakumar[1], Allison R. Greenplate[3], Sokratis A. Apostolidis[4], Seunggeun Lee[5], Yonghyun Nam[2], Dokyoon Kim[2]

[1]*Genomics and Computational Biology Graduate Group, University of Pennsylvania, Philadelphia, PA, USA*
[2]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA*
[3]*Institute for Immunology and Immune Health, University of Pennsylvania, Philadelphia, PA, USA*
[4]*Division of Rheumatology, Department of Medicine, University of Pennsylvania, Philadelphia, PA, USA*
[5]*Graduate School of Data Science, Seoul National University, Seoul, South Korea*
[†]*Equal Contribution*
Email: Yonghyun.Nam@pennmedicine.upenn.edu, Dokyoon.Kim@pennmedicine.upenn.edu

Inflammatory bowel disease (IBD), encompassing Crohn's disease (CD) and ulcerative colitis (UC), has a significant genetic component and is increasingly prevalent due to environmental factors. Current polygenic risk scores (PRS) have limited predictive power and cannot inform time of symptom onset. Circulating proteomics profiling offers a novel, non-invasive approach for understanding the inflammatory state of complex diseases, enabling the creation of proteomic risk scores (ProRS). This study utilizes data from 51,772 individuals in the UK Biobank to evaluate the unique and combined contributions of PRS and ProRS to IBD risk prediction. We developed ProRS models for CD and UC, assessed their predictive performance over time, and examined the benefits of integrating PRS and ProRS for enhanced risk stratification. Our findings are the first to demonstrate that combining genetic and proteomic data improves IBD incidence prediction, with ProRS providing time-sensitive predictions and PRS offering additional long-term predictive value. We also show that the ProRS achieves better predictive performance among individuals with high PRS. This integrated approach highlights the potential for multi-omic data in precision medicine for IBD.

Keywords: plasma proteomics; polygenic risk score; autoimmunity; multi-omics; inflammatory bowel disease.

1. Introduction

Inflammatory bowel disease (IBD) represents a chronic inflammatory condition of the gastrointestinal tract. Its subtypes, Crohn's disease (CD) and ulcerative colitis (UC) are related but unique conditions with differing properties, symptoms, and risk factors.[1] IBD affects approximately 2.4 to 3.1 million people in the United States, with most diagnoses occurring in adulthood.[2–4] Epidemiologic and genetic studies have demonstrated that these inflammatory conditions are driven by a complex interplay between genetic susceptibility and environmental factors. Genome-wide association studies (GWASs) have identified over 200 significant genetic loci,[5] and family history of the disease is the strongest risk factor.[6] Multiple lifestyle factors,[7,8] including smoking and psychological stress, as well as environmental factors[9] such as urbanization, industrialization, and westernization are also associated with the onset and progression of IBD.

Patients with IBD often develop severe complications, including strictures or fistulas in the intestine, and in extreme cases, colorectal cancer. Therefore, identifying high-risk individuals before the onset of IBD symptoms is crucial to potentially preventing or delaying irreversible bowel damage and disease progression.[10] Many studies have developed models to stratify high-risk and low-risk individuals for CD and UC using polygenic risk scores (PRSs) that incorporate GWAS summary statistics and individual genotype data.[11,12] PRSs use genetic variants to estimate an individual's susceptibility to developing a disease. However, since IBD is also influenced by non-genetic factors like lifestyle and environmental influences, accurately assessing IBD risk using models based solely on genetic data is challenging.

IBD is an autoimmune condition, so the current state of an individual's immune system provides valuable information about symptom onset.[13] While genetic data provide insights into susceptibility, they cannot predict when symptoms will appear or how the disease will progress. A PRS can identify individuals with high genetic risk for IBD, but these individuals may not necessarily develop the disease if they effectively manage factors that influence their immune system and overall health. This highlights the importance of integrating both genetic predisposition and variable non-genetic factors for a comprehensive assessment of IBD risk.

Recently, high-dimensional circulating plasma proteomics profiling has been used as a non-invasive tool to understand complex diseases on a large scale and act as endophenotypes related to disease pathogenesis and progression. Plasma proteomics provide a snapshot of an individual's current immune status, including many health-related processes and pathways. Studies have found proteins associated with the prevalence of a range of complex diseases,[14,15] including IBD.[16,17] Additionally, protein levels prior to diagnosis have been linked with subsequent disease onset,[18,19] including in IBD,[20] further motivating their use as a predictive tool. Consequently, these developments produced proteomic risk scores (ProRS), where protein signatures are consolidated into a score for the current risk of developing a disease.[21–23] Proteomic signatures are broadly more predictive of complex disease incidence and prevalence than PRS.[15,24] However, many diseases have both genetic and non-genetic components predictive of disease onset, so efforts have been made to combine scores through multi-omic integration of PRS and ProRS. Evidence suggests this combination improves the prediction of coronary artery disease,[25] coronary plaques,[26] and type 2 diabetes;[22] however, this has not been explored in IBD.

We used data from 51,772 patients in the UK Biobank (UKB) to characterize the unique contributions of polygenic risk and proteomic risk to IBD onset prediction in the largest available proteomics dataset (Figure 1). Despite the superior performance of ProRS compared to PRS for IBD risk assessment, we combined circulating plasma proteomics with genetics in two ways to leverage their gene-environment (GxE) interactions[27] and provide a more comprehensive risk assessment. We directly integrated proteomics and genetics as predictors in the same model, as well as stratified patients by PRS before assessing ProRS, showcasing the interactions between the two modalities and identifying disease-associated protein biomarkers. Since proteomics data can be obtained during a routine clinical blood test, we tested the accuracy of a personalized medicine approach through omics integration for IBD risk.

