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

(1)

Clinical value of models

DISCUSSION

TODO

MATERIALS AND METHODS

Sample collection and 16S rRNA gene amplicon sequencing (1)

Defining CDI severity IDSA definition of severe CDI based on lab values. CDC definiton of severe CDI based on disease-related complications (2).

Model training and evaluation (3)

Balanced precision

Code and data availability (4)

ACKNOWLEDGEMENTS

TODO

REFERENCES

1. **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.

2. **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.
3. **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.
4. **Sovacool K, Lesniak N, Schloss P**. 2022. Schtools: Schloss Lab Tools for Reproducible Microbiome Research doi:10.5281/zenodo.6540687.

FIGURES

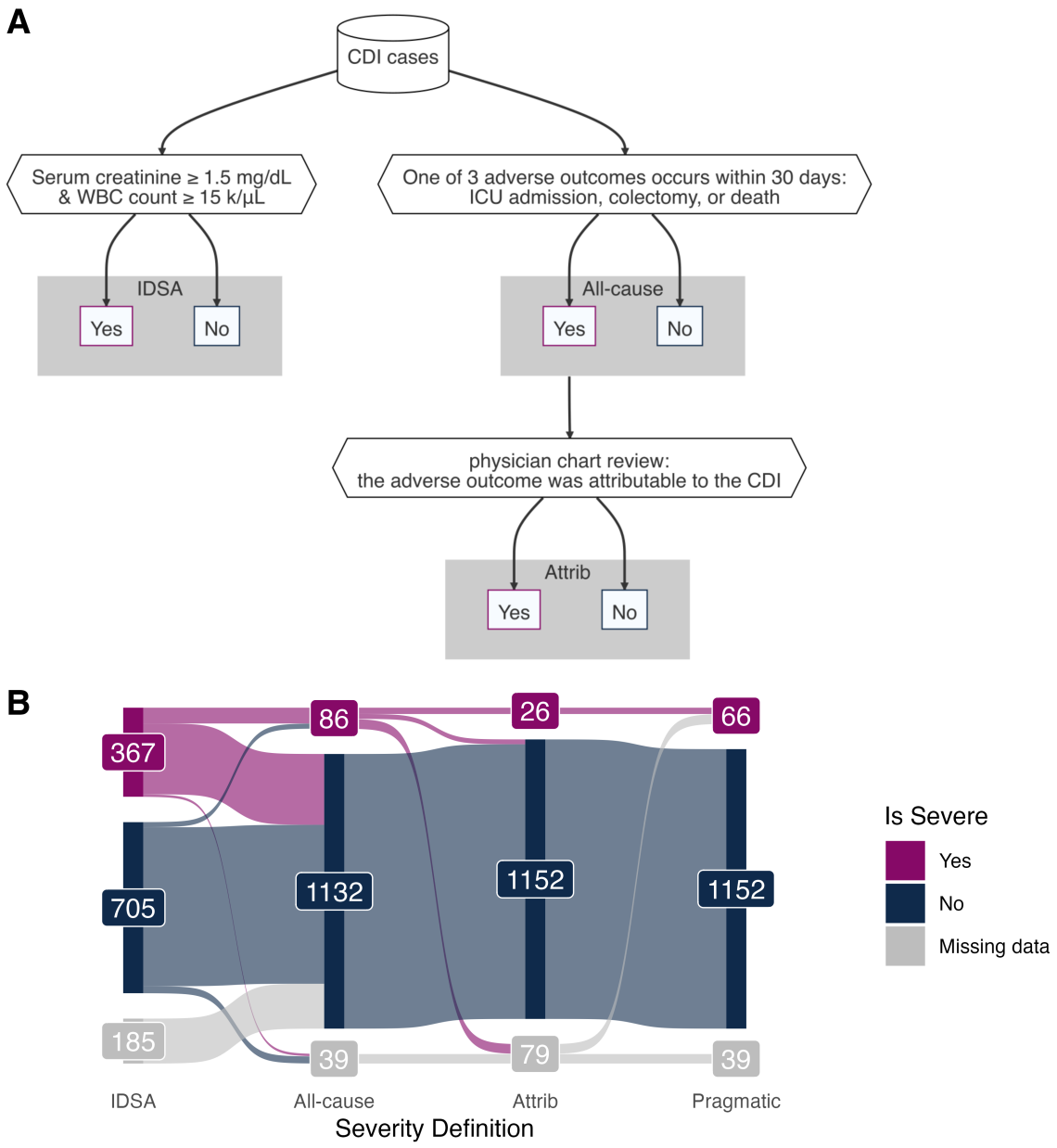

FIG 1 CDI severity definitions. A) Decision flow chart to define CDI cases as severe according to the Infectious Diseases Society of America (IDSA) based on lab values, the occurence of complications due to any cause (All-cause), and the occurence of disease-related complications confirmed as attributable 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.

