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 June 3, 2023
[†]Correspondence: pschloss@umich.edu.

ABSTRACT *Clostridioides difficile* (*C. difficile*) infection (CDI) can lead to adverse outcomes including ICU admission, colectomy, and death. The composition of the gut microbiome plays an important role in determining colonization resistance and clearance upon exposure to *C. difficile*. We investigated whether machine learning (ML) models trained on gut microbiome compositions could predict which CDI cases led to severe outcomes. We collected 1,277 stool samples from CDI patients on the day of diagnosis and characterized the gut microbiome via 16S rRNA gene amplicon sequencing. We processed and clustered the sequences into *de novo* Operational Taxonomic Units (OTUs), which we used to train ML models for predicting whether a severe outcome occurred according to three different severity definitions. We learned that our OTU-based Random Forest models perform best when predicting attributable severity, defined as an adverse outcome within 30 days of CDI diagnosis confirmed as attributable to CDI following chart review by a clinician. This finding suggests that chart review is valuable to verify the cause of severe outcomes. We used permutation importance to identify which OTUs contributed most to model performance and found that *Enterococcus*, which has previously been identified as positively correlated with *C. difficile* colonization, was the largest contributor. In all, our results show that ML models can modestly identify patients at risk for CDI, which will allow healthcare providers to personalize treatment options to protect patients likely at risk and ultimately improve CDI outcomes.

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

INTRODUCTION

prevalence of cdi. prevalence of severe cdi outcomes. antibiotics typical risk factor for cdi, but nonantibiotic medications can increase susceptibilty too (1) and non-nosocomial CDI on the rise

Numerous studies indicate that the gut microbiome may play a role in *C. diff* colonization, infection, and clearance. Contribution of the gut microbiome.

prediction models based on EHR for whether infection occurs in the first place already in use (2). so how about predicting severity of infections to guide treatment. also models on EHR to predict adverse outcomes (Li, Rao). serum-based biomarker model (Dieterle mbio 2020). OTUs vs EHRs to predict severity. CDI severity prediction models could be deployed to screen patients at risk and guide clinicians to consider prescribing a different course of treatment. When paired with treatment options that may reduce risk of severity, deploying prediction models can guide clinician decision-making to improve patient outcomes while minimizing unnecessary harms.

A few ways to define CDI severity (Figure 1).

The IDSA definition is known to be a poor predictor of adverse outcomes (3), however, it is easy to collect.

new dataset.

Two goals: investigate whether we can predict CDI severity based on OTU data to inform how the gut microbiome may modulate severity (ML-based science: good performance implies something about underlying biology), and determine whether there is potential clinical value in OTU-based models.

RESULTS

Model performance We first set out to train the best models possible for each severity definition. Not all samples have labels available for all four severity definitions due to missing data for some patient lab values and incomplete chart review (Figure 1 B), thus each severity definition has a different number of samples when using as many samples as possible (**?@tbl-counts**). We refer to these as the full datasets. Random forest models were trained on 100 splits of the datasets into training and test sets, and performance was evaluated on the held-out test set using the area under the receiver-operator characteristic curve (AUROC). Since the severity classes are highly imbalanced with different proportions of severe samples between definitions, we also calculated the balanced precision and the area under the balanced precision-recall curve (AUBPRC) as first proposed by Wu et al. (4) to describe the precision that would be expected if the outcome classes were balanced.

After training on the full datasets, the performance as measured by the AUROCs of the training set cross-validation folds were similar to those of the held-out test sets, indicating that the models are neither overfit nor underfit (Figure 2 A). As measured by AUROC on the held-out test sets, models predicting pragmatic severity performed best with a median AUROC of 0.6864548, and this was significantly different from that of the other definitions on the full datasets ($P < 0.05$). Models predicting IDSA, all-cause, and attributable severity performed similarly with median test set AUROCs of 0.6083018, 0.6320927, and 0.6143478 respectively. The test set AUROCs were not significantly different ($P > 0.05$) for attributable and IDSA nor for attributable and all-cause, but the IDSA and all-cause AUROCs were significantly different from each other ($P < 0.05$). We plotted the receiver-operator characteristic curve and found that the pragmatic severity models outperformed the others at all specificity values (Figure 2 B). For comparison, a prior study trained a logistic regression model on whole Electronic Health Record data extracted on the day of CDI diagnosis to predict attributable severity, yielding an AUROC of 0.69 (5). While our OTU-based attributable severity model did not meet this performance, the OTU-based pragmatic severity model performed just as well as the EHR-based model in terms of AUROC.

