Data Accessibility Paper

Joanna Colovas   Adena Collens   Patrick D. Schloss

Nov 18, 2025

- Abstract

- Importance

    - Incentivize authors to publish/make available their original data

    - Publishing data helps get more use out of research

    - Helps eliminate file drawer effect as it shows negative data

- Keywords

    - Data accessibility

    - Data reproducibility

Introduction

Scientific Data as a Public Good

The United States Government spent over two hundred million dollars (USD) in 2024 on research expenditures ([1]). The result of all of these investments were data, paid for by taxpayers. Therefore, data are a public good. Public goods, for example public libraries, are able to be used by anyone without barrier to entry, and without diminishing the use of others. Data are best used for the benefit of those who provided the funds for it. Generated data are useful in many ways, not only by the original generators to analyze and answer study questions, but further used to answer additional questions on the same study system, to replicate the original analyses, and for meta-studies, by

combining multiple similar datasets. A key tenet of the scientific method is this ability to replicate scientific findings to ensure that they are not due to error. Scientific findings can be replicated by re-completing the same analyses by another researcher, or by completing another type of analysis on the same data. This is only possible when the data used to complete the original analyses are available for use. Additionally, data are used to eliminate possible solutions to a problem by the publishing of negative or non-significant data. Thinking of negative data as a public good, their availability help researchers avoid sinking time and financial resources into investigation of non-viable hypotheses. Funding sources are not available to pursue the generation of negative data, and agencies look to support fruitful research. As a result, researchers have few incentives to publish negative or non-significant results. This lack of publication of non-fruitful investigation is more commonly known as the "file drawer effect" ((2), (3)). In this current time period when government funding is uncertain, it is more important than ever to pursue fruitful research.

With the latest and greatest methodologies available across fields, increasing amounts of data are being generated each day, especially in the biological sciences where large and complex datasets are the new standard. ((4), (5)). Availability of large quantities of study data and their associated metadata (data about data) are necessary resources for appropriate use and re-use of data, protocols, as well as recreation of analyses. Data availability are a deeply important component of the scientific process in the digital age, and the curation of digital records is a slowly emerging topic in data science((6)). Available data and analyses are the gold standard for recreation of studies and replication of their results. Not only is replication a worthy goal, but large datasets are often underutilized, and can continue to provide benefit and resources to researchers via their re-use towards investigating and answering further questions. As a result, the National Institutes of Health (NIH) has called for grant proposals for the creation, enhancement, and maintenance, of new and existing data repositories ((7), (8)).

There are three major databases worldwide to support sequencing and sharing efforts. The National Library of Medicine's (NLM) National Center for Biotechnology (NCBI) in the United States, the Research Organization of Information Systems' (ROIS) National Institute of Genetics (NIG) in

Japan, and the European Molecular Biology Lab's (EMBL) European Bioinformatics Institue (EBI) in Europe. These three databases are part of the International Nucleotide Sequence Database Collaboration (INSDC) ((9)). Comparative genetics and genomics would not be possible without strong community commitment to data availability.

American Society for Microbiology Journals (ASM)

The American Society for Microbiology(ASM) is the major professional body recognized by microbiologists. They have eighteen journals, thirteen primary research journals, three review journals, and two archive journals. In addition, several journals have been folded into others or renamed over time. In this study, we considered twelve of these journals, *Applied and Envrionmental Microbiology, Antimicrobial Agents and Chemotherapy, Infection and Immunity, Journal of Clinical Biology, Journal of Virology, Journal of Bacteriology, Journal of Microbiology and Biology Education, Microbiology Resource Announcments* (formerly known as *Genome Announcements*), *mSystems, mSphere, mBio, and Microbiology Spectrum.* Of note, several journals had changes to their publication goals during the 2000-2024 time period. In 2010, access to next generation sequencing (NGS) methods became commonplace, and there were a greater number of publications containing sequencing data, including full and partial genome announcments. The *Journal of Bacteriology* was the primary ASM journal to publish new genome announcements until 2013 when ASM announced journal *Genome Announcements* as a more permanent destination for this type of data. *Genome Announcements* was active from 2013 until 2018, when it was re-branded to *Microbiology Resource Announcements*, which has been active from 2018 until present. Another journal of note was *Microbiology Spectrum* and its re-brand. From 2013 until the fall of 2021, *Microbiology Spectrum* was a review journal. At this point and beyond, *Microbiology Spectrum* became a primary research journal ((10)). Several journals, *mBio* (b.2010), *Microbiology Spectrum* (b. 2013, re-brand 2021), *mSphere* (b. 2016), *mSystems* (b. 2016), and *Genome Announcements* (2013-2018) all did not span the entire study period of interest.

Current Data Availability Policies

Current data availability guidelines have been informed by a number of policies created by funding

agencies, peer-review journals, conference and special task groups, as well as community interest

groups. In 2011, after the Future of Research Communication (FoRC) conference in Germany,

scientists and others came together to establish FORCE11, a community interest group which

sought to encourage and promote data availability standards ((11)). In 2014 the FORCE11 group

published the Joint Declaration of Data Citation Principles (JDDCP), a document with continued

work towards the standardization of data citation and future availability((12)). Some of the JDDCPs

included crediting the authors of the data, providing data with unique identifiers, and the persistence

of available data.

Also in 2011, the Genomic Standards Consortium (GSC) published a set of standards in *Nature*

Biotechnology to promote the publication of the "minimum information about a marker gene

sequence" (MIMARKS) or "minimum information about x sequence" (MIxS) ((13)). These standards

are checklists usable by data generators and uploaders towards inclusion of relevant data with

sequence uploads in the International Nucleotide Sequence Database Collaboration (INSDC).

Some checklist items include if the data were published to an INSDC database and metadata about

the study systems, data collected, and authors. An important factor was the ability to link the data

to the results and to the data generators. The Findable, Accessible, Interoperable, and Reuseable

(FAIR) data science guiding principles that were put forth in 2016 in *Nature Scientific Data* urges

readers to "improve the infrastructure supporting the reuse of scholarly data" ((14)). The FAIR

principles were often cited by NIH in funding calls for strong data science practices((7)).

In 2021, a *Nature Medicine* publication put forth the "Strengthening the Organization and Reporting

of Microbiome Studies" (STORMS) checklist to help authors self-identification of report-worthy

elements of their data and metadata ((15)). Some items on the STORMS checklist included

reporting the sequencing method used in the study, the study design, and physical location of the

study. Unfortunately, none of these data availability principles or checklists were yet enforceable by

any agency.

100 The National Institutes of Health (NIH) began enforcing the "Policy for Data Management and Sharing" (NOT-OD-21-013) in January of 2023, requiring all NIH funded studies to submit a data management and sharing plan (DMS) with their funding applications, and comply with their DMS plan after generation and publication of the funded work ((16)). A DMS plan includes detailed descriptions of data that will be generated in a study, related tools, standards, and data preservation

105 plans. Non-compliance with NOT-OD-21-013 is identified by funding agencies during annual Research Performance Progress Reports (RPPRs), and may impact future funding decisions ((16)). The NIH policy for non-compliance with award terms and conditions varies due to the type of research misconduct, but is clear that the NIH will protect their own interests, including placing conditions on awards, preventing future awards, or closer monitoring of award activities. In this time

110 of uncertain funding from the NIH and other funding agencies, it is more important that ever for investigators to maintain continued compliance. This compliance starts with readily available data and manuscripts.

At time of publication, the ASM journal program required that authors "make data fully available, without restriction, except in rare circumstances" ((17)). They have adapted this policy from

115 journals *Microbial Genomics* and *PLOS*. In the ASM open data policy they described the use of a "Data Availability Statement" which includes "data description, name(s) of the repositories, and digital object identifiers (DOIs) or accession numbers" and encouraged publishing data on relevant public repositories ((17)). Consequences of non-compliance to the ASM open data policy included contacting research article authors to inform of non-compliance, publication of an

120 "Expression of Concern" for the author and their compliance issues, future sanctions on publication in ASM journals, as well as contacting the affiliated research institution and/or funding agencies of the authors ((18)). We endeavored to evaluate how well the microbiology community is using reproducible data practices as we believed that this group of researchers were early adopters of technologies available as a result of both the ASM and NIH policies towards data availability.

Historical Nucleic Acid Sequencing Efforts and Examples

Beginning in 1996 with the International Strategy Meeting on Human Genome Sequencing in Bermuda, researchers have prioritized the release of all human genome sequencing information to "maximize its benefit to society" ((19)). The meeting participants agreed that "primary sequence data should be rapidly released", with "sequence assemblies [to] be released as soon as possible, in some centres[sic], assemblies of greater than 1 kb would be released automatically on a daily basis", and that "finished annotated sequence should be submitted immediately to public databases" ((19)). In 2003, another meeting, held in Ft. Lauderdale, FL, re-affirmed the 1996 Bermuda Principles, expanded upon them to apply more broadly towards sequencing data, and called for further support of these practices ((19)).

