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 18, 2023
[†]Correspondence: pschloss@umich.edu.

ABSTRACT TODO

IMPORTANCE TODO

KEYWORDS: *C. difficile* infection, supervised machine learning, gut microbiome, amplicion sequencing

```
here() starts at /Users/sovacool/projects/schloss-lab/severe-CDI

-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.4.2     v purrr   1.0.1
v tibble  3.2.1     v dplyr   1.1.2
v tidyr   1.3.0     v stringr 1.5.0
v readr   2.1.3     v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

INTRODUCTION

A few ways to define CDI severity (Figure 1)

The IDSA definition is known to be a poor predictor of adverse outcomes (1), however, it is easy to collect.

RESULTS

Model performance Report median AUROC for training set and test set, and median AUBPRC for test set (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 1277 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 Powersoil 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 (2). 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 (v1.46) using the MiSeq SOP protocol (3, 2). Paired sequencing reads were combined and aligned with the SILVA (v132) reference database (4) and taxonomy was assigned with a modified version of the Ribosomal Database Project reference sequences (v16) (5). Sequences were clustered into *de novo* OTUs with the OptiClust algorithm in mothur (6), resulting in Samples were rarefied to 5,000 sequences per sample, repeated 1,000 times for alpha and beta diversity analysis.

Defining CDI severity We explore four different ways to define CDI cases as severe or not. The IDSA definition of severe CDI is based on lab values collected on the day of diagnosis, with a case being severe if serum creatinine level is greater than or equal to $1.5 mgdL$ and the white blood cell count is greater than or equal to $15 k\mu L$ (**?**). The remaining definitions focus on the occurrence of adverse outcomes, which may be more clinically relevant. The all-cause severity definition defines a case as severe if ICU admission, colectomy, or death occurs within 30 days of CDI diagnosis, regardless of the cause of the adverse event. The attributable severity definition is based on disease-related complications defined by the CDC, where an adverse event of ICU admission, colectomy, or death occurs within 30 days of CDI diagnosis, and the adverse event is determined to be attributable to the CDI by physician chart review (7). Finally, we defined a pragmatic severity definition that makes use of the attributable definition when available and falls back to the all-cause definition when chart review has not been completed, allowing us to use as many samples as we have available while taking physicians' expert opinions into account where possible.

Model training and evaluation Random forest models were used to examine whether OTU data collected on the day of diagnosis could classify CDI cases as severe according to four different definitions of severity. We used the mikropml R package v1.5.0 (8) implemented in a custom version of the mikropml Snakemake workflow (9) for all steps of the machine learning analysis. We have a full dataset which uses all samples available for each severity definition, and an intersection dataset which consists of only the samples that have all four definitions labelled. The intersection dataset is the most fair for comparing model performance across definitions, while the full dataset allows us to use as much data as possible for model training and evaluation. Datasets were preprocessed with the default options in mikropml to remove features with near-zero variance and scale continuous features from -1 to 1. During preprocessing, 9757 to 9760 features were removed due to having near-zero variance, resulting in datasets ranging from 179 to 182 depending on the severity definition. We randomly split the data into an 80% training and 20% test set and repeated this 100 times, followed by training models with 5-fold cross-validation. Model performance was calculated on the test set using the area under the receiver-operator characteristic curve (AUROC) and the area under the balanced precision-recall curve (AUBPRC). Permutation feature importance was then performed to determine which OTUs contributed most to model performance. We reported OTUs with a significant permutation test in at least 80 of the 100 models.

Since the severity labels are imbalanced with different frequencies of severity for each definition, we calculated balanced precision, the precision expected if the labels were balanced. The balanced precision and the area under the balanced precision-recall curve (AUBPRC) were calculated with Equations 1 and 7 from Wu et al. (10).

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 workflow was defined with Snakemake (11) and dependencies were managed with conda environments. Scripts were written in R (12), Python (13), and GNU bash. Additional software and packages used in the creation of this manuscript include cowplot (14), ggtext (15), ggsankey (16), schtools (17), the tidyverse metapackage (18), Quarto, and vegan (19).