The test set median AUBPRCs from the full datasets followed a similar pattern as the test set AUROCs with 0.595164 for IDSA severity, 0.6687631 for all-cause severity, 0.6584436 for attributable severity, and 0.7466801 for pragmatic severity. The AUBPRCs were significantly different from each other ($P < 0.05$) for each pair of severity definitions except for attributable vs all-cause. We plotted the balanced precision-recall curve and found that the IDSA definition outperformed all other models at very low recall values, but the others outperform IDSA at all other points of the curve (Figure 2 C). The 95% confidence intervals overlapped the baseline AUROC and AUBPRC for the attributable severity models, while all others did not overlap the baseline.

While it is advantageous to use as much data as available to train the best models possible, comparing performances of models trained on different subsets of the data is not entirely fair. To enable fair comparisons of the model performances across different severity definitions, we also selected the intersection of samples (n=993) that had labels for all four severity definitions and repeated the model training and evaluation process on this intersection dataset. The attributable definition is exactly the same as the pragmatic definition for the intersection dataset, as we defined pragmatic severity to use the attributable label when available. The performance results on the intersection dataset are shown in the right facets of each panel of Figure 2.

As with the full datasets, the AUROCs of the training sets and test sets were similar within each severity definition. The median test set AUROCs were 0.6026316 for IDSA severity, 0.5489418 for all-cause severity, 0.5865285 and for attributable

severity. The AUROCs on the intersection dataset were significantly different for all-cause vs attributable and all-cause vs IDSA severity ($P < 0.05$), but not for IDSA vs attributable severity ($P > 0.05$). The median test set AUBPRCs were 0.593301 for IDSA severity, 0.5523027 for all-cause severity, 0.5820498 and for attributable severity. Just as with the AUROCs, the AUBPRCs were significantly different for all-cause vs attributable and all-cause vs IDSA severity ($P < 0.05$), but not for IDSA vs attributable severity ($P > 0.05$). For all severity definitions, performance dropped between the full dataset and the intersection dataset since fewer samples are available, but this effect is least dramatic for IDSA severity as the full and intersection datasets are more similar for this definition. The 95% confidence interval overlaps with the baseline for both AUROC and AUBPRC for all definitions on the intersection dataset except for IDSA severity.

Feature importance We performed permutation feature importance to determine which OTUs contributed the most to model performance. An OTU was considered important if performance decreased when it was permuted in at least 75% of the train/test splits, with greater differences in AUROC meaning greater importance. We plotted mean decrease in AUROC alongside $\log_{10}$-transformed mean relative abundances for the top OTUs (Figure 3). There is not always a clear pattern of increased or decreased relative abundance for important OTUs in severe CDI cases, but all of the top 5 OTUs had an increased mean relative abundance in severe cases relative to not severe cases. *Enterococcus* was the most important OTU, being significantly important for all models except for attributable severity on the full dataset. *Staphylococcus* was important for the pragmatic and all-cause definitions on the full datasets, but not for models trained on the intersection dataset. *Lactobacillus* was important only for the all-cause definition on the intersection dataset. All remaining OTUs had differences in AUROC < 0.02 and were only significantly important in one or two of the models at most. TODO summarize which OTUs enriched in severe vs not severe. TODO concluding sentence.

Estimating clinical value Even if a model performs well, it may not be useful in a clinical setting unless it can guide clinicians to choose between treatment options. At this time, we are not aware of any direct evidence that a particular treatment reduces the risk of severe CDI outcomes. However, with some assumptions we offer a proof-of-concept analysis of the potential clinical value of OTU-based severity prediction models when paired with treatments that may reduce severity. When considering the suitability of a model for deployment in clinical settings, the number needed to screen (NNS) is a highly relevant metric representing how many patients must be predicted as severe by the model to identify one true positive. Similarly, the number needed to treat (NNT) is the number of true positive patients that must be treated by an intervention in order for one patient to benefit from the treatment. Multiplying NNS by NNT yields the number needed to benefit (NNB): the number of patients predicted to have a severe outcome who then benefit from

the treatment (6). Thus the NNB pairs model performance with treatment effectiveness to estimate the benefit of using predictive models in clinical practice, and are useful for comparing models and performing cost-benefit analyses.

Current clinical guidelines specify vancomycin and fidaxomicin as the standard antibiotics to treat CDI, with a preference for fidaxomicin due to its higher rate of sustained resolution of CDI and lower rate of recurrence (7). The NNTs of fidaxomicin for sustained resolution and prevention of recurrence are each estimated to be 10 (8, 9). However, fidaxomicin is considerably more expensive than vancomycin. If fidaxomicin were shown to reduce the risk of severe CDI outcomes, it could be preferentially prescribed to patients predicted to be at risk, while prescribing vancomycin to low-risk patients. If we assume that the superior efficacy of fidaxomicin for sustained resolution and reduced recurrence also translates to reducing the risk of severe outcomes, we can pair the NNT of fidaxomicin with the NNS of OTU-based prediction models to estimate the NNB.

