Mucosal microbiome is predictive of pediatric Crohn’s disease across geographic regions in North America [version 1; peer review: awaiting peer review]

Rajesh Shah\textsuperscript{1}, Kristi Hoffman\textsuperscript{2}, Lee Denson\textsuperscript{3}, Subramaniam Kugathasan\textsuperscript{4}, Richard Kellermayer\textsuperscript{2}

\textsuperscript{1}Suite 200, Baylor Health Care System, Austin, Texas, 78735, USA
\textsuperscript{2}Baylor College of Medicine, Houston, USA
\textsuperscript{3}Cincinnati Children’s Hospital Medical Center, Cincinnati, USA
\textsuperscript{4}Emory University, Atlanta, USA

\textbf{Abstract}

\textbf{Background:} Patients with Crohn’s disease (CD) have an altered intestinal microbiome, which may facilitate novel diagnostic testing. However, accuracy of microbiome classification models across geographic regions may be limited. Therefore, we sought to examine geographic variation in the microbiome of patients with CD from North America and test the performance of a machine learning classification model across geographic regions.

\textbf{Methods:} The RISK cohort included 447 pediatric patients with CD and 221 non-inflammatory bowel disease controls from across North America. Terminal ileum, rectal and fecal samples were obtained prior to treatment for microbiome analysis. We divided study sites into 3 geographic regions to examine regional microbiome differences. We trained and tested the performance of a machine learning classification model across these regions.

\textbf{Results:} No differences were seen in the mucosal microbiome of patients with CD across regions or in either the fecal or mucosal microbiomes of controls. Machine learning classification algorithms for patients with CD performed well across regions (area under the receiver operating characteristic curve [AUROC] range of 0.85-0.91) with the best results from terminal ileum.

\textbf{Conclusions:} This study demonstrated the feasibility of microbiome based diagnostic testing in pediatric patients with CD within North America, independently from regional influences.

\textbf{Keywords}
Crohn's disease, microbiome, inflammatory bowel disease, machine learning
Corresponding author: Rajesh Shah (rajesh.shah@bswhealth.org)

Author roles: Shah R: Conceptualization, Data Curation, Formal Analysis, Methodology, Project Administration, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Hoffman K: Conceptualization, Formal Analysis, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; Denson L: Writing – Original Draft Preparation, Writing – Review & Editing; Kugathasan S: Writing – Original Draft Preparation, Writing – Review & Editing; Kellermayer R: Conceptualization, Data Curation, Formal Analysis, Methodology, Project Administration, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: RK was supported by philanthropic donations through the Gutsy Kids Fund led by the Brock Wagner and Klaasmeyer families. The RISK consortium was supported by the Crohn's and Colitis Foundation.

Copyright: © 2022 Shah R et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Shah R, Hoffman K, Denson L et al. Mucosal microbiome is predictive of pediatric Crohn's disease across geographic regions in North America [version 1; peer review: awaiting peer review] F1000Research 2022, 11:156 https://doi.org/10.12688/f1000research.108810.1

First published: 08 Feb 2022, 11:156 https://doi.org/10.12688/f1000research.108810.1
Introduction

Currently, our understanding is the intestinal microbiome plays a role in the pathogenesis of inflammatory bowel disease (IBD), and specifically Crohn’s disease (CD). The RISK consortium found significant differences in the taxonomy of the mucosal and fecal microbiomes of pediatric, treatment naïve patients with CD compared to non-IBD controls. Similar results were demonstrated in a longitudinal study of adult patients with IBD with an emphasis on disruption of the microbiome during periods of disease activity.

Based on an altered microbiome composition in patients with CD, microbiome signatures may be utilized as a diagnostic biomarker. From the original RISK publication, the addition of microbiome data to clinical information improved the performance of their classification models for CD. Similarly, Pascal et al. showed microbiome classification models for CD were accurate and performed well across 4 countries in Europe (Spain, Belgium, the UK and Germany). Recent data suggested that geographic bias, however, may limit the validity of microbiome based diagnostic models. He et al. studied 7,009 individuals from 1 Chinese province with 14 districts to determine regional differences in the microbiome. They found strong associations between microbiome composition and host district, which translated into decreased model performance when classifying metabolic diseases across districts. However, they acknowledged that other diseases, such as CD, could not be studied due to a limited sample size. Therefore, we sought to examine differences in the intestinal microbiome of pediatric patients with CD by region and to determine if geographic bias hinders the performance of a machine learning classification model across regions in North America.