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

83 ## REFERENCES

1. **Stevens VW, Shoemaker HE, Jones MM, Jones BE, Nelson RE, Khader K, Samore MH, Rubin MA**. May 2020. Validation of the SHEA/IDSA Severity Criteria to Predict Poor Outcomes among Inpatients and Outpatients with Clostridioides Difficile Infection. Infect Control & Hosp Epidemiol 41 (5):510–516. doi:10.1017/ice.2020.8.
2. **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.
3. **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.
4. **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.
5. **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.
6. **Westcott SL, Schloss PD**. Mar 2017. OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units. mSphere 2 (2):e00073–17. doi:10.1128/mSphereDirect.00073-17.
7. **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.
8. **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.
9. **Sovacool K, Lapp Z, Armour C, Lucas SK, Schloss P**. Jan. 2023. Mikropml Snakemake Workflow doi:10.5281/zenodo.4759351.
10. **Wu Y, Liu H, Li R, Sun S, Weile J, Roth FP**. Oct 2021. Improved Pathogenicity Prediction for Rare Human Missense Variants. The Am J Hum Genet 108 (10):1891–1906. doi:10.1016/j.ajhg.2021.08.012.
11. **Köster J, Rahmann S**. Oct 2012. Snakemake — a Scalable Bioinformatics Workflow Engine. Bioinformatics 28 (19):2520–2522. doi: 10.1093/bioinformatics/bts480.
12. **R Core Team**. 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
13. **Van Rossum G, Drake FL**. 2009. Python 3 Reference Manual | Guide Books .
14. **Wilke CO**. 2020. Cowplot: Streamlined Plot Theme and Plot Annotations for 'Ggplot2'.
15. **Wilke CO**. 2020. Ggtext: Improved Text Rendering Support for 'Ggplot2'.
16. **Sjoberg D**. 2022. Ggsankey: Sankey, Alluvial and Sankey Bump Plots.
17. **Sovacool K, Lesniak N, Schloss P**. 2022. Schtools: Schloss Lab Tools for Reproducible Microbiome Research doi:10.5281/zenodo.6540687.
18. **Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H**. Nov 2019. Welcome to the Tidyverse. J Open Source Softw 4 (43):1686. doi:10.21105/joss.01686.
19. **Oksanen J, Simpson GL, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Solymos P, Stevens MHH, Szoecs E, Wagner H,**

Barbour M, Bedward M, Bolker B, Borcard D, Carvalho G, Chirico M, Caceres MD, Durand S, Evangelista HBA, FitzJohn R, Friendly M, Furneaux B, Hannigan G, Hill MO, Lahti L, McGlinn D, Ouellette MH, Cunha ER, Smith T, Stier A, Braak CJFT, Weedon J. 2023. Vegan: Community Ecology Package.