To calculate a clinically-relevant NNS for these models, we computed the confusion matrix at the 95th percentile of risk for each prediction model (**?@tbl-risk**). We excluded the IDSA severity models as the IDSA severity scores were calculated on the day of diagnosis, thus they are classification rather than prediction problems. Furthermore, IDSA severity scores do not correlate well with disease-related adverse events which are a more salient outcome to prevent. Among the models predicting severe outcomes, those trained on the full datasets performed best with an NNS of 4 for the all-cause definition, 6 for the attributable definition, and 3 for the pragmatic definition. Multiplying the NNS of the OTU-based models by the estimated NNT of 10 for fidaxomicin yields NNB values of 40 for all-cause severity, 60 for attributable severity, and 30 for pragmatic severity. Thus, in a hypothetical scenario where these assumptions about fidaxomicin hold true, between 30 and 60 patients would need to be predicted to experience a severe outcome and be treated with fidaxomicin in order for one patient to benefit. As the NNS values were computed at the 95th percentile of risk (where 5% of patients screened are predicted to experience severity), these NNB values mean that 600 to 1200 total CDI patients would need to be screened by an OTU-based prediction model in order for one patient to benefit. For comparison, prior studies predicted CDI-attributable severity using whole Electronic Health Record data extracted two days after diagnosis and from a smaller set of manually curated variables, achieving precision values of 0.417 (NNS = 2.3980815) for the EHR model and 0.167 (NNS = 5.988024) for the curated model at the 95th percentile of risk (5, 10). Pairing the prior EHR-based model with fidaxomicin would yield an NNB of 24 with 480 total CDI patients screened for one patient to benefit, although the EHR was extracted two days after diagnosis while OTUs in this study are from stool samples collected on the day of diagnosis. These estimates represent a proof-of-concept demonstration of the potential value and trade-offs of deploying severity

prediction models trained on microbial factors versus EHRs to guide clinicians' treatment decisions.

DISCUSSION

Performance

Discuss important OTUs. which ones concord with literature, which ones may be new. For many of the top OTUs, there is wide variance in importance, perhaps due to the imbalanced nature of the severity outcomes. Enrichment of *Enterococcus* and *Lactobacillus* in *C. difficile* infection and severity has been well-documented in prior studies, thus its importance and increase in abundance for severe cases is not surprising (11, 12, 13, 14). Abundance data are sparse, likely due to these patients being on antibiotics. Really showcases importance of having as many samples as possible when data are sparse and the outcome is low prevalence. we do not know which antibiotics were prescribed to treat these CDI cases, nor which antibiotics patients may have taken prior to the CDI. differences in microbiota between patients may be due to different abx… different antibiotics have been shown to create different forms of dysbiotic microbiota (13) and differential cdi clearance (15). vanc-resistant enterococcus new nosocomial alliance (16).

Compare to EHR-based models. in practice, EHR-based models are easier and less costly to deploy than an OTU-based model would be. further work needed to characterize the costs and benefits.

Models to guide treatment options. In the case of low-risk and non-invasive treatments such as choosing between oral antibiotics, a higher number of false positives may be tolerable as long as treatment cost is not unbearably high. However, for highly invasive and irreversibly treatments such as colectomy, false positives must be minimized. Cite studies saying fidaxomicin is cost-effective relative to vancomycin - mentioned by Johnson et al. (7), e.g. Jiang et al. (17).

It's not enough for models to perform well to justify deploying them in a clinical setting; benefit over current practices must be shown. do no harm (18). Estimating the NNB contextualizes model performance within clinical reality. Amplicon sequencing is not typically performed for CDI patients, but if there is clinical value to be gained by implementing OTU-based models, routinely sequencing and profiling the microbial communities of CDI patients could be justified. resistance to vancomycin is increasing in staph, cdi, and enterococci – even more important to find alternate treatments.

Models predicting the pragmatic definition yielded the best NNS. While the attributable definition had a worse NNS for our OTU-based models, it did not perform worse than the prior curated model, and it is the most clinically relevant as physician chart review increases confidence that positively-labelled severe outcomes are due to the CDI rather than other causes.

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 re-suspended in DNA genotek stabilization buffer and then stored in the -80°C freezer.

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 (19). 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 (SequalPrep normalization 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 (20, 19). Paired sequencing reads were combined and aligned with the SILVA (v132) reference database (21) and taxonomy was assigned with a modified version of the Ribosomal Database Project reference sequences (v16) (22). Sequences were clustered into *de novo* OTUs with the OptiClust algorithm in mothur (23), resulting in 9939 OTUs. 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.