2. Methods

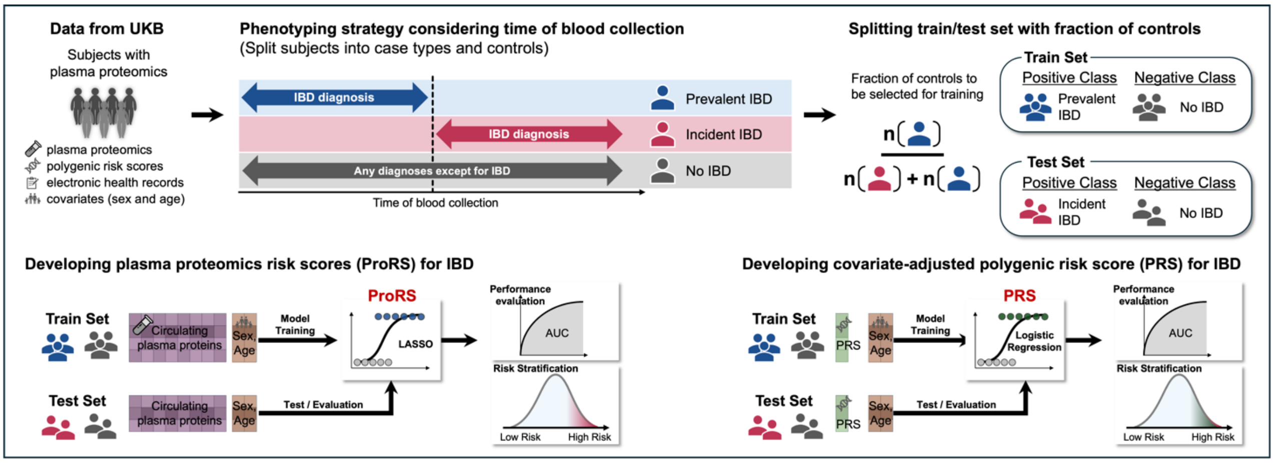

Fig. 1. Study Overview. Data was collected from UKB including plasma protein levels, disease-specific PRS models, age, sex, and ICD-10 codes. Cases were separated into prevalent and incident groups, and the fraction of cases that were prevalent determined the fraction of controls assigned to the training set. The rest were assigned to the testing set. The training set was used to create logistic regression models for the covariate-only model and covariate-adjusted PRS as well as LASSO models for the ProRS and combined score.

2.1. *Data and study participants*

2.1.1. *UK Biobank*

We used data from the UKB, a large-scale biomedical database that provides an extensive collection of genetic, health, and lifestyle information from half a million participants from the UK aged 40-69 at recruitment. With genotype information, International Classification of Diseases (ICD) codes from electronic health records, and biological samples saved for later analysis, the biobank provides the largest resource to study IBD. The breadth of data collected by the UKB and its large sample sizes enabled this project to analyze multi-omic data in a substantial sample with an adequate number of disease cases.

2.1.2. *Circulating proteomics*

In October 2023, the UKB released plasma protein levels of 53,018 blood samples from participants collected at recruitment between 2006 and 2010 as part of the Pharma Proteomics Project (UKB-PPP).[28] The circulating levels of 2,923 proteins were recorded using the Olink Explore 3072 proximity extension assay. The data had about 17.5% missingness. To preserve as many samples as possible, the missing values were imputed using the k-nearest neighbors imputation method with $k = 10$.[29] Before imputation, individuals with greater than 54% missingness ($n = 698$) and proteins with greater than 30% missingness ($n = 3$) were excluded, reducing the total missingness to 9.5%.

2.1.3. *Phenotyping*

Binary phenotypes for each IBD subtype were established using the ICD diagnosis codes K50* for CD and K51* for UC. If an individual's date of first report of disease occurrence was before their blood draw, from which circulating proteomics were profiled, they were labeled as a prevalent case. If their date of first disease occurrence was after their initial blood draw, they were labeled as an incident case. Otherwise, they were considered controls (Figure 1). Additionally, Hospital Episode Statistics were used to identify the specific ICD code for each case. A rheumatologist classified each code within K50* and K51* as an autoimmune disease or other rheumatic condition. 35 individuals with codes in the UC block (K51) that had non-autoimmune diseases (K51.4, K51.5) but no other autoimmune disease in the block, were removed from the analysis. 55 individuals had both CD and UC codes at baseline, and so were considered prevalent cases in both models. For survival analysis, individuals were considered to have the event at their date of first occurrence of the disease. Individuals were censored at their date of death if they appeared in the central death registry. To generalize the findings as much as possible, our analyses included all individuals, regardless of ancestral background. However, the vast majority of the study population self-identified as white British (n = 43,047, 83.1%).

2.2. ***Risk Scoring***

We developed a ProRS for each of UC and CD to quantify the likelihood of disease onset in undiagnosed individuals using proteomics data. To differentiate protein levels between healthy subjects and IBD patients, we stratified cases by time of disease onset (see Phenotyping) and used the prevalent cases for model development (training set). Since ProRS aims to predict future IBD onset after blood collection, the incident cases were used for model evaluation (testing set). Due to limited follow-up time in the UKB, there are fewer incident IBD cases compared to prevalent cases. This discrepancy results in an imbalance between the number of training and testing cases, which could potentially affect the accuracy and evaluation of our models by introducing bias and reducing generalizability. To address this imbalance, we randomly split controls into the training and testing sets with the same ratio as prevalent to incident cases in the data.