FIGURES

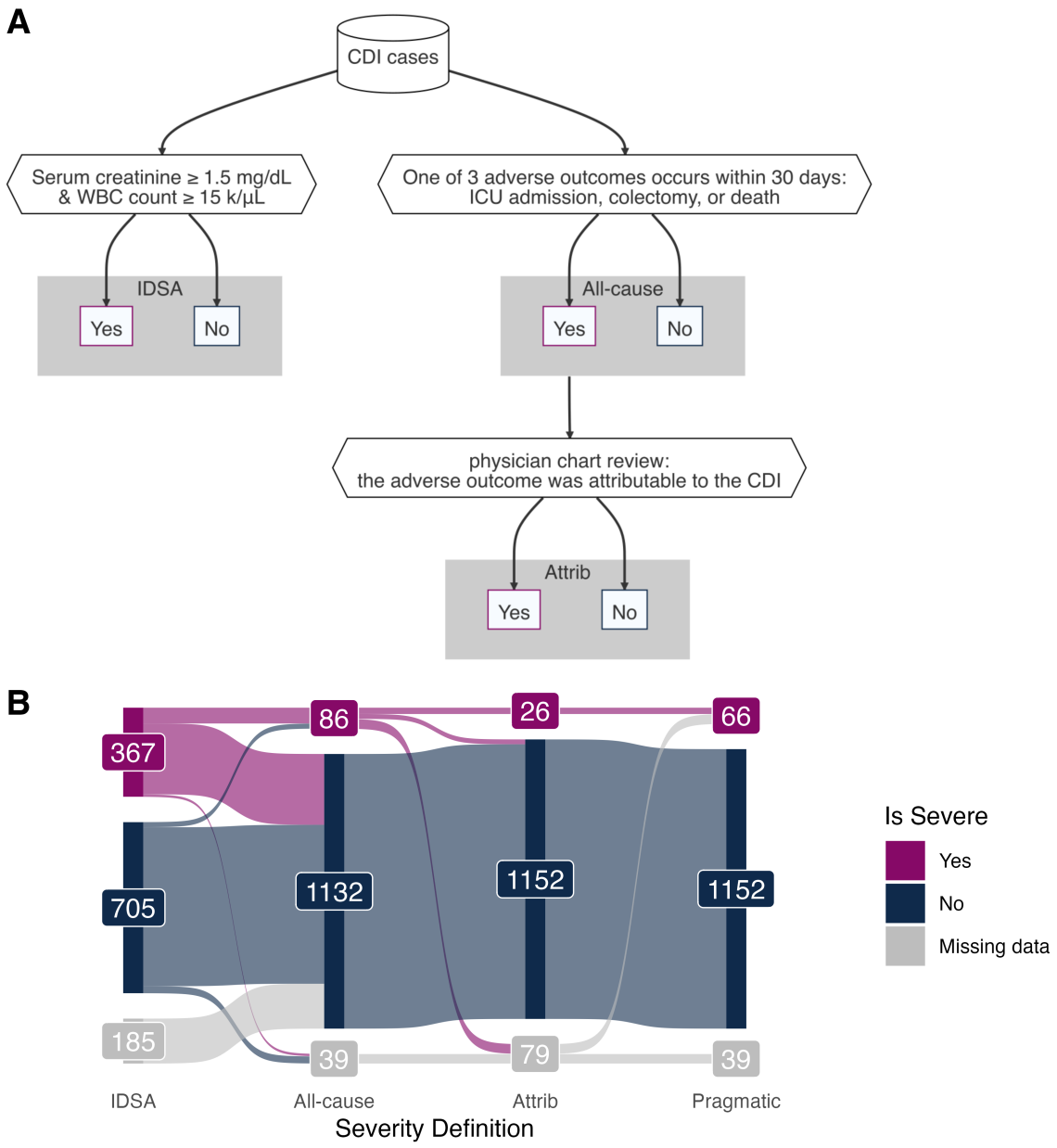

FIG 1 CDI severity definitions. A) Decision flow chart to define CDI cases as severe according to the Infectious Diseases Society of America (IDSA) based on lab values, the occurence of an adverse outcome due to any cause (All-cause), and the occurence of disease-related complications confirmed as attributable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

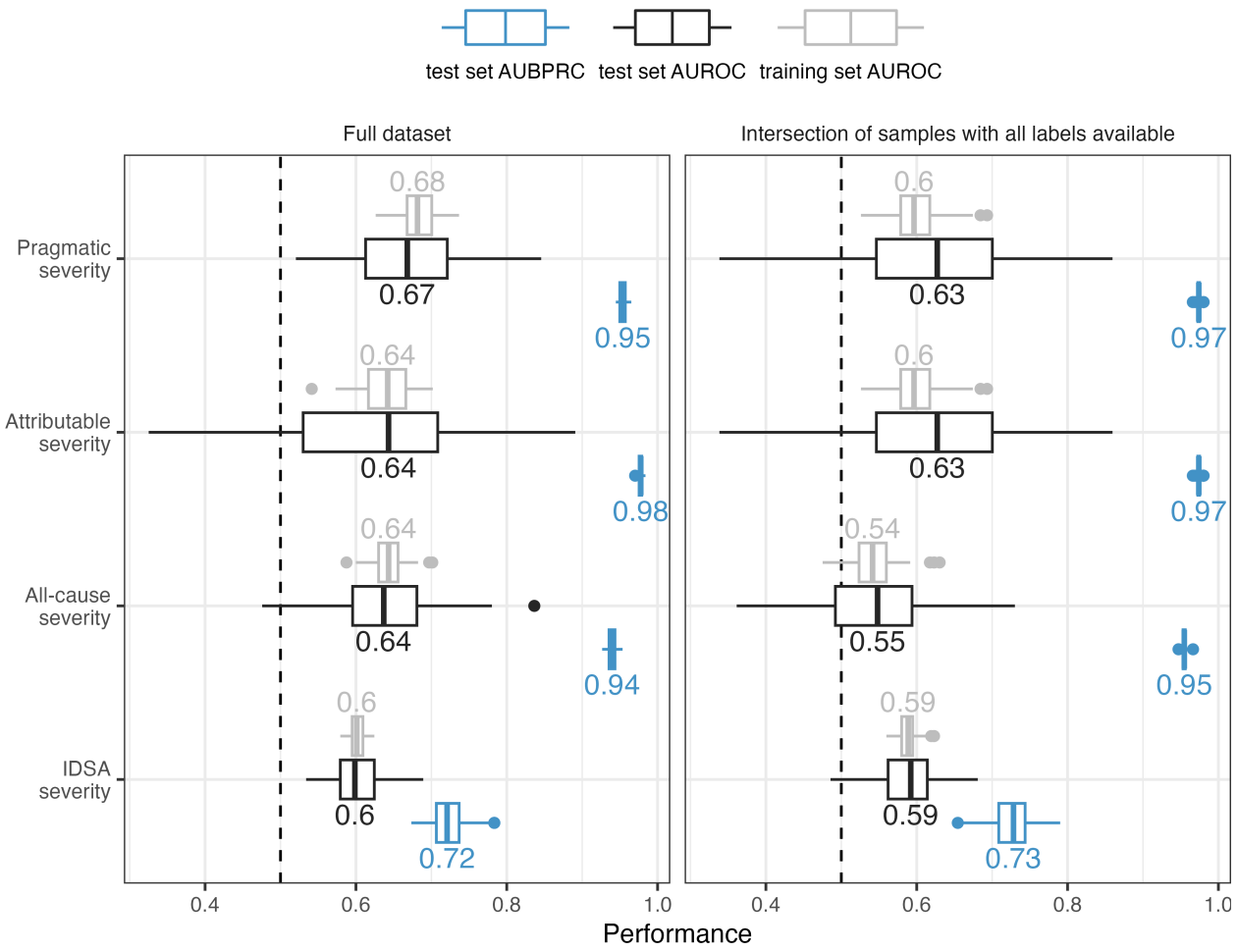

FIG 2 Performance of ML models. Area under the receiver-operator characteristic curve (AUROC) for the test sets and cross-validation folds of the training sets, and the area under the balanced precision-recall curve (AUBPRC) for the test sets. Left: models were trained on the full dataset, with different numbers of samples available for each severity definition. Right: models were trained on the intersection of samples with all labels available for each definition. Note that Attributable and Pragmatic severity are exactly the same for the intersection set.

