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

**Kelly L Sovacool[1], Sarah E Tomkovich[2], Megan L Coden[3], Krishna Rao[4], Vincent B Young[2,4], Patrick D Schloss[2,5]**

[1]Dept of Computational Medicine & Bioinformatics, [2]Dept of Microbiology & Immunology, [3]Dept of Molecular, Cellular, and Developmental Biology, [4]Dept of Internal Medicine/Div of Infectious Diseases, [5]Center for Computational Medicine and Bioinformatics, University of Michigan

## Introduction

- C. difficile infection (CDI) can lead to adverse outcomes including recurrent infections, colectomy, and death (1).
- The composition of the gut microbiome plays an important role in determining colonization resistance and clearance when exposed to C. difficile (2, 3).
- Regression models trained on Electronic Health Records extracted on the day of diagnosis perform modestly well at predicting whether the CDI resulted ICU admission, colectomy, or 30-day mortality (AUROC 0.69) (4).
- Identifying the specific microbiome features that distinguish severe CDI cases would allow clinicians to tailor interventions based on a patient's risk, ultimately leading to better health outcomes.

### Dataset

We have 16S amplicon sequence data from 1,191 CDI patient stool samples, with cases classified as severe or not severe according to three separate definitions:

- **IDSA**: the Infectious Diseases Society of America (IDSA) definition with severe CDI having a white blood cell count ≥ 15 k/µL and serum creatinine level ≥ 1.5 mg/dL (5).
- **Attributable**: the CDC definition of ICU admission, colectomy, or death occurring within 30 days of CDI, and confirmed as attributable to CDI via clinical chart review.
- **All-cause**: ICU admission, colectomy, or death occurring within 30 days of CDI, regardless of the cause.

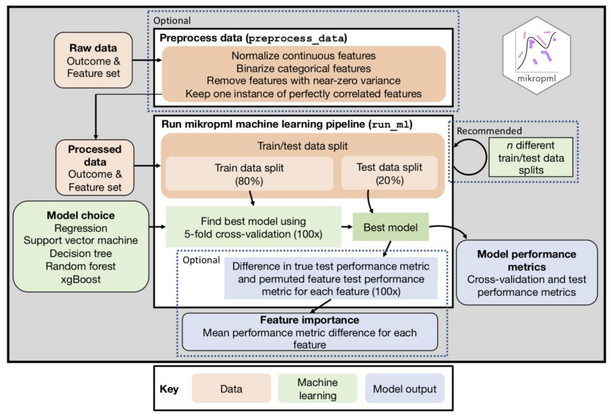
| Severe | IDSA | Attributable | All-cause |
|---|---|---|---|
| no | 649 | 513 | 1059 |
| yes | 342 | 26 | 83 |
| TOTAL | 991 | 539 | 1,142 |

The attributable severity definition requires chart review by physicians, which has been completed for about half of the cases.

## Methods

- Sequences were processed with mothur according to the MiSeq SOP and clustered into de novo OTUs at a 3% distance threshold (6, 7).
- We then trained machine learning (ML) models with OTU abundances as features to predict the IDSA severity, CDI-attributable severity, and all-cause severity of CDI cases using the mikropml R package and accompanying snakemake workflow (8, 9).

### Machine learning pipeline

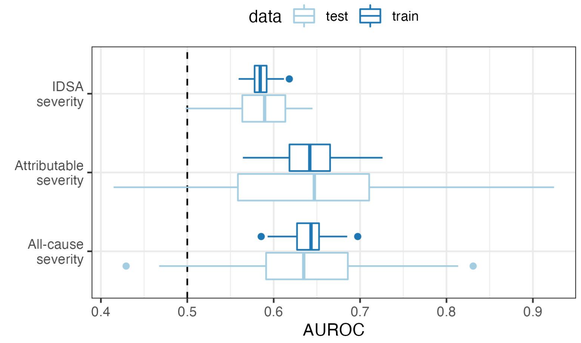



- Prior to model training, the features were pre-processed to scale and center at zero, remove features with near-zero variance, and collapse perfectly correlated features.
- The dataset was randomly split 100 times into training and testing sets with 80% of the data in the training set.
- On each partition, random forest 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 held-out testing set for the best model.
- The importances of features for model performance were found using a permutation test.

## Results

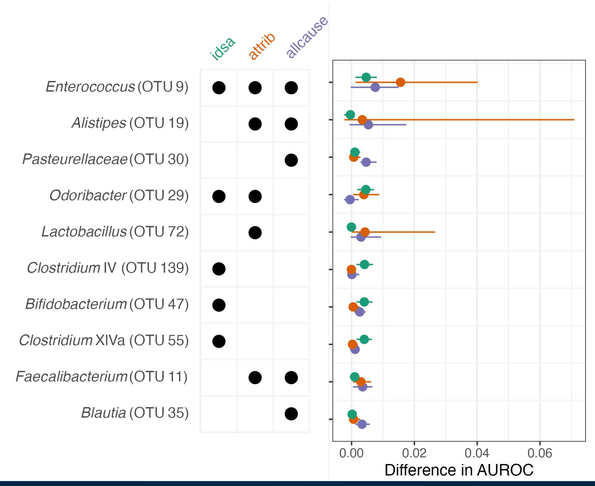
### Model performance

The models predicting CDI-attributable severity performed best (median AUROC 0.65), followed closely by those predicting all-cause severity (median AUROC 0.63). The models predicting IDSA severity performed worst (median AUROC 0.59).




### Feature importance

The top 5 most important OTUs for predicting each outcome were determined with a permutation test.




## Conclusions

- That models predicting CDI-attributable severity performed best implies that chart review by physicians is an important step to filter out other causes of complications.
- The long tails of the performance distributions for CDI-attributable and all-cause severity may reflect the rarity of severe outcomes according to these definitions.
- The poor-to-modest performance of these OTU-based models implies that the taxonomic composition of the microbiome is not the only important factor contributing to severe CDI outcomes.
- Only *Enterococcus* had a feature importance above 0.01 (diff. in AUROC). No single OTU had outsized influence on model performance, perhaps due to the interconnected nature of the microbiome.

### Future directions

- Using the precision-recall curve (AUPRC) may provide a better estimate of model performance than AUROC as the data are imbalanced.
- Training models with both EHR data and OTUs as features may improve model performance.

### Acknowledgements

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

## References

1. Kwon, J. H. et al. Infect Dis Clin North Am 29, 123–134 (2015)
2. Kociolek, L. K. et al. Nat Rev Gastroenterol Hepatol 13, 150–160 (2016)
3. Guh, A. Y. et al. Ann Intern Med 169, ITC49–ITC64 (2018)
4. Li, B. Y. et al. Open Forum Infect Dis 6, ofz186 (2019)
5. Cohen, S. H. et al. Infect Control Hosp Epidemiol 31, 431–455 (2010)
6. Kozich, J. J. et al. Appl. Environ. Microbiol. 79, 5112–5120 (2013)
7. Schloss, P. D. et al. Applied and Environmental Microbiology 75, 7537–7541 (2009)
8. Topçuoğlu, B. D. et al. JOSS 6, 3073 (2021)
9. Köster, J. et al. Bioinformatics 28, 2520–2522 (2012)

Correspondence: pschloss@umich.edu https://www.schlosslab.org