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], **Vincent B. Young**[2,4], **Krishna Rao**[4], and **Patrick D. Schloss**[2,5,†]

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

Compiled May 16, 2023
[†]Correspondence: pschloss@umich.edu.

ABSTRACT TODO

IMPORTANCE TODO

KEYWORDS: *C. difficile* infection, supervised machine learning

INTRODUCTION

A few ways to define CDI severity (Figure 1)

RESULTS

Model performance Report median AUROC for trainset and testset, and median AUBPRC for testset (Figure 2).

Feature importance Most important OTUs contributing to model performance (Figure 3)

Clinical value of severity prediction models

DISCUSSION

TODO

MATERIALS AND METHODS

Sample collection This study was approved by the University of Michigan Institutional Review Board. All patient samples were collected by the University of Michigan Health System from January 2016 through December 2017. Stool samples that had unformed stool consistency were tested for *C. difficile* by the clinical microbiology lab with a two-step algorithm that included detection of *C. difficile* glutamate dehydrogenase and toxins A and B by enzyme immunoassay with reflex to PCR for the *tcdB* gene when results were discordant. 1,517 stool samples were collected from patients diagnosed with a CDI. Leftover stool samples that were sent to the clinical microbiology lab were collected and split into different aliquots. For 16S sequencing, the aliquot of stool was resuspended in DNA genotek stabilization buffer and then stored in the -80°C freezer. Only the first CDI sample per patient was used for subsequent ML analyses such that no patient is represented more than once, resulting in a dataset of 1,191 samples.

16S rRNA gene amplicon sequencing Samples stored in DNA genotek buffer were thawed from the -80°C, vortexed, and then transferred to a 96-well bead beating plate for DNA extractions. DNA was extracted using the DNeasy Power-soil HTP 96 kit (Qiagen) and an EpMotion 5075 automated pipetting system (Eppendorf). The V4 region of the 16S rRNA gene was amplified with the AccuPrime Pfx DNA polymerase (Thermo Fisher Scientific) using custom barcoded primers, as previously described (1). Each library preparation plate for sequencing contained a negative control (water) and mock community control (ZymoBIOMICS microbial community DNA standards). The PCR amplicons were normalized (Sequal-Prep normalizatin plate kit from Thermo Fisher Scientific), pooled and quantified (KAPA library quantification kit from

KAPA Biosystems), and sequenced with the MiSeq system (Illumina).

All sequences were processed with mothur (v1.43) using the MiSeq SOP protocol (2, 1). Paired sequencing reads were combined and aligned with the SILVA (v132) reference database (3) and taxonomy was assigned with a modified version of the Ribosomal Database Project reference sequences (v16) (4). Samples were rarefied to 5,000 sequences per sample, repeated 1,000 times for alpha and beta diversity analysis.

Defining CDI severity IDSA definition of severe CDI based on lab values. CDC definiton of severe CDI based on disease-related complications (5).

Model training and evaluation mikropml R package (6)

Balanced precision

Code availability The complete workflow, code, and supporting files required to reproduce this manuscript with accompanying figures is available at https://github.com/SchlossLab/severe-CDI.

The steps for machine learning, making figures, and rendering the manuscript were defined with Snakemake (7) using a custom version of the mikropml Snakemake workflow (8). Dependencies were managed with conda environments. In addition to the software already cited above, other packages used in the creation of this manuscript include schtools (9), Quarto, ggtext (10), ggsankey (11), cowplot (12), and vegan (**?**).

Data availability The 16S rRNA sequencing data have been deposited in the National Center for Biotechnology Information Sequence Read Archive (BioProject Accession no. PRJNA729511).

ACKNOWLEDGEMENTS

TODO

REFERENCES

1. **Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD**. Sep 2013. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. Appl. Environ. Microbiol. 79 (17):5112–5120. doi: 10.1128/AEM.01043-13.

2. **Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF**. 2009. Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl Environ Microbiol 75 (23):7537–7541. doi:10.1128/AEM.01541-09.