The train/test split can be described as follows. Let $\boldsymbol{S}_{\text{train}}^{\text{case}}(\cdot)$ and $\boldsymbol{S}_{\text{test}}^{\text{case}}(\cdot)$ be the case set of patients who were diagnosed with the disease before and after blood collection respectively, where the parentheses represent the disease of interest. Given the index disease, the control set is defined as $\boldsymbol{S}^{\text{control}}(\cdot) = \{\boldsymbol{S}_{\text{train}}^{\text{case}}(\cdot) \cup \boldsymbol{S}_{\text{test}}^{\text{case}}(\cdot)\}^{c}$. We then randomly selected the training control set $\boldsymbol{S}_{\text{train}}^{\text{control}}(\cdot)$ from $\boldsymbol{S}^{\text{control}}(\cdot)$ such that the proportion of all controls that are in $\boldsymbol{S}_{\text{train}}^{\text{control}}(\cdot)$ equaled the proportion of the total number of cases ($|\boldsymbol{S}_{\text{train}}^{\text{case}}(\cdot) + \boldsymbol{S}_{\text{test}}^{\text{case}}(\cdot)|$) that are prevalent cases ($|\boldsymbol{S}_{\text{train}}^{\text{case}}(\cdot)|$). The testing control set is then the remaining set of controls:

$$\begin{cases} \boldsymbol{S}_{\text{train}}^{\text{control}}(\cdot) \subset \boldsymbol{S}^{\text{control}}(\cdot) \\ \boldsymbol{S}_{\text{test}}^{\text{control}}(\cdot) = \boldsymbol{S}^{\text{control}}(\cdot) \setminus \boldsymbol{S}_{\text{train}}^{\text{control}}(\cdot) \end{cases}$$

This approach ensures that the model is trained and evaluated on disjoint datasets with balanced case-control ratios so that the ProRS's performance can be accurately assessed despite the differences in the numbers of prevalent and incident IBD cases. The resultant case and control counts in each set for both diseases are shown in Table 1.

To evaluate the contribution of each omic level to risk prediction, four models were created for CD and UC separately: a covariate-only model, PRS, ProRS, and a combined model. The covariate model used only the age at plasma protein measurement and sex in an unpenalized logistic regression, acting as a baseline prediction. The PRS model was based on scores from Thompson et al.,[30] which were added as predictors to an unpenalized logistic regression with the covariates. After removing individuals missing a PRS, we analyzed an overall sample size of 51,772 for CD and 51,737 for UC.

Since not all 2,920 proteins are expected to be informative of disease status, we applied covariate-adjusted Least Absolute Shrinkage and Selection Operator (LASSO) models to develop the ProRS while adjusting for potential confounders (sex and age).[31] The method allows for simultaneous protein marker selection and regularization, defined by the equation:

$$\hat{\beta} = \mathrm{argmin}_{\beta} \sum_{i=1}^{n} \log\left(\exp\left(-y_i(X_i^T\beta)\right) + 1\right) \; + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \tag{1}$$

where n is the sample size, y_i is the class information (case or control) for individual i, X_i are the values of the predictors (circulating protein levels and covariates), β_j is the regression coefficient for predictor j, p is the number of predictors, and λ is the regularization parameter controlling the strength of the penalty. Protein features with non-zero coefficient ($\beta_j \neq 0$) in the trained model were considered significant proteins associated with IBD. The chosen λ was the minimum λ from 5-fold cross-validation. The combined model was created in the same fashion, with all protein values, the PRS, age, and sex as predictors in a LASSO model. Prior to LASSO, each predictor was standardized to a mean of 0 and standard deviation of 1 so that the coefficient magnitudes would be comparable. Scores for all four models were computed for each individual in the disease's testing set for further analysis and performance evaluation of disease onset prediction.

Table 1. Case-control counts and covariates by disease. The number of individuals in the case and control groups for the train and test sets for CD and UC, along with sex, mean age, and the number of individuals self-identified as white-British in each subgroup. The p-values for "Female" and "White-British" were calculated by the chi-square test, and the p-values for "Age" were calculated by the Wilcoxon signed rank test for difference between the case and control groups.

Phenotypes		Train				Test			
		Total	Case	Control	p-value	Total	Case	Control	p-value
CD	N	31,879	242	31,637	-	19,893	151	19,742	-
	Female	17,188 (54%)	124 (51%)	17,064 (54%)	0.439	10,738 (54%)	78 (52%)	10,660 (54%)	0.622
	Age	56.8 (56.7-56.9)	56.3 (55.3-57.4)	56.8 (56.7-56.9)	0.3374	56.9 (56.8-57.0)	58.5 (57.2-59.8)	56.9 (56.8-57.0)	0.01195
	White-British	26,481 (83%)	204 (84%)	26,277 (83%)	0.6699	16,566 (83%)	129 (85%)	16,437 (83%)	0.5466
UC	N	34,567	453	34,114	-	17,170	225	16,945	-
	Female	18,627 (54%)	228 (50%)	18,399 (54%)	0.1387	9,280 (54%)	118 (52%)	9,162 (54%)	0.6756
	Age	56.8 (56.7-56.9)	58.7 (58.0-59.4)	56.8 (56.7-56.9)	1.2e-6	56.8 (56.7-56.9)	57.0 (55.9-58.0)	56.8 (56.7-56.9)	0.8897
	White-British	28,740 (83%)	386 (85%)	28,354 (83%)	0.2629	14,276 (83%)	195 (87%)	14,081 (83%)	0.1833

