CONTEXTUALIZING ANTIMICROBIAL RESISTANCE DETERMINANTS

CONTEXTUALIZING ANTIMICROBIAL RESISTANCE DETERMINANTS USING DEEP-LEARNING LANGUAGE MODELS

By ARMAN EDALATMAND, BSc

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree of Master of Science

McMaster University MASTER OF SCIENCE (2022), Hamilton, Ontario (Biochemistry and Biomedical Sciences)

TITLE: Contextualizing antimicrobial resistance determinants using deep-learning language models

AUTHOR: Arman Edalatmand BSc

SUPERVISOR: Dr. Andrew G. McArthur, PhD

Number of pages: xiii, 110

Lay abstract

Antimicrobial resistance is a growing crisis worldwide, reducing the antibiotics at our disposal to treat bacterial infections. Resistance genes provide bacteria the ability to avoid the effectiveness of these antibiotics. Biomedical research publications contain information regarding these genes, informing researchers about the environmental, geographical, food, bacterial, and infection sources of these genes. Extracting this knowledge from text is an important task, typically laborious, and requires manual intervention. Using a series of machine learning techniques, I extracted this knowledge by first identifying relevant papers, selecting the associated sources and genes found in the text, and finally extracting the relationships where a gene was found in a particular source. With this information, we can summarize knowledge about resistance genes and better understand how genes move between different sources.

Abstract

Bacterial outbreak publications outline the key factors involved in uncontrolled spread of infection. Such factors include the environments, pathogens, hosts, and antimicrobial resistance genes involved. Individually, each paper published in this area gives a glimpse into the devastating impact drug resistant infections have on healthcare, agriculture, and livestock. When examined together, these papers reveal a story across time, from the discovery of new resistance genes to their dissemination to different pathogens, hosts, and environments.

My work aims to extract this information from publications by using the biomedical deep-learning language model, BioBERT. BioBERT is pre-trained on all abstracts found in PubMed and has state-of-the-art performance with language tasks using biomedical literature. I trained BioBERT on two tasks: entity recognition to identify AMR-relevant terms (i.e., AMR genes, taxonomy, environments, geographical locations, etc.) and relation extraction to determine which terms identified through entity recognition contextualize AMR genes. Datasets were generated semi-automatically to train BioBERT for these tasks. My work currently collates results from 204,094 antimicrobial resistance publications worldwide and generates interpretable results about the sources where genes are commonly found. Overall, my work takes a large-scale approach to collect antimicrobial resistance data from a commonly overlooked resource, i.e., the systematic examination of the large body of AMR literature.

Acknowledgements

I would like to sincerely thank my supervisor, Dr. Andrew G. McArthur, for allowing me to run with this project and providing an environment where I could grow as an individual and scientist. I have had the best time learning about a field I had never expected to experience without your support. Thank you to my graduate committee, Dr. Lori Burrows, Dr. John Nash, and Dr. Guillaume Paré, for your support and patience in listening as I explain natural language techniques at every committee meeting.

To all McArthur lab members, thank you for fostering such a welcoming environment since I first entered the lab and every day since. Amos, Brian, Kara, Jalees, Martins, William, Sohaib, Sally, and Rachel, thank you for being there whenever I needed to talk and supporting me throughout undergraduate school to where I am today. All of you have shaped me into who I am today, and I am unendingly grateful.

To my family and friends, thank you for your patience and support as I navigated graduate school. To my parents, Marjan and Mostafa, thank you for your unconditional love and for providing me the opportunity to pursue undergraduate and graduate school. The constant snacks were nice, too. To my sister, Roya, thanks for always being there for me. Most of all, thank you Christina for being by my side no matter what. I love you all.

Table of Contents

List of Tables

List of Figures

Abbreviations and Symbols

AMR	Antimicrobial resistance
API	Application programming interface
ARG	Antibiotic resistance genes
ARO	Antibiotic resistance ontology
BERT	Bidirectional encoder representations from transformers
BOW	Bag-of-words
CARD	Comprehensive Antibiotic Resistance Database
ENVO	Environmental ontology
FOODON	Food ontology
GAD	Genetic Association Database
GAZ	Gazetteer ontology
GEN	Generalization
GLASS	Global Antimicrobial Resistance and Use Surveillance System
IDO	Infectious disease ontology
LR	Logistic regression
MEM	Memory
NB	Naïve Bayes
NCBI	National Center for Biotechnology Information
NCBI TAXON	National Center for Biotechnology Information taxonomy
NER	Named entity recognition
NLTK	Natural language toolkit
OBO	Open Biological and Biomedical Ontology
P	Precision
PMID	PubMed ID

R	Recall
RE	Relationship extraction
RF	Random Forest
ROC	Receiver operator characteristic
SO	Sequence ontology
SVM	Support vector machine
TF-IDF	Term frequency-inverse document frequency
UBERON	Uber anatomy ontology
XGB	Extreme gradient boosting

Declaration of Academic Achievement

All natural language processing techniques and statistical analyses were performed by Arman Edalatmand.

I would like to thank Amos Raphenya, Brian Alcock, Jalees Nasir, Dr. Kara K. Tsang, Martins Oloni, William Huynh, Sohaib Syed, Marcel Jansen, Rachel Tran, and Dr. Andrew G. McArthur for validating paper classification predictions. I would like to thank Saduni Rajapaksa, Ramkrishna Upadhyaya, and Abdalmuhaymen Ibrahim for their labelling efforts in validating named-entity datasets and labelling relationship-extraction datasets. I would like to thank Pratap Ramamurthy from Google for providing insight into alternative methods of extracting relationship data.

This research was supported by an Ontario Graduate Scholarship and a McMaster University MacData Institute Graduate Fellowship.

Chapter 1: Introduction

1.0 Antimicrobial resistance databases

Antimicrobial resistance (AMR) is a growing crisis that threatens the use of life-saving therapeutics for bacterial infections and prophylactic treatment of surgical patients. AMR occurs when medicines commonly used to treat microbial infections are no longer effective due to drug-resistance. In Canada, it is estimated that ~250,000 infections were resistant to first-line antibiotics in 2018, with 5,400 deaths directly caused by AMR[1]. By 2050, if the rate of AMR remains the same at 26% or increases to 40%, it is estimated that AMR would cost the Canadian economy between $13 to $21 billion each year and take 7,000-13,700 Canadian lives[1]. Worldwide, the 2050 outlook on AMR is more grim; an estimated 10 million deaths from AMR and $100 trillion in economic losses if action is not taken to control the spread of AMR[2]. Worldwide in 2019, 1.27 million deaths were caused by AMR[3]. A significant contributor to the dissemination of resistance is bacteria's ability to transfer and acquire antimicrobial resistance genes (ARGs) via horizontal gene transfer[4]. Novel determinants of resistance can be selected for and transferred to other bacteria through this process. Yet, with improvements in DNA sequencing technologies[5], surveying and identifying novel resistance determinants contributing to AMR has improved[6].

Genomics tools have been developed to identify resistance genes and mutations found in bacterial pathogens[7]. Such tools rely on antimicrobial resistance databases to

provide reliable reference information, i.e., to know how similar a putative resistance gene found in an infection is to known ARGs at the nucleotide or protein sequence level. Antimicrobial resistance databases are resources that store information about antimicrobial resistance determinants, such as beta-lactamases or point mutations in gyrases found in bacteria. Many resources are available, but most notable are CARD[8,9] (the Comprehensive Antibiotic Resistance Database), ResFinder[10], ARG-ANNOT[11], and the National Center for Biotechnology Information (NCBI) reference gene catalogue[12]. The basic information in these databases is the genomic sequences of resistance determinants and the impacted drug classes. The pathogens that contain these determinants are also reported by CARD and NCBI. By having a unifying set of reference resistance determinant sequences in these databases, we can reliably annotate the ARGs within bacterial isolate genomes or more complex metagenomes. Both CARD and NCBI contain pathogen-tracking resources that assess the ARGs of pathogenic bacteria. NCBI goes beyond CARD by including the geographical location, upload date, host, environment type, and strain; however, they do not analyze or summarize these data and leave them in a simple table format. Overall, these tools minimally examine the contextual background of resistance determinants, yet understanding the epidemiology of resistance determinants can be informative in making clinical and public health decisions.

1.1 Context for understanding antimicrobial resistance

The spread of AMR is due to a complex intertwining of environmental, animal, food, human, and industrial sources contributing to the dissemination of resistance[13].

Understanding the sources of resistance determinants is essential when assessing and managing risk[14]. However, this epidemiological information about resistance determinants and the bacteria that contain them goes beyond the data reported in the databases. Although resistance databases do not report such contextual information, work has been done to gather these data on a broader scale. Previous work has used genome sequencing and associated metadata to analyze the geographical distribution and mobility of resistant pathogens and resistance determinants[15–18]. These studies examined small samples of resistance and did not capture the totality of resistance determinants found in CARD. However, one central resource that analyzes global AMR alongside its epidemiology is Resistome Tracker[19]. Using sequencing data from NCBI, they created visualization summaries of four bacterial groups based upon their geographical locations, the host type (i.e., food, animal), and environment (i.e., farm, wastewater, clinical). Using AMRFinderPlus[20], they annotate the resistance genes in these samples and generate visualizations showing the genes' distribution across different environments, hosts, and countries. This gives users an idea of the prevalence of resistance across different environments and hosts. The main drawback with their resource is that it is limited to isolates for only four groups of bacteria: *Escherichia coli*, Non-typhoidal *Salmonella*, *Enterococcus,* and *Campylobacter*. A similar platform that reports the geographical location of isolates along with their resistome and epidemiological metadata is Pathogenwatch[21]. With this platform they analyzed thousands of *Neisseria gonorrhoeae* isolates, their resistance distribution around the world and their susceptibility patterns[22]. Unlike Resistome Tracker, there is no gene-level analysis of resistance and only a broad

pathogen-to-geographical location analysis; aside from geographical representation, no other metadata in Pathogenwatch is examined. More recently, a study examining ~214,000 metagenomic samples explored the relative abundance of ARGs across different years, geographical locations, and hosts[23]. Currently, this data has only briefly been used to examine resistance on a broad scale. In the future, this data can be used to explore the distribution of individual ARGs.

Except for a subsection of Resistance Tracker and the metagenomic dataset, all the tools, resources, and papers mentioned previously consider AMR through a pathogen-centric view; they take genomic sequences, identify the resistance genes, and examine where isolates are located geographically and environmentally. This allows for high-resolution identification of pathogen transmission across populations. However, this does not reveal much about the long-term transmission patterns of individual ARGs, especially in the context of horizontal gene transmission. Thus, a more gene-centric view must be taken to understand how individual genes disseminate throughout the world.

An alternative to sequencing metadata is the publications associated with resistance determinants. Contained in peer-reviewed, published papers describing resistance determinants are details revealing the epidemiological background these determinants are found within. The pathogen source, geographical location, environmental source, food source, and infection sites are all details commonly mentioned in publications when describing where a resistance determinant was located. This includes publications reporting the discovery of a novel determinant, outbreak studies examining the dissemination of resistance, and review articles. This information is

waiting to be collated so that we can gain a holistic understanding of the epidemiological background of resistance determinants. In addition to the epidemiological information provided by papers, there is also a temporal aspect to papers. A paper's publication date reveals when a resistance determinant is discovered. This means we can begin to analyze trends in the dissemination of resistance solely through publications. However, we must employ several biomedical text mining approaches outlined below to extract important epidemiological information about ARGs.

1.2 Biomedical text classification

As of 2022, over 33M publications are stored across PubMed databases corresponding to over 120 gigabytes of data[24]. Compared to previous years, the number of publications added to PubMed each year has grown rapidly (Figure 1A). Compared to 2000, 3.3 times more articles were added to PubMed in 2021. Importantly, the number of AMR papers added each year has grown 3.9 times since 2000 and is growing faster than the PubMed baseline (Figure 1B). Manually reviewing every publication to extract new data is a time-consuming and expensive process[25]. We can reduce the human effort needed and off-load work to computer models using automated text mining methods. However, we must first identify the relevant sources before extracting knowledge from the literature. When dealing with millions of papers, filtering out irrelevant papers is essential to improve the speed of downstream processes. The method of identifying sets of similar papers is considered classification or clustering depending on whether the approach is supervised or unsupervised, respectively. Unsupervised methods of clustering

are exploratory approaches to understanding structure across biomedical text[26]. Although useful for cases where no desired set of similar publications is required, we will primarily focus on supervised classification methods. As the name implies, supervised classification requires human supervision by providing a set of known publications to identify similar publications using the text within those publications. Several publicly available web applications can perform text classification on PubMed articles given a set of PubMed IDs (PMIDs)[27–29]. However, neither LitSuggest nor MedlineRanker open-sources their code for reproduction. Applications of text classification in the biomedical field range from triaging publications for database curation[30–32], compiling publications for review[33–35], or general user querying.

1.3 Natural language via artificial neural networks

There are several levels of information we can retrieve from text. In this work, I use named entity recognition (NER) and relationship extraction (RE) to accurately extract epidemiological information associated with ARGs from published papers in PubMed. NER aims to extract terms that fall under the same label, e.g., extracting all gene names from a text. This has commonly been a challenging task because we must consider term synonyms and semantics; a term may reference different concepts based on context. NER has been performed on tasks of gene[36–38], chemical[39], and protein[40] name extraction from biomedical literature. Once NER is complete, RE takes terms within a piece of text and examines if they are related to one another. This method has commonly been used to identify protein-protein[41–43] and drug-drug[44,45] interactions within the published literature.

In this work, I will perform NER and RE upon the antimicrobial resistance literature via natural language deep learning methods.

Artificial neural networks are a method of predictive modelling by mimicking biological neurons[46]. Like neurons, each node in a network receives a set of inputs. Each input is assigned a weight and the total of all weights are fed into an activation function that calculates the node's output[46]. The goal is to find a set of weights that provide an output close to the desired output through model training. Back-propagation acts as a way of updating weights based on the error found in the network. Neural networks have been applied to various fields, from science and engineering to policy and insurance[47]. The general type of artificial neural network that takes sentences and converts them into another sequence of words is called Seq2Seq. Seq2seq models consist of 2 parts: an encoder and a decoder. The encoder takes a sentence input and tries to learn its context, and the decoder takes the context output of the encoder and returns its predicted output**.** For example, an encoder could take the sentence "Charlie was hungry, so he ate food," and give the context to the decoder that the pronoun "he" refers to "Charlie".

A new and robust player in the field of artificial neural networks is the transformer model. Prior models like recurrent neural networks followed a sequential understanding of language. This meant, given a sentence, the further away words are from each other, the less context they gain from each other. In transformer models, a self-attention mechanism is added that takes a word and examines which parts of the sentence are important for that word[48]. This is done through the following formula:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

The main components of this formula are Q, the query vector you give the network; K, the key vector, which contains all the words in a sentence; and V, the values associated with those keys. The attention formula calculates a score between the query word and each key by taking the dot-product between the two (QK^T), scaling it down by the shape of the key vector ($\sqrt{d_k}$) and normalizing the value via softmax. Finally, the resulting softmax result is multiplied by the value vector (V) of each key. A weighted output sum is taken to obtain each word's relative importance on another in a sentence.

The original problem transformer models aimed to solve was machine translation of human language[48], and it worked well because the context of the entire sentence is provided through encoding. However, different training methods are needed if you want to perform NER and RE tasks. The creation of BERT[49] (Bidirectional Encoder Representations from Transformers) solves these tasks through a creative method of training. Training involves a combination of "masked" word prediction alongside next sentence predictions to handle multiple NLP tasks. Masked word predictions take 15% of the words in a sentence and (1) hides, or "masks" it 80% of the time; (2) replaces the word with a random word 10% of the time; or (3) does not change the word 10% of the time. At the same time, BERT is trained to predict if sentence A is followed by sentence B. In the training data, 50% of the time B follows A, and in the remaining 50% B is a random sentence. BERT then predicts each masked term and whether the sentences are real pairs. BERT then backpropagates to optimize the weights using an algorithm called

Adam[50]. Backpropagation optimizes weights by minimizing the error in a model's predictions, also known as a model's "loss". Loss is calculated every training iteration until the maximum number of iterations is reached or when additional training does not reduce a model's loss, otherwise known as model convergence. Although a model may have reached convergence, indicating that it cannot improve loss via additional backpropagation, this does not mean that the model has converged on the global minimum loss possible, but may instead be algorithmically stuck at a local minimum.

The result of training BERT is a model that understands the context in sentences and that can be fine-tuned for specific tasks like NER and RE. The original BERT model was trained on Wikipedia and books, yielding a model with 110 million parameters. BERT can be further trained on subsequent data to make domain-specific models. In this project, I use BioBERT, a biomedically trained version of BERT[51] that takes the trained BERT model and further trains it on PubMed abstracts and full-text articles from PubMed Central. My goal will be to have a trained version of BioBERT that can accurately extract epidemiological and resistance determinant terms and relationships from biomedical publications, e.g., aph(6)-Id is associated with food-borne pathogens.

1.4 Summary of intent

This work seeks to test two hypotheses:

1. Publication mentions of ARGs and their epidemiological background accurately reflect surveillance data.
2. Gene, environment, and transmission risks can be accurately predicted using publication text.

If the null hypotheses of this project can be rejected with statistical significance, then CARD would gain an additional resource that would accurately reflect the epidemiological landscape of resistance beyond what can currently be known through sequencing metadata reports. Additionally, reporting of high-risk genes, the risk they pose to environments, and the transmission risks associated with these genes between environments would be valuable when trying to identify priority hotspots.

The main Aims of my thesis are to:

1. Use BioBERT natural language processing to annotate epidemiological and AMR terms and extract their contextual relationships based upon the published scientific literature;
2. Create a metric to assess the strength of these relationships over publication history;
3. Predict the risk associated with genes, their environments, and their risk of mobilizing to another environment.

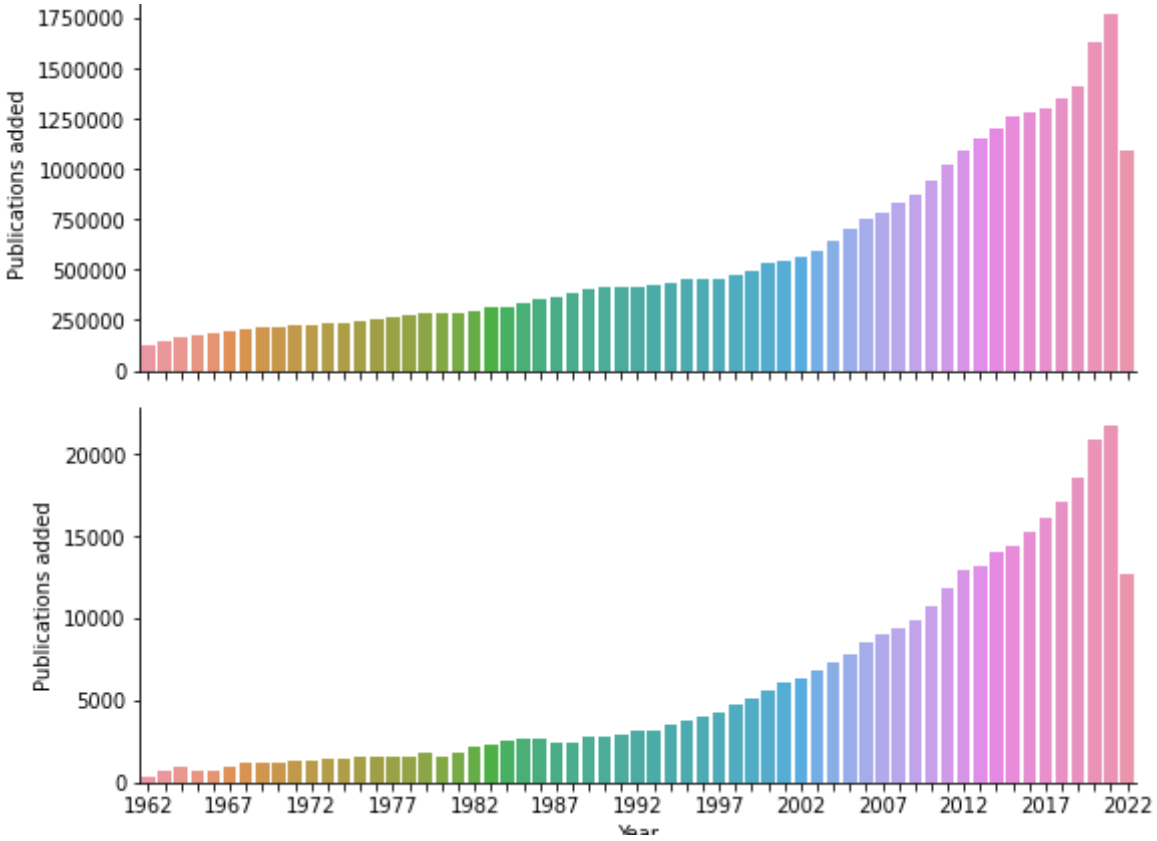


Figure 1.1 Distribution of publications added to PubMed each year. (A) Total number of publications added to PubMed since 1962, (B) subset of PubMed articles under the PubMed query “antimicrobial resistance”.

Chapter 2: Identifying relevant publications using paper classification

2.1 Introduction

Paper classification is not only an important tool for assisting researchers in identifying relevant publications in their field of work, but for biomedical databases that rely on triaging thousands of papers to identify relevant publications. Individuals may try to generate queries using keywords related to a particular topic, but on large scales where thousands of publications are added to PubMed daily, automated classification methods are needed[30,52].

Supervised labelling techniques involve using prelabeled data to generate a model that predicts labels on new data. Several models that have seen good performance in paper labelling include logistic regression, naive Bayes, random forest, boosting methods, and support vector machines[53,54]. Logistic regression is arguably the simplest model, using a logistic function to make predictions on a binary set of labelled data, i.e., relevant versus not relevant. The parameters provided to the logistic function are the words, called features, in a paper's abstract. Naïve Bayes text classifiers create predictions relative to the probability distribution of terms found across different labels, assuming that terms are independent of one another. Random forest models average several decision trees together; a decision tree model generates a flowchart of binary choices based on features and associated labels. Boosting methods are like random forest in that they combine

several weak models to generate a more robust model. Unlike random forest, however, boosting models are generated by using sequential training of weak learners where the output of one model is reweighted and passed to another. The boosting method examined in my thesis was extreme gradient boosting[55], which optimized the boosting algorithm's efficiency. Finally, support vector machines create a decision boundary to maximize the distance between two groups of data[56].

This chapter aims to explore the use of classification algorithms to accurately and reliably identify AMR-epidemiological papers in PubMed. If successful, the set of papers the classifier identifies is hypothesized to contain relevant named-entity and relationship data for ARGs from which we can begin to train NER and RE models (Chapter 3).

2.2 Methods

2.2.1 Paper retrieval

A visual flowchart of all methods (sections 2.2.1 to 2.2.7) can be found in Figure 2.1. First, to extract knowledge from papers, we must have a source of publications from which to extract and a set of terms representing the information we want to extract. Scholarly papers are available at publisher websites along with databases that compile publications from several journals. With a focus on the biomedical realm, PubMed is one of the most extensive publication databases worldwide, with over 33 million papers across >5,200 journals[57]. The two major components making up PubMed include MEDLINE and PubMed Central. MEDLINE accounts for the largest proportion of PubMed's database and involves 27 million publications. When accessing MEDLINE

papers, only their title, abstract, and publication information are available. To access full-text papers, the second-largest portion of PubMed, PubMed Central, is freely available.