Methods

Participants

A post hoc analysis of the RISK cohort was performed. The RISK cohort was a multicenter study that enrolled treatment naïve pediatric patients aged 3 to 17 years with CD and non-IBD controls from 28 sites with the United States and Canada from 2008 to 2012. All patients had symptoms suggestive of CD, including abdominal pain or diarrhea that prompted evaluation with a colonoscopy with biopsies from the terminal ileum and rectum. A subset of patients also provided fecal samples. Patients were either diagnosed with CD, based on endoscopic appearance and histology, or a non-inflammatory etiology for their symptoms, which served as the non-IBD controls. Full inclusion and exclusion criteria for the RISK cohort have been described in the original publication. In total, 447 patients with CD and 221 non-IBD controls were included in the original publication and they provided a total of 1,321 samples, including 630 ileal, 387 rectal and 304 fecal samples.

Ethical considerations

IRB approval was not required for this study, as deidentified data was used and consent was previously obtained from participants when they enrolled in the RISK cohort study.

Statistical analysis

Age at diagnosis, sex, race, disease phenotype, and treatment center were examined. To evaluate the influence of geography on microbiome composition, we grouped the treatment centers into 3 subjective regions based on overall geography (North-East, South-East and West, Figure 1A).

16S rRNA gene analysis was performed in the original cohort study using the Illumina MiSeq platform. For our analysis, the original biom table was obtained and rarefied to 3,441 sequences per sample. This rarefaction depth was chosen to retain the maximum number of samples and preserve the most amount of sequencing data per sample. The alpha and beta diversity and taxonomic composition of the terminal ileum, rectum, and fecal microbiomes were evaluated using the ATIMA interface version 1.0 available through the Baylor College of Medicine Alkek Center for Metagenomics and Microbiome Research. ATIMA is a graphic user interface that allows users to provide a biom table and mapping file for microbiome analysis. To adjust for potential confounding, MaasLin was used to control for variations in age at diagnosis, sex, race, sample type and geographic region.

Finally, we sought to develop a machine learning model to evaluate the accuracy of a microbiome model to identify patients with CD across different regions. A random forest machine learning model was trained on patients from the North-East and tested in the South-East and West using the R package healthcare.ai version 2.5.0 with the default settings. The healthcare.ai package is an open-source R package that allows for data cleaning, manipulation, imputation, tuning of models and evaluation of model performance. Visualization of model performance with AUROC metrics was done using the R package pROC version 1.18.0.

Results

Based on the terminal ileum biopsies retained after rarefaction, we included 227 patients with CD and 165 controls with a mean age of 12.2 and 12.1 years, respectively. Approximately half of patients with CD and controls were male (58.6%
and 53%, respectively) and a larger proportion of patients with CD were Caucasian compared to controls (78.9% and 68.7%, respectively). Since microbiome composition can be influenced by the presence of stricturing/fistulizing disease and these patients present less of a diagnostic challenge, they were excluded from our analysis to create a consistent population with an inflammatory phenotype. After separating into regions, 182 patients were in the North-East, 33 in the South-East and 12 in the West with CD, and 106 patients in the North-East, 43 in the South-East and 16 in the West without IBD.