These foundational "Bermuda Principles" and "Ft. Launderdale Accords" agreements set the stage for both the Human Genome Project (HGP) and the Human Microbiome Project (HMP) to generate and share massive amounts of data over the course of their studies ((20), (21), (22)). The goal of the HMP was to sequence all body sites to determine the microbes found on and in the human body. Starting with major projects such as the HGP and HMP, nucleic acid sequencing efforts have been commonly uploaded and released using public databases. This allows for researchers to use and re-use the data from the HGP and HMP. Us of HMP sequencing data by researchers has resulted in over 650 scientific publications ((23)), and the completion of metadata studies, including those efforts participated in by these authors ((24), (25), (26)).

An important tool for creating phylogenies is the NCBI Basic Local Alignment Search Tool (BLAST) ((27)), which is an essential tool for comparative research. The BLAST algorithm allows users to compare a nucleic acid or protein sequence to the NCBI database of over 1TB of data to find similar and related sequences. Without the upload of sequences to the NCBI database, the use and success of BLAST would not be possible, despite the effort required on part of the researcher to upload of sequences to one of the INSDC databases.

Availablity of data contributed to the rapid sequencing of the SARS-CoV-2 virus during the 2020

pandemic and subsequent expedition of vaccine development ((28)).

Results

Descriptive Statistics

Using the Crossref database of DOIs, with validation from the Web of Science, NCBI, and Scopus DOI databases, we downloaded 154720 unique records of papers published in ASM journals. All papers were published between 2000 and 2024. These papers came from *Applied and Envrionmental Microbiology (N = 20401), Antimicrobial Agents and Chemotherapy (N = 21943), Infection and Immunity (N = 14500), Journal of Bacteriology (N = 16821), Journal of Clinical Microbiology (N = 18450), Journal of Microbiology and Biology Education (N = 1312), Journal of Virology (N = 29797), Microbiology Resource Announcments* (formerly known as *Genome Announcements 12601),* Microbiology Spectrum (N = 6120), mBio (N = 7732), mSphere(N = 2712), and mSystems (N = 2331).*

Training Dataset

As this was a large number of papers to investigate by hand, we created a subset of data to train two machine learning models. This dataset was representative of the larger dataset in age of paper and journal of origin and initially contained 500 papers. During iterative model training, a subset of papers were hand validated after each completed cycle. These subsets were created using dplyr's slice_n() to obtain papers from each journal. Additional hand selected papers were validated to ensure robust model training. Validated papers were added to the training set as we identified gaps in the dataset. This provided a total of 1045 papers.

Random Forest Modeling

Two random forest models were trained to predict if published scientific papers "contained new sequence data", and if the paper "had data available", one model for each variable. Each model

was trained using the same set of data, the normalized number of times a word or set of words appears in each paper in the set. Briefly, the HTML content of each paper was cleaned to removed non-meaningful words such as "a, an, the", and separated into tokens, meaningful units of 1 - 3 words. These tokens were also modified to eliminate issues with word tense. For example "interest_importance" was a two word token that appeared in the model. These tokens were summed and normalized to each other and were presented to the model as a table with each row as a paper, and each column as the frequency of a specific token. See methods for more information on this process.

Other models such as generalized linear regression (GLM) and boosted trees (XGBoost) were evaluated, but were ultimately rejected in favor of the random forest model due to fit (XXdata not shown- need to re-generateXX). Random forest models were chosen by their high Area Under the Receiver Operator Curve (AUROC) value to aid in this classification problem as the creation of many decision trees helps to improve accuracy and precision.

These model trainings resulted in two different models to answer two different questions. The new sequencing data model used an mtry value of 200 had an Area Under the Curve(AUC) of 0.96 and an accuracy of 0.9. The sensitivity of the new sequencing data model was 0.94, and the specificity of the model was 0.86. The data availability model used an mtry value of 300 had an AUC of 0.99 and an accuracy of 0.95. The sensitivity of the data availability model was 0.96, and the specificity of the model was 0.92 (See supplement for more information on trained machine learning models). This showed us that the models fit the data well, and could provide classifications on new data with an expected error rate of less than 10%. We deemed this as acceptable, accounting for variability in papers and data, as well as the large size of the dataset on which we deployed the models.

Deploying on 150K+ Papers

After downloading the HTML content of each paper, we cleaned the HTML content and readied it to apply our machine learning models to classify each paper, in brief, by removing non-meaningful words, followed by counting and filtering tokens, and finishing in model-compatible matrix for-

mat. Overall, 26.94% of papers had new sequencing data. The journal with the highest rate of new sequencing data was *Genome Announcements* at 98.72%, and the lowest was *Journal of Microbiology & Biology Education* at 0.31%. This was expected as *Genome Announcements* and *Microbiology Resource Announcements* were the primary places for the publication of new sequence data for ASM journals. In 2013, *Genome Announcements* was created to house papers describing new genomic sequencing efforts, and these publications were redirected from the *Journal of Bacteriology*. In 2018, *Genome Announcments* was re-branded to *Microbiology Resource Announcements*, the permanent home of new sequencing efforts. This change in journal scope for the *Journal of Bacteriology* explained the change in percentage of new sequencing data found in the journal over time. The *Journal of Microbiology and Biology Education*, as expected, contained the lowest percentage of new sequencing data. The scope of this journal is mainly resources for educators at the high school and college levels, and we did not expect papers to contain new sequencing data. In many of the ASM journals, the percentages of new sequencing data did not change significantly over time as their scope did not change. See Figure XX for percentages of new sequencing data for each journal, and figure XXX for the trends over time. (XX? Put figures in panel together for nsd overall and nsd over time?) XX Other questions: XX Why did AEM go up so much in 2013ish? XX Should we discuss the IAI 2012ish blurb? XX
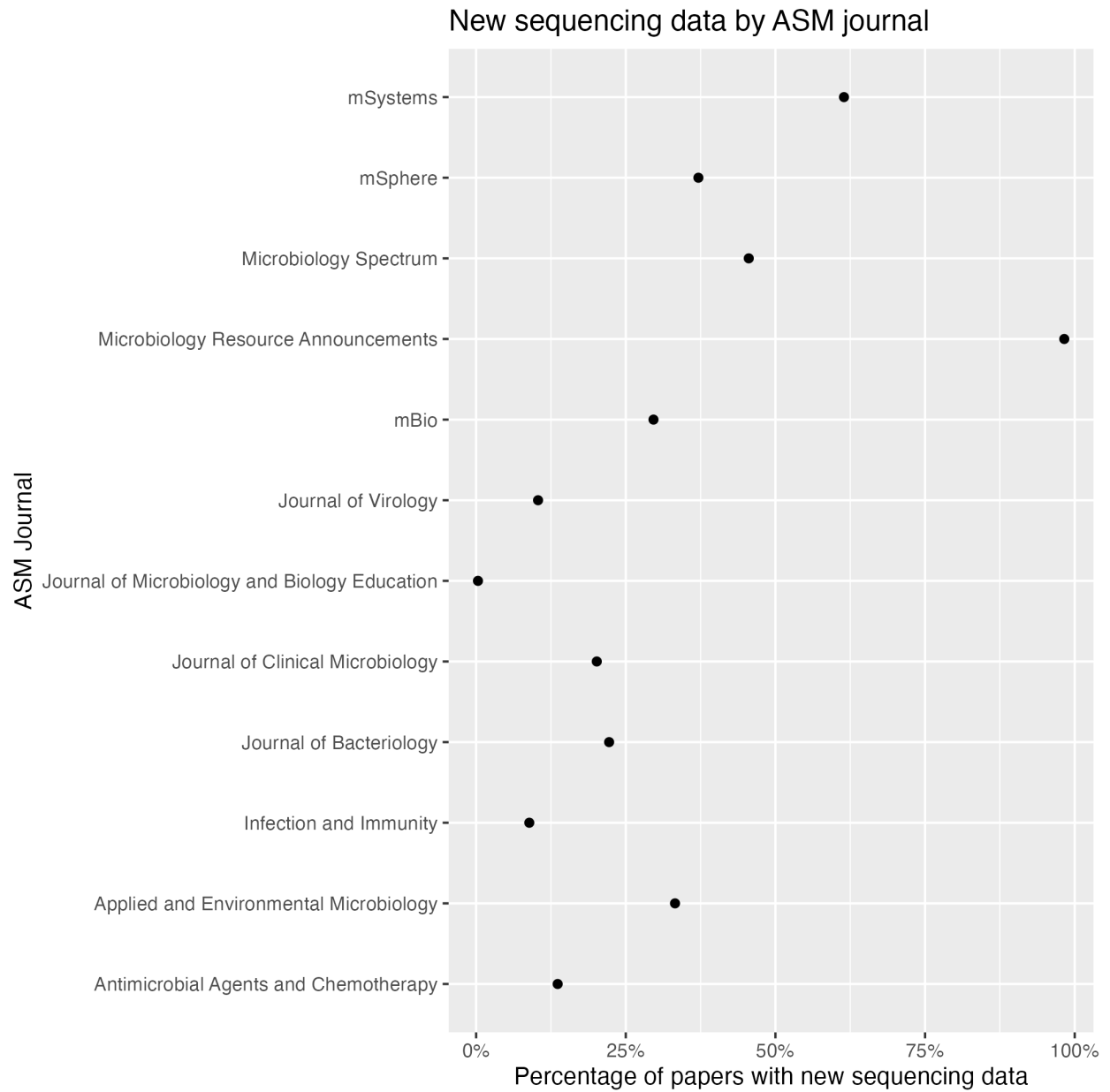
9