PubMed allows users to access data through a web view or an application programming interface (API). APIs allow for high-throughput programmatic communication between a service and end-users. The API provided by PubMed allows for the rapid downloading of publications. Information retrievable from the PubMed API includes the publication date, journal name, abstract, and, for publications found in PubMed Central, the body of a paper. Using custom Python scripts and the Entrez library to communicate with PubMed, I downloaded over 5 million publications from PubMed between 2017 to 2020 to build my classifier.

2.2.2 Paper preprocessing

Biomedical text is unstructured and heterogeneous, requiring preprocessing to standardize input data for machine-learning applications. Text preprocessing helps normalize and reduce redundancy within the text. Preprocessing is an important step in improving model performance by removing non-essential terms and digits from the text, removing punctuation, and converting terms to their base form by removing suffixes. My custom preprocessing pipeline, written in Python, for all papers includes:

1. Converting sentences to lowercase;
2. Removing punctuation using regular expressions;
3. Converting words into their base form using a Porter stemmer;
4. Removing stop words;

5. Removing digits from abstracts

The Natural Language Toolkit (NLTK)[58], regular expressions, and Python's built-in string library allow for this preprocessing to occur. The NLTK provides access to the Porter stemmer[59], which converts words to their root form. For example, the word "relational" would be converted to "relate." Additionally, NLTK provides a list of stop words that can filter overly common and uninformative words in sentences, e.g., "the". Removing digits and lowercase conversion is easily performed using Python's built-in string function string.lower() and string.isdigit(). Regular expressions, which are standard computer science tools to find matches to search terms in text, were used to remove punctuation characters.

2.2.3 Feature extraction

Once preprocessed, the procedure of converting data into usable vectors is called feature extraction. Feature extraction aims to reduce the dimensionality of data for algorithmic use. Two major types of dimensionality reduction applied to text include bag-of-words (BOW) and term frequency-inverse document frequency (TF-IDF). BOW methods count the total number of term occurrences across all papers, e.g., the *CTX-M-15* gene name is found 208 times in 10,783 downloaded PubMed abstracts. The TF-IDF method scales BOW counts based on how uncommon a word is across all documents, with the TF-IDF formula for a word (t) in document (d) found below.

$$\text{TF-IDF} = TF * IDF$$

$$IDF = \log\left(\frac{1+n}{1+df(t)}\right) + 1$$

The term-frequency (TF) is a simple BOW count of all term occurrences. The inverse document frequency (IDF) takes the total number of documents (n) and the number of documents the term appears in (df(t)). A value of 1 in the numerator and denominator of the formula prevents zero divisions from occurring. The 1 addition outside the logarithm allows terms that appear in all documents to have a weight of one. From the formula, very common terms are scaled less than uncommon terms. In my work, two types of TF-IDF were performed: TF-IDF on unigrams (i.e., individual words) and TF-IDF on bi- and tri-grams (i.e., pairs and triplets of words). For example, in the sentence "Charlie is a cat," the possible bigrams are ("Charlie", "is"), ("is", "a"), and ("a", "cat").

2.2.4 Model cross-validation

To avoid examining non-informative papers for extraction of AMR gene epidemiological information, I explored five different machine learning models to see which could best predict which papers had the highest value in terms of AMR gene epidemiological information. Every combination of machine learning model and feature extraction method was explored. To train a machine learning model to identify papers that contain epidemiological information, I used all papers found under the "'Drug Resistance, Bacterial/genetics' [Mesh]" query in PubMed and filtered them down to contain papers with a country mention. This set would serve as our positive set for model training/testing. My assumption is that papers that mention drug resistance and a country

name will contain additional epidemiological information. For the negative set, I randomly generated PubMed paper IDs and downloaded those papers. This yielded 10,532 negative and 3,843 positive papers (Table 1.1) for training and testing. 75% of these papers were used for cross-validation, training, and tuning, as outlined below. The remaining 25% were held out for a final evaluation of the models' performance on data they did not use during training.

To test the various machine learning models against each other, stratified 5-fold cross-validation was performed. The purpose of cross-validation is to reduce the probability that a model will overfit on a dataset. Overfitting occurs when a model is trained for too long on a dataset and cannot generalize well to unseen data because the model learned the "noise" associated with the training dataset, i.e., it can only be accurate on the data upon it was trained. By splitting a dataset into 5 'folds', we reduce the innate bias an entire dataset contains with a potential increase in variance, i.e., each fold contains $1/5^{th}$ of the papers in the training set. For each split, four parts of the data are used in generating features via feature extraction - these features are then used for model training. The remaining one part is used as testing, i.e., does the model trained on the other 4 folds accurately classify the papers in the 5^{th} fold. This process is repeated for the number of k-folds. Through stratified cross-validation, we ensure that the data is represented equally by selecting a portion of each label based on its relative proportion. For example, given 100 green apples and 10 red apples, 5-fold stratified cross-validation would select 20 green apples and 2 red apples for each fold of validation. This prevents sampling 22 green apples and 0 red apples if simply choosing at random.

2.2.5 Curator validation

The above model development was based on a comparison of random papers against papers tagged with AMR MeSH terms. With these models in place, we would like to evaluate model performance using unbiased real-world data, as model validation is an important step to verify the performance of a model against data it has never seen. After cross-validation, models were used to make predictions on papers published in September and November 2019, while a 430-paper subset from September and November 2019 were additionally and independently examined by human curators in the McArthur lab. For the human curators, 100 papers were randomly selected from logistic regression model predictions of September papers, with an even proportion of positive to negative predictions (i.e., 50 positive predictions and 50 negative predictions), with the remaining 330 papers selected from November predictions. Half the November papers were selected based on naïve Bayes predictions and half from random forest model predictions (both with an even 50/50 split between positive and negative predictions). These three models used for paper sub-setting had been trained on TF-IDF unigram features of preprocessed abstracts. The human curators were given three true/false questions on whether: (1) the paper contained an AMR gene reference, (2) if a geographical location was mentioned in the abstract, and (3) if any other epidemiological information was mentioned. Each paper was assigned two curators; if curator results disagreed, the paper was ignored. These results were compared against computational model predictions for an independent real-world evaluation.

2.2.6 Hyperparameter tuning through grid-search cross-validation

Once a final model has been confirmed to perform well through curator validation, the model's hyperparameters must be tuned. Hyperparameters are options that can be changed to fine-tune the performance of a model. The model was tuned based on hyperparameters found in Table 2.5. Grid-search cross-validation performs 3-fold cross-validation across every hyperparameter combination to identify the best-performing combination. Once the best hyperparameter set has been found, it then refits the model on the entire dataset.

2.2.7 Holdout validation

Once a final model and feature extractor pair was obtained and optimal hyperparameters determined, the model was then tested against the holdout set of 3,594 papers not used for model training, cross-validation, or tuning.

2.3 Results

Five models were tested through cross-validation to identify which one can reliably identify papers containing AMR and epidemiological information. Cross-validation resulted in a high receiver operator characteristic (ROC) curve area under the curve of >97% (Figure 2.2, 2.3). ROC curves measure a model's ability to distinguish between positive and negative papers. The larger the ROC area under the curve, the better a model can distinguish between negative and positive papers. These curves typically over-predict performance in unbalanced datasets (i.e., datasets with a large difference between positive and negative items). Thus, a better indication of performance would be

precision, recall, and F1 scores. Precision represents a model's ability to reliably identify positive papers without mistake, i.e., few false positives, while recall reveals a model's ability to identify all positive papers correctly, i.e., not miss any relevant papers. Optimally, we would desire a model with both high recall and high precision and the F1 score is a combined measure of precision and recall reflective of a model's overall accuracy. When examining these metrics, the models still perform quite well (>90% F1 score) except for the support vector machine (Table 2.2). Overall, there seemed to be no significant difference between models generated on preprocessed text compared to non-preprocessed text (i.e., removing punctuation, converting words to their base, etc.). When making predictions on papers from September and November 2019, you can see the impact low precision has on the number of positively predicted papers (Table 2.3). The lower the precision value, the more papers a model will believe are positive, opening the possibility of false predictions and noise in downstream steps.

Since all models performed well, we wanted to tease apart the difference between each model. Thus, we created an external validation set to assess model performance using volunteer human curators. The performance of the machine learning models varied much more in this validation. Yet again, the performance difference between models provided by preprocessed text compared to non-preprocessed text is minor (Table 2.4). Of all the models, logistic regression and naïve Bayes with the TF-IDF bi and tri-gram feature extraction method performed the best with an F1 score of 56% and 60%, respectively (Table 2.4). However, when looking back at Table 2.2, the recall value of both these models was low at 87% and 88%. Low recall values are undesirable as we

would like to capture as many true positives as possible. The best-performing model from validation with a high recall value in Table 2.2 is logistic regression using TF-IDF unigram features. Performance-wise, there is no major difference between this model's preprocessed and non-preprocessed versions (Tables 2.2 and 2.4). The only difference lay in the number of positive predictions made on the September and November 2019 papers as the preprocessed version selected 142 more positive papers (Table 2.3). As such, the logistic regression model using TF-IDF unigram on preprocessed features was selected as the final model for the possibility of identifying papers missed by the non-preprocessed version.

The logistic regression hyperparameters were fine-tuned using the parameters in Table 2.5. This generated 2,880 combinations across 3-fold cross-validation for a total of 8,640 iterations. When testing the final model on the remaining 25% holdout validation set, this model performed exceptionally well with precision, recall, and F1 values of 99%.

2.4 Discussion

The goal of paper classification was to create a model that can identify as many publications containing AMR gene epidemiology information as possible, without concern for false positives. Overall, the evidence suggests that the logistic regression model trained on preprocessed abstracts with TF-IDF unigram features will flag nearly all papers with AMR gene epidemiological content in PubMed. These false positives will be removed during subsequent labelling steps if they are found to not contain relevant information.

With this classifier complete and deemed reliable, the next task is to train BioBert models to perform NER and RE on PubMed. As outlined in the next chapter, I selected all AMR papers from 2019 as the BioBert training set. To prepare this training set, the logistic regression classifier was used to make predictions on all 1,079,050 million papers added to PubMed for 2019. This yielded 10,784 papers the model believed to contain epidemiological information about AMR. Using this set of papers, in the next chapter I create datasets to train and test NER and RE models in the task of identifying ARGs and epidemiology terms and extracting the relationships that exist between these terms.

Table 2.1: Paper dataset used for text classification training/testing. The PubMed query "'Drug Resistance, Bacterial/genetics'[Mesh]" was used to extract 16,198 positive papers. The negative paper set was obtained by randomly generating PubMed Ids and querying them. Positive papers were filtered out if they did not contain a country name from the international standard for country names and codes (ISO 3116).

	All papers	**Filtered papers**
Negative	10,531	/
Positive	16,198	3,843

Table 2.2: Precision, recall, and F1 performance of classification models during cross-validation. Highlighted in yellow is the final model selected for future use.

Model	Feature extraction method	Processed text	Precision	Recall	F1
Extreme gradient boosting	TF-IDF Bi/tri-gram	F	0.97	0.89	0.93
		T	0.98	0.85	0.91
	TF-IDF Word	F	0.97	0.98	0.97
		T	0.97	0.97	0.97
	Word count	F	0.97	0.97	0.97
		T	0.97	0.97	0.97
Logistic regression	TF-IDF Bi/tri-gram	F	0.99	0.88	0.93
		T	0.99	0.87	0.92
	TF-IDF Word	F	0.98	0.96	0.97
		T	0.98	0.96	0.97
	Word count	F	0.98	0.97	0.97
		T	0.98	0.97	0.97
Naive Bayes	TF-IDF Bi/tri-gram	F	0.99	0.88	0.93
		T	0.98	0.96	0.97
	TF-IDF Word	F	0.98	0.95	0.97
		T	0.98	0.96	0.97
	Word count	F	0.95	0.99	0.97
		T	0.95	0.99	0.97
Random forest	TF-IDF Bi/tri-gram	F	0.96	0.93	0.94
		T	0.97	0.93	0.95
	TF-IDF Word	F	0.97	0.96	0.96
		T	0.97	0.95	0.96
	Word count	F	0.97	0.95	0.96
		T	0.97	0.95	0.96
Support vector machine	TF-IDF Bi/tri-gram	F	NaN	0.00	0.00
		T	NaN	0.00	0.00
	TF-IDF Word	F	NaN	0.00	0.00
		T	NaN	0.00	0.00
	Word count	F	0.98	0.75	0.85
		T	0.98	0.82	0.89

Table 2.3: Positive and negative predictions of models for September and November 2019 papers.

Model	Feature extraction method	Processed text	Negative	Positive
Extreme gradient boosting	TF-IDF Bi/tri-gram	F	239,250	2,555
		T	239,508	2,297
	TF-IDF Word	F	238,494	3,311
		T	238,818	2,987
	Word count	F	238,747	3,058
		T	238,905	2,900
Logistic regression	TF-IDF Bi/tri-gram	F	240,376	1,429
		T	240,381	1,424
	TF-IDF Word	F	240,035	1,770
		T	239,893	1,912
	Word count	F	239,379	2,426
		T	239,309	2,496
Naive Bayes	TF-IDF Bi/tri-gram	F	240,723	1,082
		T	239,896	1,909
	TF-IDF Word	F	240,156	1,649
		T	240,053	1,752
	Word count	F	237,198	4,607
		T	237,252	4,553
Random forest	TF-IDF Bi/tri-gram	F	238,020	3,785
		T	238,470	3,335
	TF-IDF Word	F	238,347	3,458
		T	238,850	2,955
	Word count	F	238,330	3,475
		T	239,035	2,770
Support vector machine	TF-IDF Bi/tri-gram	F	241,805	0
		T	241,805	0
	TF-IDF Word	F	241,805	0
		T	241,805	0
	Word count	F	240,510	1,295
		T	239,920	1,885

Table 2.4: September and November validation performance of classification models. Predictive performance for each model against a set of 381 human-curated papers. Highlighted in yellow is the final model selected for future use chosen based on recall performance during cross-validation seen in Table 2.2 and overall performance during curator validation.

Model	Feature extraction method	Processed text	FN	FP	TN	TP	Precision	Recall	F1
Extreme gradient boosting	TF-IDF Bi/tri-gram	F	1	66	289	25	0.27	0.96	0.43
		T	2	53	302	24	0.31	0.92	0.47
	TF-IDF Word	F	0	86	269	26	0.23	1.00	0.38
		T	0	79	276	26	0.25	1.00	0.40
	Word count	F	0	79	276	26	0.25	1.00	0.40
		T	0	71	284	26	0.27	1.00	0.42
Logistic regression	TF-IDF Bi/tri-gram	F	1	49	306	25	0.34	0.96	0.50
		T	1	38	317	25	0.40	0.96	0.56
	TF-IDF Word	F	0	58	297	26	0.31	1.00	0.47
		T	0	59	296	26	0.31	1.00	0.47
	Word count	F	1	70	285	25	0.26	0.96	0.41
		T	1	73	282	25	0.26	0.96	0.40
Naive Bayes	TF-IDF Bi/tri-gram	F	1	33	322	25	0.43	0.96	0.60
		T	1	59	296	25	0.30	0.96	0.45
	TF-IDF Word	F	1	58	297	25	0.30	0.96	0.46
		T	0	64	291	26	0.29	1.00	0.45
	Word count	F	0	122	233	26	0.18	1.00	0.30
		T	0	124	231	26	0.17	1.00	0.30
Random forest	TF-IDF Bi/tri-gram	F	0	84	271	26	0.24	1.00	0.38
		T	0	82	273	26	0.24	1.00	0.39
	TF-IDF Word	F	0	81	274	26	0.24	1.00	0.39
		T	1	73	282	25	0.26	0.96	0.40
	Word count	F	0	86	269	26	0.23	1.00	0.38
		T	0	81	274	26	0.24	1.00	0.39
Support vector machine	TF-IDF Bi/tri-gram	F	26	0	355	0	NA	0.00	0.00
		T	26	0	355	0	NA	0.00	0.00
	TF-IDF Word	F	26	0	355	0	NA	0.00	0.00
		T	26	0	355	0	NA	0.00	0.00
	Word count	F	2	45	310	24	0.35	0.92	0.51
		T	2	61	294	24	0.28	0.92	0.43

Table 2.5: Hyperparameters tuned for logistic regression. To identify the best hyperparameter values, all combinations of the values column were tested through 3-fold cross-validation. The best-performing parameters are listed under the "Final parameter" column.

	Parameter	Values	Final parameter
Logistic regression	Penalty	L1, L2	L1
	Solver	liblinear, saga, newton-cg, lbfgs, sag	Liblinear
	C	0, 0.001, 0.006, 0.046, 0.359,2.783, 21.544, 166.810, 1,291.55, 10000	10000
	max_iter	10,50,100,150	50
	random_state	100	100
Count vectorizer	max_df	0.5,0.75,1	0.75
	max_features	None, 5000, 10000, 50000	None

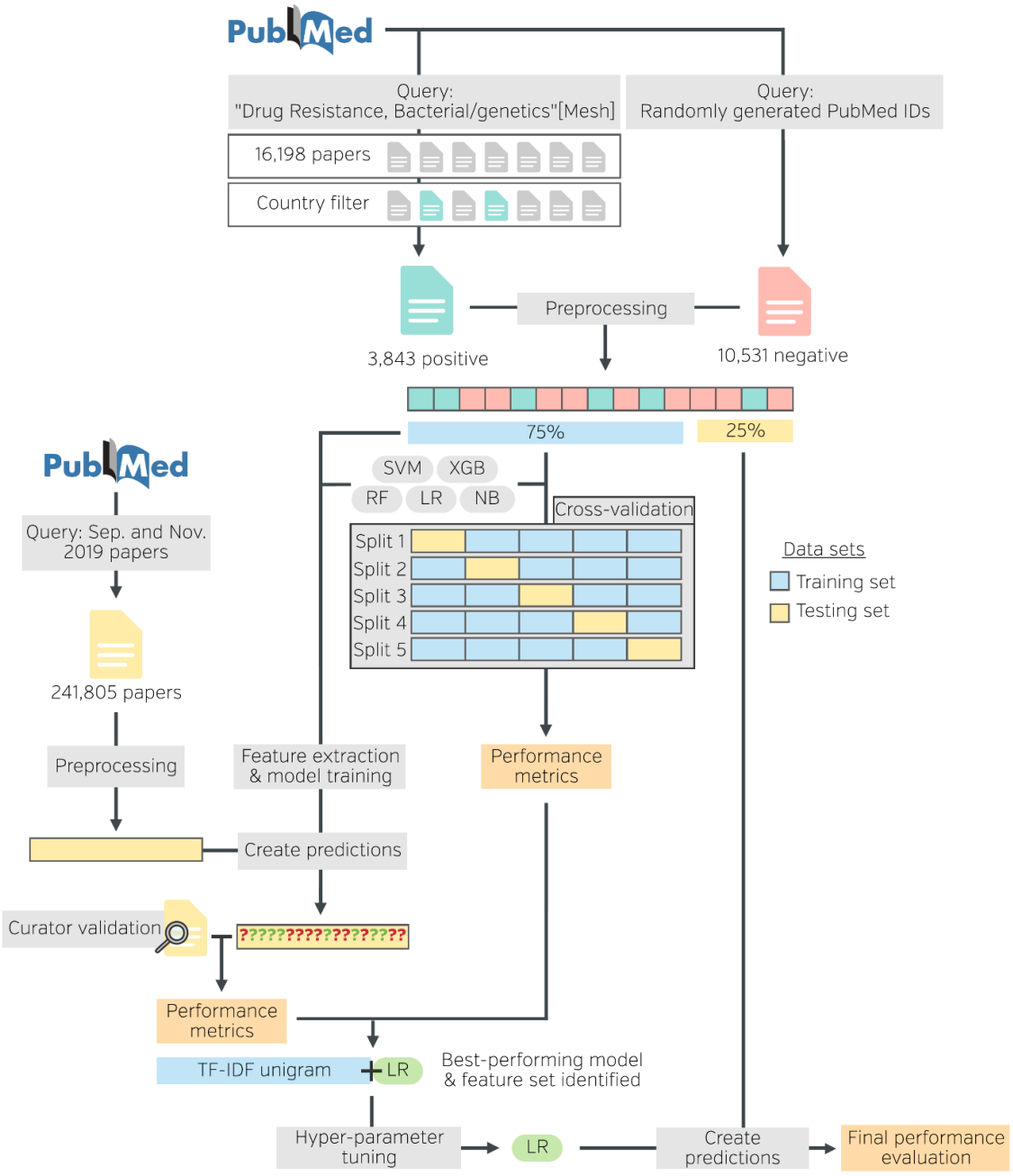


Figure 2.1: Flowchart of generating a paper classification model. Papers are first obtained via PubMed queries for the purpose of training machine learning models. A positive set of papers is obtained using the "'Drug Resistance, Bacterial/genetics'[Mesh]" query and filtered down to contain only papers mentioning a country name. Negative papers are

obtained by randomly generating PubMed IDs. These papers were preprocessed and split 75:25 between a training set and a final evaluation set. All combinations of five models and three feature extraction methods were examined through cross-validation. For each cross-validation split, four folds of the data, shown in blue, are processed via a feature extraction method to generate a series of vectors on which the model is trained. The remaining fold, shown in yellow, is tested upon by the model. This continues for the remaining four splits to gain an idea of model performance. Each model-feature extraction pair was also trained on the original 75% of the data, which a feature extraction method converts to features. These models were used to predict papers published in September and November 2019. A group of curators validated a 430-paper subset of the 240 thousand predicted papers to generate additional performance metrics. Based on the performance metrics provided by cross-validation and curator validation, logistic regression trained on TF-IDF unigram features was chosen as the final model. The parameters of this model were tuned and used to create predictions on the 25% holdout set for a final performance evaluation. The models shown here are support vector machine (SVM), extreme gradient boosting (XGB), random forest (RF), logistic regression (LR), and naïve Bayes (NB). The three feature extraction methods used were BOW, TF-IDF unigram, and TF-IDF bi/trigram.