2.3. *Statistical Analyses*

2.3.1. *Risk prediction evaluation*

All data analyses were performed in R 4.4.0. All models were adjusted for age and sex. Area under the receiver operating characteristic curve (AUC) and Nagelkerke's R^2 were used as evaluation metrics to assess the classification ability of each quantitative score.[32] DeLong's test was used to compare AUCs and establish confidence intervals[33] with the pROC R package.[34] This nonparametric approach is suitable for comparing AUCs of two correlated receiver operating characteristic curves, especially when the models are built from the same samples. The CD and UC ProRS models had more proteins with non-zero coefficients than the combined models. In order to evaluate their genetic backing, SNP-based heritability estimates were established for the circulating levels of each protein from pQTL summary statistics of European ancestry individuals[28] using LD score regression of roughly 1.2 million HapMap3 SNPs.[35] Gene set enrichment analysis was then used to test if the heritability estimates were higher in the sets of removed proteins than expected by chance. This analysis was run with the clusterProfiler R package[36] using the heritability estimates of all 2,923 proteins as the background set. Kaplan-Meier cumulative incidence curves were constructed to visualize and test the cumulative incidence of each disease using the survminer R package.[37]

2.3.2. *Longitudinal Analyses*

As protein levels in an individual are dynamic while genotypes are static, performances of the PRS and ProRS models were evaluated in the short term (5 years) and in the long term (10 years) after the blood draw. In these experiments, individuals were only considered incident cases if they were diagnosed with the disease within that time frame (five or ten years). Otherwise, they were considered controls.

To test the relationship between the ProRS and time to diagnosis, mean scores were calculated on a backward timescale for each year leading up to the diagnosis date. Those who would go on to develop IBD were tested against those who did not. Using the approach described in Guo, You, Zhang et al., a nested case-control study was implemented to match individuals with incident diagnosis events to healthy controls.[38] Individuals were matched based on age and sex, with a 1:5 case-control ratio. The event date for matched controls was set to their corresponding case, and incident cases past 14 years were set to have an event date of 14 years. Mean values at each time point were fitted using locally weighted smoothing curves ($\alpha = 0.8$). The Mann-Kendall trend test was used to compare differences in ProRS between cases and controls longitudinally.