New sequencing data by ASM journal

Figure 1: Percentage of papers with new sequencing data

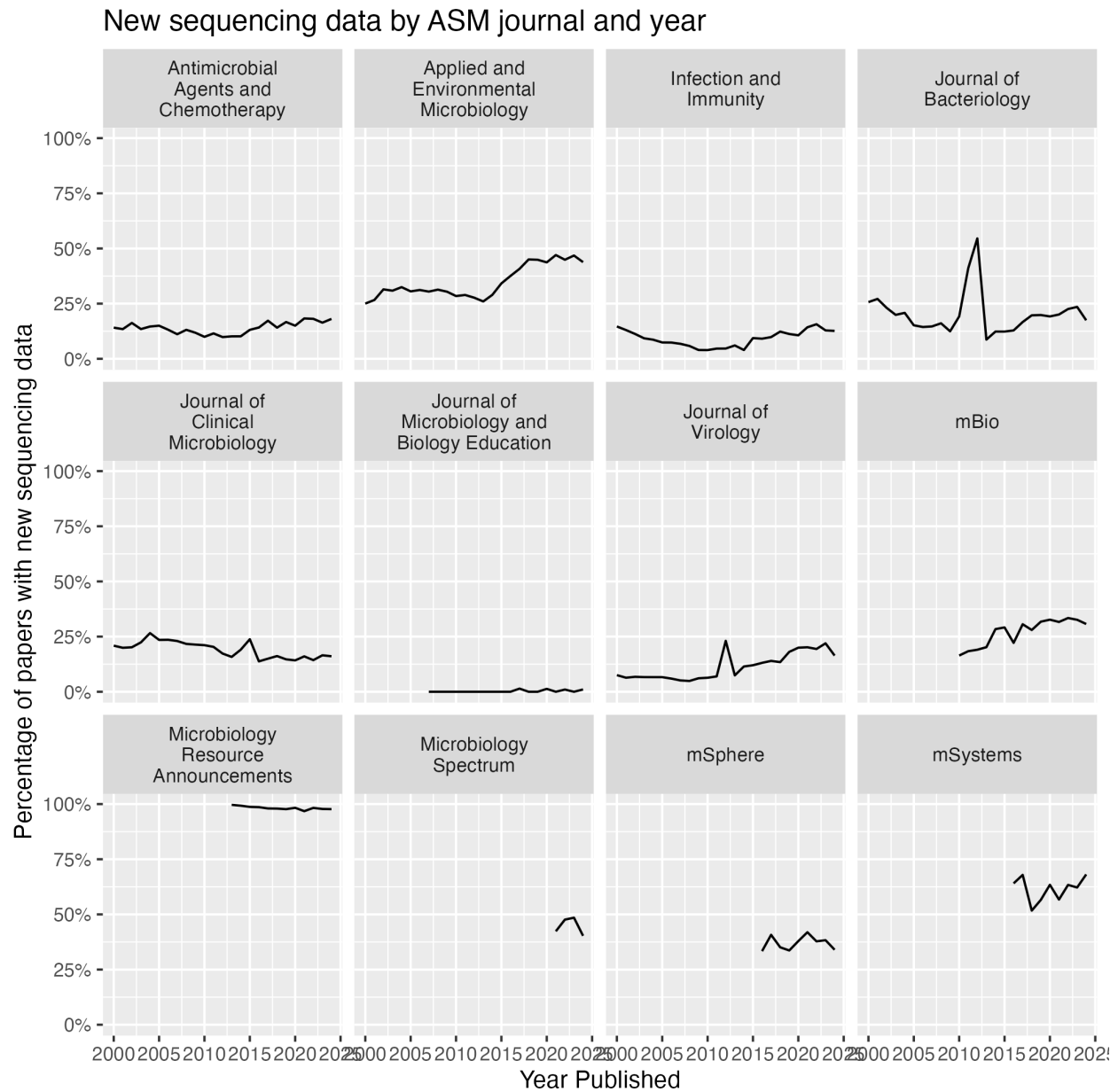## New sequencing data by ASM journal and year

Figure 2: Percentage of papers with new sequencing data over time

Of all papers with new sequencing data, 58.86% of papers had data available. The journal with the highest rate of data availability was *Microbiology Resource Announcements* at 99.83%, and the lowest was *Journal of Clinical Microbiology* at 13.44%. Similarly to the new sequencing data, this was expected as *Microbiology Resource Announcements* publishes new genomic sequence data for ASM and makes the data available per journal requirements. While overall data availability

rates appeared low, these values took into account all papers published over the entire time period of interest. When plotted over time, data availability rates trended upwards over time, with a greater fraction of papers with new sequencing data having made their data available. As before, the *Journal of Bacteriology* had an upward trajectory until the redirection of papers to *Genome Announcements* and *Microbiology Resource Announcements* in 2013 and beyond. In 2023, the NIH policy requiring a data management and sharing plan (DMS) went into effect. All of the ASM journals show an upward trend in data availability in 2023 and 2024 to comply with funding agency policy. The *Journal of Microbiology and Biology Education* published so few papers during this period with new sequencing data, that they did not publish any papers with data available. See figures XXX and XXX for percentages of new sequencing data papers with data availability for each journal and over time. XX??Want to investigate some of the weirdly high years for major disease outbreaks/policy changes/editor changesXXX
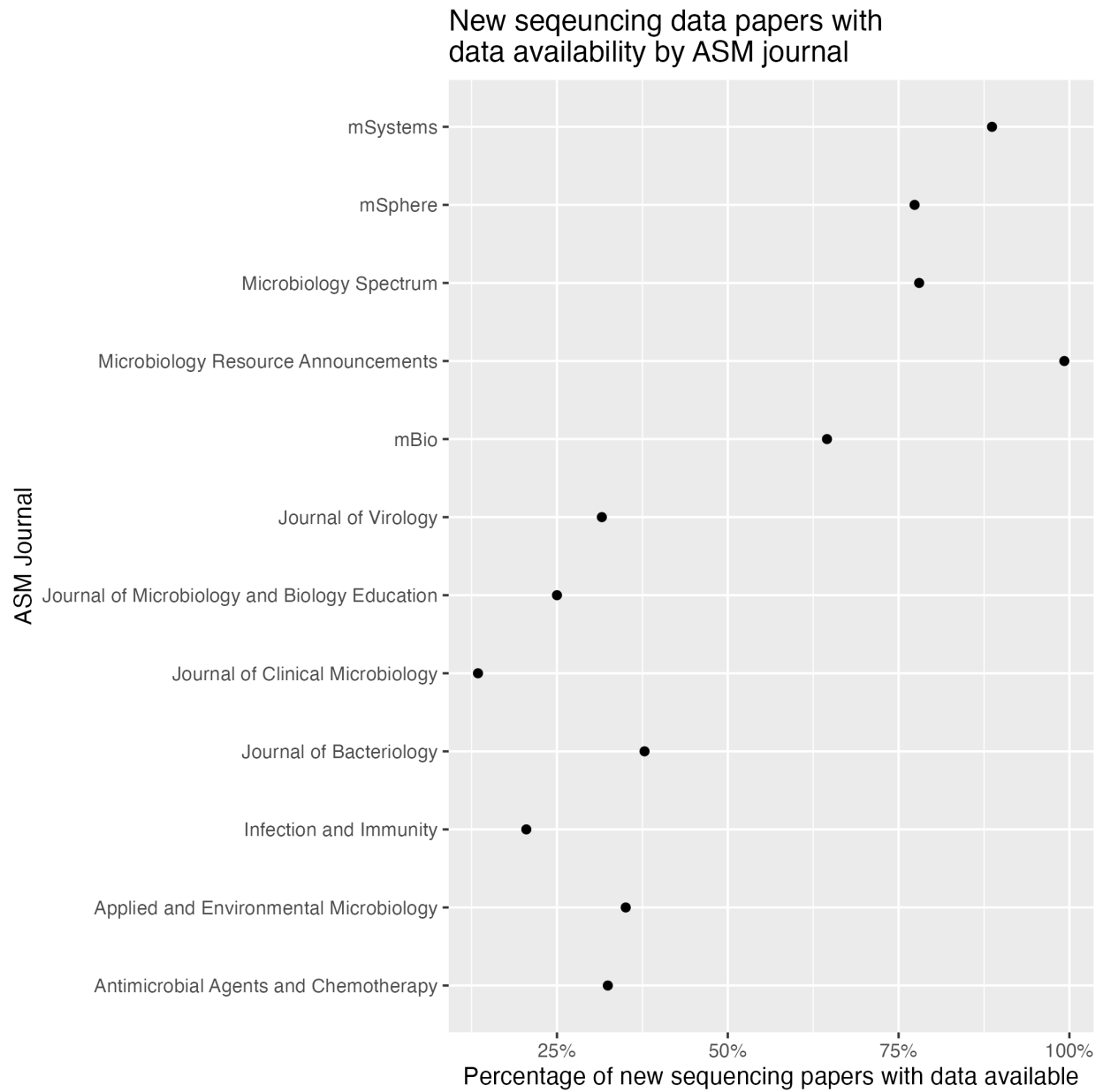
12