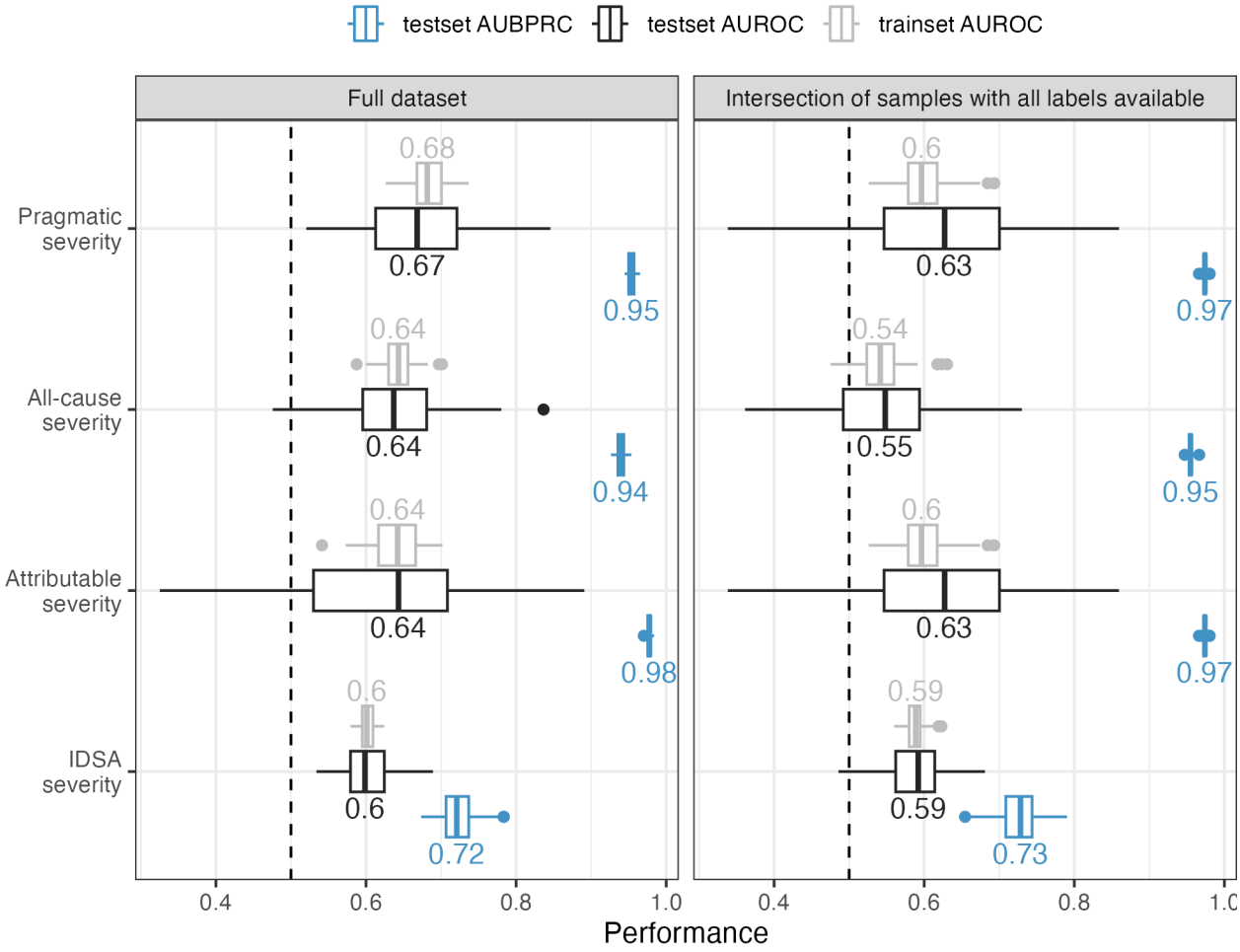

FIG 2 Performance of ML models. Area Under the Receiver-Operator Characteristic Curve (AUROC) for the cross-validation trainsets and testsets, and the Area Under the Balanced Precision-Recall Curve (AUPRC) for the testsets. Left: models were trained on the full dataset, with different numbers of samples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