3. Results

3.1. *Genomics and proteomics uniquely predict IBD incidence*

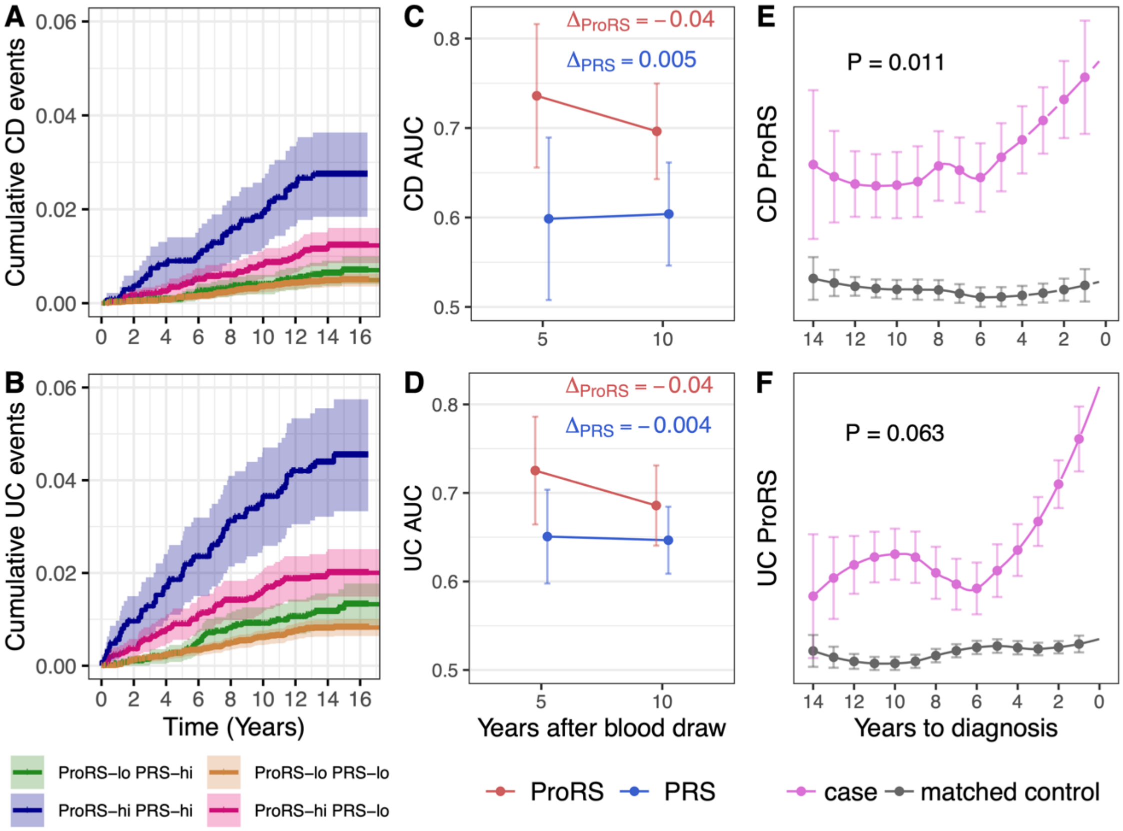

Fig. 2. Longitudinal Analysis. (**A** + **B**) Kaplan Meier curves of time to disease onset, stratified by PRS and ProRS. Individuals are considered high risk when at the 75th percentile or higher. (**C** + **D**) AUC of risk score models at five and ten years after blood draw. (**E** + **F**) ProRS of disease cases compared to age- and sex-matched controls using locally weighted smoothing curves, where the x-axis represents the time after blood draw that individuals were diagnosed. (**A**,**C**,**E**) for Crohn's disease, (**B**,**D**,**F**) for ulcerative colitis.

3.1.1. *PRS and ProRS both effectively stratify individuals at risk for IBD*

ProRS models for CD and UC were developed using LASSO, selecting 216 proteins and 338 proteins, respectively, to predict disease onset. Although age and sex were included as input variables, neither the CD nor the UC model included these covariates as significant features, aligning with the known lack of a sex bias in these diseases.[39] Consistent with other studies, both PRS and ProRS effectively stratified individuals at high risk for disease (Supplemental Figure 1). We also observed that high ProRS was more distinguishing than high PRS. To assess their combined predictive utility, we stratified individuals based on both polygenic risk and proteomic risk. This yielded a cumulative incidence curve with four strata (Figure 2A-B), where high risk was defined as greater than the 75th percentile for each score, and low risk as all others. Interestingly, polygenic risk further stratified individuals within the proteomic risk categories, suggesting PRS can offer additional information on time to disease onset beyond what ProRS can provide.

3.1.2. *ProRS are time-sensitive and reduce in predictive ability over time*

Since circulating protein signatures indicate current health status, we hypothesized that the ProRS predictive accuracy is higher closer to disease onset, while stable for PRS. We tested the models at 5 years and 10 years post blood draw, finding that the ProRS model had an AUC reduction of 0.04 in both CD and UC (CD: 0.74→0.70, UC: 0.73→0.69). The PRS AUC, however, remained similar in both diseases (Figure 2C-D). This reduction may be explained by the observation that the ProRS for both CD and UC increased dramatically in the ~5 years preceding disease diagnosis, whereas matched controls demonstrated little difference in risk over time (Figure 2E-F). The increasing difference in mean ProRS between cases and controls at each time point indicates a likely increase in IBD protein signatures in the years leading up to disease onset (CD: $p = 0.011$, UC: $p = 0.063$).

3.2. *Genomics and proteomics in combination improve IBD prediction*

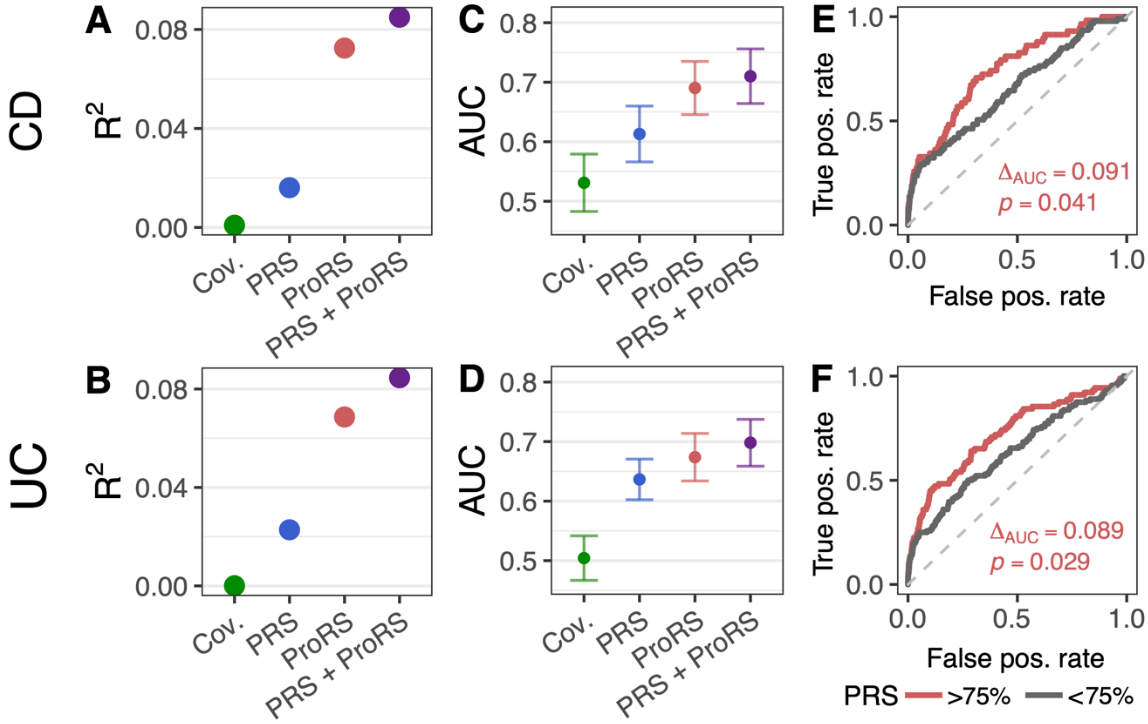

Fig. 3. Combining PRS and ProRS. (A + B) R^2 estimate for disease incidence variance in the covariate-only model and the adjusted PRS, ProRS, and combined models. **(C + D)** AUC comparison of risk score models to predict disease incidence. **(E + F)** Performance of the ProRS model in high disease-risk individuals (>75 percentile) and low disease-risk individuals (<75 percentile). (**A**,**C**,**E**) for Crohn's disease, (**B**,**D**,**F**) for ulcerative colitis.