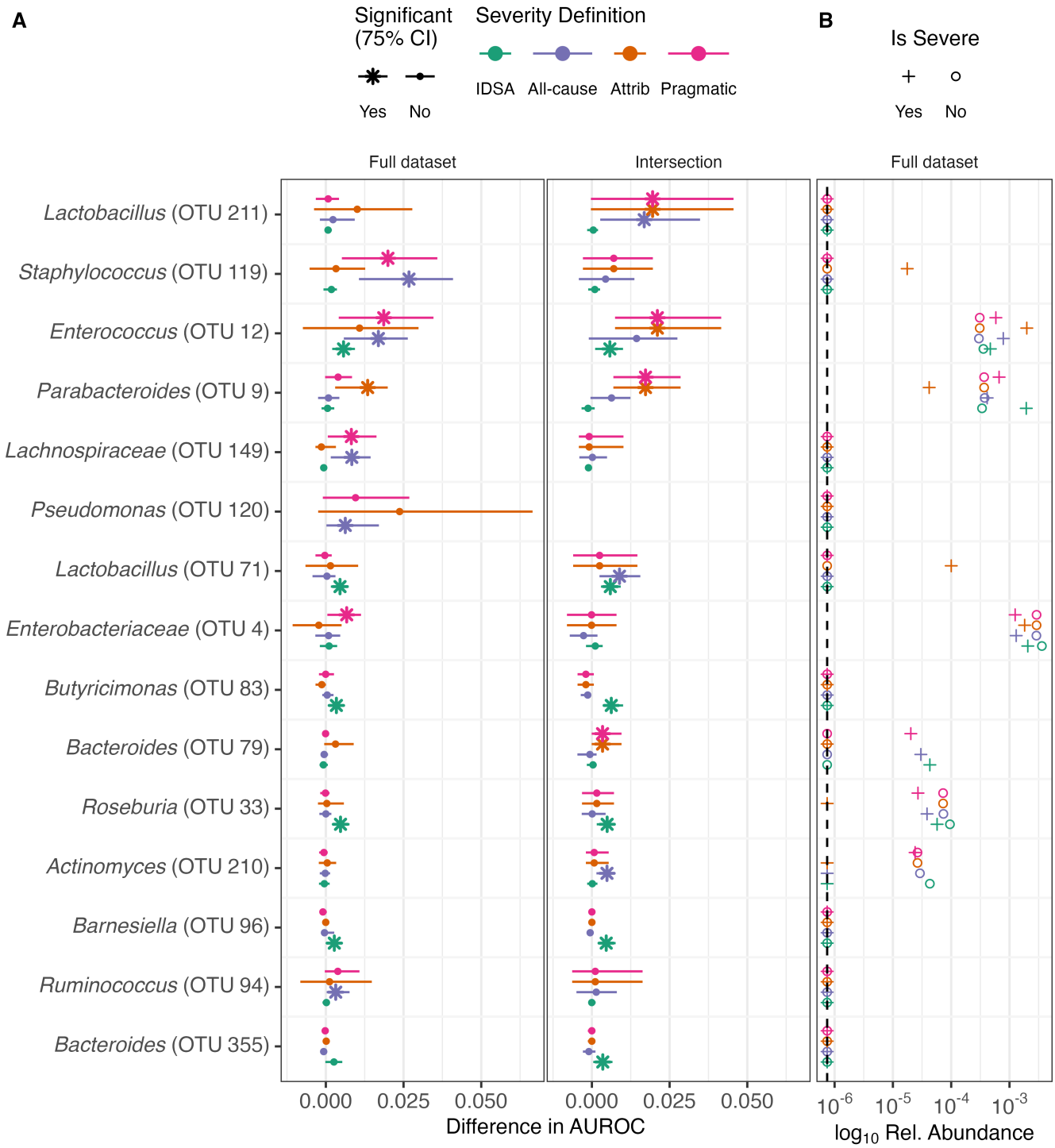

FIG 3 Feature importance. A) Feature importance via permutation test. For each OTU, the order of samples was randomized in the test set 100 times and the performance was re-calculated to estimate the permutation performance. An OTU was considered important if the performance decreased when the OTU was permuted in at least 75% of the models. OTUs with a greater difference in AUROC (actual performance minus permutation performance) are more important. Left: models were trained on the full datasets, 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. Note that Attributable and Pragmatic severity are exactly the same for the intersection set. OTU 120 (*Pseudomonas*) is not shown for the full data set with IDSA severity on the full dataset because it was removed during pre-processing due to having near-zero variance. **B)** Log_{10} transformed median relative abundances of the most important OTUs on the full datasets, grouped by severity (shape).