New seqeuncing data papers with
data availability by ASM journal

Figure 3: Percentage of new sequencing papers with data available

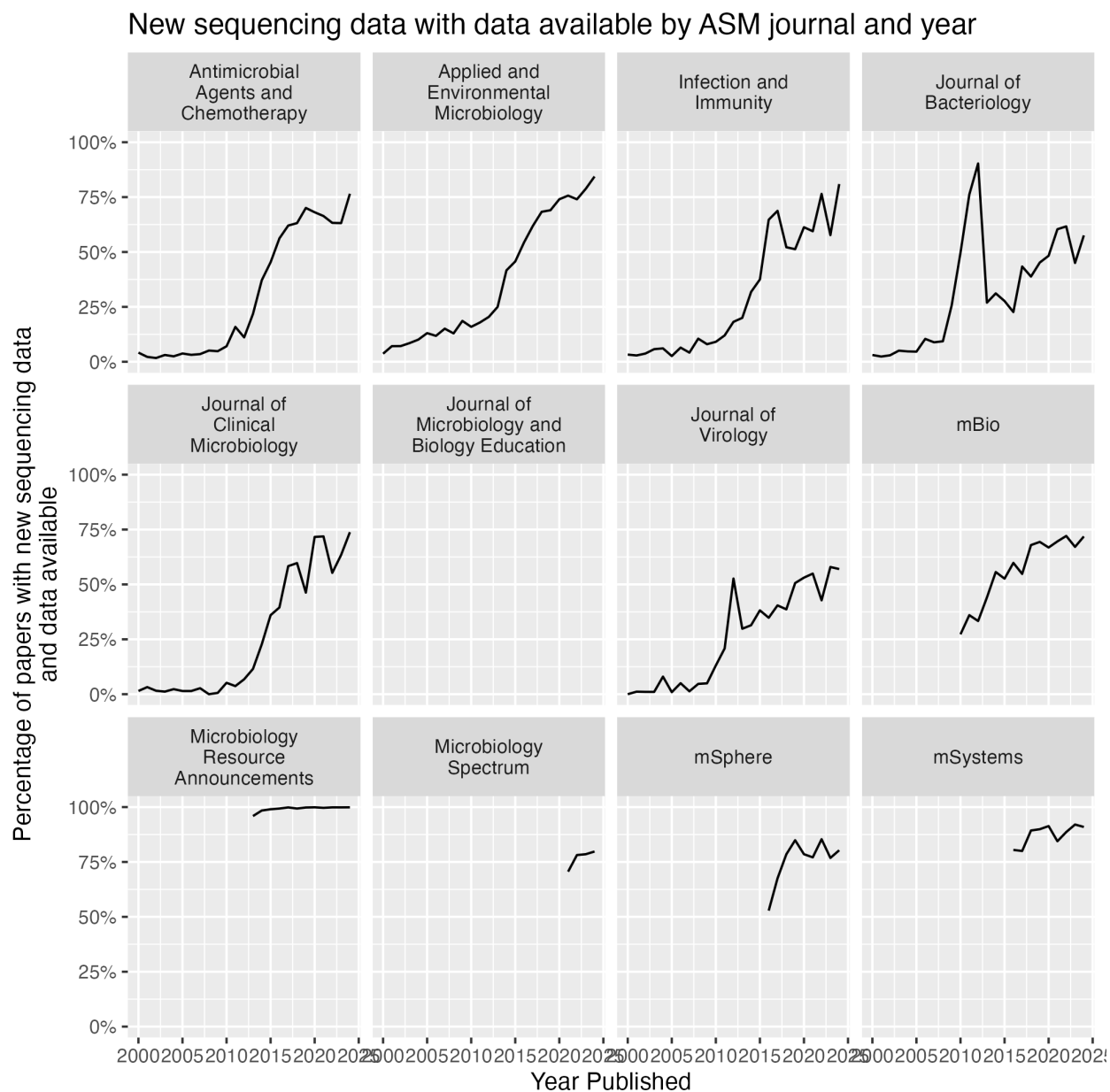New sequencing data with data available by ASM journal and year

Figure 4: Percentage of new sequencing papers with data available over time
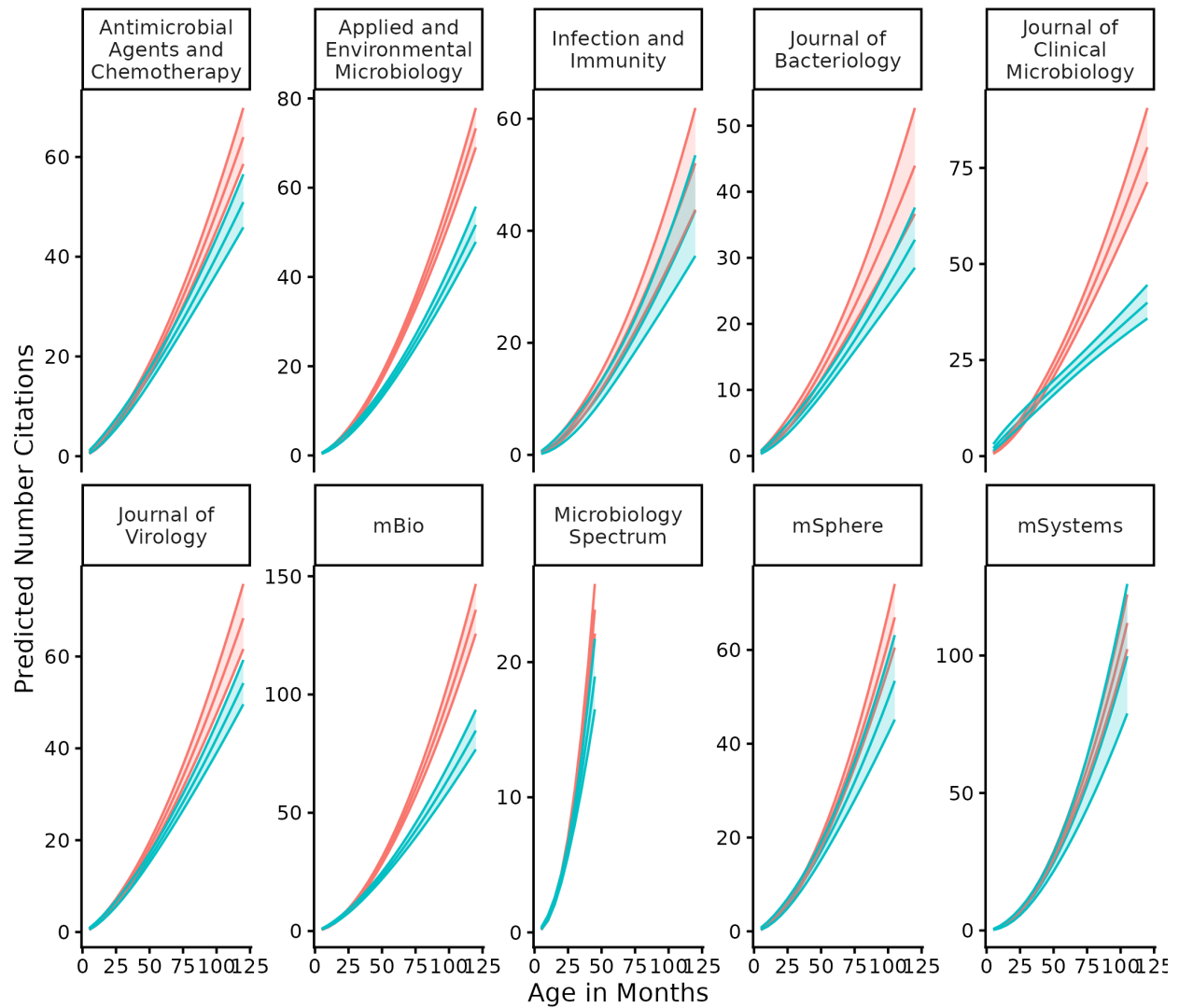
Regression Model using Negative Binomial Models

