Impacts of Medication Use on Child Gut Microbiota Outcomes in the CHILD Cohort Study

by

Emma Garlock

B.Sc. (Biochemistry), Mount Allison University, 2019

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Molecular Biology and Biochemistry
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2022

Declaration of Committee

Name:	**Emma Garlock**
Degree:	**Master of Science**
Title:	**Impacts of Medications Use on Child Gut Microbiota Outcomes in the CHILD Cohort Study**
Committee:	**Chair:** **Amy Lee** Assistant Professor, Molecular Biology and Biochemistry
	Fiona Brinkman Supervisor Professor, Molecular Biology and Biochemistry
	Lisa Craig Committee Member Professor, Molecular Biology and Biochemistry
	William Hsiao Committee Member Associate Professor, Molecular Biology and Biochemistry
	Pablo Nepomnaschy Committee Member Professor, Health Sciences
	Ryan Morin Examiner Associate Professor, Molecular Biology and Biochemistry

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

The gut microbiota is a complex ecosystem playing a role in health, primarily colonized in the first year of life. Some medication use, such as antibiotics, can cause microbial dysbiosis (unusual microbiome composition) and is associated with development of pathologies such as asthma. However, the impact of other medication use on the gut microbiome is poorly characterized. In this thesis, I use the rich data collected in the CHILD Cohort Study, to develop models examining the impacts of medication practices on infant microbial dysbiosis and associated child outcomes. Certain medication use was associated with changes in the gut microbiome, but the results suggest medication use may be a proxy for other lifestyle factors that make the child more prone to microbial dysbiosis. This work forms the base for further characterizations of the links between medication use, metabolism, and lifestyle, to identify the most effective intervention points for preventing microbial dysbiosis.

Keywords: Microbiota; Microbial dysbiosis; Medication; Machine learning; Lifestyle

Acknowledgements

The first and biggest thank you goes to my family. I would not have even started this degree without all of you supporting me when I decided to go so far away from home. COVID made the distance feel much farther than it was but thank goodness you were all only a phone call away when I needed you. I would need another thesis to fully thank you for everything but, I only have one page and there are some other people I need to thank.

That being said, thank you to my supervisor, Dr. Fiona Brinkman, for her enthusiasm during this project and all her support as I have worked to find the right career path for myself. Thank you to my committee members, Dr. Lisa Craig, Dr. Will Hsiao and Dr. Pablo Nepomnaschy. Your diverse expertise and feedback have made me a better researcher. Thank you to my fellow Brinkman Lab members and especially the other grad students: Kristen, Justin, Venus, and Geoff. Lab meetings would have been a lot less fun without getting to hang out with all of you. As a fellow CHILD Cohort Study researcher, a special thank you to Geoff for always being so helpful, I don't think it is an exaggeration to say that my project would not have been possible without Geoff and all the work he put in to support the CHILD Cohort Study before I even arrived.

When I left, I was nervous to move away from my MtA friends, I didn't know what I would do without them. Thankfully, Kat has continued to be my grad school ally. I am so glad I had you to remind me to take a step back and confirm that I wasn't going crazy. To Joe, I appreciate your irregularly scheduled phone calls so much. I can't list all the ways they have helped keep me sane. To Sarah and Nick, thank you both for struggling though stats questions with me even though we are 4000km apart.

I was very lucky to find a wonderful support system in my new friends in Burnaby. Kristen gets a second shout-out for all the fun we have outside of the lab too. I am so glad I found someone who gets as excited about knitting conventions and period dramas as I do. For the rest of the group, Matt, Samantha, and Stephen: Thank you so much for so many fun DnD sessions and Great Burnaby Bake-Offs. Meeting all of you is one of the highlights of my time on the West Coast. To Rachel, thank goodness our book shopping schedules aligned. It's always a joy to take a break, fangirl, and explore coffee shops with you. Finally, shout out to my cat, Nova, for always walking across my keyboard and giving me someone to blame my typos on.

Table of Contents

List of Tables

List of Figures

List of Acronyms

%IncMSE	Decrease in Mean Squared Error
ANCOM-BC	Analysis of Compositions of Microbiomes with Bias Correction
BMI	Body Mass Index
CINECA	Common Infrastructure for National Cohorts in Europe, Canada, and Africa
cv.error	Cross validation error
DOID	Human Disease Ontology
EB_1y	Enterobacteriaceae-to-Bacteroidaceae at 1 Year
EB_3m	Enterobacteriaceae-to-Bacteroidaceae at 3 Months
FB_1y	Firmicutes-to-Bacteroidetes at 1 year
FB_3m	Firmicutes-to-Bacteroidetes at 3 Months
GBM	Gradient Boosting Machine
HP	Human Phenotype Ontology.
MICE	Multiple Imputation by Chained Equations
NCIT	National Cancer Institute Thesaurus
OAE	Adverse Event Ontology
OTC	Over the Counter
RelInf	Relative influence Score
RF	Random Forest
RMSE	Root Mean Squared Error
SCFA	Short Chain Fatty Acid
SEM	Structural Equation Modelling
SYMP	Symptom Ontology