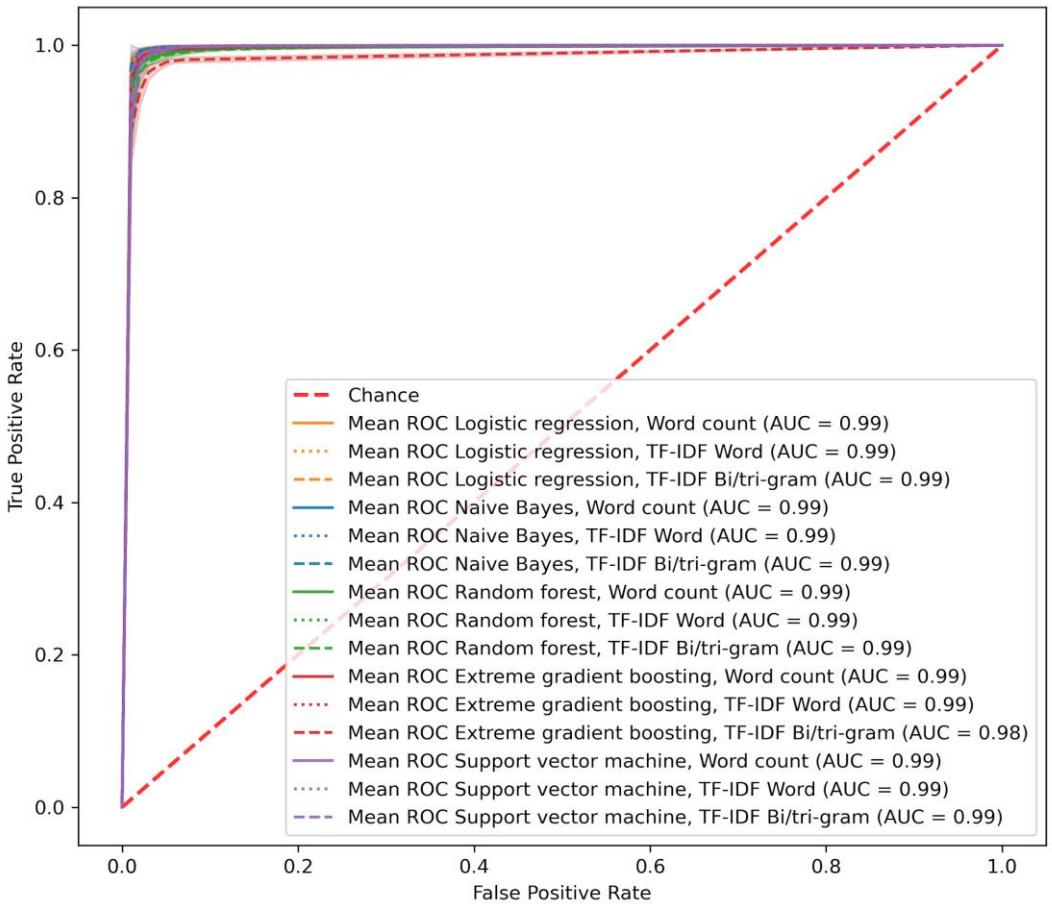


Figure 2.2: Receiver operator characteristic curve from 5-fold cross-validation of classification models trained on preprocessed abstracts using three feature extraction methods. Results from all five cross-validation tests were averaged to produce a single curve. Shadows around each line are +/- 1 standard deviation from the mean. Models were trained on abstracts that had been preprocessed.

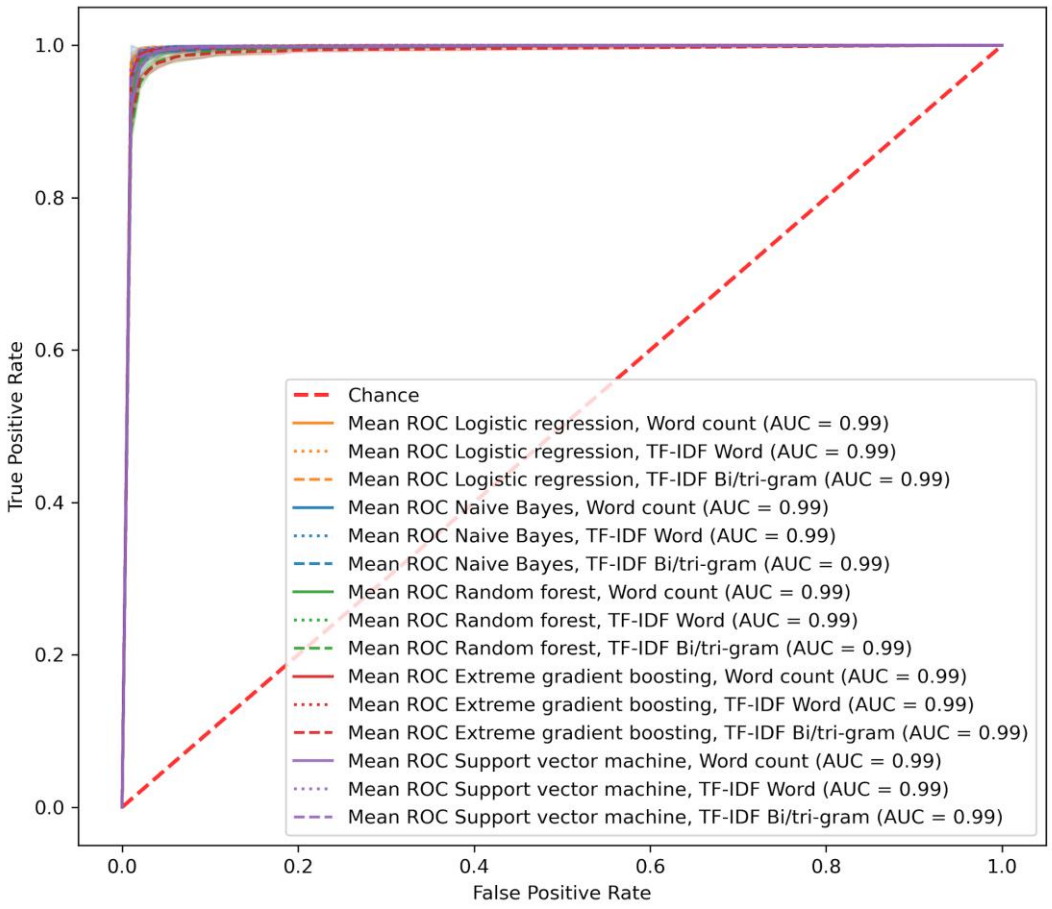


Figure 2.3: Receiver operator characteristic curve from 5-fold cross-validation of classification models trained on non-preprocessed abstracts using three feature extraction methods. Results from all five cross-validation tests were averaged to produce a single curve. Shadows around each line are +/- 1 standard deviation from the mean. Models were trained on abstracts that had not been preprocessed.

Chapter 3: Generating gold-standard entity recognition and relationship extraction training datasets

3.1 Introduction

Datasets are arguably the most important aspect of machine learning applications; while algorithms in the machine learning space improve with time, high-quality datasets are required to consistently train, evaluate, and benchmark models. Datasets that have been manually reviewed and verified by humans are considered "gold standard". However, with the advent of crowdsourcing methods like those provided by Amazon's mechanical Turk[60], researchers have found that not all crowdsourcing communities create concordant datasets[61]. Dataset labels created by subject matter experts differed greatly from those made by mechanical Turk workers, resulting in different algorithmic performances[61].

As the biomedical field is so diverse, many datasets are available for different NER and RE tasks. Broadly, biomedical concepts extracted via NER include identifying diseases[62,63], chemicals[64], genes[65,66], and taxonomy[67,68], while RE datasets have focused primarily on identifying protein-protein[69], gene-disease[70,71], or protein-chemical[72] relationships. To orient text mining researchers towards creating the best machine learning models to complete a common task, the BioCreative workshop has held seven challenges that provide reference datasets for various biomedical natural language tasks[32,62,65,73,74].

To generate NER and subsequently RE training datasets, we need standard sets of terms that represent concepts that can be identified within biomedical literature. This is necessary when normalizing predictions from NER models back to a standard set of terms. There are two major types of these lexicons: dictionaries and taxonomies. Dictionaries contain terms and their definitions, while taxonomies create hierarchies of these terms. The most well-known taxonomy is the Linnaean taxonomy which classifies organisms into parent-child relationships, i.e., each child term is a sub-class of their respective parent, e.g., *Homo sapiens* is a child term of *Homo*. In biomedical research, "ontologies" are frequently used taxonomies that allow for multiple relationship types and more complex hierarchies. Also known as "controlled vocabularies", the major benefit of ontologies is their ability to interoperate with each other and reference terms outside their domain without duplicating data unnecessarily. For example, the food ontology (FOODON) contains structured information about food products including mammalian species names[75]. Instead of recreating a mammalian taxonomy within FOODON, they cross-reference the National Center for Biotechnology Information taxonomy (NCBI TAXON) to integrate this information[76]. This allows FOODON to build upon NCBI TAXON's knowledge network for terms. For example, in NCBI TAXON, the term "*Bos taurus*" has no child terms and is not described in a food context. This is reasonable since NCBI TAXON strives only to understand the nomenclature of organisms. But within FOODON, we learn that "*Bos taurus*" has the child terms "cattle bull", "ox", and "dairy cow". Thus, by using FOODON, we can identify and classify cattle into more nuanced levels than what was possible with only NCBI TAXON.

This chapter outlines how I use eight existing ontologies to generate 15 gold-standard datasets based on 10,784 AMR-epidemiological papers from 2019 (Chapter 2). I subsequently use these datasets to train and evaluate 8 NER and 7 RE BioBert models, with the best performing models used to mine all of PubMed to make predictions about the risk and transmission of antimicrobial resistance genes (Chapter 4).

3.2 Methods

3.2.1 Lexicons and Ontologies

The resource used to gather, find, and download ontologies was the Open Biological and Biomedical Ontology (OBO) Foundry[77]. The OBO Foundry stores hundreds of biologically and clinically relevant ontologies that are interoperable, non-redundant, and well-maintained. Seven ontologies obtained from the OBO Foundry for my work include the sequence ontology (SO)[78], uber anatomy ontology (UBERON)[79], infectious disease ontology (IDO)[80], antibiotic resistance ontology (ARO)[8,9], environmental ontology (ENVO)[81], gazetteer ontology (GAZ)[82], and the food ontology (FOODON)[75]. Terms were converted from obo to json format through ROBOT[83]. The eighth and final lexicon, National Center for Biotechnology Information taxonomy (NCBI TAXON), was generated using the taxonomy database available at the National Center for Biotechnology Information file-transfer protocol server[84].

Many of the ontologies contained terms that were irrelevant to the entity type we would like to identify. For example, our goal for the use of ARO was to identify ARG terms in the literature, which can be found under the “determinant of antibiotic

resistance" (ARO:3000000) branch of the ontology, e.g., APH(4)-Ib (ARO:3002656). But, for the ARO to describe the antibiotics that these determinants resist or the mechanisms they utilize, other branches are included within the ARO, adding irrelevant terms to our lexicon, e.g., clarithromycin (ARO: 0000065). Thus, to generate NER training/testing datasets that only contain relevant concepts, the ontologies were filtered using parent terms that captured the concepts I wanted to identify in the literature. Since ontologies are hierarchical, all child terms to the parent term were retained while all other terms were discarded. A detailed breakdown of the lexicons used, the concepts they represent, and the parent term used as the primary filter can be found in Table 3.1 below.

3.2.2 Generation of a named entity recognition (NER) training and testing set

Given an ontology of terms and a computation model that can reliably identify papers believed to be rich in AMR gene epidemiological information (section 2.2.1-2.2.7), my goal was then to download all papers added to PubMed in 2019 and filter them based upon the Chapter 2 classifier's predictions (section 2.3) to build a BioBert training set. With these 2019 AMR papers in hand and as a step towards NER and RE, I determined which terms from the lexicons above were present in the abstract of these papers. To this end, I used regular expression algorithms for text matching such that each mention of a term found in the lexicons outlined above would be identified in a paper.

I created a pipeline that goes through a series of steps to resolve the major problems faced when generating annotations. The pipeline has two main sections: a

method that automatically filters annotations based on rules and a series of steps that require manual review.

Automatic filtering criteria:

- Remove terms fewer than 4 characters in length;
- Remove any "stop words" or terms that are extremely common in the English language (i.e., "and", "a", "the", etc.);
- Remove terms that nest inside another larger term (i.e. "*Escherichia*" would be removed if located inside "*Escherichia coli*");
- Remove duplicate annotations with the same term ID (i.e., both IDO and UBERON have the UBERON term "blood" and it is thus double counted; in another case, "hot dog" has two synonyms, "wiener" and "Wiener" that match the same positions during annotation);
- Remove duplicate terms from the same lexicon (i.e., the two terms "bone element" and "bone tissue" both have the synonym "bone" in UBERON);
- Remove duplicate terms from different lexicons (i.e., "kidney" appears in both FOODON and UBERON, but most commonly refers to the anatomy term and not the food product)

Every term used in the annotation of the 2019 papers (6,839 in total across lexicons) was manually reviewed by me and given a label of "keep", "remove", or "review" based on if I thought they were appropriate for their lexicon and their likelihood to mismatch. So, terms with only one meaning (i.e., bacterial taxonomy, genes, food products) were given the label of "keep"; terms that were almost always going to mismatch because their

ontology meaning would rarely be used in biomedical text (i.e., the abbreviation “as” for “American Samoa”) were labelled “remove”; and terms that fell into neither category were labelled “review”. Any terms labelled “review” and unresolved terms resulting from the automatic section went through a series of manual filters:

- Filter all annotations for a term after reviewing a randomly drawn sample set of five sentences;
- Terms unresolved after viewing a sample set of sentences were examined on a sentence-to-sentence basis

After the combination of automated annotation of terms and manual curation, the NER training/testing dataset was then generated, and one final review was performed to add and remove any annotations.

3.2.3 Generation of a relationship extraction (RE) training and testing set

Using the set of filtered annotations, the RE training/testing dataset was generated by taking all ARO:other lexicon pair combinations found in a sentence. These pairings were manually examined by me and my volunteers and given a label of 0/1 if the two terms were describing each other (1) or not (0). BioBert requires a sampling of both to suppress false positive and false negative predictions of relationships. For example, consider the following sentence:

> *This study provides the first report of bla NDM-1-positve **K. pneumoniae** along with ST268 as well as the spread of nosocomial infections with six different STs*

*harboring bla **NDM-1** and other resistance genes in **hospital** settings especially neonatal intensive care unit.*[44]

The second mention of "*NDM-1*", highlighted in bold, is not related to "*K. pneumoniae*" but is described as found in a "hospital". Thus the *NDM-1*:*K. pneumoniae* pair is labelled 0, and the *NDM-1*:hospital pair is labelled 1.

3.3 Results

3.3.1 NER training/testing datasets

From regular expression matching of exact terms and synonyms to 10,784 biomedical abstracts from 2019 (Chapter 2), 730,538 annotations and 6,839 unique terms were identified (Table 3.2). Since pattern-matching algorithms do not consider that a term may have multiple meanings, many annotations result in a mismatch between the annotated term and the corresponding lexicon term's meaning. For example, the GAZ term "American Samoa" has the synonym "as", referring to the geographical location. However, this annotation process erroneously labelled 10,437 instances of the term.

Filtering methods limited these mismatches by reducing the total number of annotations from ~730,000 to ~148,000 (Table 3.3). Despite removing so many annotations, the number of unique terms did not decrease dramatically, only reducing to 4,903 from a pre-filtered level of 6,839 (Table 3.4). The ontologies contributing the most erroneous annotations were GAZ, followed by SO, and UBERON. These lexicons tended to have very short abbreviations that would consistently mismatch with common English

words. This is revealed by their dramatic drop in the number of annotations during the length filtering step (Table 3.5).

While some lexicons mismatched extensively, others did not match enough. During the final manual labelling step, many terms that conceptually belonged to a lexicon had to be identified and manually added. During this process, 64 ENVO, 116 FOODON, and 311 NCBI TAXON terms were manually identified (Table 3.4). These new terms corresponded with manually adding 2,154, 5,421, and 17,349 annotations to ENVO, FOODON, and NCBI TAXON, respectively (Table 3.3). These terms were not identified from the initial annotation process because they were not found in the original ontology, e.g., manual review flagged the term "pork" even though it did not exist in FOODON.

This work created eight gold-standard datasets for training BioBERT models to identify terms for ARGs, environments, food products, geographical locations, infectious disease terms, bacterial taxonomy, sequencing terms, and anatomy terms.

3.3.2 RE training/testing datasets

Using the annotations generated from the NER training/testing datasets, 12,588 relationships were extracted based on ARG-epidemiology terms that appear within the same sentence (Table 3.5). Of these ~12,000 relationships, 10,329 were labelled as genuine relationships (1), while the remaining 2,254 found no relationship (0) between the ARG and epidemiology term. The greatest number of relationships to ARO terms were found among IDO, NCBI TAXON, and SO terms.

This work created seven gold-standard datasets for training BioBERT models to identify relationships between ARGs and environments, food products, geographical locations, infectious disease terms, bacterial taxonomy, sequencing terms, and anatomy terms.

Table 3.1 Breakdown of ontologies used, concepts they represent, and parent term used as primary filter.

Lexicon	Main concept	Parent filter
ARO	ARG names	Determinant of antibiotic resistance (ARO:3000000)
FOODON	Food products	Food product (FOODON:00001002)
UBERON	Anatomy parts	Anatomical entity (UBERON:0001062)
ENVO	Environments	Construction (ENVO:01001813)
NCBI TAXON	Bacterial taxonomy	Bacteria (0)
GAZ	Geographical locations	No filter
SO	Genomic terms	No filter
IDO	Infectious disease terms	Continuant (BFO:0000002)

Table 3.2: Term, annotation, and paper counts of lexicons. The term count is the total number of words in a lexicon, not including synonyms. Annotation count is the total number of matches found across all abstracts. The paper count is the number of unique papers in which lexicon terms appear. The number of unique papers with annotations is 10,782 prior to filtering.

Lexicon	Term count	Annotation count	Paper count
ARO	3,850	10,118	2,689
FOODON	10,815	15,529	6,682
UBERON	13,847	113,456	10,646
ENVO	217	4,866	2,633
NCBI TAXON	466,787	43,566	8,597
GAZ	565,454	361,759	10,780
SO	2,597	146,299	10,719
IDO	274	34,945	8,793

Table 3.3: Automatic and manual filtering initial annotation results on an annotation level. All terms were reviewed and assigned a label of "review", "keep", or "remove". When generating the final NER dataset, the terms labelled "remove" were taken out. All terms/annotations went through the following filtering steps: (1) length filter removing terms fewer than four characters in length; (2) remove any stop words or common English words; (3) remove terms that nest inside another term; (4) remove duplicate terms with the same term ID; (5) remove annotation duplicates under the same lexicon; (6) remove annotation duplicates under different lexicons; (7) manual filtering by examining 5 sample sentences on terms unresolved from steps 1-6; (8) manual filtering by examining 5 sample sentences on terms labelled "review"; (9) sentence-level filtering of annotations of terms unresolved from step 7 and 8; (10) final manual review of generated NER dataset. Values under steps 1-10 represent the number of annotations removed after each filter. Negative values indicate that annotations were added.

Lexicon	Label	No filter	(1) Length	(2) Stop word	(3) Nested	(4) Dup.	(5) Same lexicon	(6) Diff. lexicon	(7) 5 sen.	(8) 5 sen. "review"	(9) All sen.	(10) NER dataset	Results	% Change
ARO	Review	1398	156		6			709	1	512	2		12	-99%
	Keep	8720	632		184		65	1	15			2	7821	-10%
ENVO	Review	805	24		38	31	10	60	144	340			158	-80%
	Remove	4												-100%
	Keep	4057			69	2	8		22			-2154	6110	51%
FOODON	Review	10441	6623		94			22		3463	59	-2	182	-98%
	Remove	107	1			1		96						-100%
	Keep	4981	79		268		392	825	31			-5421	8807	77%
GAZ	Review	38765	27270	7599	277		26	19	188	3156	24	10	196	-99%
	Remove	303960	275903	5748	1066		1029	38	1803					-100%
	Keep	19034	557		1581		484	5	205			196	16006	-16%
IDO	Remove	2475			46									-100%
	Keep	32470			770							4	31696	-2%
NCBI TAXON	Review	204	17		123			13					51	-75%
	Remove	4170			152			2380	43					-100%
	Keep	39192			12933		126		4199			-17349	39283	0%
SO	Review	10723	4194		267		38	1016	163	2431	87		2527	-76%
	Remove	109500	103682		416		976	383	146					-100%
	Keep	26076			2421		30	2	31			-549	24141	-7%
UBERON	Review	4102	838		310		164	239	330	2023	9		189	-95%
	Remove	92593	85619		1062		13	1549	332					-100%
	Keep	16761	1624		989	959	918	495	1130			-620	11266	-33%

Table 3.4: Automatic and manual filtering initial annotation results on a term level. The methods used to generate the table are the same as in Table 3.3 but on a term basis.

Lexicon	Label	No filter	(1) Length	(2) Stop word	(3) Nested	(4) Dup.	(5) Same lexicon	(6) Diff. lexicon	(7) 5 sen.	(8) 5 sen. "review"	(9) All sen.	(10) NER dataset	Results	% Change
ARO	Review	10	2					2	1	3			2	-80%
	Keep	676	25		1		8	1	4				637	-6%
ENVO	Review	28	2				1	4	3	11			7	-75%
	Remove	1												-100%
	Keep	51					2		4			-64	109	114%
FOODON	Review	41	4					4		15	3	1	14	-66%
	Remove	7	1					1						-100%
	Keep	250	5		3		6	9	8			-116	335	34%
GAZ	Review	742	131	3	19		9	3	19	476		6	76	-90%
	Remove	547	261	6	14		11	2	10					-100%
	Keep	1231	4		42		35	1	5			2	1142	-7%
IDO	Remove	8												-100%
	Keep	72											72	0%
NCBI TAXON	Review	18	1		2			3					12	-33%
	Remove	5						2	1					-100%
	Keep	1542			102		8		9			-311	1734	12%
SO	Review	191	63		1		1	5	6	43			72	-62%
	Remove	69	39				1	1	2					-100%
	Keep	212			1		1	1	3			-1	207	-2%
UBERON	Review	188	32		9		15	7	22	71			32	-83%
	Remove	380	327		4		4	5	6					-100%
	Keep	570	19		23		56	3	21			-4	452	-21%

Table 3.5: Relationship counts within the RE dataset. ARO to lexicon relationships were generated using sentences that contain at least one ARO term and another term from a different lexicon. The total number of ARO:lexicon relationships available can be seen in the "All" column. Relationships were manually labelled and given a 0/1 depending on whether the ARO:lexicon terms contextualize one another (1) or not (0). The subset of labeled relationships (1) were used to train the RE models.

Lexicon	All	Labelled 0	Labelled 1	Unique Papers
ENVO	386	25	361	153
FOODON	438	49	389	142
GAZ	791	62	729	339
IDO	1,460	273	1,187	549
NCBI TAXON	3,347	583	2,763	968
SO	5,766	1,204	4,558	1,149
UBERON	400	58	342	203
Total	12,588	2,254	10,329	1,664

3.4 Discussion

3.4.1 Ontologies vary in their text-mining usefulness

To generate NER datasets, we use ontologies with terms we can normalize back to after making predictions. General dictionaries work but lack hierarchy which reveals how terms are related. Additionally, simple dictionaries do not contain descriptions of a particular term, reducing their usefulness for others trying to interpret complicated relationships. Another benefit of ontologies is that they are largely continuously maintained and updated, i.e., new genes and species will be added to ontologies while dictionaries are static. Overall, ontologies are vital resources for providing terms that represent concepts we would like to extract for NER tasks. However, from my time filtering the initial annotations, I found that two major factors impacted how functional these ontologies are for text-mining purposes: (i) how polysemantic the ontology terms are, and (ii) how comprehensive the ontology is and the depth/nuance of terms within.