3.2.1. *PRS adds complementary predictive information to ProRS*

We evaluated each risk score individually and in combination to test their unique and combined contributions to IBD risk prediction. As previously observed, the ProRS had a much higher R^2 (Figure 3A-B) and AUC (Figure 3C-D) than the PRS for predicting IBD subtype incidence, indicating that ProRS more meaningfully stratifies patients at high risk for the disease. The combined model, however, outperforms either omic modality alone. Compared to the ProRS model, the combined model's R^2 increased by 0.012 in CD and 0.016 in UC, while AUC increased by 0.020 in CD and 0.024 in UC. This emphasizes the importance of both genetic and proteomic screening in the clinic to identify patients likely to develop CD or UC soon.

3.2.2. *Adding PRS to ProRS removes more heritable proteins*

In the construction of the ProRS, LASSO selected 216 predictors for CD and 338 for UC. When the PRS was added as a predictor, it became a significant predictor with the 9th largest coefficient in CD and the 4th largest in UC. The number of predictors with non-zero coefficients decreased to 203 in CD and 284 in UC. We hypothesized that the PRS might replace proteins whose levels are influenced by genetics. To test this, we used LD score regression to estimate the heritability for each protein and performed gene set enrichment analysis to see if the heritabilities for the removed proteins were significantly higher than expected by chance. These sets consisted of 24 proteins for CD and 66 proteins for UC. With a p-value of 0.004 for CD and 0.08 for UC, there is evidence that the PRS accounts for heritable differences in protein levels.

3.2.3. *High PRS for CD and UC is associated with better incident disease prediction accuracy*

It is thought that genetically susceptible individuals develop IBD due to specific environmental or lifestyle triggers. We hypothesize that protein measurements can reflect when such conditions are met. To test this, we stratified individuals into high (>75 percentile) and low (<75 percentile) PRS groups and evaluated the accuracy of ProRS (Figure 3E-F). Compared to the low PRS group, we observed that the AUC in the high PRS group is 0.091 higher in CD (p = 0.041), and 0.089 higher in UC (p = 0.029). This suggests that an IBD-related inflammatory state from the ProRS model is more predictive in those already known to be at higher risk. This substantial difference in ProRS classification may be explained by higher false positive rates in the low PRS group, resulting from inflammatory states not caused by IBD.

4. Discussion

We evaluated the predictive ability of circulating plasma proteins and genetics for IBD risk and their interactions. Our study highlights three novel findings with implications for their clinical utility. Firstly, combining proteomic and genomic information enabled more precise patient stratification into risk groups. This approach yielded better predictive performance, as indicated by higher AUC and R^2 values, and improved survival analysis for predicting time-to-disease onset. Secondly, stratifying patients by PRS revealed substantial differences in the ProRS model performance for predicting later onset of CD and UC. This may indicate that the inflammatory protein signature is more likely to be an accurate marker of the disease in individuals with high PRS, as opposed to being a confounding condition in low PRS individuals. Thirdly, we found that ProRS prediction accuracy decreases over time, whereas the performance of PRS remained stable. This is likely because ProRS, based on dynamic circulating plasma protein levels, becomes less distinguishing over time, while the static nature of PRS maintains its predictive power.

IBD is a highly polygenic and heritable disease with a significant environmental component. A leading theory of IBD pathogenesis is that environmental exposures in life may trigger inflammatory bowel disease in genetically susceptible individuals.[40] Although this exposure component is difficult to measure, the genetic component is increasingly measurable. Additionally, circulating proteins can act as an early endophenotype to indicate whether the exposure has happened, and autoimmunity initiated. Our demonstration that the performance of ProRS to predict onset of IBD subtypes is increased in high PRS individuals provides further support to this theory.

Advancements in proteomic technologies have enabled biobanks to generate large-scale data for analyzing the circulating proteome, with many new projects already underway.[41,42] Thus, the utility of risk scoring for precision medicine in both clinical and research settings is becoming more realistic. With increasingly affordable genotyping technologies, it is plausible that lifetime polygenic risk for diseases could be part of a patient's health history available to clinicians. If circulating plasma proteomics were measured in a patient and a ProRS developed, the additional insight from a PRS could help refine this risk. For example, higher PRS could indicate higher confidence in the estimated probability of developing IBD. Additionally, the falling accuracy of ProRS over time suggests scores from older data should be analyzed with skepticism. Given the difference in the cost of genotyping a patient and generating proteomics panels, we suggest an initial assessment with a cheaper genomics approach may be more efficient. If a patient is at high genetic risk for IBD, regularly generating proteomics panels may be necessary.