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 $1.5 mg/dL$ and the white blood cell count is greater than or equal to $15 k/\mu L$ (24). 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 (25). 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 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 (26) implemented in a custom version of the mikropml Snakemake workflow (27) for all steps of the machine learning analysis. We have full datasets which use 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 pre-processed with the default options in mikropml to remove features with near-zero variance and scale continuous features from -1 to 1. During pre-processing, 9757 to 9760 features were removed due to having near-zero variance, resulting in datasets having 179 to 182 features depending on the severity definition. No features had missing values and no features were perfectly correlated. 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 evaluation Model performance was calculated on the held-out test sets 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 75 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. (4).

Number needed to benefit For the severity prediction models (which excludes the IDSA definition), we set out to estimate the potential benefit of deploying models in clincal settings. We determined the decision threshold at the 95th percentile of risk for each model, which corresponds to 5% of cases being predicted by the model to experience a severe outcome. At this threshold we computed the number needed to screen (NNS), which is the reciprocal of precision and represents the number of cases that must be predicted as severe to identify one true positive (28). The number needed

to treat (NNT) is the number of true positive patients that must be treated by an intervention in order for one patient to benefit, and is calculated from the reciprocal of absolute risk in randomized controlled trials (29). Multiplying the NNS of a model by the NNT of a treatment yields the number needed to benefit (NNB) - the number of patients that must be predicted to have a severe outcome and undergo a treatment to benefit from it (6). NNB encapsulates the benefit of pairing a predictive model with a treatment in a clinical setting, with lower NNB numbers being better.

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 (30) and dependencies were managed with conda environments. Scripts were written in R (31), Python (32), and GNU bash. Additional software and packages used in the creation of this manuscript include cowplot (33), ggtext (34), ggsankey (35), schtools (36), the tidyverse metapackage (37), Quarto, and vegan (38).

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

223 ## REFERENCES

1. **Tomkovich S, Taylor A, King J, Colovas J, Bishop L, McBride K, Royzenblat S, Lesniak NA, Bergin IL, Schloss PD**. 2021. An Osmotic Laxative Renders Mice Susceptible to Prolonged Clostridioides Difficile Colonization and Hinders Clearance. mSphere 0 (0):e00629–21. doi: 10.1128/mSphere.00629-21.
2. **Ötleş E, Balczewski EA, Keidan M, Oh J, Patel A, Young VB, Rao K, Wiens J**. Apr 2023. Clostridioides Difficile Infection Surveillance in Intensive Care Units and Oncology Wards Using Machine Learning. Infect Control & Hosp Epidemiol p 1–6. doi:10.1017/ice.2023.54.
3. **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.
4. **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.
5. **Li BY, Oh J, Young VB, Rao K, Wiens J**. May 2019. Using Machine Learning and the Electronic Health Record to Predict Complicated Clostridium Difficile Infection. Open Forum Infect Dis 6 (5):ofz186. doi:10.1093/ofid/ofz186.
6. **Liu VX, Bates DW, Wiens J, Shah NH**. Dec 2019. The Number Needed to Benefit: Estimating the Value of Predictive Analytics in Healthcare. J Am Med Informatics Assoc 26 (12):1655–1659. doi:10.1093/jamia/ocz088.
7. **Johnson S, Lavergne V, Skinner AM, Gonzales-Luna AJ, Garey KW, Kelly CP, Wilcox MH**. Sep 2021. Clinical Practice Guideline by the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology of America (SHEA): 2021 Focused Update Guidelines on Management of *Clostridioides Difficile* Infection in Adults. Clin Infect Dis 73 (5):e1029–e1044. doi:10.1093/cid/ciab549.
8. **Long B, Gottlieb M**. 2022. Oral Fidaxomicin versus Vancomycin for Clostridioides Difficile Infection. Acad Emerg Med 29 (12):1506–1507. doi: 10.1111/acem.14600.
9. **Tashiro S, Mihara T, Sasaki M, Shimamura C, Shimamura R, Suzuki S, Yoshikawa M, Hasegawa T, Enoki Y, Taguchi K, Matsumoto K, Ohge H, Suzuki H, Nakamura A, Mori N, Morinaga Y, Yamagishi Y, Yoshizawa S, Yanagihara K, Mikamo H, Kunishima H**. Nov 2022. Oral Fidaxomicin versus Vancomycin for the Treatment of Clostridioides Difficile Infection: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. J Infect Chemother 28 (11):1536–1545. doi:10.1016/j.jiac.2022.08.008.
10. **Rao K, Micic D, Natarajan M, Winters S, Kiel MJ, Walk ST, Santhosh K, Mogle JA, Galecki AT, LeBar W, Higgins PDR, Young VB, Aronoff DM**. Jul 2015. Clostridium Difficile Ribotype 027: Relationship to Age, Detectability of Toxins A or B in Stool With Rapid Testing, Severe Infection, and Mortality. Clin Infect Dis 61 (2):233–241. doi:10.1093/cid/civ254.
11. **Schubert AM, Rogers MAM, Ring C, Mogle J, Petrosino JP, Young VB, Aronoff DM, Schloss PD**. Jul 2014. Microbiome Data Distinguish Patients with Clostridium Difficile Infection and Non-C. Difficile-Associated Diarrhea from Healthy Controls. mBio 5 (3). doi:10.1128/mBio.01021-14.
12. **Antharam VC, Li EC, Ishmael A, Sharma A, Mai V, Rand KH, Wang GP**. Sep 2013. Intestinal Dysbiosis and Depletion of Butyrogenic Bacteria in Clostridium Difficile Infection and Nosocomial Diarrhea. J Clin Microbiol 51 (9):2884–2892. doi:10.1128/JCM.00845-13.
13. **Berkell M, Mysara M, Xavier BB, van Werkhoven CH, Monsieurs P, Lammens C, Ducher A, Vehreschild MJGT, Goossens H, de Gunzburg J, Bonten MJM, Malhotra-Kumar S**. Apr 2021. Microbiota-Based Markers Predictive of Development of Clostridioides Difficile Infection. Nat Commun 12 (1):2241. doi:10.1038/s41467-021-22302-0.