235 In this study we sought to investigate the effect of new sequencing data and data availability on

the number of citations received by a given paper. We focused on new sequencing data papers to

determine the effect of having data available. On average, papers in the dataset had a median of

25 citations/article. The journal with the highest median rate of citations/article was *Applied and*

14

Environmental Microbiology at 38, and the lowest was *Microbiology Resource Announcements*

at 1. To further investigate the effects of time (age in months), the journal that a given paper was

published in, and the data availability status, a negative binomial regression model was used to best

describe our data (see methods). We also limited our analysis to the last ten years of data (age in

months <=120) to exclude the change in scope for *Journal of Bacteriology* and removed papers

from before the re-brand of *Microbiology Spectrum*. Our model has shown that over time, new

sequencing papers with data available have a greater number of citations than those that do not

have their data available(fig XXX). This relationship holds true in each journal publishing primary

research articles. In journals *mSystems and Infection and Immunity*, the confidence intervals

overlap, but that phenomenon was attributed to the low number of papers from these two journals

with no data available.

In figure XX, we calculated the ratio of number of citations for papers of similar age containing

data vs those that did not contain data by journal. Over time, papers with data available received

more citations than those without up to well over 2 times more citations in the *Journal of Clinical*

Microbiology. In all journals, papers with data availability had a greater number of citations at

time point 60 months after publication, if not sooner. During the time interval this ratio grew,

demonstrating increased citations for papers with data available available over time compared to

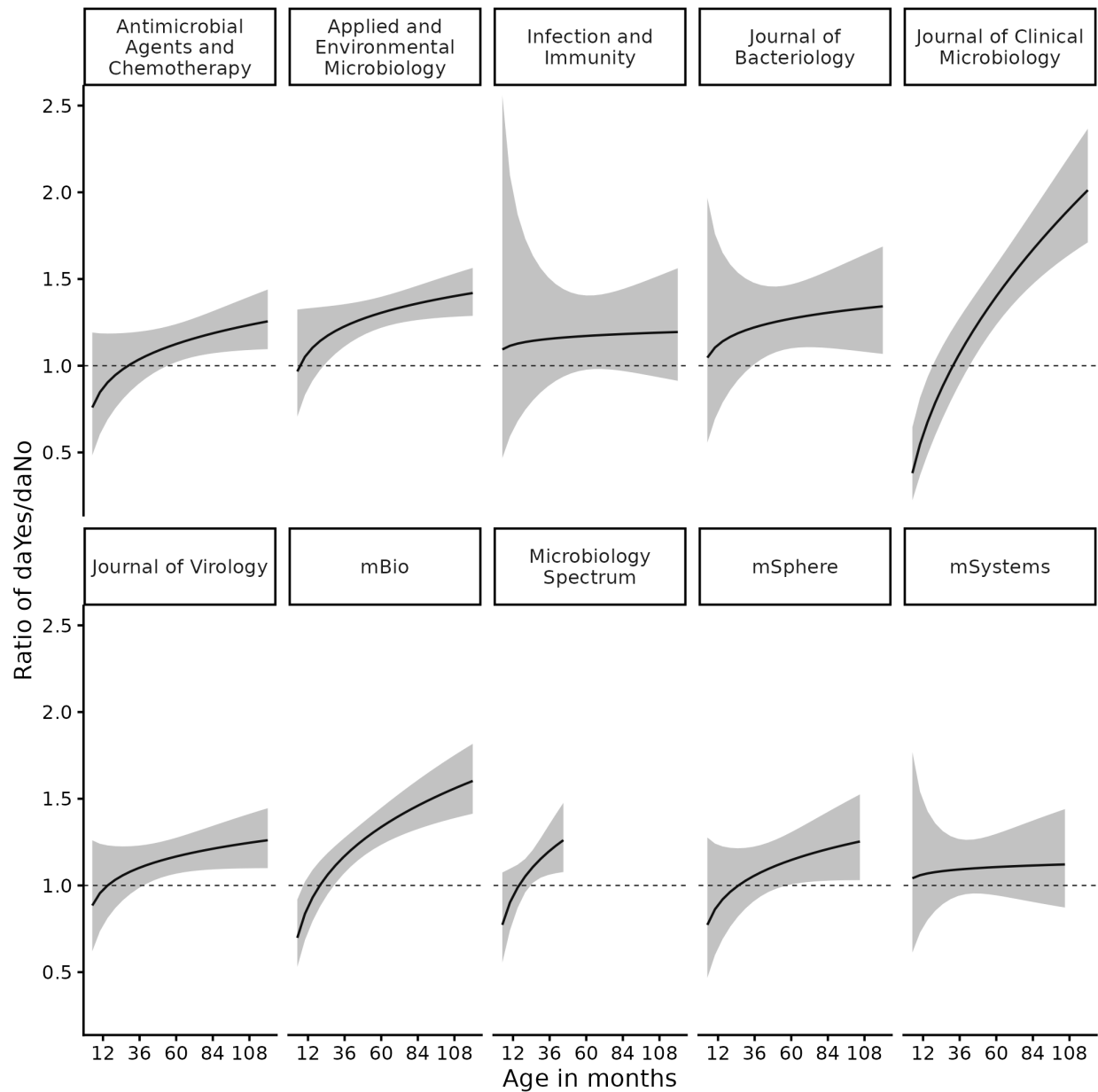those without, and the difference grew over the time interval.

Predicted Number of Citations from GLM.NB

Discussion

Our study sought to investigate the compliance and impact of data availability and associated policies on the ASM journal program. After training machine learning models to classify papers to determine if they contained new sequencing data and if each new sequencing paper had data available, we found that on average, more than half of papers with new sequencing data had data availability (58.86%). More often than not, authors of new sequencing data papers were

making their data publicly available. This was in line with 2023 NIH policy requiring that data be made available using a Data Management and Sharing Plan (DMS plan) outlined in the NIH's NOT-OD-21-013 ((16)). XXDo we want percent for 2024?XX

Expectledly, *Microbiology Resource Announcements* had the highest rates of data availability in new sequencing papers. This journal published new genomic sequences and was required by ASM to make data available. Journals such as the *Journal of Microbiology and Biology Education* and *Infection and Immunity* which published fewer new sequencing data papers due to their specific subject matters, had lower rates of new sequencing data and data availability in their journals.

Once these questions were answered and papers were classified, we moved to statistical analyses to answer the further question, "how does making data available impact citation metrics over time?" We were interested in citation metrics as a concrete metric to examine how making data availability benefits researchers and as a possible incentive towards making data availability. Our hypotheses were that microbiologists would have fairy high rates of data availability given the field's reliance on comparative research, and that other fields such as immunology would have lowered rates of data availability, as well as the hope that papers with data available would provide a greater number of citations over time compared to those without data available.

Using a negative binomial model we found that over time, papers with data available received more citations than those without. For each journal, this pattern held true. The difference in citations between those with and without data available climbed to well over 1.5x the amount of citations in some journals (Fig XXXX. *Journal of Clinical Microbiology*). These differences in ratio of papers by journal may have been due to the scope of papers found in each journal.

We found that the effect of data availability providing greater citations intensifies over time, with the greatest differences in citations occurring at the 108 months since publication time point. This was great news to us and to other manuscript authors, that simply making their data available could provide as much as 50% increase in citations over time. We hope that these data will help to incentivize authors to make data available.

18

We acknowledge the limitations of our study data, that by focusing only on papers published in the ASM journal program, our results will not be as generalizable. We understand that this relationship between data availability and citation metrics may not be as strong in other journal programs, but we hope with the newest NIH funding policies, these trends will continue to improve data availability.

