Characterizing very early onset inflammatory bowel disease by leveraging machine learning-based data extraction and unsupervised clustering

Julia Smith[1,6], Xiaoxuan Liu[1], Nicholas Denson[1], Aldo Majluta-Yeb[1], Merissa Maccani[1], Lauren Erdman[2,4], Oscar Lopez-Nuñez[3,5], Jasbir Dhaliwal[1,4]

[1]Division of Gastroenterology, Hepatology and Nutrition, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio; [2]Cincinnati Children's Hospital James M. Anderson Center for Health Systems Excellence, Cincinnati, Ohio; [3]Division of Pathology and Laboratory Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio; Department of Pediatrics, [4]University of Cincinnati College of Medicine, Cincinnati, Ohio; [5]Department of Pathology and Laboratory Medicine, University of Cincinnati College of Medicine, Cincinnati, Ohio; [6]The Ohio State University College of Medicine, Columbus, Ohio

Introduction: Approximately 25% of all patients with inflammatory bowel disease (IBD) are diagnosed during childhood, with 3-5% diagnosed before the age of 6, classified as Very Early Onset IBD (VEO-IBD). It can be challenging to diagnose VEO-IBD, and to further discern between Crohn's disease (CD) and ulcerative colitis (UC). There is a paucity of data that effectively phenotype VEO-IBD patients with the intent to understand biologic disease mechanisms and long-term outcomes.

Methods: Baseline demographic and phenotypic information were extracted from the electronic medical record for 105 VEO-IBD patients, including endoscopic disease location, physician global assessment (PGA) severity scores, and genetic testing for drivers of disease. Quality control was implemented to remove categorical features with >70% missing values, and patients with >15% missing values across all features. We then applied natural language processing (NLP) techniques to extract relevant insights from unstructured diagnostic pathology reports of 105 VEO-IBD patients, which identified 116 feature-location pairs with ≥5 occurrences in the cohort. Following histologic feature extraction, we defined phenotypic subgroups by integrating histologic and endoscopic (disease location) data, applying K-means unsupervised clustering with Principal Component Analysis (PCA) for visualization of engineered features. Clusters were further characterized using disease phenotype data, along with endoscopic and histologic features that distinguished the groups. Significant histologic features within the two clusters were selected using a Chi-squared test with significance defined as $p<0.05$.

Results: K-means clustering of the histological feature-location pairs with PCA demonstrated two patient cluster groups using the histologic feature pairs and 14 macroscopic disease location patterns (Group 1, n=14; Group 2, n=91, explained variance=0.21). There were no demographic differences between the two cluster groups related to age, sex, race, ethnicity, monogenic etiology, NOD2 risk allele presence, or PGA clinical scores at diagnosis. Group 1 demonstrated greater proportions of acute (neutrophilic) involvement throughout the colon ($p<0.001$), terminal ileum (TI) ($p=0.002$), and stomach ($p=0.003$) than Group 2. Lamina propria lymphocytes were significantly more frequent in Group 1 in the stomach (71.4% vs. 5.5%), terminal ileum (TI) (42.9% vs. 2.2%), ascending colon (78.6% vs. 3.3%) and the remainder of the left colon (93% vs 2-3%, all $p<0.001$). Similar patterns were observed with lamina propria plasma cells, with greater proportions in Group 1 across the stomach, TI, and colon (all $p<0.001$), consistent with chronic inflammation. Colonic granulomas occurred more frequently in Group 1 (21.4% vs. 4.4%, $p=0.05$).

Conclusion: We undertook histologic feature extraction and unsupervised clustering to evaluate VEO-IBD disease phenotypes. Preliminary analyses identified two distinct clusters defined by macroscopic disease location, further stratified by histologic feature profiles. Future analyses employing similarity network fusion (SNF) may yield a more robust understanding of VEO-IBD phenotypes, and their potential relationship to long-term clinical outcomes.

Acknowledgements: This study was supported in part by NIH grant T35 DK060444.