14. **Lesniak NA, Schubert AM, Flynn KJ, Leslie JL, Sinani H, Bergin IL, Young VB, Schloss PD**. Jul 2022. The Gut Bacterial Community Potentiates Clostridioides Difficile Infection Severity. mBio 13 (4):e01183–22. doi:10.1128/mbio.01183-22.
15. **Tomkovich S, Stough JMA, Bishop L, Schloss PD**. Oct 2020. The Initial Gut Microbiota and Response to Antibiotic Perturbation Influence Clostridioides Difficile Clearance in Mice. mSphere 5 (5). doi:10.1128/mSphere.00869-20.
16. **Poduval RD, Kamath RP, Corpuz M, Norkus EP, Pitchumoni CS**. Dec 2000. Clostridium Difficileand Vancomycin-Resistant Enterococcus: The New Nosocomial Alliance. Off journal Am Coll Gastroenterol | ACG 95 (12):3513. doi:10.1111/j.1572-0241.2000.03291.x.
17. **Jiang Y, Sarpong EM, Sears P, Obi EN**. Feb 2022. Budget Impact Analysis of Fidaxomicin Versus Vancomycin for the Treatment of Clostridioides Difficile Infection in the United States. Infect Dis Ther 11 (1):111–126. doi:10.1007/s40121-021-00480-0.
18. **Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, Jung K, Heller K, Kale D, Saeed M, Ossorio PN, Thadaney-Israni S, Goldenberg A**. Sep 2019. Do No Harm: A Roadmap for Responsible Machine Learning for Health Care. Nat Med 25 (9):1337–1340. doi: 10.1038/s41591-019-0548-6.
19. **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.
20. **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.
21. **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.
22. **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.
23. **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.
24. **McDonald LC, Gerding DN, Johnson S, Bakken JS, Carroll KC, Coffin SE, Dubberke ER, Garey KW, Gould CV, Kelly C, Loo V, Shaklee Sammons J, Sandora TJ, Wilcox MH**. Mar 2018. Clinical Practice Guidelines for Clostridium Difficile Infection in Adults and Children: 2017 Update by the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology of America (SHEA). Clin Infect Dis 66 (7):e1–e48. doi:10.1093/cid/cix1085.
25. **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.
26. **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.
27. **Sovacool K, Lapp Z, Armour C, Lucas SK, Schloss P**. Jan. 2023. Mikropml Snakemake Workflow doi:10.5281/zenodo.4759351.
28. **Rembold CM**. Aug 1998. Number Needed to Screen: Development of a Statistic for Disease Screening. BMJ 317 (7154):307–312. doi: 10.1136/bmj.317.7154.307.
29. **Laupacis A, Sackett DL, Roberts RS**. 1988. An Assessment of Clinically Useful Measures of the Consequences of Treatment. New Engl J Med

318 (26):1728–1733. doi:10.1056/NEJM198806303182605.