Another limitation is the availability of paper metadata from various databases, with the Crossref database having provided the most complete metadata available for each paper. We were also not able to track citation metrics for individual papers over time. Our database sources only had available citation metrics at the time of dataset download (February 2025), with no intermediate time points for any given paper. This left us unable to understand trends over time due to popularity or obsolescence of technique or results.

While we found that making data available in publications gave a citation advantage, we hope that is not the only reason authors choose to make their data available. The need for reproducibility in science and demand for large datasets are additional reasons, as well as the NIH funding requirement. We hope that authors recognize the need to contribute to the data "commons", to further work done by others, and that even their negative results have value and power to stop others from continuing down dead ends.

Materials and Methods

- need to put this stuff somewhere

To avoid overfitting the models, we trained each model multiple times, performing validations on a subset of data after each iteration. This allowed us to have a greater number of papers in the training dataset by adding these iteratively validated papers, as well as to have great accuracy and precision within our models.

19

Preparation of the Larger Experimental Dataset

To fully answer our research questions, we created a dataset with papers curated from reference

315 databases Crossref, NCBI, Scopus, and the Web of Science ((29), (30) , (31), (32)). These papers

span all twelve ASM journals of interest from January 1st, 2000 to December 31st, 2024. The

ASM Journals of interest were *Applied and Envrionmental Microbiology; Antimicrobial Agents*

and Chemotherapy; Infection and Immunity; Journal of Clinical Biology; Journal of Virology;

Journal of Bacteriology; Journal of Microbiology and Biology Education; Microbiology Resource

320 *Announcments* (formerly known as *Genome Announcements*); *mSystems; mSphere; mBio; and*

Microbiology Spectrum. The data was updated as of February 10th, 2025 with all citation counts

frozen at that date. See results for number of papers from each journal of interest. With a dataset

this large, we needed machine learning tools to classify papers as new seqeuncing papers and to

determine if those new seqeuncing papers contained available data.

325 **Cleaning data for modeling**

To perform the computational steps required for these experiments, we used the python tool

Snakemake ((33)), and the University of Michigan's high performance computing cluster (see

acknowledgements). Using our selected papers, we downloaded the entirety of each paper's

source HTML and used the command line tool wget. This allowed us to use the same source HTML

330 multiple times for updated analyses without needing to re-query the ASM webservers numerous

times. Next, we filtered the HTML using R packages rvest ((34)), textstem((35)), and xml2 ((36))

to get the desired portions of the paper from the HTML including the abstract, the body of paper,

all tables and figures with captions, as well as the side panels for all papers, but especially those

that contained the data availability statements in papers published after a 2023 change in webpage

335 format. Towards the goal of the fewest meaningful number of words, we used the 'Snowball' ((37))

dictionary of 'stop words' with R packages tm(text manipulation)((38), (39)) and textstem ((35)) to

remove non-meaningful words such as articles 'a', 'an', and 'the', and other unnecessary words

in each paper. We also converted all text to lowercase, concurrent with the removal of digits and non-alphabetic characters such as whitespace. To have the fewest number of unique words, we lemmatized (sorted words by grouping inflected or variant forms of the same word) words to trace them back to their root words and eliminated any possible issues with word tense. After this, we created and counted our 'tokens', phrases of up to 1-3 consecutive words from the text of the paper using R package tokeinziers ((40)). We removed the 'space' character with an underscore in multi-word tokens for ease of processing, and created a count table for the number of tokens in each paper.

Once the tokens in each paper were counted, we transformed the data into a sparse matrix format useable by the R package mikropml ((41)), using R packages caret and dplyr ((42), (43), (44)). Tokens were filtered to those which appear in greater than one paper. This allows comparison between papers by the model. We removed near zero variants (tokens with frequency very close to zero) as well as collapsing perfectly correlated tokens (tokens that always appear together) using R packages caret and mikropml to reduce model complexity. The data was then simplified to keep only the following variables; tokens, frequency, journal information, and hand identified new sequencing data and data availability variables. This simiplified sparse matrix data had the mean and standard deviation calculcated and saved for the frequency of each token to later apply a z-scoring method to future data to be predicted by the model.

Creation of the Training set

To train our random forest machine learning models, we first created an appropriate training data set. For our initial training set, we chose a set of papers from across each journal and the time period of interest, adding special emphasis to include papers that were part of our desired set of interest (i.e. contained published data) to ensure that our two models could adequately characterize each paper as a new sequencing paper and if it published raw sequencing data or not. These papers were chosen with dplyr's slice_n(). After creating our initial dataset, it was necessary for us to to identify the status of both variables by hand and determined if each paper contained "new

21

sequencing data", and if each one had "data available". This was completed by opening each paper in an internet browser window, and searching for a "data availability" or similar statement. See Table 1XXX for specific cases and how each of these cases were identified for the purpose of this study.

Table 1: Possible Data Scenarios

Scenario	new sequencing data status	data availability status
Paper is not about generating new sequencing data	No	No
Paper is about generating new sequencing data but has no data available	Yes	No
Paper is about generating new sequencing data and has data available	Yes	Yes
Paper uses sequencing as a confirmation of experimental technique (i.e. confirmation of plasmid insertion)	No	No
Paper discusses new computational or experimental tools	No	No
Paper has microarray data	No	No
Papers using MLST ONLY	No	No
Papers using qPCR ONLY	No	No
Papers about protein sequencing that have nucleotide sequencing	Yes	Yes/No depending on data availability
Papers using iRNA techniques	No	No
Papers using pyrosequencing/454 techniques	Yes	Yes/No depending on data availability

Training of the data availability and new sequencing data Models

We trained two random forest machine learning models using mikropml's "run_ml" function, one to determine if a paper contained new sequence data, and another to determine if the paper had data available. The mikropml "run_ml" function uses methodology described by *Topcuoglu et al.*((45)) to split data for model training. Random forest models have one hyperparameter to tune, the mtry value. We began with mtry values of 100, 200, 300, 400, 500, and 600, to find peak hyperparameter performance given *N tokens*. We trained the models multiple times in accordance with existing methodologies, first to find the optimal Area Under the Receiver-Operator Curve (AUROC) value for each model with 100 seeds. Then to find the best mtry performance for each model, with 1 seed. Finally, the models were trained with 1 seed each and generated each final model for use on experimental data.

Adding Additional Training Set Papers

After initial training of our random forest models, a random sampling of papers was collected for each journal using dplyr's slice_n() to audit the efficacy of the models ((43)). When auditing the efficacy of the models, we hand identified the status of both variables of interest, new sequencing data and data availability. We looked for weaknesses in the models, and updated methodology to reflect important areas of interest. For example, in 2023 the ASM journals changed their formatting to include the data availability statement of a paper in a sidebar of the webpage. We identified this by noticing that all papers from journal *Microbiology Resource Announcements* from 2023-2024 were incorrectly characterized by the model as data availability = No. The sidebar of the webpage was not included in the text the model was considering, and code had to be updated to include all sidebar data for all papers. These improvements to the model created a larger and more comprehensive training set of N = 1045. These validations allowed us to create confusion matrices for each model.

Deploying the RF Models

Once our RF models were ready we applied the same steps to ready papers for application of machine learning models as the model training dataset. See above for descriptions of web scraping HTML, cleaning HTML, removing unnecessary text, and creation of token count table for application in each of the machine learning models to determine the new sequencing data and data availability statuses for each paper. Once the frequency count tables were prepared for each paper, a z-score was applied using the saved data from each model appropriately, using the formula XX $((Observed token frequency\ Model token frequency)/(model token frequency sd))$. This z-scoring formula was applied to standardize the frequency of each token. Only tokens included in the machine learning models were retained in experimental datasets. Finally, each model was deployed on each paper to determine if it contained new sequencing data and if it had data available.

Statistical Methodology

To investigate how data availability impacts citation metrics, we utilized a negative binomial regression model using the R package MASS ((46)). All papers included in the regression data had new sequencing data and were filtered to only those papers less than or equal to 120 months of age. We focused on the continuous outcome of "number of citations" with predictor variables journal (categorical), age in months (continuous), and data availability status (dichotomous). Due to the number of citations being bell shaped with a long right tail (very few papers at advanced age with many citations, producing a flat but non-zero line), the model that best described our data was a negative binomial regression model. A negative binomial model is appropriate for data that begins at zero and has a long 'tail' of data, as well as has differing means per group. This model also includes a dispersion parameter, to describe the spread of the data. We applied a log transformation to our age variable (age.in.months) to make linear the relationship between time and number of citations received to help better describe the model relationship. After applying the negative binomial model, we calculated ratios of the estimated number of citations per paper over

24

time by data availability status using R package emmeans ((47)). We also used R package sjPlot to estimate the 95% CI for estimated citations over time for each journal by data availability status ((48)).