3. **Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO**. Jan 2013. The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools. Nucleic Acids Res 41 (D1):D590–D596. doi:10.1093/nar/gks1219.
4. **Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM**. Jan 2014. Ribosomal Database Project: Data and Tools for High Throughput rRNA Analysis. Nucl. Acids Res. 42 (D1):D633–D642. doi:10.1093/nar/gkt1244.
5. **McDonald LC, Coignard B, Dubberke E, Song X, Horan T, Kutty PK**. 2007. Recommendations for Surveillance of Clostridium Difficile–Associated Disease. Infect Control & Hosp Epidemiol 28 (2):140–145. doi:10.1086/511798.
6. **Topçuoğlu BD, Lapp Z, Sovacool KL, Snitkin E, Wiens J, Schloss PD**. May 2021. Mikropml: User-Friendly R Package for Supervised Machine Learning Pipelines. JOSS 6 (61):3073. doi:10.21105/joss.03073.
7. **Köster J, Rahmann S**. Oct 2012. Snakemake — a Scalable Bioinformatics Workflow Engine. Bioinformatics 28 (19):2520–2522. doi:10.1093/bioinformatics/bts480.
8. **Sovacool K, Lapp Z, Armour C, Lucas SK, Schloss P**. Jan. 2023. Mikropml Snakemake Workflow doi:10.5281/zenodo.4759351.
9. **Sovacool K, Lesniak N, Schloss P**. 2022. Schtools: Schloss Lab Tools for Reproducible Microbiome Research doi:10.5281/zenodo.6540687.
10. **Wilke CO**. 2020. Ggtext: Improved Text Rendering Support for 'Ggplot2'.
11. **Sjoberg D**. 2022. Ggsankey: Sankey, Alluvial and Sankey Bump Plots.
12. **Wilke CO**. 2020. Cowplot: Streamlined Plot Theme and Plot Annotations for 'Ggplot2'.

FIGURES

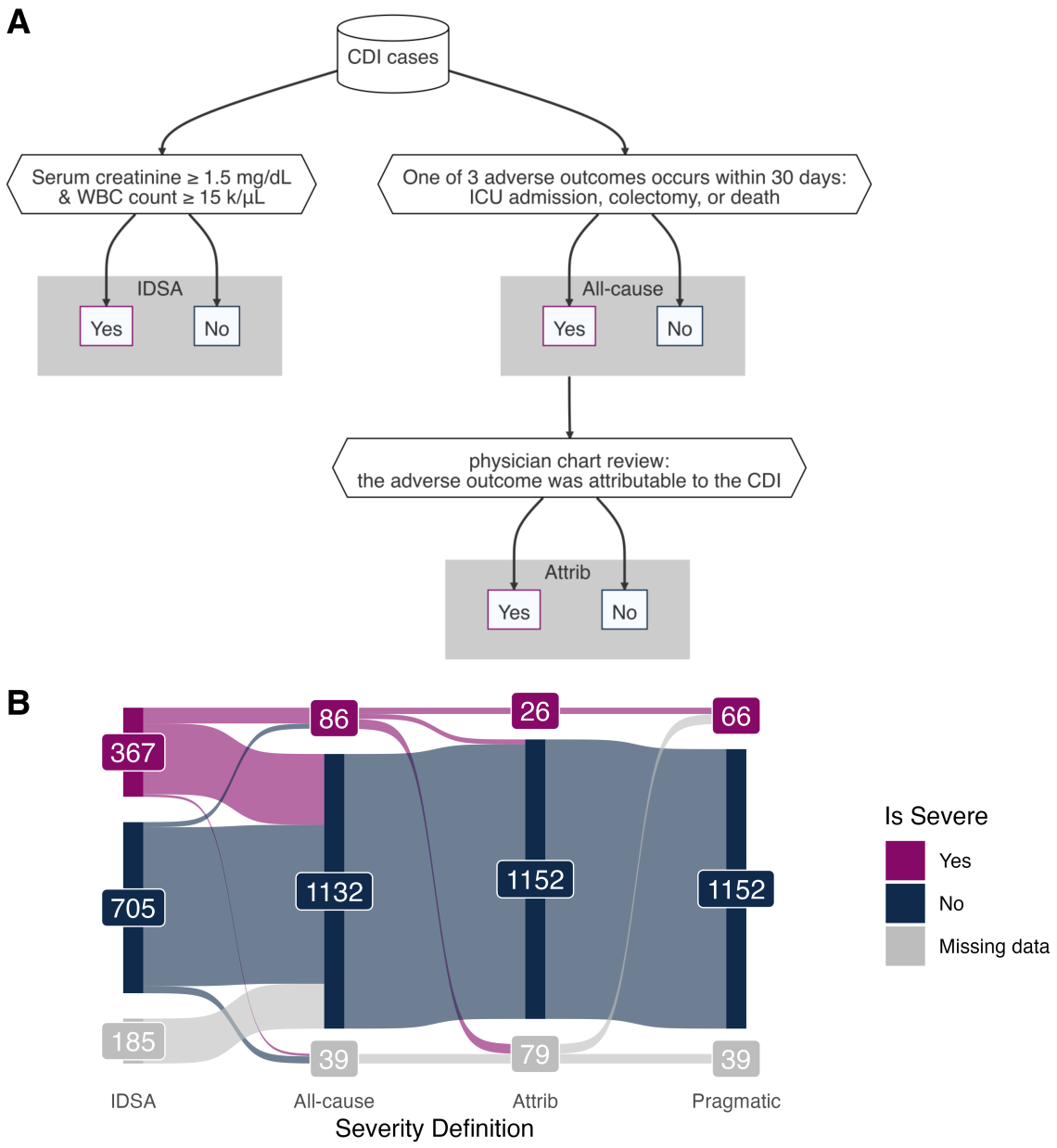

FIG 1 CDI severity definitions. A) Decision flow chart to define CDI cases as severe according to the Infectious Diseases Society of America (IDSA) based on lab values, the occurence of complications due to any cause (All-cause), and the occurence of disease-related complications confirmed as attributable to CDI with chart review (Attrib). B) The proportion of severe CDI cases labelled according to each definition. An additional 'Pragmatic' severity definition uses the Attributable definition when possible, and falls back to the All-cause definition when chart review is not available.