30. **Köster J, Rahmann S**. Oct 2012. Snakemake — a Scalable Bioinformatics Workflow Engine. Bioinformatics 28 (19):2520–2522. doi: 10.1093/bioinformatics/bts480.
31. **R Core Team**. 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
32. **Van Rossum G, Drake FL**. 2009. Python 3 Reference Manual | Guide Books .
33. **Wilke CO**. 2020. Cowplot: Streamlined Plot Theme and Plot Annotations for 'Ggplot2'.
34. **Wilke CO**. 2020. Ggtext: Improved Text Rendering Support for 'Ggplot2'.
35. **Sjoberg D**. 2022. Ggsankey: Sankey, Alluvial and Sankey Bump Plots.
36. **Sovacool K, Lesniak N, Lucas S, Armour C, Schloss P**. 2022. Schtools: Schloss Lab Tools for Reproducible Microbiome Research doi: 10.5281/zenodo.6540686.
37. **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.
38. **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.

TABLES

FIGURES

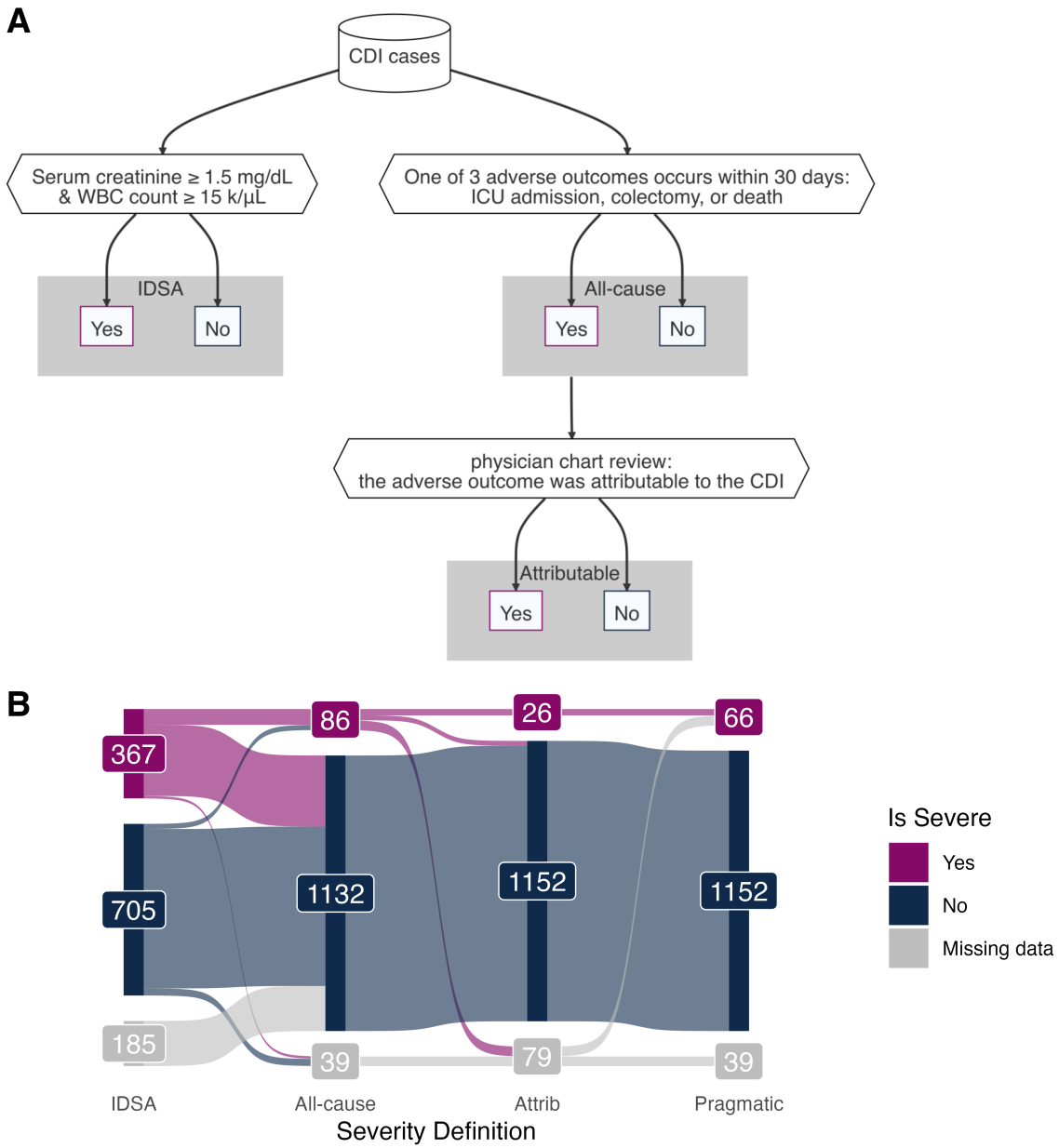

FIG 1 CDI severity definitions. A) Decision flow chart to define CDI cases as severe according to the Infectious Diseases Society of America (IDSA) based on lab values, the occurrence of an adverse outcome due to any cause (All-cause), and the occurrence of disease-related complications confirmed as attributable to CDI with chart review (Attributable). **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. See **?@tbl-counts** for sample counts and proportions of severe cases across severity definitions.