Inherently, the ontologies most impacted by filtering were those that mismatched the most. This includes GAZ, UBERON, and SO, which had hundreds of terms removed after filtering (Table 3.4). The primary cause of these mismatches was because these terms have multiple meanings but the most common meaning used in biomedical text does not correspond to the one used in the ontology. For example, the term “an” in FOODON refers to a sweet bean paste, but this is not the most common meaning for the term in biomedical text. As a result, this term mismatches thousands of times. Ontologies with terms that have only one definition mismatched much less often. This can be seen

with ARO, NCBI TAXON, and IDO, which lost very few annotations/terms to filtering and whose terms have one meaning.

While polysemantic terms frequently caused mismatches, the inability to identify relevant terms within a text is another area of concern. There are three reasons we cannot identify terms within a text: (i) the ontology is not comprehensive enough and lacks relevant terms, (ii) the ontology terms are too verbose, and (iii) the ontology does not capture enough nuance within the terms. From the manual step of filtering, many terms were added to ENVO, FOODON, and NCBI TAXON, indicating that the ontologies alone could not capture the extent of terms belonging to their overall concept. For NCBI TAXON, the ontology was not comprehensive and lacked species abbreviations. These abbreviations were added retroactively to capture all bacterial taxonomy. While comprehensive, FOODON had verbose terms and went into too much detail to be helpful for a text-mining application. For example, the term “poultry” is not in FOODON, but the terms “poultry product”, “poultry (frozen)”, “poultry (raw)” and many other verbose terms for poultry food items can be found in the ontology. While these terms are necessary to understand the nuances between food items, none of these terms were identified in the literature because of their verbosity and specificity. ENVO perhaps takes the opposite direction as FOODON and has too few details. For example, they have the term “intensive care unit” but do not have “neonatal intensive care unit”. As of writing, ENVO had updated their ontology to include “neonatal intensive care unit facility” in their development branch of the ontology seven months ago[85]. However, they have yet to release a new stable version of the ontology including these new terms. Despite adding

this new term, it is now too verbose as most mentions of such an environment are typically “neonatal intensive care unit”, not “facility”. Additionally, although they are rapidly expanding the ontology, they do not release stable versions very often, as the last stable release was on May 14, 2021[86].

Both mismatches and the lack of matches stem from an ontology’s structure and the terms contained within. Thankfully, these annotation problems were corrected via manual filtering steps aimed at improving the resulting NER training/testing datasets. However, manual review is time-consuming, requiring attention to detail and patience. Although a large initial effort was required to generate these gold-standard datasets, minimal human intervention is needed to maintain their long-term value. With constant improvements in natural language model performance and generalization, gaps in datasets are largely overcome. Areas lacking generalization can be improved by creating and incorporating small, targeted datasets into the training process. These targeted datasets will be small in scale (i.e., tens of labelled examples) and will improve the performance of identifying specific terms or relationships.

3.4.2 NER training/testing datasets

The 8 NER training and testing datasets I have created provide a resource of annotations in biomedical literature identifying ARGs, environments, food products, geographical regions, infectious disease terms, bacterial taxonomy, sequencing terms, and anatomy. While other NER datasets have been created using biomedical literature, they primarily focus on identifying genes[65,66], chemicals[62,64], diseases[62,63], and species[67,68]. Additionally, NER datasets with geographical terms, food terms, or environment terms

use news articles, recipes, or encyclopedia pages rather than biomedical text to extract their terms[87–89]. While datasets are available for food, environments, geographies, and other terms, no ARO-epidemiology relationship datasets are available. To generate these datasets, I would still need to go through the process of generating NER and subsequently RE datasets to capture entities in a biomedical context. Thus, generating the data myself was the simplest and most effective method.

3.4.3 RE training/testing datasets

The 7 RE training and testing datasets my volunteers and I have created are foundational for training BioBERT models to identify ARO-epidemiology relationships more broadly across biomedical literature. Of the ~12,000 relationships labelled, most were positive (1) relationships, indicating that when two terms are co-mentioned in the same sentence, they are most likely related to each other. This intuitively makes sense, as abstracts commonly report positive associations (i.e., an ARG was found in a particular environment) rather than negative associations. Most negative associations came from cases where multiple ARGs are described alongside multiple epidemiology terms yet only specific pairs of terms are related.

While SO and IDO contain many biomedically-relevant terms, these terms are uninformative for creating relationships with ARGs. Except for the term “plasmid” which is essential for understanding which genes are plasmid-borne, SO contains redundant terms like “gene”, “polypeptide”, or “amino acid” which are uninformative in understanding more about ARG epidemiology. Similarly, for IDO, the most common relationships are “bacteria”, “infectivity”, “antimicrobial”, “virulence”, “host”, and “cell”.

In future iterations of these datasets, aside from plasmid-associated relationships, I would remove SO and IDO from RE tasks. Except for SO and IDO, relationships were highly informative for understanding the sources in which genes are found.

3.4.4 Challenges associated with creating gold-standard training and testing datasets

Creating gold-standard datasets is a manual process involving the removal or addition of terms. Removing terms is relatively easy, as most terms and annotations were removed from automated filtering steps like filtering by length, removing stopwords, dropping duplicated terms, and removing nested terms (Table 3.3 and 3.4). However, removing the remainder of those mismatched terms requires one to review annotations sentence-by-sentence, to identify all irrelevant terms and remove them. To speed the process up, I selected all unique terms annotated across the ~10,000 papers and labelled them as “review”, “keep”, or “remove” based on their likelihood to mismatch and how appropriate they were for their ontology. This sped up the process considerably by allowing me to review only 1,218 of the 6,839 lexicon terms, which corresponded with reviewing 66,438 of the 730,538 annotations. However, the process was anything but fast and took several weeks to complete. After filtering, I identified terms that were missing but were associated with the concept we would like to extract. This process was more laborious than filtering as I had to go through the raw sentences, not just the annotations, to identify terms that were not picked up during annotations or were mistakenly removed during filtering. This is extremely important when generating gold-standard NER datasets, as it is important that we capture all relevant terms associated with a concept. The ontologies that required the most additional labelling were NCBI TAXON, ENVO,

and FOODON due to the complications discussed in Section 3.4.1. Overall, the process of filtering and identifying missing annotations took months.

The process of labelling relationships was similarly time-consuming but straightforward as no filtering was needed. The main difficulty associated with labelling relationships was that we had to read most of a sentence to identify if a relationship existed between an ARG and epidemiology terms. Across ~12,000 relationships, 3,410 sentences were reviewed to identify if associations existed.

To generate high-quality relationship training/testing data, manual curation is ideal. A highly-cited database for gene-disease relationships called Genetic Association Database (GAD)[70] has recently come under scrutiny. Their semi-automatic method of labelling relationships led to the generation of questionable labels[90,91]. Without expert manual curation to verify your data, you can never truly trust the results you generate.

Overall, the efforts taken by myself and my volunteers have generated gold-standard datasets that identify epidemiology concepts alongside ARGs and the relationships between them. Other NLP researchers can use these data to train new models and benchmark current models for understanding how these terms interact in biomedical literature, among many other applications. For my work, these datasets are used to train and evaluate BioBERT models for the task of NER and RE to develop a better understanding of ARGs transmission and risk (Chapter 4).

Chapter 4: Understanding risk and transmission of antimicrobial resistance genes

4.1 Introduction

Deploying effective public health interventions to control the spread of AMR requires an understanding of where ARGs or pathogen transmission is most likely to occur and how much risk is associated with that transmission event[14,92]. To monitor the spread of AMR across various environments, surveillance programs are in place that monitor resistant pathogens in high-risk environments as well as rates of treatment failure. Canada has two major national surveillance programs: the Canadian Nosocomial Infection Surveillance Program surveys healthcare-associated infections and the Canadian Integrated Program for Antimicrobial Resistance surveys food-borne, environmental, and zoonotic infections[93,94]. Smaller but more specific surveillance initiatives in Canada focus on tuberculosis, gonococcal, or streptococcal infections[95–97]. Like Canada, many countries conduct their own surveillance monitoring and participate in a global surveillance effort organized by the World Health Organization to understand AMR's drivers better, called the Global Antimicrobial Resistance and Use Surveillance System (GLASS)[98]. Ideally, with these surveillance programs in place, we can generate up-to-date estimates of the prevalence of AMR in different environments and monitor the performance of mitigation strategies. Unfortunately, we cannot currently estimate this[92]. Additionally, Canada lacks surveillance data for community settings, domestic animals, wildlife, soil, and water samples[1,92]. In addition, most surveillance is phenotypic and lacks information on

individual ARGs. Challenges stemming from the lack of wholistic surveillance make understanding AMR transmission at a national level difficult and limits our ability to conduct risk assessments[14].

Text mining processes can complement phenotypic and genotypic surveillance programs and help overcome challenges faced by current surveillance programs by collecting information from publications worldwide, from sources beyond sentinel-based surveillance efforts, thus helping researchers and public health officials better understand the large-scale dynamics of resistance. By collecting epidemiological NER and RE information related to ARGs, we can begin to analyze resistance trends and understand how likely transmission is to occur based on environment similarity metrics.

4.2 Methods

4.2.1 Training BioBERT

As a step toward the prediction of transmission and risk, a BioBERT model was trained on the 2019 training datasets (section 3.2.1-3.2.3) for the task of NER (8 training datasets) and RE (7 training datasets) for all of PubMed. During this process, 25% of data was held out for final evaluation, with the remaining 75% used for training (80% of the 75%) and hyperparameter tuning (20% of the 75%). Combinations of two hyperparameters, batch size and learning rate, were adjusted to identify the best-performing model. Batch size adjusts the number of sentence examples a model is trained upon for each iteration, and learning rate impacts the size of the weight adjustment made each time back propagation occurs. A larger learning rate results in a more significant

adjustment of the weight parameters in the model and can help avoid local minima loss values. However, they may result in an unstable training process and a failure to converge on a minimum loss. A lower learning rate will have a better chance of converging but may get stuck in a local minimum. The other main hyperparameter, epochs, remained at a value of four as recommended by the BERT paper[49]. Epochs adjust the times the dataset is passed through the model during training. Table 4.1 shows the hyperparameter values examined.

4.2.2 Term normalization

To generate interpretable results, annotations gathered from BioBERT NER must be mapped to a set of standardized terms. For example, UBERON has multiple terms describing stool, including: “stool”, “feces”, “fecal”, and “faeces”. Matches for these term variants must be mapped back to the standard “stool” term. Using the ontological lexicons as a source of standard terms, I mapped all of BioBERT’s NER annotations using Boolean logic or fuzzy string searching. First, all annotations were standardized by removing punctuation, lowercasing, and stemming. Afterward, annotations were checked via Boolean logic to see if the term is an extract match to any terms within the lexicons. If there was no perfect match, a fuzzy search was conducted by calculating the Levenshtein distance between the annotation and every lexicon term and synonym. Levenshtein distance measures the number of character changes needed to convert one word into another. An arbitrary cut-off of 95% was used when assigning annotations to lexicon terms. Terms with no perfect match that fell below the 95% cut-off were passed on to one final fuzzy-matching algorithm called token_set_ratio, which calculates three Levenshtein

distances using the sorted intersection and difference of token sets between two terms. For example, given two terms, “Minnesota” and “State of Minnesota”, we tokenize them first into sorted sets (i.e., “Minnesota” becomes [“Minnesota”] and “State of Minnesota” becomes [“Minnesota”, “of”, “State”]). Then three datasets are generated based on the following.

- Set 1: $A \cap B$ (intersection) = [“Minnesota”]
- Set 2: $A - B =$ [“Minnesota”] – [“State”, “of”, “Minnesota”] = empty set
- Set 3: $B - A =$ [“State”, “of”, “Minnesota”] – [“Minnesota”] = [“State”, “of”]

From these sets, three strings are created for comparison:

- String 1: Set 1 alone = “Minnesota”
- String 2: Set 1 + Set 2 = “Minnesota”
- String 3: Set 1 + Set 3 = “Minnesota of State”

Levenshtein distances are then calculated between strings 1 and 2, 1 and 3, and 2 and 3, taking the maximum of the three scores. In the case of “Minnesota”, the score would be 100%. I selected the first ontology term for these annotations that yielded a 100% score through token_set_ratio. Any terms that were unsuccessful in being normalized were considered “non-normalized” but still included in downstream analyses. For example, the term “broiler houses” was annotated across 56 papers but did not normalize to any term in ENVO.

4.2.3 Similarity scores

To assess the similarity between epidemiology terms, similarity metrics were calculated using RE prediction results for all AMR-epidemiology papers. For calculating similarity, each epidemiology term was represented by the set of ARGs related to that term. For example, the term "chicken" was related to 110 ARG terms from the ARO, including *MCR-1.1*, *CMY-2*, *vanA*, and *CTX-M-1*, among others. Metrics were then based on the overlap in ARGs between terms, e.g., the ARG terms shared or not shared by "chicken" and "farm".

The first similarity metric calculated was Jaccard similarity, which relies on the presence/absence of ARGs. We thus represent A and B as sets of ARGs related to two epidemiology terms:

$$Jaccard = \frac{|A \cap B|}{|A \cup B|}$$

Jaccard similarity is measured as the shared number of ARGs between two epidemiology terms divided by the total number of unique ARGs related to both terms. In contrast, Bray-Curtis is a more quantitative method that considers the number of papers each epidemiology term and ARG are found to have a relationship in:

$$Bray - Curtis = \frac{2S_{AB}}{P_A + P_B}$$

For this metric, S_{AB} takes the shared ARGs between two terms and sums the minimum papers associated with each ARG-epidemiology relationship. P_A and P_B are the total

numbers of papers associated with epidemiology terms A and B. For example, given the following terms:

- "chicken" with relationships with *NDM-1* found across 10 papers, *MCR-5* across 5 papers, and *CTX-M-15* across 3 papers;
- "*Bos taurus*" with relationships with *NDM-1* found across 4 papers, *MCR-5* across 8 papers, and vanA across 4 papers

S_{AB} is calculated by taking *NDM-1* and *MCR-5* since they are shared between both terms, and then taking the sum of the minimum number of papers (i.e., 4 for *NDM-1* and 5 for *MCR-5*). P_A and P_B is the sum of papers associated with "chicken" and "*Bos taurus*", respectively (i.e., 10+5+3 for "chicken", and 4+8+4 for "*Bos taurus*"). The resulting score would be (4+5)/(18+16) = 0.26.

The final similarity metric examined, co-occurrence affinity, is insensitive to the prevalence of ARGs associated with one epidemiology term compared to another[99]. Co-occurrence affinity is measured based on the probability that for each ARG, term A is associated with the ARG with a probability of p_1 if term B is present but with p_2 if term B is absent. Using these probabilities, we calculate the log odds ratio:

$$\alpha = \log\left(\frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2}\right)$$

For a particular ARG, α reflects the similarity between two epidemiology terms. The affinity metric is estimated across all possible ARGs using maximum likelihood to calculate a similarity metric.

4.3 Results

4.3.1 Performance and generalizability of models for named entity recognition (NER)

A generalized NER model can identify terms belonging to a lexicon while having never previously seen those terms during training. To assess how well a model generalizes, I set aside 15% of all terms in each lexicon within the NER dataset for testing. This method of splitting resulted in roughly half the annotations in the testing sets measuring memorization, i.e., recognizing terms included in the training, and the other half testing generalizability (Table 4.2). Only recall was measured for memorization and generalizability as there were an unknown number of false positives, ruling out the calculation of precision. For lexicons with very few unique terms like ENVO, IDO, SO, and UBERON, generalization performance fell below 60% (Table 4.2); as the number of unique terms in lexicons increased, the better the models generalized (Figure 4.1). As a result, ARO, FOODON, GAZ, and NCBI TAXON had recall values greater than 70% (Table 4.2). For downstream predictions, the best NER model was selected for each lexicon based on the highest generalization recall value.

4.3.2 Model performance for relation extraction (RE)

Fine-tuning models for RE took half a day to complete as the dataset was much smaller than the NER dataset. Models were evaluated based on loss, accuracy, and F1 values (Table 4.3). Loss measures the difference between raw predictions and the label. A loss closer to 0 results in a better-performing model. Accuracy is calculated as the total true positive predictions divided by the total number of predictions, while F1 is the

weighted average between precision and recall. As the evaluation datasets were relatively small for some lexicons (i.e., the smallest dataset, ENVO, only contains 97 examples), accuracy and F1 values across different learning rates and batch size combinations stayed relatively similar. Additionally, there was a disproportionately more significant number of positive relationship labels than negative relationships found in the training/testing dataset (Table 3.5). This unbalanced dataset could lead to bias in the trained models. To better understand the impact of this potential bias, these models must be evaluated relative to a naïve approach of relationship prediction that assumes any two terms appearing in an abstract are related. Nevertheless, for now, the best-performing models were selected based on the highest accuracy and F1, followed by the lowest loss score if there was a tie.

4.3.3 NER and RE results on 204k papers

Using the logistic regression paper classification model described in Chapter 2, 204k papers were identified to bear AMR-relevant information from the entirety of PubMed. The NER models then annotated these papers, generating over 2.2 million annotations (Table 4.4). Terms from these annotations were normalized using the Boolean and fuzzy logic described in Section 4.2.2, yielding 20,055 unique terms. Of these terms, 6,002 could not be normalized and were labelled “non-normalized”. Under ARO, most non-normalized terms turned out to be other bacterial genes misannotated as ARGs.

The annotations generated from NER model predictions are largely promising. The NER models identified thousands of gene, environment, food, taxonomy, and geographical mentions over years of published research (Figure 4.2). However, the

ENVO annotations were severely skewed towards the term “hospital”. Out of the 78,298 ENVO annotations identified across 40,654 papers, 42,114 are the term “hospital”, with the second most common term, “farm”, annotated only 14,261 times. This skewed data reflects the level of importance the scientific community has given to each environment. What is also interesting to note, it seems that only since the year 2000 has research picked up in examining wastewater treatment plants (Figure 4.2).

The distribution of annotated terms across papers can be seen in Figure 4.3. Only five ARO terms (*CTX-M-15*, *TEM-1*, beta-lactamase, *mecA*, and *vanA*) were found in at least 1,000 papers. Additionally, no ARO, FOODON, or GAZ terms were found in more than 10,000 papers (Figure 4.3). The relative proportion of NCBI TAXON terms compared to other lexicons across papers stayed relatively the same, while the proportion of GAZ and ARO terms decreased (Figure 4.3B). In contrast, the proportion of SO, UBERON, IDO, and ENVO terms increased, indicating they are more common in thousands and tens of thousands of papers relative to all lexicons. Overall, most annotated terms are found in only one paper, while a small group of terms are found across thousands of papers.

To generate RE data, all ARO:other lexicon annotations found in the same sentence were masked and given to the RE models for prediction, generating 110,369 positive relationships across 33,580 relationship pairs (Table 4.5). Most terms found in these relationship pairs were successfully normalized (Table 4.6). Of the 110,369 relationships, nearly 100,000 relationships existed between two normalized terms, while only 91 annotations were between non-normalized terms. Thus, despite one-quarter of the

NER terms being non-normalized, very few of them formed relationships with ARO terms. This might mean that many non-normalized terms are unrelated to ARO terms and are false positives generated by the NER models, or the RE models are biased against previously unseen terms and cannot generalize. A combination of these two is most likely.

4.3.4 Affinity co-occurrence should be used to compare similarity between two epidemiology terms

Besides exploring the relationships between ARO terms and epidemiology terms, the relationship data I generated can be used to compare different epidemiology terms to explore their genetic similarity. We can see which genes are shared between different terms by taking the ARO relationships associated with 1,341 epidemiology terms. Of the 1,341 terms, 33 are ENVO, 65 FOODON, 380 GAZ, 37 IDO, 512 NCBI TAXON, 202 SO, and 112 UBERON. The four similarity metrics compared are outlined in Section 4.2.4 and include Jaccard, Bray-Curtis, co-occurrence affinity, and gene overlap. Although the magnitude of these similarity metrics varies, they share similar hotspots where terms are highly related (Figure 4.4). Notably, Jaccard and Bray-Curtis share the same hotspots, with the primary difference being that Bray-Curtis resulted in higher similarity scores for these hotspot regions (Figure 4.4A and 4.4B). Gene overlap scores show that most term pairs share zero genes (499,848, or 55.6% of the 898,470 total term pairs), and the most prominent hotspots occur where more than four terms are shared between term pairs (Figure 4.4D). The most notable metric is affinity co-occurrence,

which combines all hotspots across the other three similarity metrics and dramatically increases the magnitude of similar terms relative to other metrics (Figure 4.4C).

Exploring the top 10 lexicon-lexicon term similarities based on their genetic overlap, we can identify term pairs that share the most ARGs and explore the nuances between the similarity metrics (Table 4.7). For example, while the top GAZ-GAZ term similarity score based on gene overlap is between “Japan” and “China” with 57 genes shared between them, all similarity metrics agree that the most similar GAZ-GAZ terms within the top 10 is between “Spain” and “Europe” despite only sharing 37 genes. Important to note is the magnitude difference between affinity co-occurrence values and the other similarity score values.

Of term pairs with at least one gene shared between them, affinity co-occurrence has a median score of 0.71. In contrast, Jaccard similarity and Bray-Curtis have median scores of 0.053 and 0.096, respectively (Figure 4.5). Exploring the distributions of the similarity scores, we can see that Jaccard appears not to follow any distribution, Bray-Curtis follows a lognormal distribution, and affinity co-occurrence follows a normal distribution (Figure 4.5). Out of the 398,622 term pairs with at least one gene overlap, a perfect similarity score (i.e., a score of 1) occurred 195, 17, and 31,266 times for Jaccard similarity, Bray-Curtis, and affinity co-occurrence, respectively. By examining the top 10 similarity scores sorted based on the gene overlap, you can see that perfect similarity scores with Jaccard similarity and Bray-Curtis result in terms with very few gene overlaps (Table 4.7). Affinity co-occurrence, however, ranks terms with more gene overlap higher in comparison.

For this work, since affinity co-occurrence can capture the strengths of Jaccard, Bray-Curtis, and gene overlap, while maintaining a normal distribution of scores that are not skewed towards zero, it was used as the primary metric when exploring similarity scores between America and Canada described in Section 4.3.5 below.

4.3.5 Exploring transmission between Canada and America

By selecting all ARO-epidemiology relationships that appeared in the same abstract as terms associated with America and/or Canada (i.e., all child terms to "United States of America" and "Canada"), I calculated the co-occurrence affinity between the American- and Canadian-associated epidemiology terms, i.e., for the term "chicken" I selected all relationships between ARO and "chicken" for abstracts where American and/or Canadian geographical terms were mentioned and calculated the similarity between American- and Canadian-associated "chicken". A total of 3,140 papers contained Canadian geographical annotations, while 8,556 contained American annotations. From the selection process, 2,417 American-associated relationships were found in 815 papers compared to 1,253 Canadian-associated relationships found across 353 papers. Stratifying these results by lexicon type, GAZ, NCBI TAXON, and SO contribute the greatest number of papers, relationships, and terms towards these country-associated relationships (Figure 4.6).