For patients with CD, no significant differences were found in alpha and beta diversity of the ileal and rectal mucosal microbiome by geography. However, PCoA plots of unweighted and weighted beta diversity (Figure 1B) determined through the Bray Curtis metric revealed significant differences in fecal samples. In controls, no significant differences were found in alpha and beta diversity of the ileum, rectum or fecal samples. In the South-East, patients with CD had a relative increase in Fusobacteria and Bacteroidetes with a decrease in Actinobacteria and Firmicutes in fecal samples compared to the other 2 regions. This corresponded to an increase in the genera Bacteroides and Fusobacterium with a decrease in Bifidobacterium and Lactobacillus. However, after adjustment with MaasLin, Erwinia was the only genus associated with geographical variation in patients with CD. Specifically, fecal samples from CD patients in the South-East had increased abundance of Erwinia compared to other geographic regions in North America (q=0.04).

Random forest models across sample types performed well (Figure 1C). The best performance occurred with ileal samples (North-East AUROC 0.89, South-East AUROC 0.85 and West AUROC 0.91). The rectal (North-East AUROC 0.87, South-East AUROC 0.83, West AUROC 0.76) and fecal (North-East AUROC 0.82, South-East AUROC 0.85, West AUROC 0.74) samples performed well, but experienced decreased performance in the West. Comparing the models, those for ileum and rectum shared OTUs discriminating CD, which included members of the Lachnospiraceae and Clostridiaceae families and the genus Blautia. Intriguingly, ileal biopsies and fecal samples shared top CD-discriminating OTUs from the Erysipelotrichaceae family and Haemophilus genus, which were not present between rectal biopsies and fecal samples.

Discussion
Our results indicate that CD influences mucosal microbiome composition to a greater extent than geography in pediatric patients from North America. Machine learning classification models performed well across the regions, despite minor differences in the fecal microbiome of CD patients. Differences in microbiome composition are known to vary across populations in healthy cohorts and in patients with metabolic syndrome. Yatsunenko et al. showed Westernization may influence fecal microbiome composition by comparing samples from subjects in the US, Venezuela and Malawi. Similar patterns were seen by Pasolli et al. when they examined metagenomes from 9,428 samples from 32 countries and
noted significant differences in the metagenomes of Western populations. Together, these studies demonstrated microbiome composition varies across populations, however, they did not address microbiome differences within countries. To that end, He et al. studied a single province in China and noted differences in microbiome composition between its districts. This suggested, as has been previously reviewed, that a vast number of environmental factors may play a role in shaping the microbiome and may limit the accuracy of microbiome classification models.

Overall, our classification models performed well across regions and is consistent with prior reports. Using 2,045 fecal samples taken from patients with IBD and non-IBD controls across 4 European countries, Pascal et al. showed that a microbial signature could be used to discriminate patients with CD from non-IBD controls with an overall sensitivity of 80% and specificity of 94%. In a separate cohort, Franzosa et al. used metagenomics and metabolomics to distinguish IBD patients from non-IBD patients also with high accuracy. Our findings in pediatric CD are consistent with these results and demonstrate the feasibility of using microbiome classification models to accurately diagnose CD without geographic bias within North America.

Despite the limitations of our study, our classification models performed well. We were unable to adjust for additional confounders of microbiome composition, such as diet and supplement intake. However, even without this information, our models based on ileal biopsies performed well. Additionally, we noted a decrease in model performance for fecal samples and in the West, but this may be linked to a smaller sample size, which is known to hinder the performance of machine learning models. Further work with larger cohorts and different control groups will be needed to fully determine whether microbiome machine learning models can support the diagnosis of CD in children without geographical bias, and if non-invasive testing with fecal samples is feasible.

In summary, machine learning models can distinguish patients with CD from non-IBD controls without geographic bias in North America. Further development of microbiome machine learning models to diagnose CD may be warranted.

Data availability
Underlying data

The underlying clinical data used for this study is available through the RISK consortium. Consortium approval was required to access de-identified patient data and requests can be placed through the Crohn’s and Colitis Foundation IBD Plexus Initiative (www.crohnscolitisfoundation.org/research/grant-s-fellowships/ibd-plexus).

References
The benefits of publishing with F1000Research:

• Your article is published within days, with no editorial bias
• You can publish traditional articles, null/negative results, case reports, data notes and more
• The peer review process is transparent and collaborative
• Your article is indexed in PubMed after passing peer review
• Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com