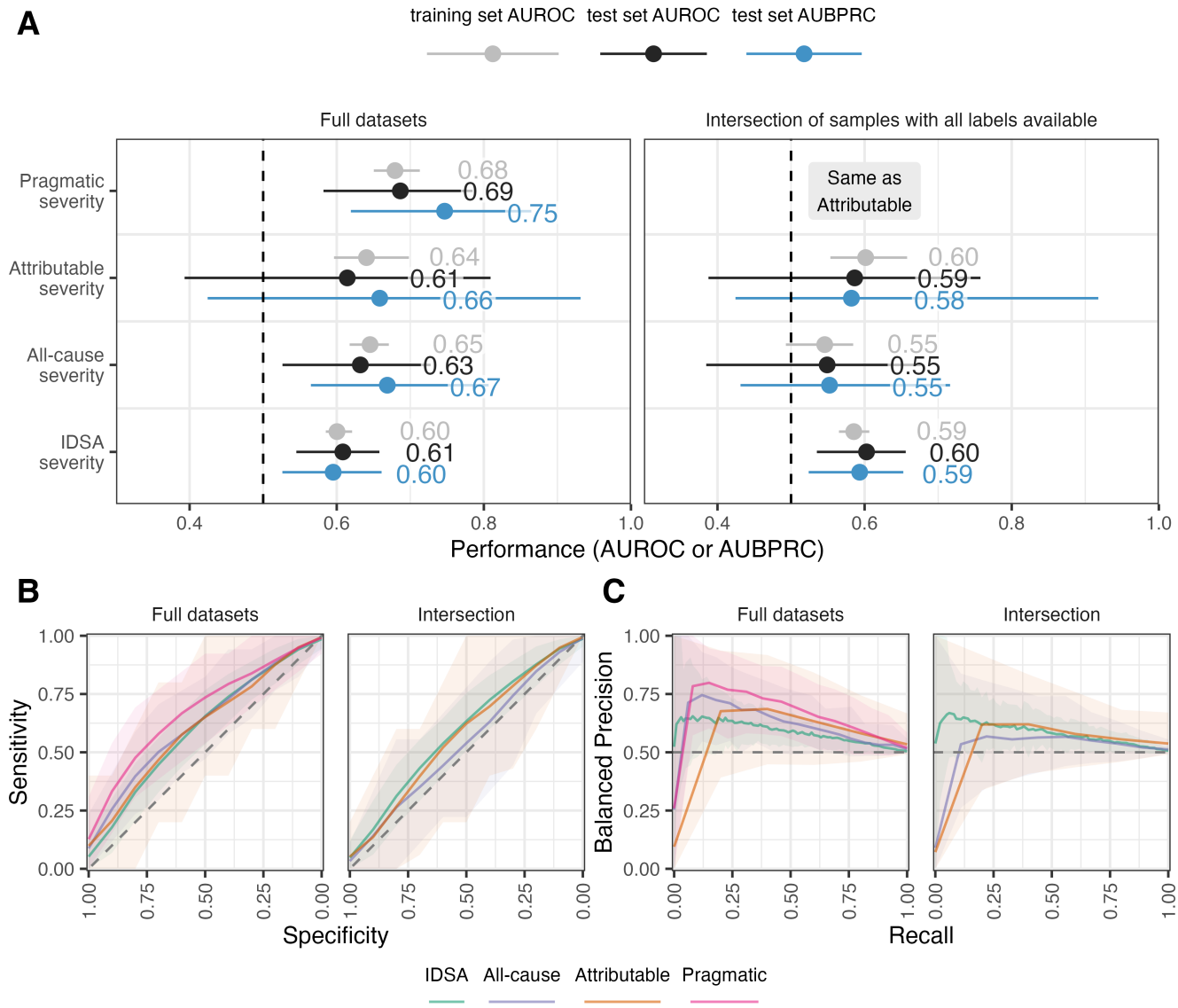

FIG 2 Performance of ML models. In the left facets, models were trained on the full datasets, with different numbers of samples available for each severity definition. In the right facets, models were trained on the same dataset consisting of the intersection of samples with labels available for all definitions. Note that the intersection dataset has exactly the same labels for attributable and pragmatic severity, thus these have identical performance. **A)** 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. Each point is annotated with the median performance across 100 train/test splits with tails as the 95% CI. **B)** Receiver-operator characteristic curves for the test sets. Mean specificity is reported at each sensitivity value, with ribbons as the 95% CI. **C)** Balanced precision-recall curves for the test sets. Mean balanced precision is reported at each recall value, with ribbons as the 95% CI. Original unbalanced precision-recall curves are shown in Figure S1.