Glossary

CHILD	A prospective, longitudinal birth cohort study with study centres located in Vancouver, Edmonton, Winnipeg and Toronto. The study was launched in 2009 with 3500 pregnant mothers that gave birth between 2009 and 2012.
Child	free text term was more specific than the assigned ontology
Domains	For correlation analyses, a set of variables that have been determined to be related.
Free Text	String based data collected from individuals who are allowed to write in any answer they see fit. There are no predetermined responses.
Gut Microbiome	The collection of genetic information from all individuals found (ie the microbiota) in the gut
Gut Microbiota	The entire collection of microorganisms that inhabits the gut.
Match	the assigned term was an appropriate term for the free text entry
Microbial dysbiosis	An imbalance it the gut microbiota generally associated with disease that is different from expected development.
No match	the assigned term was a not appropriate term for the free text entry
Off-Label	Usage of a medication that is different from the approved indication.
Ontology	Standard terms that contain their properties and relationships to other terms within the defined scope.
Parent	free text term was less specific than the ontology term

Chapter 1.

Introduction

1.1. Composition and Function of The Gut Microbiota

The gut microbiota is defined as the population of bacteria and archaea that inhabit the gastrointestinal tract (Thursby & Juge, 2017). These microbes coexist with their host and tremendously impact the host's metabolic pathways and immune response (Zou et al., 2019). In turn, the environment of the host influences the gut microbiota. While there is no one “healthy” gut microbiota, a healthy gut microbiota is generally high in species diversity, maintains homeostasis and is compatible with the host’s diet, allowing an individual to thrive in their unique environment (McBurney et al., 2019.; Yatsunenko et al., 2012). However, several phyla are consistently identified. Using data from King et al., Figure 1.1 shows proportions of common phyla in adult gut microbiomes, and Figure 1.2 shows the breakdown of these phyla and some examples of their family membersFigure 1.2 (Rinninella et al., 2019).

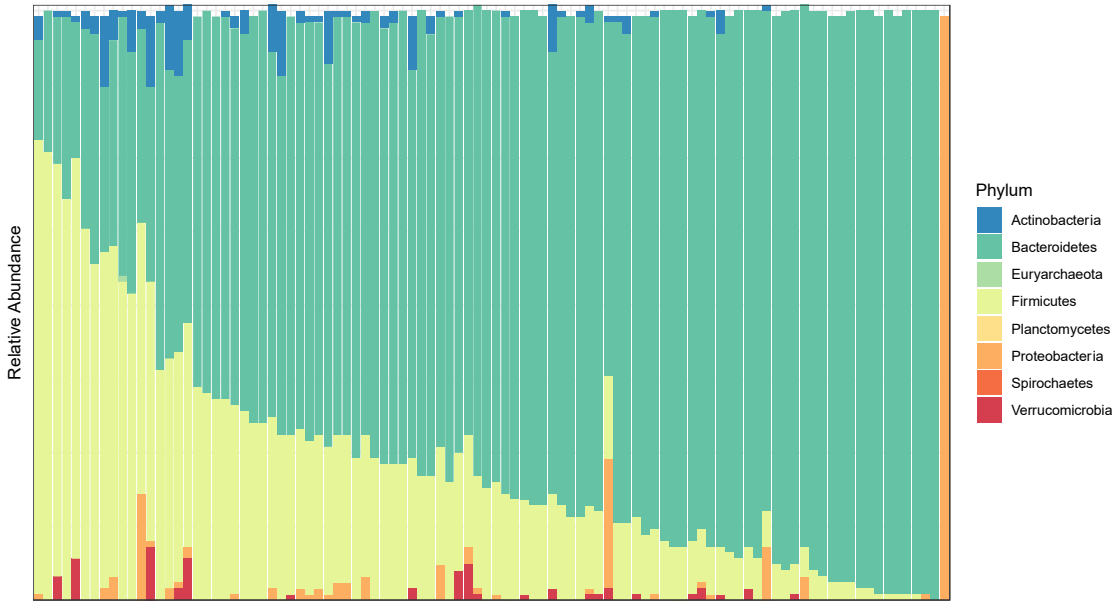

Figure 1.1. Common proportions of taxa found in the gut microbiota from an adult cohort (n=98). Taxa are shown at the phylum level with the corresponding colours indicated in the legend. For ease of visibility, the plot is sorted with the highest proportion of Firmicutes on the left with the lowest on the right. Data obtained from King et al., 2019.