- Supplemental Material file list (where applicable)

420 - Acknowledgments

1. Federal Research and Development (R&D) Funding: FY2024. legislation.

2. **Moniz P**, **Druckman JN**, **Freese J**. 2025. The file drawer problem in social science survey experiments. Proceedings of the National Academy of Sciences **122**:e2426937122. doi:10.1073/pnas.2426937122.

3. **Rosenthal R**. 1979. The file drawer problem and tolerance for null results. Psychological bulletin **86**:638.

4. **Li Y**, **Chen L**. 2014. Big Biological Data: Challenges and Opportunities. Genomics, Proteomics & Bioinformatics **12**:187–189. doi:10.1016/j.gpb.2014.10.001.

430 5. **Pal S**, **Mondal S**, **Das G**, **Khatua S**, **Ghosh Z**. 2020. Big data in biology: The hope and present-day challenges in it. Gene Reports **21**:100869. doi:10.1016/j.genrep.2020.100869.

6.   **Howe D**, **Costanzo M**, **Fey P**, **Gojobori T**, **Hannick L**, **Hide W**, **Hill DP**, **Kania R**, **Schaeffer M**, **St Pierre S**, **Twigger S**, **White O**, **Yon Rhee S**. 2008. The future of biocuration. Nature **455**:47–50. doi:10.1038/455047a.

7.   Expired NOT-OD-24-096: Notice of special interest (NOSI): Promoting data reuse for health research.

8.   PAR-23-236: Early-stage biomedical data repositories and knowledgebases (R24 clinical trial not allowed).

9.   International nucleotide sequence database collaboration.

10.   **Cuomo CA**. 2021. The Relaunch of Microbiology Spectrum. Microbiology Spectrum **9**:10.1128/spectrum.00396–21. doi:10.1128/spectrum.00396-21.

11.   About FORCE11 – FORCE11.

12.   **Altman Director of Research and Head/Scientist, Micah**, **Borgman Professor and Presidential Chair, Christine**, **Crosas Director of Data Science M**, **Matone Co-Director M**. 2015. An introduction to the joint principles for data citation. Bulletin of the Association for Information Science and Technology **41**:43–45. doi:10.1002/bult.2015.1720410313.

13. **Yilmaz P**, **Field D**, **Knight R**, **Cole JR**, **Amaral-Zettler L**, **Gilbert JA**, **Karsch-Mizrachi I**, **Johnston A**, **Cochrane G**, **Vaughan R**, **Hunter C**, **Park J**, **Morrison N**, **Rocca-Serra P**, **Sterk P**, **Arumugam M**, **Bailey M**, **Baumgartner L**, **Birren BW**, **Blaser MJ**, **Bonazzi V**, **Booth T**, **Bork P**, **Bushman FD**, **Buttigieg PL**, **Chain PSG**, **Charlson E**, **Costello EK**, **Huot-Creasy H**, **Dawyndt P**, **DeSantis T**, **Fierer N**, **Fuhrman JA**, **Gallery RE**, **Gevers D**, **Gibbs RA**, **Gil IS**, **Gonzalez A**, **Gordon JI**, **Guralnick R**, **Hankeln W**, **Highlander S**, **Hugenholtz P**, **Jansson J**, **Kau AL**, **Kelley ST**, **Kennedy J**, **Knights D**, **Koren O**, **Kuczynski J**, **Kyrpides N**, **Larsen R**, **Lauber CL**, **Legg T**, **Ley RE**, **Lozupone CA**, **Ludwig W**, **Lyons D**, **Maguire E**, **Methé BA**, **Meyer F**, **Muegge B**, **Nakielny S**, **Nelson KE**, **Nemergut D**, **Neufeld JD**, **Newbold LK**, **Oliver AE**, **Pace NR**, **Palanisamy G**, **Peplies J**, **Petrosino J**, **Proctor L**, **Pruesse E**, **Quast C**, **Raes J**, **Ratnasingham S**, **Ravel J**, **Relman DA**, **Assunta-Sansone S**, **Schloss PD**, **Schriml L**, **Sinha R**, **Smith MI**, **Sodergren E**, **Spor A**, **Stombaugh J**, **Tiedje JM**, **Ward DV**, **Weinstock GM**, **Wendel D**, **White O**, **Whiteley A**, **Wilke A**, **Wortman JR**, **Yatsunenko T**, **Glöckner FO**. 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology **29**:415–420. doi:10.1038/nbt.1823.

14. **Wilkinson MD**, **Dumontier M**, **Aalbersberg IjJ**, **Appleton G**, **Axton M**, **Baak A**, **Blomberg N**, **Boiten J-W**, **Silva Santos LB da**, **Bourne PE**, **Bouwman J**, **Brookes AJ**, **Clark T**, **Crosas M**, **Dillo I**, **Dumon O**, **Edmunds S**, **Evelo CT**, **Finkers R**, **Gonzalez-Beltran A**, **Gray AJG**, **Groth P**, **Goble C**, **Grethe JS**, **Heringa J**, **Hoen PAC 't**, **Hooft R**, **Kuhn T**, **Kok R**, **Kok J**, **Lusher SJ**, **Martone ME**, **Mons A**, **Packer AL**, **Persson B**, **Rocca-Serra P**, **Roos M**, **Schaik R van**, **Sansone S-A**, **Schultes E**, **Sengstag T**, **Slater T**, **Strawn G**, **Swertz MA**, **Thompson M**, **Lei J van der**, **Mulligen E van**, **Velterop J**, **Waagmeester A**, **Wittenburg P**, **Wolstencroft K**, **Zhao J**, **Mons B**. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**:160018. doi:10.1038/sdata.2016.18.

440 15. **Mirzayi C**, **Renson A**, **Zohra F**, **Elsafoury S**, **Geistlinger L**, **Kasselman LJ**, **Eckenrode K**, **Wijgert J van de**, **Loughman A**, **Marques FZ**, **MacIntyre DA**, **Arumugam M**, **Azhar R**, **Beghini F**, **Bergstrom K**, **Bhatt A**, **Bisanz JE**, **Braun J**, **Bravo HC**, **Buck GA**, **Bushman F**, **Casero D**, **Clarke G**, **Collado MC**, **Cotter PD**, **Cryan JF**, **Demmer RT**, **Devkota S**, **Elinav E**, **Escobar JS**, **Fettweis J**, **Finn RD**, **Fodor AA**, **Forslund S**, **Franke A**, **Furlanello C**, **Gilbert J**, **Grice E**, **Haibe-Kains B**, **Handley S**, **Herd P**, **Holmes S**, **Jacobs JP**, **Karstens L**, **Knight R**, **Knights D**, **Koren O**, **Kwon DS**, **Langille M**, **Lindsay B**, **McGovern D**, **McHardy AC**, **McWeeney S**, **Mueller NT**, **Nezi L**, **Olm M**, **Palm N**, **Pasolli E**, **Raes J**, **Redinbo MR**, **Rühlemann M**, **Balfour Sartor R**, **Schloss PD**, **Schriml L**, **Segal E**, **Shardell M**, **Sharpton T**, **Smirnova E**, **Sokol H**, **Sonnenburg JL**, **Srinivasan S**, **Thingholm LB**, **Turnbaugh PJ**, **Upadhyay V**, **Walls RL**, **Wilmes P**, **Yamada T**, **Zeller G**, **Zhang M**, **Zhao N**, **Zhang M**, **Zhao L**, **Zhao L**, **Bao W**, **Culhane A**, **Devanarayan V**, **Dopazo J**, **Fan X**, **Fischer M**, **Jones W**, **Kusko R**, **Mason CE**, **Mercer TR**, **Sansone S-A**, **Scherer A**, **Shi L**, **Thakkar S**, **Tong W**, **Wolfinger R**, **Hunter C**, **Segata N**, **Huttenhower C**, **Dowd JB**, **Jones HE**, **Waldron L**. 2021. Reporting guidelines for human microbiome research: The STORMS checklist. Nature Medicine **27**:1885–1892. doi:10.1038/s41591-021-01552-x.

16. NOT-OD-21-013: Final NIH Policy for Data Management and Sharing.

17. Open Data Policy. ASM Journals.

18. Publishing Ethics Policies and Procedures. ASM Journals.

19. **Maxson Jones K**, **Ankeny RA**, **Cook-Deegan R**. 2018. The Bermuda Triangle: The Pragmatics, Policies, and Principles for Data Sharing in the History of the Human Genome Project. Journal of the History of Biology **51**:693–805. doi:10.1007/s10739-018-9538-7.