Calculating the similarity scores on a year-over-year basis, we can potentially see how transmission may take place. However, there was not enough information for many years and epidemiology terms to calculate a similarity score (Figure 4.7). The terms with the most similar information were bacterial taxonomy terms, geographical terms, and the

environment term “hospital” (Figure 4.7). Although, even with terms discussed very often like “hospital”, only 14 ARGs were Canadian-associated and 46 American-associated. Year over year, this information becomes sparser, making it difficult to assess the similarity between even one of the most common terms like “hospital”.

Table 4.1: Hyperparameters tuned for NER and RE. All value combinations were evaluated to identify the best epoch, learning rate, and batch size combination for NER and RE tasks.

Parameter	Value(s)
Epoch	4
Learning rate	1e-5, 3e-5, 5e-5
Batch size	8, 16, 32, 64

Table 4.2: NER model performance and generalization ability. Precision (P), recall (R), and F1 scores were calculated on the testing NER dataset for the respective lexicon. Additionally, recall was measured on two subsets of the training data: memory (MEM) and generalization (GEN). MEM counts were generated based on terms in both the training and testing set. GEN counts were generated when neither the term nor their synonym appeared in the training set but appeared in the testing set. NER models were trained on a combination of three learning rates (1e-5, 3e-5, and 5e-5) and four batch sizes (8, 16, 32, 64). All models were trained using an epoch of 4. Boxed in red are the best-performing models based on generalizability and overall performance used for downstream NER annotations of PubMed. Green color scales were applied column-wise within lexicon boundaries to indicate the best-performing metric.

						MEM		GEN	
Lexicon	**Learning rate**	**Batch size**	**P**	**R**	**F1**	**R**	**Count**	**R**	**Count**
ARO	1.00E-05	8	0.984	0.851	0.913	0.966	879	0.749	963
		16	0.969	0.840	0.900	0.972	879	0.723	963
		32	0.964	0.852	0.905	0.958	879	0.758	963
		64	0.961	0.848	0.901	0.948	879	0.760	963
	3.00E-05	8	0.980	0.802	0.882	0.961	879	0.659	963
		16	0.975	0.803	0.881	0.974	879	0.649	963
		32	0.952	0.818	0.880	0.975	879	0.676	963
		64	0.963	0.850	0.903	0.966	879	0.746	963
	5.00E-05	8	0.974	0.810	0.884	0.952	879	0.682	963
		16	0.959	0.818	0.883	0.966	879	0.685	963
		32	0.979	0.807	0.885	0.969	879	0.661	963
		64	0.973	0.853	0.909	0.976	879	0.744	963
ENVO	1.00E-05	8	0.984	0.512	0.674	0.989	88	0.362	279
		16	0.969	0.512	0.670	0.977	88	0.366	279

		32	0.949	0.512	0.665	0.989	88	0.362	279
		64	0.959	0.512	0.668	0.977	88	0.366	279
		8	0.974	0.510	0.669	0.989	88	0.358	279
	3.00E-05	16	1.000	0.512	0.677	0.989	88	0.362	279
		32	0.995	0.512	0.676	0.977	88	0.366	279
		64	0.979	0.507	0.668	0.977	88	0.358	279
		8	0.995	0.510	0.674	0.966	88	0.366	279
		16	0.974	0.512	0.671	0.989	88	0.362	279
	5.00E-05	32	0.979	0.518	0.677	0.966	88	0.376	279
		64	0.979	0.501	0.663	0.966	88	0.355	279
		8	1.000	0.832	0.908	0.964	664	0.643	465
		16	0.999	0.843	0.915	0.973	664	0.658	465
	1.00E-05	32	0.999	0.826	0.904	0.946	664	0.654	465
		64	0.998	0.756	0.860	0.956	664	0.471	465
		8	0.999	0.761	0.864	0.967	664	0.467	465
FOODON		16	0.999	0.863	0.926	0.979	664	0.697	465
	3.00E-05	32	0.998	0.788	0.881	0.970	664	0.529	465
		64	1.000	0.834	0.910	0.968	664	0.643	465
		8	0.997	0.801	0.888	0.925	664	0.624	465
		16	1.000	0.859	0.924	0.955	664	0.723	465
	5.00E-05	32	0.998	0.843	0.914	0.965	664	0.669	465
		64	0.993	0.764	0.864	0.980	664	0.456	465
		8	0.987	0.892	0.937	0.988	1353	0.813	1645
		16	0.984	0.897	0.938	0.987	1353	0.824	1645
	1.00E-05	32	0.983	0.907	0.943	0.984	1353	0.843	1645
		64	0.980	0.902	0.939	0.978	1353	0.840	1645
		8	0.992	0.877	0.931	0.982	1353	0.793	1645
GAZ		16	0.994	0.881	0.934	0.984	1353	0.796	1645
	3.00E-05	32	0.993	0.886	0.936	0.987	1353	0.804	1645
		64	0.986	0.891	0.936	0.982	1353	0.816	1645
		8	0.992	0.880	0.932	0.987	1353	0.793	1645
		16	0.988	0.868	0.924	0.990	1353	0.768	1645
	5.00E-05	32	0.988	0.905	0.945	0.992	1353	0.835	1645
		64	0.990	0.872	0.927	0.983	1353	0.781	1645
		8	0.998	0.344	0.512	1.000	2164	0.000	4118
		16	0.997	0.345	0.513	1.000	2164	0.001	4118
	1.00E-05	32	0.997	0.345	0.513	1.000	2164	0.001	4118
IDO		64	0.993	0.345	0.512	1.000	2164	0.001	4118
	3.00E-05	8	0.999	0.344	0.512	1.000	2164	0.000	4118

		16	1.000	0.344	0.512	1.000	2164	0.000	4118
		32	0.996	0.344	0.512	1.000	2164	0.000	4118
		64	0.999	0.345	0.513	1.000	2164	0.001	4118
		8	0.991	0.344	0.511	1.000	2164	0.000	4118
	5.00E-05	16	0.992	0.345	0.512	1.000	2164	0.000	4118
		32	0.986	0.344	0.510	1.000	2164	0.000	4118
		64	1.000	0.345	0.513	1.000	2164	0.000	4118
		8	0.991	0.803	0.887	0.983	3197	0.628	3287
	1.00E-05	16	0.986	0.867	0.922	0.982	3197	0.755	3287
		32	0.981	0.772	0.864	0.978	3197	0.572	3287
		64	0.980	0.761	0.857	0.977	3197	0.551	3287
		8	0.994	0.889	0.939	0.982	3197	0.800	3287
NCBI TAXON	3.00E-05	16	0.985	0.821	0.896	0.986	3197	0.662	3287
		32	0.989	0.881	0.932	0.985	3197	0.780	3287
		64	0.983	0.738	0.843	0.981	3197	0.503	3287
		8	0.995	0.872	0.929	0.984	3197	0.764	3287
	5.00E-05	16	0.993	0.866	0.926	0.981	3197	0.756	3287
		32	0.991	0.864	0.923	0.985	3197	0.747	3287
		64	0.983	0.871	0.924	0.982	3197	0.765	3287
		8	0.976	0.546	0.700	0.979	1609	0.275	2570
	1.00E-05	16	0.965	0.663	0.786	0.976	1609	0.468	2570
		32	0.969	0.557	0.708	0.985	1609	0.290	2570
		64	0.978	0.572	0.722	0.979	1609	0.318	2570
		8	0.980	0.659	0.788	0.976	1609	0.460	2570
SO	3.00E-05	16	0.960	0.636	0.765	0.961	1609	0.435	2570
		32	0.973	0.539	0.694	0.973	1609	0.268	2570
		64	0.984	0.570	0.722	0.975	1609	0.317	2570
		8	0.911	0.519	0.661	0.976	1609	0.233	2570
	5.00E-05	16	0.932	0.555	0.696	0.990	1609	0.284	2570
		32	0.981	0.546	0.702	0.971	1609	0.281	2570
		64	0.979	0.548	0.703	0.977	1609	0.281	2570
		8	0.977	0.682	0.803	0.962	396	0.540	798
	1.00E-05	16	0.974	0.717	0.826	0.960	396	0.589	798
		32	0.966	0.683	0.801	0.949	396	0.543	798
UBERON		64	0.956	0.670	0.788	0.955	396	0.519	798
		8	0.990	0.663	0.795	0.944	396	0.531	798
	3.00E-05	16	0.989	0.627	0.767	0.957	396	0.472	798
		32	0.991	0.707	0.826	0.955	396	0.576	798
		64	0.977	0.697	0.813	0.967	396	0.551	798

5.00E-05	**8**	0.988	0.615	0.758	0.947	396	0.457	798
	16	0.995	0.682	0.809	0.962	396	0.548	798
	32	0.985	0.667	0.796	0.960	396	0.518	798
	64	0.991	0.672	0.801	0.962	396	0.525	798

Table 4.3: RE model performance. Loss, accuracy, and F1 values were calculated based on predictions made on the evaluation dataset. The number of relationships found in each lexicon's RE evaluation dataset is shown. RE models were trained on a combination of three learning rates (1e-5, 3e-5, and 5e-5) and four batch sizes (8, 16, 32, 64). All models were trained using an epoch of 4. Boxed in red are the best-performing models based on the lowest loss and overall performance used for downstream RE predictions. Green color scales were applied column-wise within lexicon boundaries to indicate the best-performing metric.

Lexicon	Learning rate	Batch size	# Examples	Loss	Accuracy	F1	Accuracy and F1
ENVO	1.00E-05	8	97	0.289	0.897	0.945	0.921
		16		0.345	0.897	0.945	0.921
		32		0.347	0.897	0.945	0.921
		64		0.403	0.897	0.945	0.921
	3.00E-05	8		0.339	0.897	0.943	0.920
		16		0.334	0.897	0.945	0.921
		32		0.368	0.897	0.945	0.921
		64		0.335	0.897	0.945	0.921
	5.00E-05	8		0.475	0.914	0.952	0.933
		16		0.334	0.897	0.943	0.920
		32		0.338	0.897	0.945	0.921
		64		0.353	0.897	0.945	0.921
FOODON	1.00E-05	8	110	0.284	0.909	0.952	0.930
		16		0.275	0.909	0.952	0.930
		32		0.295	0.909	0.952	0.931
		64		0.336	0.909	0.952	0.931
	3.00E-05	8		0.342	0.894	0.942	0.918
		16		0.284	0.864	0.926	0.895
		32		0.220	0.909	0.952	0.931
		64		0.198	0.909	0.952	0.931
	5.00E-05	8		0.421	0.894	0.942	0.918
		16		0.295	0.879	0.932	0.905
		32		0.256	0.864	0.926	0.895
		64		0.188	0.909	0.952	0.931
GAZ	1.00E-05	8	198	0.538	0.866	0.928	0.897

		16		0.408	0.874	0.933	0.903
		32		0.392	0.874	0.933	0.903
		64		0.369	0.874	0.933	0.903
		8		0.546	0.874	0.932	0.903
	3.00E-05	16		0.452	0.866	0.927	0.896
		32		0.388	0.866	0.928	0.897
		64		0.393	0.874	0.933	0.903
		8		0.554	0.874	0.932	0.903
	5.00E-05	16		0.479	0.874	0.932	0.903
		32		0.384	0.891	0.940	0.915
		64		0.390	0.866	0.928	0.897
		8		0.408	0.872	0.926	0.899
	1.00E-05	16		0.415	0.845	0.912	0.878
		32		0.434	0.831	0.906	0.868
		64		0.497	0.817	0.899	0.858
		8		0.520	0.904	0.943	0.924
IDO	3.00E-05	16	365	0.398	0.904	0.943	0.924
		32		0.329	0.881	0.930	0.906
		64		0.459	0.845	0.911	0.878
		8		0.449	0.900	0.940	0.920
	5.00E-05	16		0.545	0.881	0.931	0.906
		32		0.405	0.872	0.924	0.898
		64		0.432	0.845	0.912	0.879
		8		0.326	0.912	0.947	0.930
	1.00E-05	16		0.262	0.914	0.948	0.931
		32		0.298	0.896	0.939	0.918
		64		0.327	0.875	0.928	0.901
		8		0.274	0.936	0.961	0.949
NCBI TAXON	3.00E-05	16	837	0.231	0.940	0.964	0.952
		32		0.271	0.912	0.947	0.930
		64		0.293	0.886	0.932	0.909
		8		0.275	0.946	0.967	0.957
	5.00E-05	16		0.213	0.944	0.966	0.955
		32		0.282	0.924	0.955	0.940
		64		0.230	0.920	0.952	0.936
		8		0.305	0.921	0.950	0.936
	1.00E-05	16		0.253	0.919	0.949	0.934
		32		0.242	0.906	0.942	0.924
		64		0.265	0.908	0.942	0.925
SO		8	1,441	0.350	0.927	0.954	0.940
	3.00E-05	16		0.297	0.925	0.952	0.938
		32		0.220	0.929	0.955	0.942
		64		0.217	0.926	0.953	0.939
	5.00E-05	8		0.331	0.933	0.957	0.945

		16		0.334	0.921	0.950	0.936
		32		0.239	0.935	0.959	0.947
		64		0.247	0.928	0.954	0.941
UBERON	1.00E-05	8	100	0.322	0.850	0.919	0.884
		16		0.330	0.883	0.938	0.911
		32		0.354	0.883	0.938	0.911
		64		0.403	0.883	0.938	0.911
	3.00E-05	8		0.487	0.833	0.902	0.868
		16		0.330	0.833	0.907	0.870
		32		0.335	0.867	0.929	0.898
		64		0.346	0.883	0.938	0.911
	5.00E-05	8		0.509	0.850	0.911	0.880
		16		0.456	0.817	0.893	0.855
		32		0.305	0.833	0.907	0.870
		64		0.339	0.883	0.938	0.911

Table 4.4: NER predictions on 204k PubMed paper abstracts. The total number of annotations, unique number of terms, non-normalized terms, and papers associated with each lexicon. Non-normalized terms could not be mapped back to standardized terms found in a lexicon. Each lexicon's best-performing NER model generated annotations on 204k PubMed abstracts. Prior to normalization, annotations resulted in 28,438 unique terms. After normalization, there were 2,253,704 annotations and 20,055 unique terms from 201,791 unique papers. Of the 20,055 unique terms, 6,002 terms could not be normalized back to an ontology. Non-normalized terms accounted for 37,865 annotations while normalized terms account for 2,215,839 annotations.

	Annotations	Total terms	Non-normalized terms	Papers
ARO	112,059	4,194	2,433	29,151
ENVO	78,298	188	140	40,654
FOODON	111,696	587	191	32,673
GAZ	223,487	5,672	996	80,475
IDO	400,711	197	129	133,420
NCBI TAXON	681,153	6,847	1,443	163,418
SO	461,144	678	313	105,236
UBERON	185,156	1,692	357	78,257
Total unique	2,253,704	20,055	6,002	201,791

Table 4.5: Number of ARO-lexicon relationships generated from 204k papers. The number of unique relationship term pairs, ARO terms, total relationships, and papers associated with each lexicon is shown. Each lexicon's best-performing RE model generated relationships on 204k PubMed abstracts. In total, there are 33,580 unique pairs of ARO:lexicon terms and 110,369 relationships from 24,235 unique papers. Relationships that appeared multiple times in a single paper were counted as one relationship.

Lexicon	Relationship pairs	Relationships	Papers	Unique epi terms	Unique ARO terms
ENVO	793	2,768	1,819	48	390
FOODON	1,185	2,742	1,255	93	373
GAZ	4,233	7,706	4,000	650	775
IDO	2,829	9,931	5,620	45	1,148
NCBI TAXON	8,454	32,833	15,090	883	2,232
SO	14,240	50,338	16,171	253	2,960
UBERON	1,846	4,051	2,583	175	692
Total	33,580	110,369	24,235	2,147	3,528

Table 4.6: Relationships containing normalized and non-normalized terms from 204k papers. If both the ARO and the other lexicon terms are non-normalized, they fall under (False, False). If the ARO term is non-normalized but the other lexicon term is normalized, they fall under (False, True). If the ARO term is normalized but the other is not, they fall under (True, False). If both terms in the relationship are normalized, they fall under (True, True). Relationships that appeared multiple times in a single paper were counted as one relationship.

	Relationships (ARO, OTHER)			
Lexicon	**(False, False)**	**(False, True)**	**(True, False)**	**(True, True)**
ENVO	13	61	430	2,264
FOODON	4	153	11	2,574
GAZ	2	380	40	7,284
IDO	25	933	84	8,889
NCBI TAXON	34	2,814	237	29,748
SO	10	5,653	49	44,626
UBERON	3	475	8	3,565
Total	91	10,469	859	98,950

Table 4.7: Top 10 lexicon-lexicon term combinations based on genetic overlap. Term combinations are sorted in descending order based on the gene overlap between two epidemiology terms. Green color scales were applied column-wise within lexicon boundaries to indicate each metric's most similar term pair.

Lexicon-Lexicon	Term 1	Term 2	Jaccard similarity	Bray-Curtis	Affinity Co-occurrence	Gene overlap count
GAZ-GAZ	Japan	China	0.214	0.313	0.648	57
	[Former] State of Korea	China	0.229	0.349	0.674	54
	Europe	China	0.205	0.339	0.659	50
	Italy	China	0.186	0.293	0.650	46
	Kingdom of Spain	China	0.182	0.291	0.664	42
	[Former] State of Korea	Japan	0.240	0.413	0.676	42
	India	China	0.190	0.361	0.683	42
	China	Brazil	0.172	0.384	0.658	40
	Iran	China	0.177	0.254	0.670	40
	Kingdom of Spain	Europe	0.276	0.541	0.701	37
GAZ-ENVO	China	hospital	0.276	0.340	0.649	100
	Japan	hospital	0.201	0.485	0.643	65
	Europe	hospital	0.212	0.582	0.671	63
	China	farm	0.244	0.405	0.669	60
	[Former] State of Korea	hospital	0.195	0.417	0.664	58
	Italy	hospital	0.185	0.522	0.655	56
	Kingdom of Spain	hospital	0.185	0.553	0.680	53
	India	hospital	0.179	0.481	0.696	50
	Iran	hospital	0.173	0.439	0.683	49
	China	reservoir	0.192	0.290	0.658	46
ENVO-ENVO	reservoir	hospital	0.193	0.465	0.668	57
	hospital	farm	0.174	0.313	0.638	55
	intensive care unit	hospital	0.156	0.648	0.698	43
	reservoir	farm	0.225	0.443	0.674	36
	large river delta biome	hospital	0.099	0.239	0.650	28
	large river delta biome	farm	0.202	0.375	0.697	26
	waste treatment plant	hospital	0.089	0.250	0.699	24
	reservoir	large river delta biome	0.205	0.352	0.696	23
	large river delta biome	landlocked sea	0.415	0.635	0.810	22
	reservoir	intensive care unit	0.182	0.389	0.677	22
FOODON-ENVO	swine	farm	0.380	0.674	0.719	70
	swine	hospital	0.188	0.267	0.627	65
	water food product	hospital	0.191	0.325	0.637	63
	chicken	hospital	0.195	0.315	0.649	61
	chicken	farm	0.337	0.568	0.707	55
	water food product	farm	0.297	0.467	0.690	54
	poultry drumstick	farm	0.378	0.579	0.736	51
	poultry drumstick	hospital	0.171	0.310	0.665	50
	Bos taurus	farm	0.313	0.493	0.710	45
	Bos taurus	hospital	0.150	0.265	0.645	45
FOODON-FOODON	water food product	swine	0.280	0.485	0.676	60
	swine	chicken	0.293	0.564	0.687	58
	swine	poultry drumstick	0.310	0.536	0.716	53
	poultry drumstick	chicken	0.372	0.631	0.734	51
	swine	*Bos taurus*	0.290	0.451	0.704	51
	water food product	*Bos taurus*	0.267	0.405	0.693	44
	water food product	poultry drumstick	0.256	0.402	0.691	42
	meat (skinless)	chicken	0.289	0.564	0.707	41

	poultry drumstick	meat (skinless)	0.373	0.607	0.738	41
	swine	meat (skinless)	0.217	0.446	0.679	39
UBERON-FOODON	feces	swine	0.361	0.549	0.704	73
	animal hemisphere	swine	0.327	0.538	0.691	69
	feces	water food product	0.311	0.464	0.690	61
	animal hemisphere	poultry drumstick	0.377	0.588	0.742	58
	animal hemisphere	chicken	0.305	0.556	0.692	57
	animal hemisphere	water food product	0.266	0.422	0.672	55
	feces	poultry drumstick	0.353	0.575	0.730	54
	feces	chicken	0.285	0.517	0.685	53
	membrane organ	swine	0.136	0.178	0.600	51
	animal hemisphere	*Bos taurus*	0.295	0.472	0.705	49
UBERON-UBERON	feces	animal hemisphere	0.322	0.555	0.692	64
	proliferative region	membrane organ	0.184	0.265	0.647	60
	membrane organ	feces	0.133	0.261	0.604	48
	membrane organ	animal hemisphere	0.122	0.199	0.596	45
	membrane organ	manus	0.133	0.189	0.630	43
	renal system	animal hemisphere	0.235	0.332	0.684	40
	urine	feces	0.236	0.428	0.685	39
	blood	animal hemisphere	0.203	0.389	0.661	38
	renal system	feces	0.220	0.404	0.677	37
	feces	blood	0.189	0.446	0.655	35
NCBI TAXON-ENVO	*Escherichia coli*	hospital	0.169	0.130	0.591	169
	Klebsiella pneumoniae	hospital	0.254	0.140	0.623	139
	Pseudomonas aeruginosa	hospital	0.222	0.177	0.610	116
	Enterobacteriaceae	hospital	0.254	0.235	0.637	97
	Enterobacter cloacae	hospital	0.244	0.277	0.638	89
	Acinetobacter baumannii	hospital	0.196	0.179	0.611	81
	Escherichia coli	farm	0.086	0.157	0.610	80
	Salmonella	hospital	0.184	0.198	0.604	79
	Citrobacter freundii	hospital	0.198	0.404	0.642	64
	Escherichia coli	reservoir	0.068	0.084	0.603	63
GAZ-FOODON	China	swine	0.288	0.401	0.670	77
	China	chicken	0.273	0.376	0.681	66
	China	water food product	0.212	0.270	0.646	57
	China	poultry drumstick	0.255	0.324	0.704	56
	[Former] State of Korea	swine	0.259	0.351	0.683	49
	China	*Bos taurus*	0.192	0.245	0.664	45
	China	meat (skinless)	0.183	0.292	0.667	42
	Europe	swine	0.204	0.327	0.658	41
	[Former] State of Korea	poultry drumstick	0.308	0.435	0.711	40
	Europe	water food product	0.211	0.346	0.662	39
NCBI TAXON-FOODON	*Escherichia coli*	swine	0.112	0.153	0.607	106
	Escherichia coli	water food product	0.091	0.143	0.593	86
	Escherichia coli	chicken	0.092	0.130	0.619	85
	Salmonella	swine	0.230	0.290	0.646	73
	Klebsiella pneumoniae	swine	0.136	0.130	0.600	68
	Enterobacteriaceae	swine	0.220	0.210	0.642	65
	Escherichia coli	*Bos taurus*	0.068	0.083	0.619	63
	Escherichia coli	poultry drumstick	0.064	0.110	0.613	59
	Escherichia coli	meat (skinless)	0.063	0.093	0.624	58
	Salmonella	chicken	0.195	0.264	0.648	58
NCBI TAXON-GAZ	*Escherichia coli*	China	0.142	0.273	0.601	137
	Klebsiella pneumoniae	China	0.218	0.265	0.626	111
	Enterobacteriaceae	China	0.275	0.406	0.651	89
	Salmonella	China	0.207	0.265	0.623	76
	Escherichia coli	Japan	0.080	0.144	0.579	76
	Enterobacter cloacae	China	0.240	0.331	0.642	75
	Escherichia coli	Europe	0.081	0.139	0.621	75
	Escherichia coli	[Former] State of Korea	0.080	0.133	0.628	74
	Pseudomonas aeruginosa	China	0.144	0.169	0.587	72
	Acinetobacter baumannii	China	0.196	0.192	0.619	70