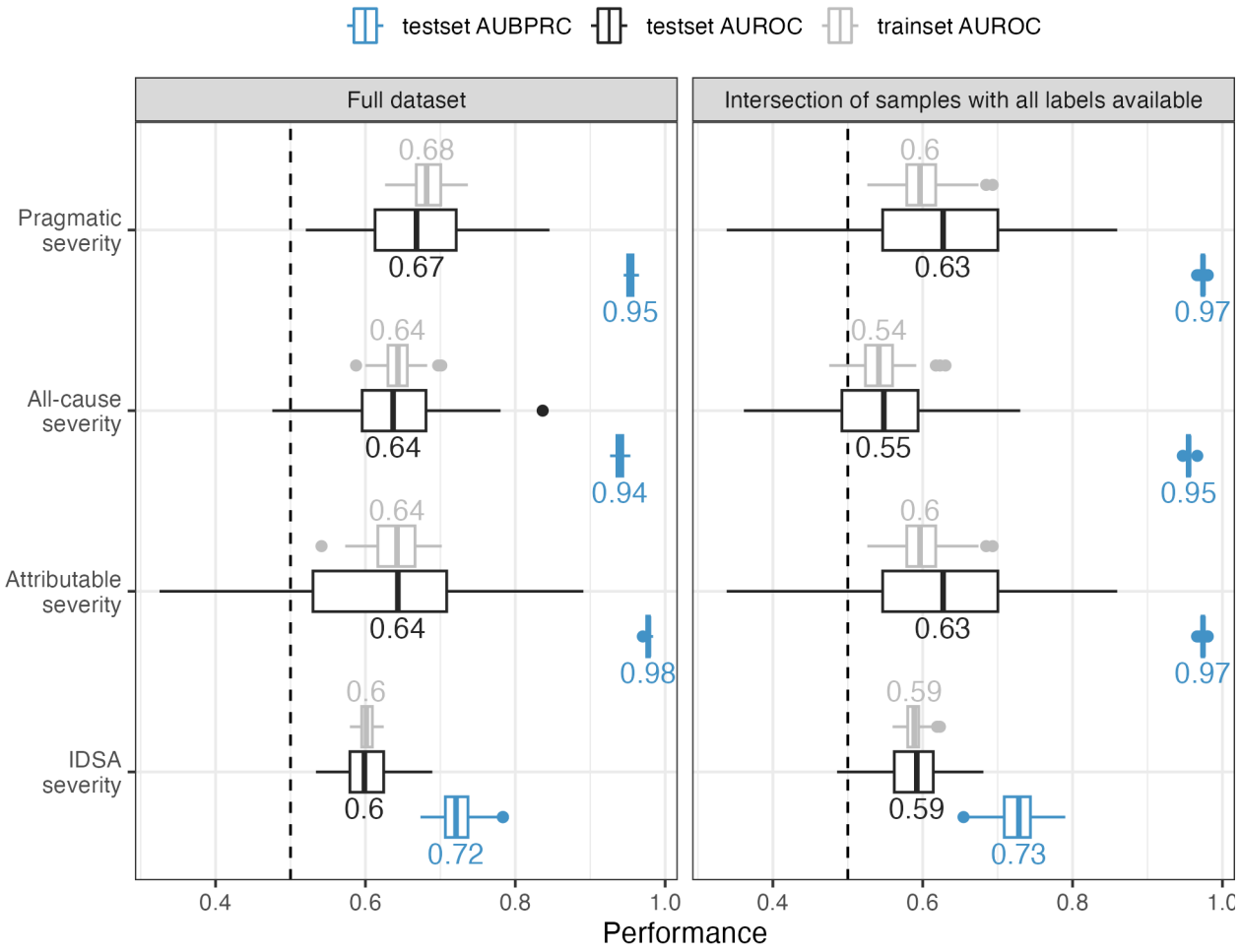

FIG 2 Performance of ML models. Area Under the Receiver-Operator Characteristic Curve (AUROC) for the cross-validation trainsets and testsets, and the Area Under the Balanced Precision-Recall Curve (AUPRC) for the testsets. Left: models were trained on the full dataset, with different numbers of samples available for each severity definition. Right: models were trained on the intersection of samples with all labels available for each definition.

TODO insert figure here

FIG 3 Feature importance.