445   20.   **Proctor LM**, **Creasy HH**, **Fettweis JM**, **Lloyd-Price J**, **Mahurkar A**, **Zhou W**, **Buck GA**, **Snyder MP**, **Strauss JF**, **Weinstock GM**, **White O**, **Huttenhower C**, **The Integrative HMP (iHMP) Research Network Consortium**. 2019. The Integrative Human Microbiome Project. Nature **569**:641–648. doi:10.1038/s41586-019-1238-8.

21.   **Gevers D**, **Pop M**, **Schloss PD**, **Huttenhower C**. 2012. Bioinformatics for the Human Microbiome Project. PLOS Computational Biology **8**:e1002779. doi:10.1371/journal.pcbi.1002779.

22.   **Group JCHMPDGW**. 2012. Evaluation of 16S rDNA-Based Community Profiling for Human Microbiome Research. PLOS ONE **7**:e39315. doi:10.1371/journal.pone.0039315.

23.   Human Microbiome Project (HMP) | NIH Common Fund.

24.   **Ding T**, **Schloss PD**. 2014. Dynamics and associations of microbial community types across the human body. Nature **509**:357–360. doi:10.1038/nature13178.

450   25.   **Abubucker S**, **Segata N**, **Goll J**, **Schubert AM**, **Izard J**, **Cantarel BL**, **Rodriguez-Mueller B**, **Zucker J**, **Thiagarajan M**, **Henrissat B**, **White O**, **Kelley ST**, **Methé B**, **Schloss PD**, **Gevers D**, **Mitreva M**, **Huttenhower C**. 2012. Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. PLOS Computational Biology **8**:e1002358. doi:10.1371/journal.pcbi.1002358.

26. **Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, Giglio MG, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi V, Paul Brooks J, Buck GA, Buhay CJ, Busam DA, Campbell JL, Canon SR, Cantarel BL, Chain PSG, Chen I-MA, Chen L, Chhibba S, Chu K, Ciulla DM, Clemente JC, Clifton SW, Conlan S, Crabtree J, Cutting MA, Davidovics NJ, Davis CC, DeSantis TZ, Deal C, Delehaunty KD, Dewhirst FE, Deych E, Ding Y, Dooling DJ, Dugan SP, Michael Dunne W, Scott Durkin A, Edgar RC, Erlich RL, Farmer CN, Farrell RM, Faust K, Feldgarden M, Felix VM, Fisher S, Fodor AA, Forney LJ, Foster L, Di Francesco V, Friedman J, Friedrich DC, Fronick CC, Fulton LL, Gao H, Garcia N, Giannoukos G, Giblin C, Giovanni MY, Goldberg JM, Goll J, Gonzalez A, Griggs A, Gujja S, Kinder Haake S, Haas BJ, Hamilton HA, Harris EL, Hepburn TA, Herter B, Hoffmann DE, Holder ME, Howarth C, Huang KH, Huse SM, Izard J, Jansson JK, Jiang H, Jordan C, Joshi V, Katancik JA, Keitel WA, Kelley ST, Kells C, King NB, Knights D, Kong HH, Koren O, Koren S, Kota KC, Kovar CL, Kyrpides NC, La Rosa PS, Lee SL, Lemon KP, Lennon N, Lewis CM, Lewis L, Ley RE, Li K, Liolios K, Liu B, Liu Y, Lo C-C, Lozupone CA, Dwayne Lunsford R, Madden T, Mahurkar AA, Mannon PJ, Mardis ER, Markowitz VM, Mavromatis K, McCorrison JM, McDonald D, McEwen J, McGuire AL, McInnes P, Mehta T, Mihindukulasuriya KA, Miller JR, Minx PJ, Newsham I, Nusbaum C, O'Laughlin M, Orvis J, Pagani I, Palaniappan K, Patel SM, Pearson M, Peterson J, Podar M, Pohl C, Pollard KS, Pop M, Priest ME, Proctor LM, Qin X, Raes J, Ravel J, Reid JG, Rho M, Rhodes R, Riehle KP, Rivera MC, Rodriguez-Mueller B, Rogers Y-H, Ross MC, Russ C, Sanka RK, Sankar P, Fah Sathirapongsasuti J, Schloss JA, Schloss PD, Schmidt TM, Scholz M, Schriml L, Schubert AM, Segata N, Segre JA, Shannon WD, Sharp RR, Sharpton TJ, Shenoy N, Sheth NU, Simone GA, Singh I, Smillie CS, Sobel JD, Sommer DD, Spicer P, Sutton GG, Sykes SM, Tabbaa DG, Thiagarajan M, Tomlinson CM, Torralba M, Treangen TJ,**

27. **Altschul SF**, **Gish W**, **Miller W**, **Myers EW**, **Lipman DJ**. Basic Local Alignment Search Tool.

28. **Maxmen A**. 2021. One million coronavirus sequences: popular genome site hits mega milestone. Nature **593**:21–21. doi:10.1038/d41586-021-01069-w.

29. **Chamberlain S**, **Zhu H**, **Jahn N**, **Boettiger C**, **Ram K**. 2025. Rcrossref: Client for various 'CrossRef' 'APIs'.

30. **Sayers EW**, **Beck J**, **Brister JR**, **Bolton EE**, **Canese K**, **Comeau DC**, **Funk K**, **Ketter A**, **Kim S**, **Kimchi A**, **Kitts PA**, **Kuznetsov A**, **Lathrop S**, **Lu Z**, **McGarvey K**, **Madden TL**, **Murphy TD**, **O'Leary N**, **Phan L**, **Schneider VA**, **Thibaud-Nissen F**, **Trawick BW**, **Pruitt KD**, **Ostell J**. 2020. Database resources of the national center for biotechnology information. Nucleic Acids Research **48**:D9–D16. doi:10.1093/nar/gkz899.

31. **Muschelli J**. 2025. Rscopus: Scopus database 'API' interface.

32. Accessed 2025. Web of science. Internet Database, Clarivate.

33. **Mölder F**, **Jablonski KP**, **Letcher B**, **Hall MB**, **Tomkins-Tinch CH**, **Sochat V**, **Forster J**, **Lee S**, **Twardziok SO**, **Kanitz A**, **Wilm A**, **Holtgrewe M**, **Rahmann S**, **Nahnsen S**, **Köster J**. Sustainable data analysis with Snakemake. doi:10.12688/f1000research.29032.2.

34. **Wickham H**. 2025. Rvest: Easily harvest (scrape) web pages.

35. **Rinker TW**. 2018. textstem: Tools for stemming and lemmatizing text. Buffalo, New York.

36.   **Wickham H**, **Hester J**, **Ooms J**. 2025. xml2: Parse XML.

37.   **Benoit K**, **Muhr D**, **Watanabe K**. 2021. Stopwords: Multilingual stopword lists.

38.   **Feinerer I**, **Hornik K**, **Meyer D**. 2008. Text mining infrastructure in r. Journal of Statistical Software **25**:1–54. doi:10.18637/jss.v025.i05.

39.   **Feinerer I**, **Hornik K**. 2025. Tm: Text mining package.

40.   **Mullen LA**, **Benoit K**, **Keyes O**, **Selivanov D**, **Arnold J**. 2018. Fast, consistent tokenization of natural language text. Journal of Open Source Software **3**:655. doi:10.21105/joss.00655.

41.   **Topçuolu BD**, **Lapp Z**, **Sovacool KL**, **Snitkin E**, **Wiens J**, **Schloss PD**. 2021. mikropml: User-friendly r package for supervised machine learning pipelines. Journal of Open Source Software **6**:3073. doi:10.21105/joss.03073.

42.   **Kuhn**, **Max**. 2008. Building predictive models in r using the caret package. Journal of Statistical Software **28**:1–26. doi:10.18637/jss.v028.i05.

43.   **Wickham H**, **François R**, **Henry L**, **Müller K**, **Vaughan D**. 2025. Dplyr: A grammar of data manipulation.

44.   **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**. 2019. Welcome to the tidyverse. Journal of Open Source Software **4**:1686. doi:10.21105/joss.01686.

465

470   45.   **Topçuolu BD**, **Lapp Z**, **Sovacool KL**, **Snitkin E**, **Wiens J**, **Schloss PD**. 2021. Mikropml: User-Friendly R Package for Supervised Machine Learning Pipelines. Journal of Open Source Software **6**:3073. doi:10.21105/joss.03073.

46.   **Venables WN**, **Ripley BD**. 2002. Modern applied statistics with sFourth. Springer, New York.

47.   **Lenth RV**. 2025. Emmeans: Estimated marginal means, aka least-squares means.

48.   **Lüdecke D**. 2024. sjPlot: Data visualization for statistics in social science.