NCBI TAXON-NCBI TAXON	*Klebsiella pneumoniae*	*Escherichia coli*	0.262	0.587	0.600	275
	Pseudomonas aeruginosa	*Escherichia coli*	0.180	0.332	0.566	195
	Salmonella	*Escherichia coli*	0.187	0.410	0.615	181
	Escherichia coli	*Enterobacteriaceae*	0.163	0.530	0.611	157
	Escherichia coli	*Enterobacter cloacae*	0.155	0.389	0.623	147
	Klebsiella pneumoniae	*Enterobacteriaceae*	0.297	0.620	0.657	146
	Klebsiella pneumoniae	*Enterobacter cloacae*	0.283	0.454	0.663	135
	Pseudomonas aeruginosa	*Klebsiella pneumoniae*	0.195	0.348	0.582	130
	Escherichia coli	*Acinetobacter baumannii*	0.122	0.206	0.565	123
	Salmonella	*Klebsiella pneumoniae*	0.208	0.301	0.607	115
UBERON-NCBI TAXON	membrane organ	*Escherichia coli*	0.135	0.206	0.560	141
	membrane organ	*Pseudomonas aeruginosa*	0.200	0.255	0.598	109
	animal hemisphere	*Escherichia coli*	0.102	0.143	0.604	96
	feces	*Escherichia coli*	0.101	0.176	0.609	95
	membrane organ	*Klebsiella pneumoniae*	0.138	0.150	0.566	85
	membrane organ	*Salmonella*	0.159	0.290	0.591	72
	renal system	*Escherichia coli*	0.078	0.138	0.690	71
	proliferative region	*Escherichia coli*	0.076	0.082	0.592	71
	animal hemisphere	*Klebsiella pneumoniae*	0.142	0.130	0.610	69
	membrane organ	*Acinetobacter baumannii*	0.156	0.222	0.592	69

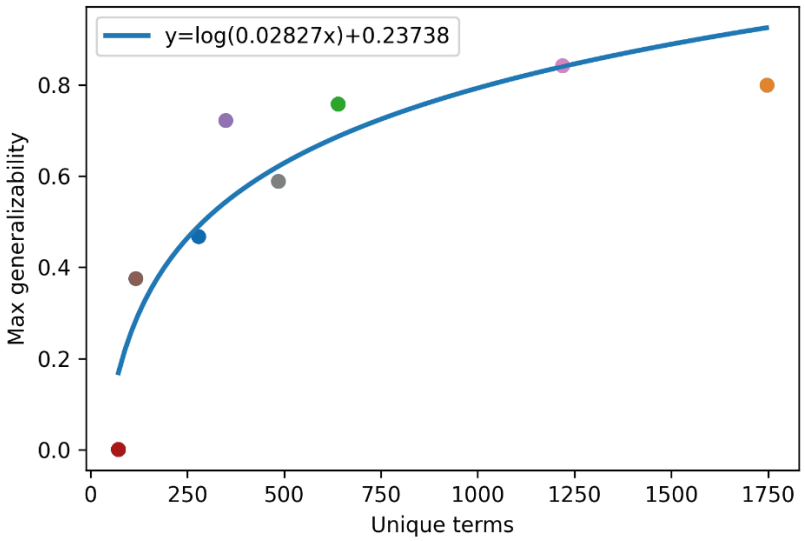
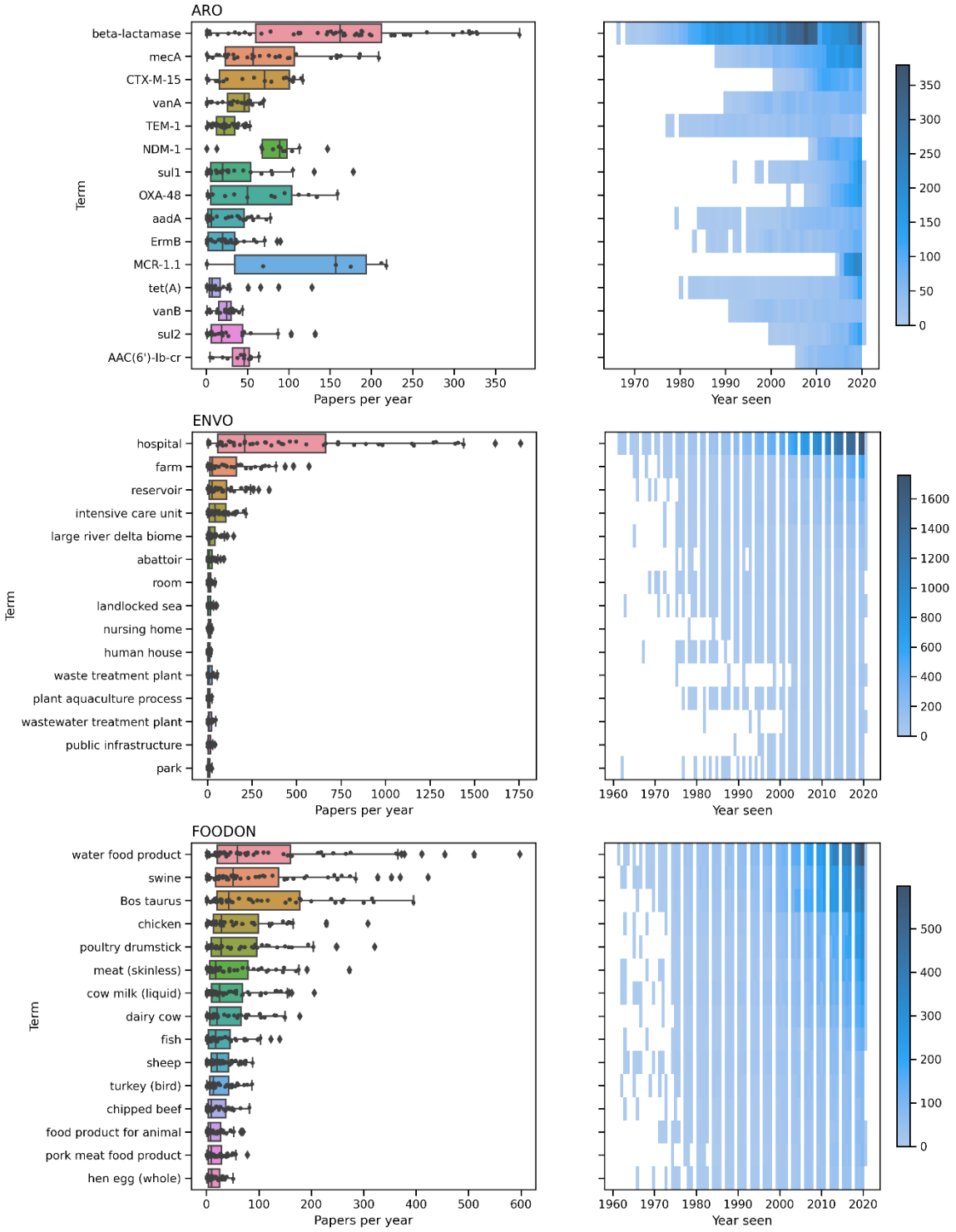
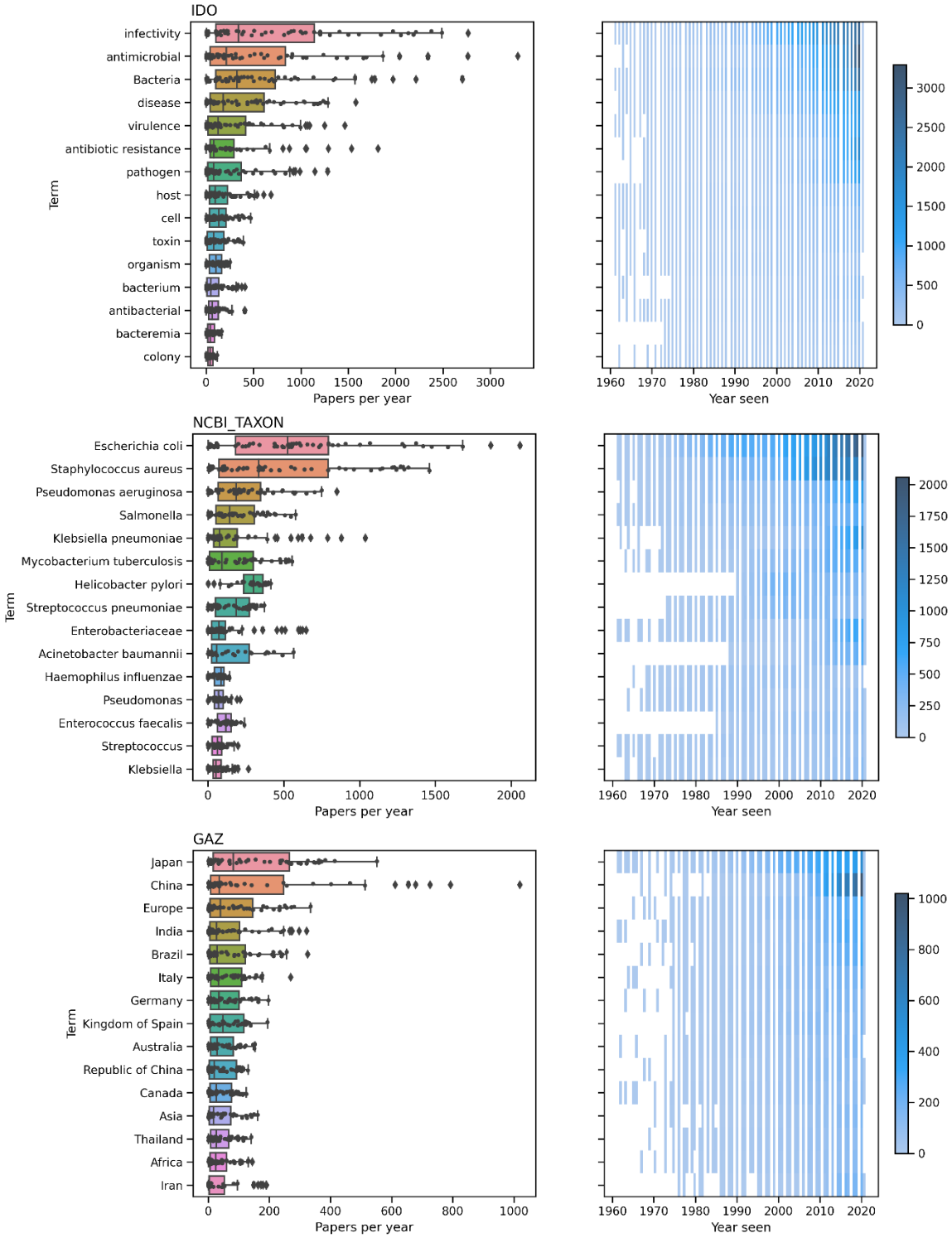


Figure 4.1: Strength of generalizability linked with dataset size. Maxiumim (max) generalizability refers to the recall value associated with each of the best-performing models outlined in Table 4.2. Unique terms refer to the number of unique terms contained in each lexicon dataset outlined in Table 3.4. The positively correlated trendline indicates that a larger dataset can achieve better recall.


ARO
beta-lactamase
mecA
CTX-M-15
vanA
TEM-1
NDM-1
sul1
OXA-48
aadA
ErmB
MCR-1.1
tet(A)
vanB
sul2
AAC(6')-Ib-cr
Term
0 50 100 150 200 250 300 350
Papers per year
1970 1980 1990 2000 2010 2020
Year seen
0 50 100 150 200 250 300 350
ENVO
hospital
farm
reservoir
intensive care unit
large river delta biome
abattoir
room
landlocked sea
nursing home
human house
waste treatment plant
plant aquaculture process
wastewater treatment plant
public infrastructure
park
Term
0 250 500 750 1000 1250 1500 1750
Papers per year
1960 1970 1980 1990 2000 2010 2020
Year seen
0 200 400 600 800 1000 1200 1400 1600
FOODON
water food product
swine
Bos taurus
chicken
poultry drumstick
meat (skinless)
cow milk (liquid)
dairy cow
fish
sheep
turkey (bird)
chipped beef
food product for animal
pork meat food product
hen egg (whole)
Term
0 100 200 300 400 500 600
Papers per year
1960 1970 1980 1990 2000 2010 2020
Year seen
0 100 200 300 400 500


IDO
infectivity
antimicrobial
Bacteria
disease
virulence
antibiotic resistance
pathogen
host
cell
toxin
organism
bacterium
antibacterial
bacteremia
colony
Term
0 500 1000 1500 2000 2500 3000
Papers per year
1960 1970 1980 1990 2000 2010 2020
Year seen
3000 2500 2000 1500 1000 500 0
NCBI_TAXON
Escherichia coli
Staphylococcus aureus
Pseudomonas aeruginosa
Salmonella
Klebsiella pneumoniae
Mycobacterium tuberculosis
Helicobacter pylori
Streptococcus pneumoniae
Enterobacteriaceae
Acinetobacter baumannii
Haemophilus influenzae
Pseudomonas
Enterococcus faecalis
Streptococcus
Klebsiella
Term
0 500 1000 1500 2000
Papers per year
1960 1970 1980 1990 2000 2010 2020
Year seen
2000 1750 1500 1250 1000 750 500 250 0
GAZ
Japan
China
Europe
India
Brazil
Italy
Germany
Kingdom of Spain
Australia
Republic of China
Canada
Asia
Thailand
Africa
Iran
Term
0 200 400 600 800 1000
Papers per year
1960 1970 1980 1990 2000 2010 2020
Year seen
1000 800 600 400 200 0

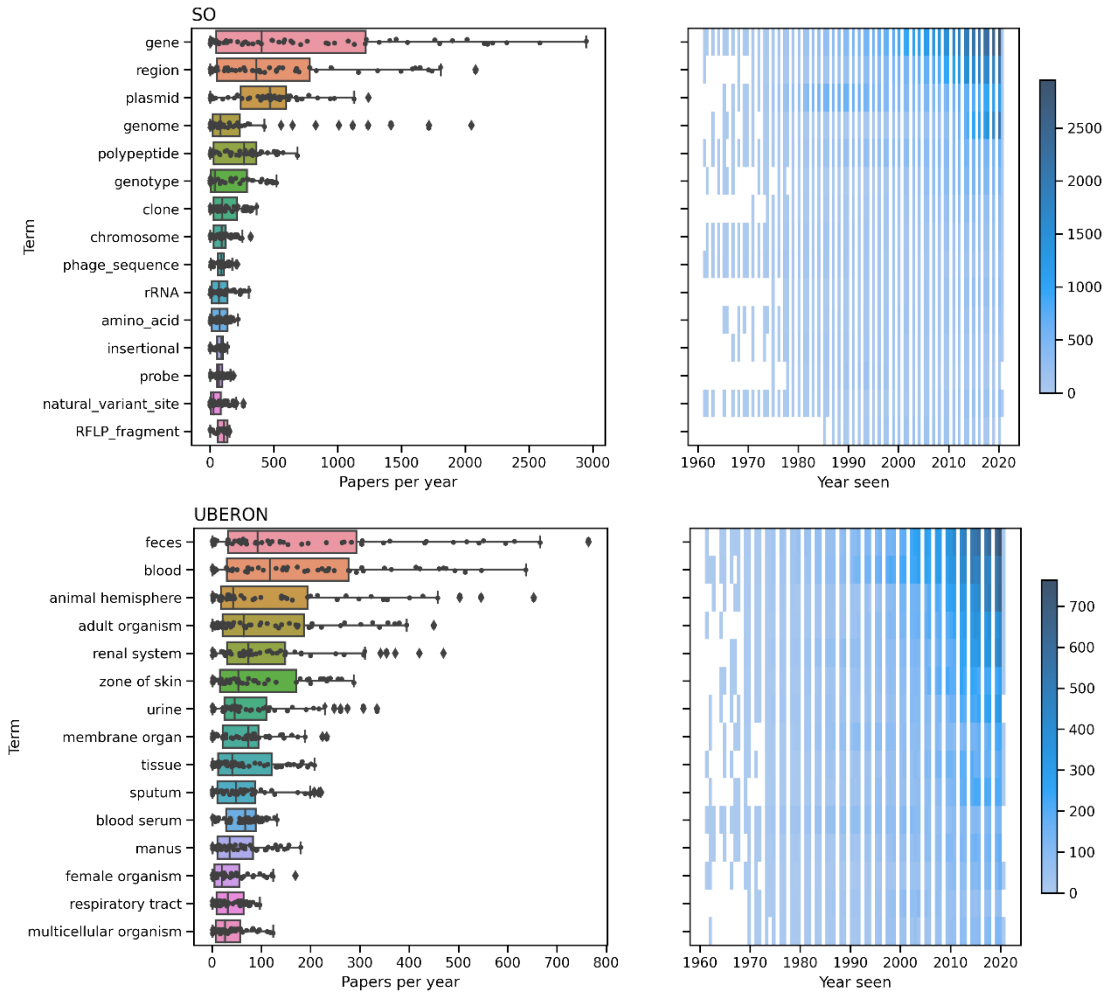


Figure 4.2: Top 15 terms in each lexicon annotated from 204k papers. A boxplot and heatmap represent the number of papers each term appears in per year, sorted in descending order by the total number of annotations.

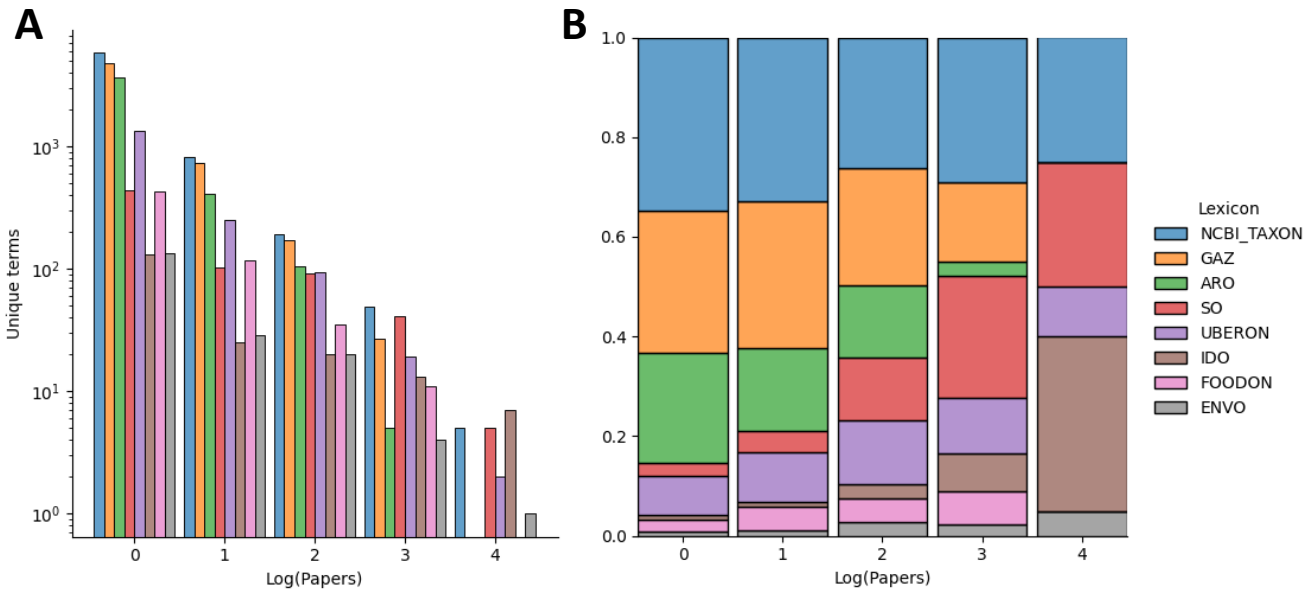


Figure 4.3: Distribution of lexicon annotations across 204k papers.

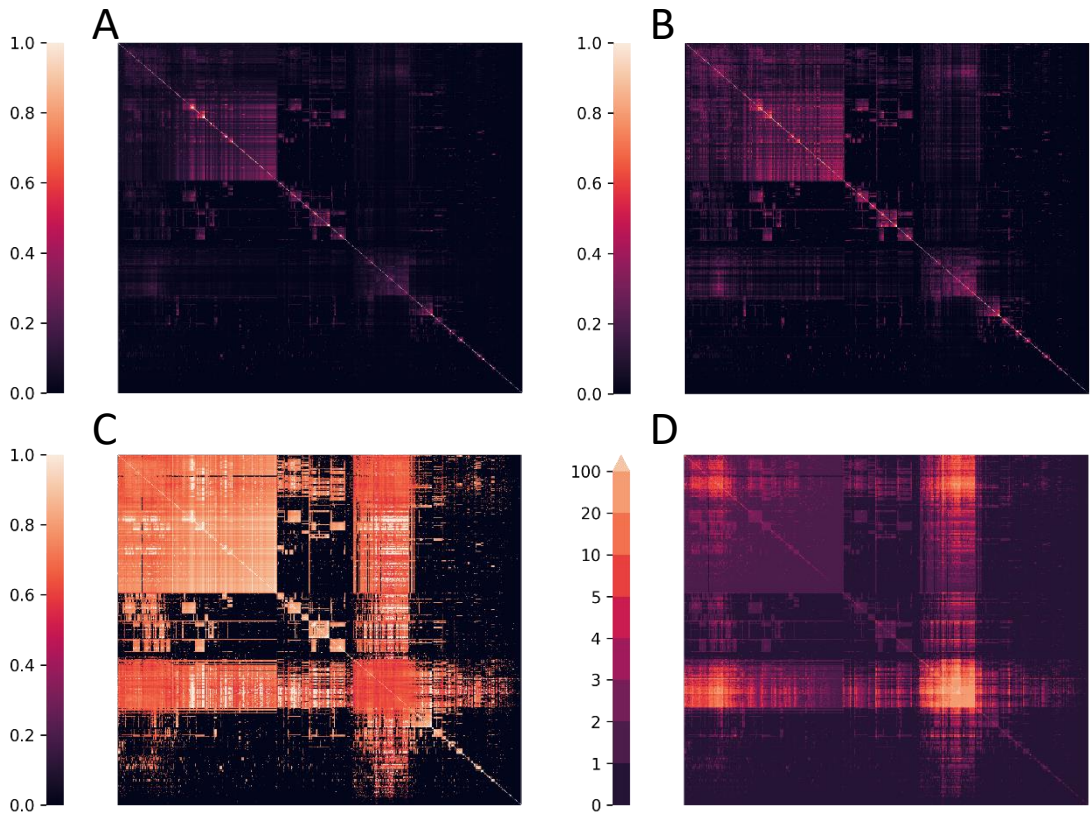


Figure 4.4: Genetic similarity between epidemiology terms. Heatmaps were generated by taking the ARO relationships with each epidemiology term and calculating a similarity metric between all epidemiology term combinations. Out of 2,147 epidemiology terms, 806 terms with only one gene relationship were not included. A) Jaccard similarity, B) Bray-Curtis, C) co-occurrence affinity, D) gene overlap. Calculations for all similarity metrics can be found in section 4.3.4. In total, there are 1,341 rows and columns, representing 1,341 different epidemiology terms for a total of 1,798,281 total term-term scores. Of the 1,341 epidemiology terms, 33 are ENVO, 65 FOODON, 380 GAZ, 37 IDO, 512 NCBI TAXON, 202 SO, and 112 UBERON.