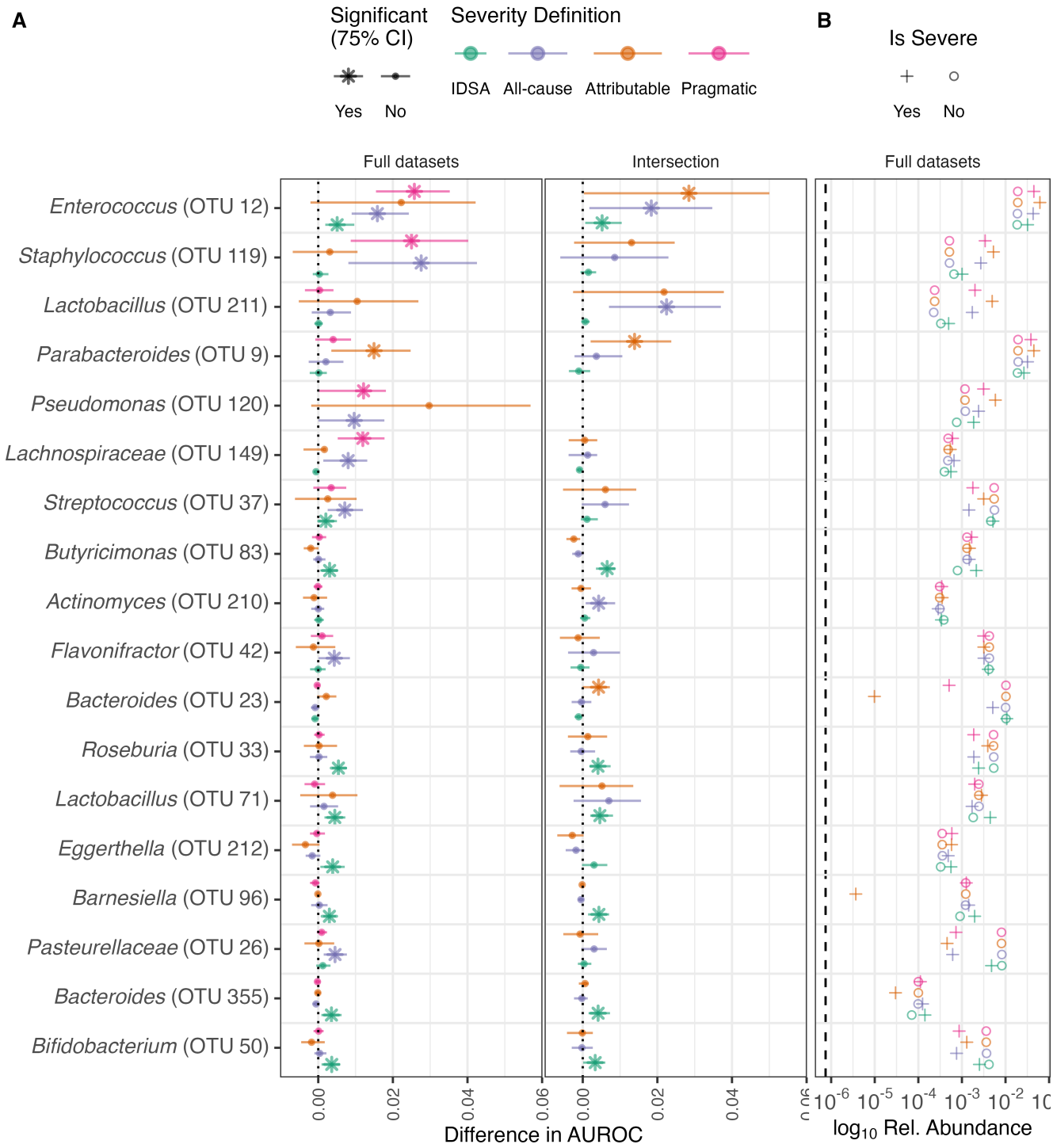

FIG 3 Feature importance. A) Feature importance via permutation test. For each OTU, the order of samples was randomized in the test set 100 times and the performance was re-calculated to estimate the permutation performance. An OTU was considered important if the performance decreased when the OTU was permuted in at least 75% of the models. Mean difference in AUROC and the 75% confidence interval (CI) is reported for each OTU, with starred OTUs being significant for the 75% CI. 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 dataset. *Pseudomonas* (OTU 120) is not shown for IDSA severity in the full datasets nor in the intersection dataset because it was removed during pre-processing due to having near-zero variance. **B)** Log_{10}-transformed mean relative abundances of the most important OTUs on the full datasets, grouped by severity (shape). The vertical dashed line is the limit of detection.