Predicting the severity of *C. difficile* infections from the taxonomic composition of the gut microbiome

Kelly L. Sovacool[1], Sarah Tomkovich[2], Megan L. Coden[3], Patrick D. Schloss[2,4,]

1 Department of Computational Medicine and Bioinformatics, University of Michigan

2 Department of Microbiology and Immunology, University of Michigan

3 Department of Molecular, Cellular, and Developmental Biology, University of Michigan

4 Center for Computational Medicine and Bioinformatics, University of Michigan

To whom correspondence should be addressed: pschloss@umich.edu

Abstract

C. difficile infection (CDI) can lead to adverse outcomes including recurrent infections, colectomy, and death. The composition of the gut microbiome plays an important role in determining colonization resistance and clearance when exposed to *C. difficile*. We have 16S amplicon sequence data from CDI patient stool samples, with 691 samples classified as severe CDI and 468 as not severe according to the Infectious Diseases Society of America (IDSA) definition. IDSA defines severe CDI cases based on a white blood cell count 15 k/{}L and serum creatinine level 1.5 mg/dL. Sequences were processed with mothur according to the MiSeq SOP and clustered into *de novo* OTUs at a 3% distance threshold. We then trained machine learning (ML) models with OTU abundances as features to predict the IDSA severity of CDI cases using the mikropml R package. The dataset was randomly split into training and testing sets with 80% of the data in the training set, then models were trained with 5-fold cross-validation repeated 100 times, and performance as the area under the receiver-operator curve (AUROC) was measured on the testing set for the best model. This was repeated for 100 different random seeds and three different ML methods: logistic regression, random forest, and support vector machines with a radial basis kernel. This process yielded median AUROC values of 0.61 for logistic regression, 0.60 for random forest, and 0.58 for support vector machines. Feature importance was determined with a permutation test for the best random forest model, revealing that the top 5 OTUs that contributed the most to model performance were *Clostridiales* (OTU 195), *Lachnospiraceae* (OTU 45), *Bacteroides* (OTU 7), *Enterococcus* (OTU 9), and *Pasteurellaceae* (OTU 30). The modest performance may be improved in future work by training to predict clinically confirmed adverse patient outcomes rather than IDSA severity, such as recurrence, admission to intensive care, colectomy, or death. Predicting a patient's risk of experiencing a severe CDI and identifying the specific microbiome features that distinguish severe CDI cases will allow clinicians to tailor interventions based on each patient's individual microbiome, ultimately leading to better health outcomes.

Acknowledgements

This research was supported by National Institutes of Health grants U01AI124255 and the Michigan Institute for Clinical and Health Research Postdoctoral Translational Scholars Program (UL1TR002240 from the National Center for Advancing Translational Sciences).