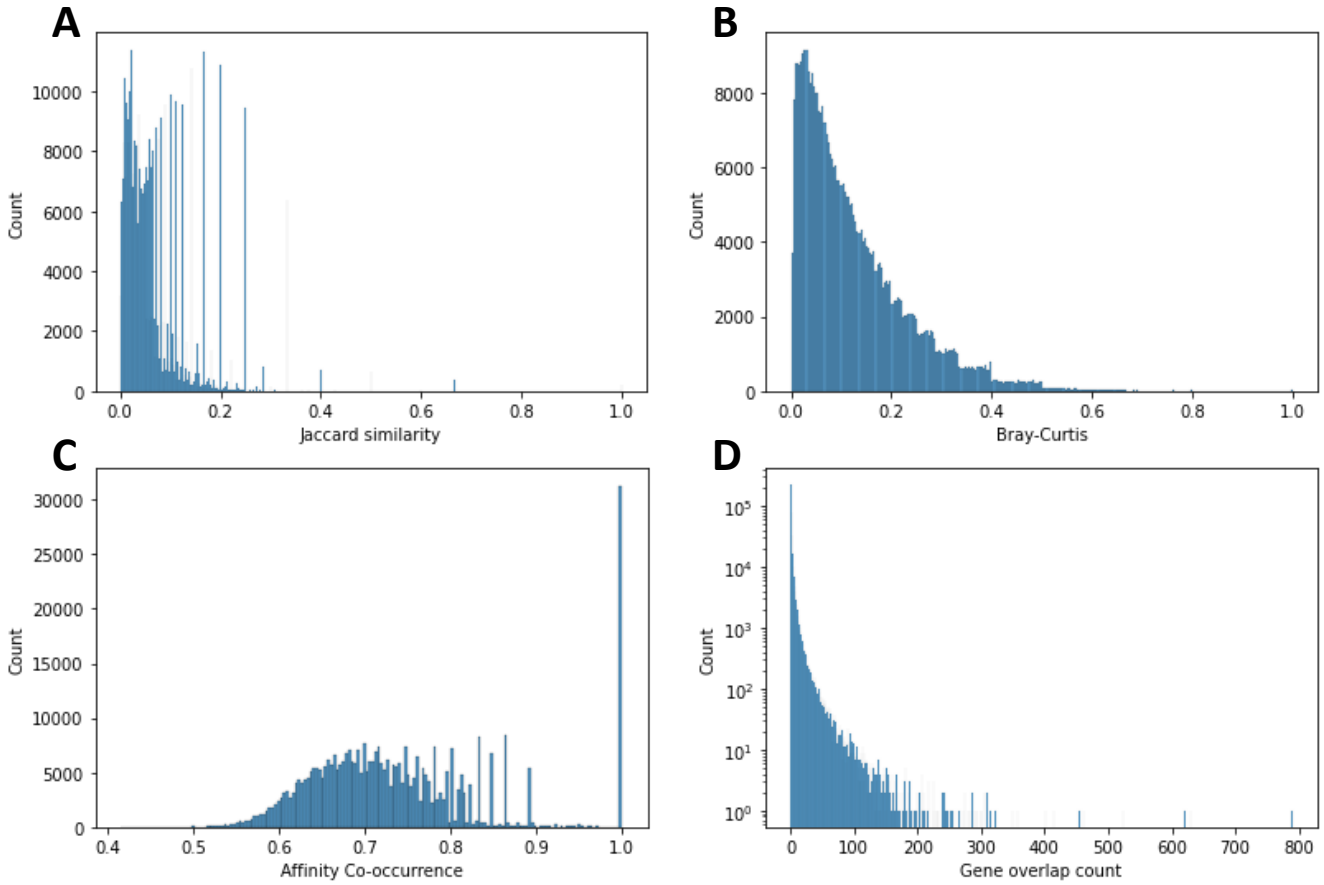


Figure 4.5: Distribution of similarity scores. Excluding duplicate term pairs or pairs between the same term, there are 398,622 pairs with at least one gene shared out of 898,470 pairs. For Jaccard, Bray-Curtis, and affinity co-occurrence, 195, 17, and 31,266 term pairs have perfect similarity scores (i.e., score of 1), respectively.

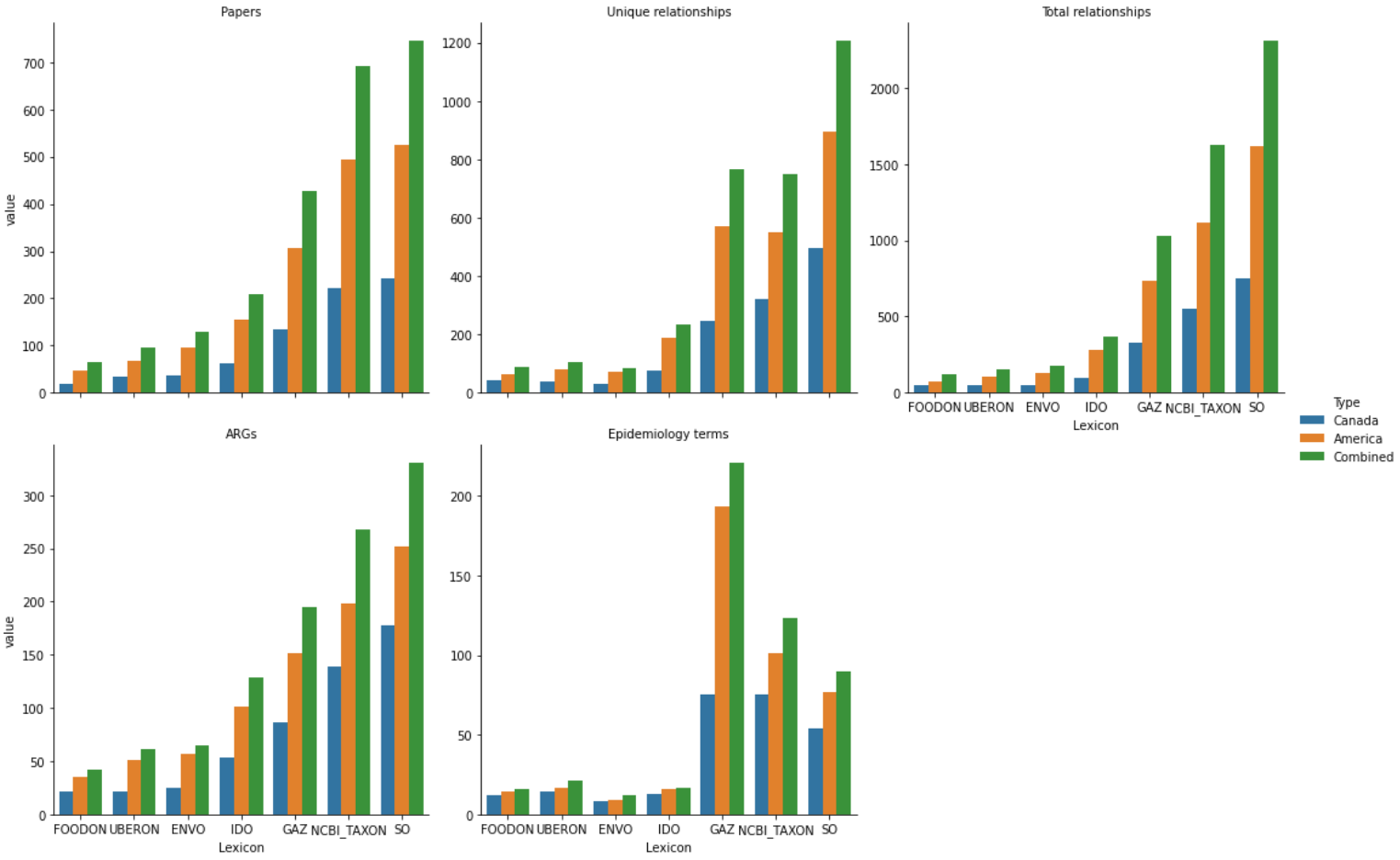


Figure 4.6: Epidemiology relationships associated with Canada and America. Papers mentioning Canada/America and all their child terms found in GAZ were selected, and the epidemiology-ARG relationships across these papers were extracted.

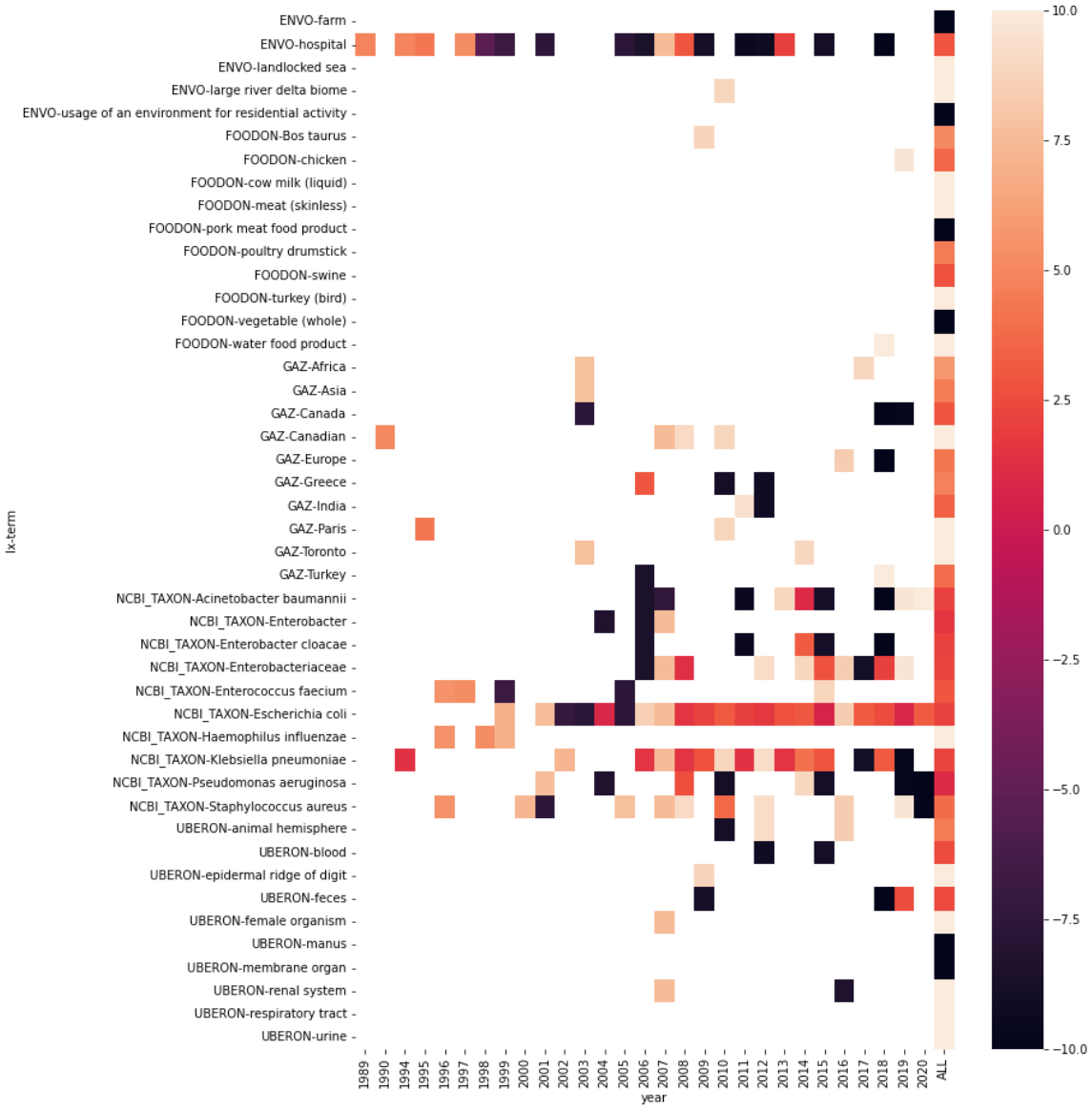


Figure 4.7: Canada-America epidemiology similarity across time. All papers mentioning “Canada” (GAZ:00002560) and/or “United States of America” (GAZ:00002459) or their child terms were selected. From these papers, all associated ARO-epidemiology terms were selected. Using these relationships, co-occurrence affinity scores were calculated between the Canadian- and American-associated epidemiology

terms. Similarity scores were calculated year-over-year, and the similarity score across all years is shown in the final column. Regions with no color indicate insufficient data to calculate a similarity score. The top ten epidemiology terms were selected for each lexicon based on the years a similarity score could be calculated. Relationships with IDO and SO terms were excluded. Only five ENVO terms were found to have enough data to calculate similarity scores.

4.4 Discussion

4.4.1 Named entity recognition (NER) models generalize well

Generalization in NER reveals a model's ability to identify terms it has never seen before as belonging to the same entity. Generalization is one of the most important aspects of a model's performance, ensuring long term value of a model based on a one-time training set. As ontologies are updated and changed, we must capture terms that previously did not exist in the ontology. As shown in Table 4.2, I have created 8 different BioBERT models that can identify different entity types corresponding to 8 ontologies. Memorization across all models was extremely high; however, generalization performance scaled with the number of unique terms in the underlying training dataset (Figure 4.1). For the three smallest ontologies with the fewest unique terms in the training datasets (IDO, ENVO, and SO), the recall values for generalization fell below 0.5 (Figure 4.1). However, as the number of unique terms increased, so did generalization performance. Generalization performance is crucial for ARO and NCBI TAXON as new ARGs are added to ARO monthly and bacterial taxonomy changes often. The ARO model was perhaps too good at generalizing as it could capture not just ARGs but 2,433 other genes that could not be normalized back to the ARO. This generalization is most likely caused since genes (ARG or other) are mentioned in similar contexts, making it challenging to create a model that can differentiate between ARGs or other genes. However, one ontology that suffers from poor generalizability is ENVO. Although the last stable release of ENVO was quite a while ago, on May 14, 2021, the ontology is under active development[100]. Thus, generalization must improve for this model until the

development of the ontology slows down and the number of new terms lessens. Once the addition of new terms to an ontology slows down dramatically, like that of GAZ, IDO, and SO, memorization metrics are enough to evaluate model performance. However, until ENVO makes a new release, the current ENVO model identifies environment terms in biomedical text well. The other ontology that generalizes poorly is UBERON. Despite having more unique terms than FOODON, the model generalizes much worse. This might be because the terms within UBERON are used in many different contexts. In biomedical text, anatomy parts are used to describe diseases (e.g., acute kidney injury) and sites of infection. Interestingly, despite generalizing worse than FOODON, the number of unique terms identified when applying the model on ~204,000 papers were more than FOODON. The FOODON model identified 587 unique terms where 191 could not be normalized. In comparison, UBERON identified 1,692 unique terms, and only 357 could not be normalized despite only being trained on 484 unique terms (Table 4.6). This indicates that perhaps the UBERON performed poorly at generalizing on the testing dataset because of some peculiarities within the testing dataset.

4.4.2 Imbalance of relationship training/testing data did not negatively impact relationship extraction (RE) performance

Annotating entities within biomedical text using NER models provides the foundation for understanding the dynamics between terms. To understand these dynamics, we must identify the relationships between terms by training RE models. Using the 7 RE training and testing datasets outlined in Chapter 3, 7 different models were trained and evaluated on their ability to identify whether relationships exist between

epidemiology and ARO terms. An area of concern with these datasets is the imbalanced number of positive relationships compared to negative ones. Positive relationships are significantly overrepresented in the RE training and testing datasets, with 82% of relationships being positive labels. While in biomedical literature, this makes sense as negative results are unlikely to be reported in an abstract, there is the potential that models trained on such data will overfit and overpredict positive relationships, impacting their performance. However, when evaluating model performance on previously unseen relationships, all models perform exceptionally well with high F1 and accuracy scores (Table 4.3). Thus, it is unlikely that the models are overfitting on the training data. Even if the models were to overpredict, since many relationships are indeed positive relationships, the impact on interpreting predictions will be negligible on a broad analysis of the data as the noise will be outweighed by correct data. However, noise generated by improper labelling of relationships will have the most impact on a gene-by-gene basis, where we would examine an individual ARG to understand the related epidemiology. Overpredicting in these situations would result in interpreting ARG-epidemiology relationships inaccurately as the number of relationships is now reduced significantly. Consequently, the impact of a single mislabel is more impactful. Overall, however, the models perform extremely well at identifying relationships between ARGs and epidemiology terms.

4.4.3 Ontologies are necessary to organize biomedical knowledge

To correctly interpret the results produced by models that annotate biomedical papers and identify relationships between these annotations, the terms identified must be

normalized back to a standard set of terms. Normalization is arguably the most important step before analyzing NER and RE results as it provides the foundation for examining the underlying data using known resources as references. Additionally, by normalizing annotations back to a standard set, terms unable to be normalized indicate that either the model is capturing terms that do not belong to the ontology or that the ontology is missing terms that should be added. For this work, the ontologies used to generate NER and RE training and testing data are the same used for normalization.

Normalization successfully limited the number of unique terms generated via the annotation process. Out of 28,438 unique terms identified by NER models, 20,055 unique terms remained after normalization (Table 4.4). Of these terms, roughly one-quarter could not be normalized. While one-quarter appears to be a lot, these terms account for only 37,865 annotations of a total of 2,215,839 total annotations or ~2% of all annotations.

While the normalization process did well in reducing the number of unique terms annotated, certain aspects can improve. Looking at the top annotations generated by the NER models, a few annotations seem out of place for the number of times they are discussed in the literature (Figure 4.2). For example, “poultry drumstick” is the fifth most discussed food term each year. The underlying annotation normalizing to “poultry drumstick” was “poultry”. However, since the word “poultry” does not appear within FOODON, the method of partial matching was used. This matching method is helpful when dealing with verbose ontology terms while the predicted annotation is short, but this is not without its drawbacks. In cases where the query is one word long, any ontology term containing that word will yield a 100% score. For example, by using this method,

the term “poultry” has a perfect 100% score with over 60 FOODON terms, including “poultry meat (dried, cooked)”, “poultry by-product”, and “poultry drumstick”. Since the method selects the first 100% match, the resulting normalized annotation is now “poultry drumstick”. While this method is vital for adequately normalizing thousands of terms, there is leniency built into it that can allow improper normalization to occur. Since normalization is such an integral part of the analysis, a manual review must be conducted to ensure that the correct terms are referenced to the correct ontology term. NER and RE validation should also be integrated during this process to update and train models continuously.

4.4.4 Similarity scores have the potential to contribute to understanding transmission but are hindered by data sparsity and differential publishing rates

Using the relationship predictions made by BioBERT NER and RE models, I wanted to see if it was possible to better understand transmission between epidemiology terms using similarity metrics as a surrogate measure. The assumption is that the more similar the two environments are based on their shared ARGs, the more likely transmission has occurred. Similarity metrics have been used extensively in ecology, computer science, and genomics to compare sets of data[101–104]. Comparing Jaccard similarity, Bray-Curtis, affinity co-occurrence, and raw gene overlap measures, I found that co-occurrence affinity could capture the strengths of all the metrics. For one, the distribution of scores for Jaccard similarity, Bray-Curtis, and gene overlap scores are skewed towards zero, which makes it difficult to compare scores across different epidemiology term pairs (Figure 4.5). In comparison, with a normal distribution, affinity

co-occurrence is more interpretable and captures regions of high similarity found in Jaccard similarity, Bray-Curtis, and gene overlap.

However, there are limitations when calculating similarity scores with epidemiology terms. A major factor influencing the strength of these similarity metrics is the number of ARGs related to each epidemiology term. This is influenced by the number of papers in which the term appears. For example, suppose we were comparing a highly mentioned term like “China”, which has 198 gene relationships, to “Province of Ontario”, which only has 13 ARG relationships. In that case, the similarity score will be low for both Jaccard similarity and Bray-Curtis, even if all 13 ARGs were shared. In this case, where 7 genes are shared, the Jaccard similarity and Bray-Curtis scores are 0.034 and 0.061, respectively. For affinity co-occurrence, the score is 0.65. Affinity co-occurrence overcomes this limitation by not being influenced by the prevalence of ARGs[99]. However, future metrics should consider the number of papers epidemiology terms and ARGs appear in when calculating similarity scores.

Additionally, the level of analysis based on ontology parentage is important to consider. While the specific term “Canada” may only have relationships with 35 unique ARGs, by selecting all ARG relationships associated with child terms to “Canada” in GAZ, we now have a set of 218 unique ARGs related to the country. By collating this information, the sparsity in these data decreases dramatically. While ontologies like GAZ are hindered by sparseness in relationship data, ontologies like ENVO are hindered due to their structure and inability to capture nuance. As mentioned in Section 3.4.1, ENVO does not contain enough terms to capture the nuance of existing environments. As a

result, there are only 33 ENVO terms with similarity scores. Consequently, these terms are very broad (i.e., “farm”, “hospital”, “intensive care unit”, “reservoir”, “large river delta biome”), and extracting meaningful similarity scores is problematic. One way to capture meaningful relationships between broad terms is to take a subset of relationships associated with a particular epidemiology term. For example, in this work, I take subsets of ARO-epidemiology relationships found in the same abstract as Canadian and American GAZ terms. With these sets of terms, we can calculate similarity scores within sets (i.e., similarity scores between Canadian-associated epidemiology terms) or between sets (i.e., similarity scores between Canadian- and American-associated epidemiology terms).

Similarity results between American- and Canadian-associated terms indicate that there is not enough information gathered through this process to assess the transmission dynamics between countries accurately. Limitations in how relationships are only predicted for terms that appear in the same sentence contribute to this lack of information. Additionally, the publication output between countries and their level of surveillance impacts these results. Countries that conduct deep surveillance would identify more ARGs in various epidemiological sources relative to countries that lack the surveillance capacity to do the same depth of analysis. For example, in Canada, we conduct limited environmental surveillance of soil and water[105]. Thus, environmental sources related to Canada are much less common in the literature, making it difficult to compare results with other countries that conduct more rigorous surveillance in these environments. The number of ARGs subsequently identified would impact similarity metrics. With this

sparse and heterogeneous data, similarity analysis should be conducted on a country-by-country basis or on a country-agnostic scale.