There are several limitations in our study motivating future work. We used a simple linear model with an L1 penalty to generate the ProRS, but such models may oversimplify the complex biological interactions between circulating proteins and genetic factors. Although preliminary evidence suggests that ensemble methods for proteomic scores perform equally to linear methods when predicting cardiovascular events[31], linear models inherently cannot capture higher-order interactions that might be important for predicting disease risk. In future studies, more sophisticated computational methodologies should be explored for predictive capacity, such as graph machine learning algorithms that might better represent the relationships between biological entities. Another limitation is that this study was only performed in one biobank, with no external validation. Given the uniqueness of the UKB proteomics dataset, it is not possible to replicate the results on a large scale, but more datasets will soon be available for validation. This single biobank also means that results can only be interpreted for a British population. The effect of ancestry could not be sufficiently evaluated in this study due to power constraints. However, protein risk scores have been reported to be transferable across populations with no heterogeneity in effect, even with models trained on much smaller sample sizes.[31] Nonetheless, we acknowledge the need for more diverse cohorts in multi-omic studies. A further limitation of the UKB is the well documented challenge of using mapped ICD-10 codes for phenotyping.[43] Studies suggest positive predictive values of >70% for mapping electronic health records to stroke[44] and acute myocardial infarction,[45] however further work is needed to evaluate their accuracy in phenotyping IBD onset.

This study demonstrates the predictive nature of genetic risk scores, proteomic risk scores, and especially their combination, on IBD incidence. Future work involves using large biobank proteomics to predict IBD progression and prognosis, as shown in smaller studies.[46,47] There is also evidence that proteomics[48] and genomics[49] can be employed to subtype IBD, and their integration may be useful to further distinguish disease types to inform the best clinical care. Our approach is appropriate to analyze any heritable condition that can arise throughout life and would be valuable to apply to more autoimmune and neurodegenerative diseases. These results offer hope for successfully integrating biological data to improve risk prediction.

5. Acknowledgments

This work was supported by NIGMS R01 GM138597 and has been conducted using the UK Biobank Resource under Application Number 45227.

6. Code Availability

No previously unreported algorithm was used to generate results central to the conclusions. Any additional information required to re-analyze the data reported in this work paper is available from the lead contact upon request.

7. Supplemental Materials

All supplemental materials are available at https://s3.amazonaws.com/biomedinfolab/supp/ibd.pdf.

References

1. Seyedian, S. S., Nokhostin, F. & Malamir, M. D. A review of the diagnosis, prevention, and treatment methods of inflammatory bowel disease. *J. Med. Life* **12**, 113–122 (2019).
2. Lewis, J. D. *et al.* Incidence, Prevalence, and Racial and Ethnic Distribution of Inflammatory Bowel Disease in the United States. *Gastroenterology* **165**, 1197-1205.e2 (2023).
3. Xu, F. Health-Risk Behaviors and Chronic Conditions Among Adults with Inflammatory Bowel Disease — United States, 2015 and 2016. *MMWR Morb. Mortal. Wkly. Rep.* **67**, (2018).
4. Dahlhamer, J. M. Prevalence of Inflammatory Bowel Disease Among Adults Aged ≥18 Years — United States, 2015. *MMWR Morb. Mortal. Wkly. Rep.* **65**, (2016).
5. El Hadad, J., Schreiner, P., Vavricka, S. R. & Greuter, T. The Genetics of Inflammatory Bowel Disease. *Mol. Diagn. Ther.* **28**, 27–35 (2024).
6. Santos, M. P. C., Gomes, C. & Torres, J. Familial and ethnic risk in inflammatory bowel disease. *Ann. Gastroenterol.* **31**, 14–23 (2018).
7. Mawdsley, J. E. & Rampton, D. S. Psychological stress in IBD: new insights into pathogenic and therapeutic implications. *Gut* **54**, 1481–1491 (2005).
8. Parkes, G. C., Whelan, K. & Lindsay, J. O. Smoking in inflammatory bowel disease: Impact on disease course and insights into the aetiology of its effect. *J. Crohns Colitis* **8**, 717–725 (2014).
9. Ng, S. C. *et al.* Geographical variability and environmental risk factors in inflammatory bowel disease. *Gut* **62**, 630–649 (2013).
10. Noor, N. M., Sousa, P., Paul, S. & Roblin, X. Early Diagnosis, Early Stratification, and Early Intervention to Deliver Precision Medicine in IBD. *Inflamm. Bowel Dis.* **28**, 1254–1264 (2022).
11. Khunsriraksakul, C. *et al.* Construction and Application of Polygenic Risk Scores in Autoimmune Diseases. *Front. Immunol.* **13**, 889296 (2022).
12. Gettler, K. *et al.* Common and Rare Variant Prediction and Penetrance of IBD in a Large, Multi-ethnic, Health System-based Biobank Cohort. *Gastroenterology* **160**, 1546–1557 (2021).
13. de Souza, H. S. P. & Fiocchi, C. Immunopathogenesis of IBD: current state of the art. *Nat. Rev. Gastroenterol. Hepatol.* **13**, 13–27 (2016).
14. Jacobs, B. M. *et al.* Plasma proteomic profiles of UK Biobank participants with multiple sclerosis. *Ann. Clin. Transl. Neurol.* **11**, 698–709 (2024).
15. Smelik, M. *et al.* An interactive atlas of genomic, proteomic, and metabolomic biomarkers promotes the potential of proteins to predict complex diseases. *Sci. Rep.* **14**, 12710 (2024).
16. Di Narzo, A. F. *et al.* High-Throughput Identification of the Plasma Proteomic Signature of Inflammatory Bowel Disease. *J. Crohns Colitis* **13**, 462–471 (2019).
17. Drobin, K. *et al.* Targeted Analysis of Serum Proteins Encoded at Known Inflammatory Bowel Disease Risk Loci. *Inflamm. Bowel Dis.* **25**, 306–316 (2019).
18. Papier, K. *et al.* Identifying proteomic risk factors for cancer using prospective and exome analyses of 1463 circulating proteins and risk of 19 cancers in the UK Biobank. *Nat. Commun.* **15**, 4010 (2024).

19. Tran, D. *et al.* Plasma Proteomic Signature Predicts Myeloid Neoplasm Risk. *Clin. Cancer Res.* OF1–OF9 (2024) doi:10.1158/1078-0432.CCR-23-3468.
20. Torres, J. *et al.* Serum Biomarkers Identify Patients Who Will Develop Inflammatory Bowel Diseases Up to 5 Years Before Diagnosis. *Gastroenterology* **159**, 96–104 (2020).
21. You, J. *et al.* Plasma proteomic profiles predict individual future health risk. *Nat. Commun.* **14**, 1–13 (2023).
22. Gadd, D. A. *et al.* Blood protein levels predict leading incident diseases and mortality in UK Biobank. 2023.05.01.23288879 Preprint at https://doi.org/10.1101/2023.05.01.23288879 (2023).
23. Ganz, P. *et al.* Development and Validation of a Protein-Based Risk Score for Cardiovascular Outcomes Among Patients With Stable Coronary Heart Disease. *JAMA* **315**, 2532–2541 (2016).
24. Carrasco-Zanini, J. *et al.* Proteomic signatures improve risk prediction for common and rare diseases. *Nat. Med.* 1–10 (2024) doi:10.1038/s41591-024-03142-z.
25. Møller, P. L. *et al.* Combining Polygenic and Proteomic Risk Scores With Clinical Risk Factors to Improve Performance for Diagnosing Absence of Coronary Artery Disease in Patients With de novo Chest Pain. *Circ. Genomic Precis. Med.* **16**, 442–451 (2023).
26. Møller, P. L. *et al.* Predicting the presence of coronary plaques featuring high-risk characteristics using polygenic risk scores and targeted proteomics in patients with suspected coronary artery disease. *Genome Med.* **16**, 40 (2024).
27. Yang, A. Z. & Jostins-Dean, L. Environmental variables and genome-environment interactions predicting IBD diagnosis in large UK cohort. *Sci. Rep.* **12**, 10890 (2022).
28. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
29. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**, 2825–2830 (2011).
30. Thompson, D. J. *et al.* UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. 2022.06.16.22276246 Preprint at https://doi.org/10.1101/2022.06.16.22276246 (2022).
31. Helgason, H. *et al.* Evaluation of Large-Scale Proteomics for Prediction of Cardiovascular Events. *JAMA* **330**, 725–735 (2023).
32. Nagelkerke, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691–692 (1991).
33. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
34. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
35. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
36. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).
37. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *J. Am. Stat. Assoc.* **53**, 457–481 (1958).
38. Guo, Y. *et al.* Plasma proteomic profiles predict future dementia in healthy adults. *Nat. Aging* **4**, 247–260 (2024).
39. Ngo, S. T., Steyn, F. J. & McCombe, P. A. Gender differences in autoimmune disease. *Front. Neuroendocrinol.* **35**, 347–369 (2014).
40. Borowitz, S. M. The epidemiology of inflammatory bowel disease: Clues to pathogenesis? *Front. Pediatr.* **10**, 1103713 (2023).
41. Verma, A. *et al.* The Penn Medicine BioBank: Towards a Genomics-Enabled Learning Healthcare System to Accelerate Precision Medicine in a Diverse Population. *J. Pers. Med.* **12**, 1974 (2022).
42. Sun, B. B., Suhre, K. & Gibson, B. W. Promises and Challenges of populational Proteomics in Health and Disease. *Mol. Cell. Proteomics* **23**, (2024).
43. Stroganov, O. *et al.* Mapping of UK Biobank clinical codes: Challenges and possible solutions. *PLOS ONE* **17**, e0275816 (2022).
44. Woodfield, R., Grant, I., Group, U. B. S. O., Group, U. B. F.-U. and O. W. & Sudlow, C. L. M. Accuracy of Electronic Health Record Data for Identifying Stroke Cases in Large-Scale Epidemiological Studies: A Systematic Review from the UK Biobank Stroke Outcomes Group. *PLOS ONE* **10**, e0140533 (2015).
45. Rubbo, B. *et al.* Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: A systematic review and recommendations. *Int. J. Cardiol.* **187**, 705–711 (2015).

46. Ungaro, R. C. *et al.* Machine learning identifies novel blood protein predictors of penetrating and stricturing complications in newly diagnosed paediatric Crohn's disease. *Aliment. Pharmacol. Ther.* **53**, 281–290 (2021).
47. Kalla, R. *et al.* Serum proteomic profiling at diagnosis predicts clinical course, and need for intensification of treatment in inflammatory bowel disease. *J. Crohns Colitis* **15**, 699–708 (2020).
48. Fabian, O. *et al.* A Current State of Proteomics in Adult and Pediatric Inflammatory Bowel Diseases: A Systematic Search and Review. *Int. J. Mol. Sci.* **24**, 9386 (2023).
49. Voskuil, M. D. *et al.* Genetic Risk Scores Identify Genetic Aetiology of Inflammatory Bowel Disease Phenotypes. *J. Crohns Colitis* **15**, 930–937 (2021).