Overall, many aspects of the generated NER and RE predictions influence the similarity scores between epidemiology terms that should be considered when conducting future examinations of the data. The quantity of ARO relationships, number of publications, lexicon category, and depth of ontology parentage should be considered when calculating and evaluating similarity scores.

Chapter 5: Discussion and future directions

Biomedical publications contain an untapped wealth of knowledge that is impossible to extract through slow and laborious manual curation efforts. To extract this knowledge accurately and reliably, manually reviewed gold-standard datasets must be created to train and evaluate machine learning models. Throughout this work, I show the process of identifying relevant publications (Chapter 2), creating 15 gold-standard datasets to train NER and RE models (Chapter 3), and the creation, evaluation, and application of such models on ~204,000 papers (Chapter 4). With this work, we can better understand the epidemiology associated with ARGs, the similarity between epidemiology terms, and in the future, use these data to assess the risk associated with ARGs.

Understanding the risk and transmission of AMR is a challenging task. Recently, gene-level risk assessment has been explored using selection criteria depending on if the ARG is found in a human environment, if a gene is mobile, whether it is found on a human pathogen, and, in one study, how many antibiotics the ARG confers resistance towards[106,107]. While one method aimed to assess risk qualitatively[106], the other quantitatively measured gene risk[107]. While these studies do not evaluate the risk posed by AMR or the risk associated with transmission, they provide important frameworks for assessing the risk of individual ARGs. Future work should apply these frameworks to assess how publication data compares to metagenomics data.

The “Confusogram” is a graph that provides a One Health perspective on the spread of bacteria by examining transmission between humans, animals, and the

environment (Figure 5.1)[108]. While one of the main aims of this project was to assign metrics to the transmission events that take place on this graph, I could not complete this aim in time because of the challenges faced by exploring similarity metrics. I believe a variant of the Confusogram is possible that includes similarity scores between epidemiology terms at different ontology levels to capture the differences between macro- and micro-associated relationships. Additionally, to avoid the noise generated by the thousands of relationships and to simplify the analysis, relationships associated with a single country should be considered. Since transmission events are dynamic and can change with time, it is important to consider the temporal aspect of relationships and the publications in which they are found. Preferably, temporal dynamics should be considered on large time scales of years rather than months because of the reporting delay between data collection, data analysis, and publication. Some programs report data quicker than others, so comparisons between results should be taken on longer time scales to account for the variability in reporting speed. By zooming out on a timeline, the timing variability becomes less influential.

Aside from risk assessment and Confusogram metrics, there are many future directions for this project, including:

I) Open-sourcing the 15 gold-standard datasets and BioBERT models that were trained and evaluated. I am open to dataset name suggestions.

II) Incorporating human-in-the-loop intervention to evaluate model performance continuously. Having human evaluators examine a subset of model predictions to evaluate if the model is performing well has the

simultaneous benefit of expanding the training/testing datasets for future training. As new stable releases of ontologies are made public, datasets should be updated to accommodate the new terms and BioBERT models retrained and evaluated.

III) Applying BioBERT models against full-text articles and the entirety of PubMed to capture any overlooked publications.

IV) Since most relationships in the RE testing/training datasets can be guessed correctly based on whether they appear in the same sentence, we can use distant supervision techniques for training RE models[109]. Distant supervision uses thousands of abstracts that are weakly labelled using rules-based methods without care for the false positives or negatives produced. Since the number of correct relationships will overwhelmingly outweigh the number of false positives/negatives, models can accurately identify relationships[109]. Since this is a less laborious annotation task, its performance should be evaluated compared to gold-standard datasets. This would inform us if there were a need to manually label relationship datasets when ontologies are updated.

Overall, I have created a text classification algorithm from this work that can identify publications containing AMR-epidemiology information. With this classifier, I selected ~10,000 papers to create 15 gold-standard training and testing datasets: 8 NER datasets that contain manually filtered ontology labels, and 7 RE datasets that identify ARO-epidemiology relationships. With these datasets, I trained and evaluated 8 NER

BioBERT models for identifying ARO and epidemiology terms and 7 RE BioBERT models to identify relationships between identified ARO-epidemiology terms. I applied these models on a set of ~200,000 papers identified by the text classification algorithm to generate thousands of annotations and relationships. By exploring the similarity metrics between these epidemiology terms, I identified limitations and aspects to consider when conducting future similarity analyses.

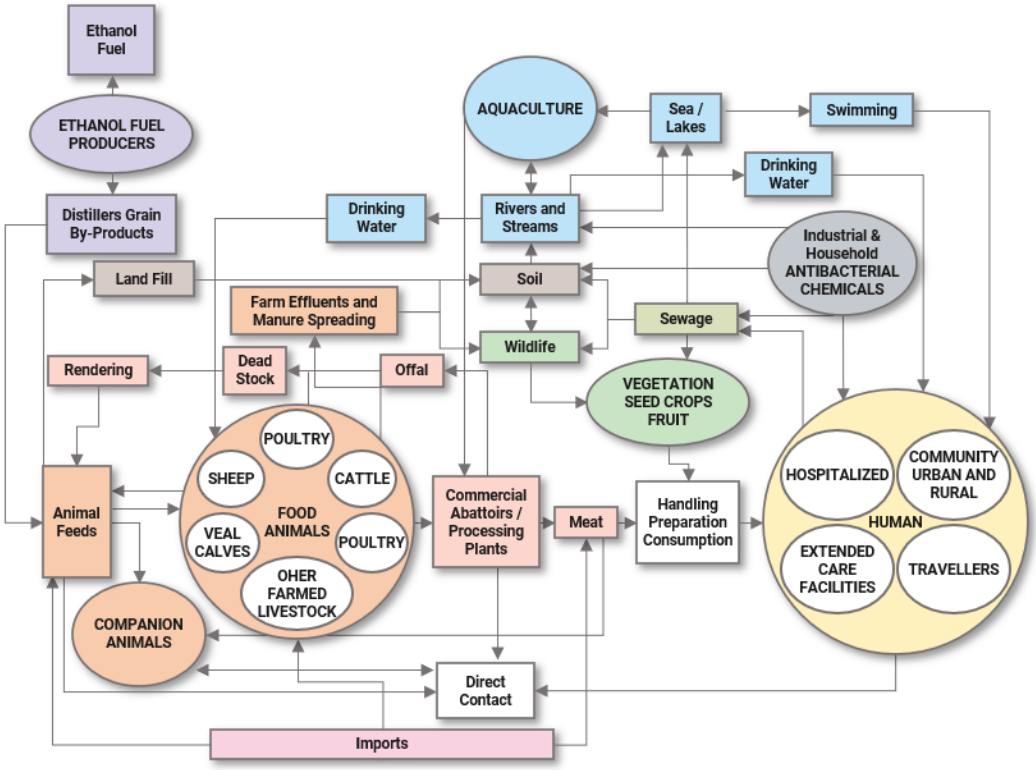


Figure 5.1: "Confusogram" revealing the pathways bacteria can spread between human, animal, and environmental sources. Adapted from Canadian Food Inspection Agency, Genome Canada. 2016. Workshop Report - Forum on Genomics and Antimicrobial Resistance[108].

Bibliography

1. Finlay, B. B. *et al.* When Antibiotics Fail: The Expert Panel on the Potential Socio-Economic Impacts of Antimicrobial Resistance in Canada. (2019).

2. O'Neill, J. Tackling Drug-Resistance Infections Globally: Final Report and Recommendations. *Rev Antimicrobial Resist* (2016).

3. Murray, C. J. *et al.* Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet* **399**, 629–655 (2022).

4. von Wintersdorff, C. J. H. *et al.* Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer. *Front. Microbiol.* **7**, 173 (2016).

5. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).

6. Boolchandani, M., D'Souza, A. W. & Dantas, G. Sequencing-based methods and resources to study antimicrobial resistance. *Nat. Rev. Genet.* **20**, 356–370 (2019).

7. Hendriksen, R. S. *et al.* Using Genomics to Track Global Antimicrobial Resistance. *Front. Public Health* **7**, 242 (2019).

8. Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525 (2019).

9. Jia, B. *et al.* CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45**, D566–D573 (2017).

10. Bortolaia, V. *et al.* ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob. Chemother.* **75**, 3491–3500 (2020).

11. Gupta, S. K. *et al.* ARG-ANNOT, a New Bioinformatic Tool To Discover Antibiotic Resistance Genes in Bacterial Genomes. *Antimicrob. Agents Chemother.* **58**, 212–220 (2014).

12. Reference Gene Catalog - Pathogen Detection - NCBI. *National Library of Medicine (US), National Center for Biotechnology Information* https://www.ncbi.nlm.nih.gov/pathogens/refgene/# (2021).

13. Holmes, A. H. *et al.* Understanding the mechanisms and drivers of antimicrobial resistance. *The Lancet* **387**, 176–187 (2016).

14. Pires, S. M., Duarte, A. S. & Hald, T. Source Attribution and Risk Assessment of Antimicrobial Resistance. *Microbiol. Spectr.* **6**, 6.3.04 (2018).

15. Lin, Y. *et al.* Metadata Analysis of mcr-1-Bearing Plasmids Inspired by the Sequencing Evidence for Horizontal Transfer of Antibiotic Resistance Genes Between Polluted River and Wild Birds. *Front. Microbiol.* **11**, 352 (2020).

16. Stokes, H. W. & Gillings, M. R. Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens. *FEMS Microbiol. Rev.* **35**, 790–819 (2011).

17. Munoz-Price, L. S. *et al.* Clinical epidemiology of the global expansion of Klebsiella pneumoniae carbapenemases. *Lancet Infect. Dis.* **13**, 785–796 (2013).

18. Dortet, L., Poirel, L. & Nordmann, P. Worldwide Dissemination of the NDM-Type Carbapenemases in Gram-Negative Bacteria. *BioMed Res. Int.* **2014**, e249856 (2014).

19. Global Resistome Data - Resistome Tracker. *FDA* https://www.fda.gov/animal-veterinary/national-antimicrobial-resistance-monitoring-system/global-resistome-data (2021).

20. Feldgarden, M. *et al.* Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrob. Agents Chemother.* **63**, e00483-19 (2019).

21. Pathogenwatch | A Global Platform for Genomic Surveillance. https://pathogen.watch/.

22. Sánchez-Busó, L. *et al.* A community-driven resource for genomic surveillance of Neisseria gonorrhoeae at Pathogenwatch. *bioRxiv* 2020.07.03.186726 (2020) doi:10.1101/2020.07.03.186726.

23. Martiny, H.-M., Munk, P., Brinch, C., Aarestrup, F. M. & Petersen, T. N. *A curated data resource of 214K metagenomes for characterization of the global resistome*. http://biorxiv.org/lookup/doi/10.1101/2022.05.06.490940 (2022) doi:10.1101/2022.05.06.490940.

24. 2022 MEDLINE/PubMed Baseline: 33,405,863 Citations Found. https://www.nlm.nih.gov/bsd/licensee/2022_stats/2022_LO.html.

25. Bourne, P. E., Lorsch, J. R. & Green, E. D. Perspective: Sustaining the big-data ecosystem. *Nature* **527**, S16–S17 (2015).

26. Suominen, A. & Toivanen, H. Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *J. Assoc. Inf. Sci. Technol.* **67**, 2464–2476 (2016).

27. Fontaine, J.-F. *et al.* MedlineRanker: flexible ranking of biomedical literature. *Nucleic Acids Res.* **37**, W141-146 (2009).

28. Allot, A., Lee, K., Chen, Q., Luo, L. & Lu, Z. LitSuggest: a web-based system for literature recommendation and curation using machine learning. *Nucleic Acids Res.* **49**, W352–W358 (2021).

29. Simon, C. *et al.* BioReader: a text mining tool for performing classification of biomedical literature. *BMC Bioinformatics* **19**, 57 (2019).

30. Poux, S. *et al.* On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics* **33**, 3454–3460 (2017).

31. McQuilton, P. & the FlyBase Consortium. Opportunities for text mining in the FlyBase genetic literature curation workflow. *Database* **2012**, bas039 (2012).

32. Mao, Y. *et al.* Overview of the gene ontology task at BioCreative IV. *Database J. Biol. Databases Curation* **2014**, bau086 (2014).

33. Aum, S. & Choe, S. srBERT: automatic article classification model for systematic review using BERT. *Syst. Rev.* **10**, 285 (2021).

34. Cohen, A. M., Hersh, W. R., Peterson, K. & Yen, P.-Y. Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. *J. Am. Med. Inform. Assoc. JAMIA* **13**, 206–219 (2006).

35. Frunza, O., Inkpen, D. & Matwin, S. Building Systematic Reviews Using Automatic Text Classification Techniques. in *Coling 2010: Posters* 303–311 (Coling 2010 Organizing Committee, 2010).

36. Wei, C.-H., Kao, H.-Y. & Lu, Z. GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. *BioMed Res. Int.* **2015** (2015).

37. Settles, B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* **21**, 3191–3192 (2005).

38. Wei, C.-H., Harris, B. R., Kao, H.-Y. & Lu, Z. tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics* **29**, 1433–1439 (2013).

39. Leaman, R., Wei, C.-H. & Lu, Z. tmChem: a high performance approach for chemical named entity recognition and normalization. *J. Cheminformatics* **7**, S3 (2015).

40. Egorov, S., Yuryev, A. & Daraselia, N. A Simple and Practical Dictionary-based Approach for Identification of Proteins in Medline Abstracts. *J. Am. Med. Inform. Assoc. JAMIA* **11**, 174–178 (2004).

41. Raja, K., Subramani, S. & Natarajan, J. PPInterFinder—a mining tool for extracting causal relations on human proteins from literature. *Database* **2013**, bas052 (2013).

42. Donaldson, I. *et al.* PreBIND and Textomy – mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics* **4**, 11 (2003).

43. Friedman, C., Kra, P., Yu, H., Krauthammer, M. & Rzhetsky, A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* **17**, S74–S82 (2001).

44. Chowdhury, F. M. & Lavelli, A. FBK-irst : A Multi-Phase Kernel Based Approach for Drug-Drug Interaction Detection and Classification that Exploits Linguistic Information. **2**, 351–355 (2013).

45. Wu, H.-Y., Chiang, C.-W. & Li, L. Text mining for drug-drug interaction. *Methods Mol. Biol. Clifton NJ* **1159**, 47–75 (2014).

46. Krogh, A. What are artificial neural networks? *Nat. Biotechnol.* **26**, 195–197 (2008).

47. Abiodun, O. I. *et al.* State-of-the-art in artificial neural network applications: A survey. *Heliyon* **4**, e00938 (2018).

48. Vaswani, A. *et al.* Attention Is All You Need. *ArXiv170603762 Cs* (2017).

49. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv181004805 Cs* (2019).

50. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs* (2017).

51. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).

52. Chen, Q., Allot, A. & Lu, Z. LitCovid: an open database of COVID-19 literature. *Nucleic Acids Res.* **49**, D1534–D1540 (2021).

53. Kowsari, K. *et al.* Text Classification Algorithms: A Survey. *Information* **10**, 150 (2019).

54. Allahyari, M. *et al.* A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *ArXiv170702919 Cs* (2017).

55. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD 16* 785–794 (2016) doi:10.1145/2939672.2939785.

56. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).

57. MEDLINE PubMed Production Statistics. https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html.

58. Loper, E. & Bird, S. NLTK: The Natural Language Toolkit. *arXiv:cs/0205028* (2002).

59. Porter, M. F. An algorithm for suffix stripping. in *Readings in information retrieval* 313–316 (Morgan Kaufmann Publishers Inc., 1997).

60. Amazon Mechanical Turk. https://www.mturk.com/.

61. Sen, S. *et al.* Turkers, Scholars, 'Arafat' and 'Peace': Cultural Communities and Algorithmic Gold Standards. in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* 826–838 (ACM, 2015). doi:10.1145/2675133.2675285.

62. Li, J. *et al.* BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database* **2016**, baw068 (2016).

63. Doğan, R. I., Leaman, R. & Lu, Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. *J. Biomed. Inform.* **47**, 1–10 (2014).

64. Krallinger, M. *et al.* The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J. Cheminformatics* **7**, 1–17 (2015).

65. Smith, L. *et al.* Overview of BioCreative II gene mention recognition. *Genome Biol.* **9**, 1–19 (2008).

66. Collier, N. & Kim, J.-D. Introduction to the Bio-entity Recognition Task at JNLPBA. in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)* 73–78 (COLING, 2004).

67. Gerner, M., Nenadic, G. & Bergman, C. M. LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinformatics* **11**, 85 (2010).

68. Pafilis, E. *et al.* The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. *PLoS ONE* **8**, e65390 (2013).

69. Ding, J., Berleant, D., Nettleton, D. & Wurtele, E. Mining Medline: Abstracts, Sentences, or Phrases? *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* **7**, 326–337 (2001).

70. Bravo, À., Piñero, J., Queralt-Rosinach, N., Rautschka, M. & Furlong, L. I. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics* **16**, 55 (2015).

71. van Mulligen, E. M. *et al.* The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. *J. Biomed. Inform.* **45**, 879–884 (2012).

72. Peng, Y., Rios, A., Kavuluru, R. & Lu, Z. Extracting chemical–protein relations with ensembles of SVM and deep learning models. *Database* **2018**, bay073 (2018).

73. Arighi, C. N. *et al.* Overview of the BioCreative III Workshop. *BMC Bioinformatics* **12**, 1–9 (2011).

74. Islamaj Doğan, R. *et al.* Overview of the BioCreative VI Precision Medicine Track: mining protein interactions and mutations for precision medicine. *Database* **2019**, bay147 (2019).

75. Dooley, D. M. *et al.* FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *Npj Sci. Food* **2**, 1–10 (2018).

76. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **40**, D136–D143 (2012).

77. Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**, 1251–1255 (2007).

78. Eilbeck, K. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* **6**, R44 (2005).

79. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* **13**, 1–20 (2012).

80. Cowell, L. Infectious Disease Ontology. http://purl.obolibrary.org/obo/ido.owl (2017).

81. Buttigieg, P. L. *et al.* The environment ontology: contextualising biological and biomedical entities. *J. Biomed. Semant.* **4**, 43 (2013).

82. Schriml, L. & Ashburner, M. Gazetteer. http://purl.obolibrary.org/obo/gaz.owl (2018).

83. Jackson, R. C. *et al.* ROBOT: A Tool for Automating Ontology Workflows. *BMC Bioinformatics* **20**, 407 (2019).

84. Index of /pub/taxonomy/new_taxdump. https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump/.

85. fix: remove exact synonym spring · EnvironmentOntology/envo@64e46ef · GitHub. https://github.com/EnvironmentOntology/envo/commit/64e46ef5fcf86fcde0f6e6032d4c823d282cbe58#diff-978de7b7280c6684ab5cc6cc010fc87496c9195077f1721c1585a42ce90418e9R2327.

86. Release 2021-05-14 release · EnvironmentOntology/envo · GitHub. https://github.com/EnvironmentOntology/envo/releases/tag/v2021-05-14.

87. Popovski, G., Seljak, B. K. & Eftimov, T. FoodBase corpus: a new resource of annotated food entities. *Database* **2019**, baz121 (2019).

88. Pafilis, E. *et al.* ENVIRONMENTS and EOL: identification of Environment Ontology terms in text and the annotation of the Encyclopedia of Life. *Bioinformatics* **31**, 1872–1874 (2015).

89. Sang, E. F. T. K. & De Meulder, F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Preprint at http://arxiv.org/abs/cs/0306050 (2003).

90. RE datasets · Issue #162 · dmis-lab/biobert. *GitHub* https://github.com/dmis-lab/biobert/issues/162.

91. Questionable training labels in GAD RE dataset · Issue #153 · dmis-lab/biobert. *GitHub* https://github.com/dmis-lab/biobert/issues/153.

92. McCubbin, K. D. *et al.* Knowledge Gaps in the Understanding of Antimicrobial Resistance in Canada. *Front. Public Health* **9**, 1523 (2021).

93. Rudnick, W. *et al.* Antimicrobial use among adult inpatients at hospital sites within the Canadian Nosocomial Infection Surveillance Program: 2009 to 2016. *Antimicrob. Resist. Infect. Control* **9**, 32 (2020).

94. Deckert, A. *et al.* CIPARS: A One-Health Approach to Antimicrobial Resistance Surveillance. *Online J. Public Health Inform.* **7**, e68 (2015).

95. Canadian Lung Association, Canadian Thoracic Society, Public Health Agency of Canada, & Centre for Communicable Diseases and Infection Control (Canada). *Canadian tuberculosis standards*. (2014).

96. Public Health Agency of Canada. National surveillance of antimicrobial susceptibilities of Neisseria gonorrhoeae annual summary 2019. https://www.canada.ca/en/services/health/publications/drugs-health-products/national-

surveillance-antimicrobial-susceptibilities-neisseria-gonorrhoeae-annual-summary-2019.html (2021).

97. Public Health Agency of Canada. National laboratory surveillance of invasive streptococcal disease in Canada - Annual summary 2019. https://www.canada.ca/en/public-health/services/publications/drugs-health-products/national-laboratory-surveillance-invasive-streptococcal-disease-canada-annual-summary-2019.html (2021).

98. World Health Organization. *Global antimicrobial resistance and use surveillance system (GLASS) report: 2021*. (World Health Organization, 2021).

99. Mainali, K. P., Slud, E., Singer, M. C. & Fagan, W. F. A better index for analysis of co-occurrence and similarity. *Sci. Adv.* **8**, eabj9204 (2022).

100. The Environment Ontology. https://github.com/EnvironmentOntology/envo (2022).

101. Janson, S. & Vegelius, J. Measures of ecological association. *Oecologia* **49**, 371–376 (1981).

102. Smith, T. F. & Waterman, M. S. Comparison of biosequences. *Adv. Appl. Math.* **2**, 482–489 (1981).

103. De Mántaras, R. L. A Distance-Based Attribute Selection Measure for Decision Tree Induction. *Mach. Learn.* **6**, 81–92 (1991).

104. Niwattanakul, S., Singthongchai, J., Naenudorn, E. & Wanapu, S. Using of Jaccard Coefficient for Keywords Similarity. in *Proceedings of the international multiconference of engineers and computer scientists* vol. 1 380–384 (2013).

105. Progress on Integrated Antimicrobial Resistance and Antimicrobial Use Surveillance in Canada (2014-2019). *National Collaborating Centre for Infectious Diseases* https://nccid.ca/publications/progress-on-integrated-antimicrobial-resistance-and-antimicrobial-use-surveillance-in-canada/ (2021).

106. Zhang, A.-N. *et al.* An omics-based framework for assessing the health risk of antimicrobial resistance genes. *Nat. Commun.* **12**, 4765 (2021).

107. Zhang, Z. *et al.* Assessment of global health risk of antibiotic resistance genes. *Nat. Commun.* **13**, 1553 (2022).

108. Forum on Genomics and Antimicrobial Resistance – Workshop Report | Genome Canada. https://www.genomecanada.ca/en/news/forum-genomics-and-antimicrobial-resistance-workshop-report.

109. Mintz, M., Bills, S., Snow, R. & Jurafsky, D. Distant supervision for relation extraction without labeled data. in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - ACL-IJCNLP '09* vol. 2 1003 (Association for Computational Linguistics, 2009).