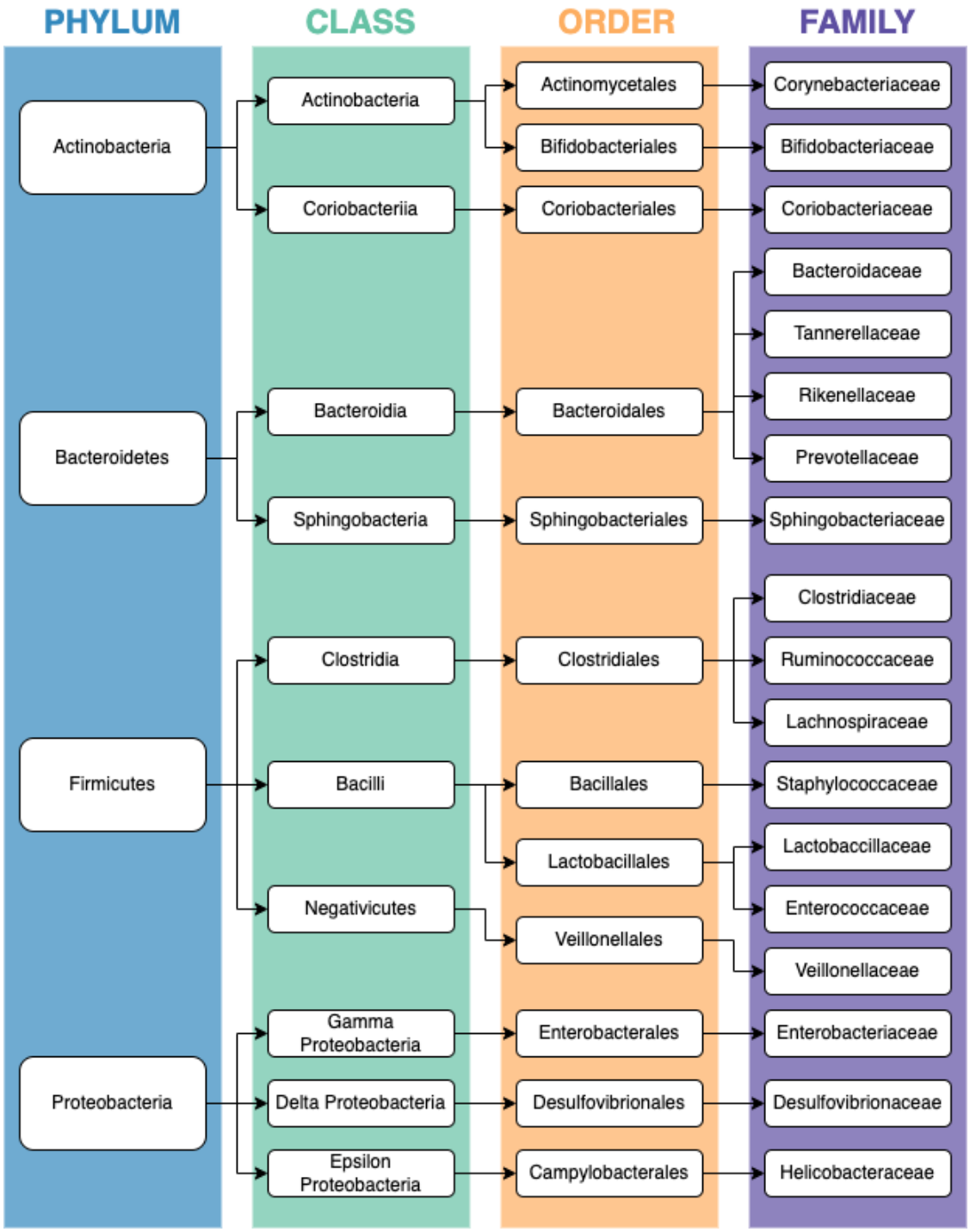

Figure 1.2. **Taxonomic Breakdown of Dominant Phyla in the Human Gut Microbiome. Taxa are organized alphabetically by phylum from top to bottom. Arrows originating in the Phylum column indicate which lower taxonomic labels are members of their rank. The taxa shown in this figure are non-exhaustive, and the figure is adapted from Rinniella et al., 2019. Levels are shown for phylum, class, order and family, using the official taxonomic names from before the NCBI update in 2021.**

The composition and abundance of species vary from person to person and can change over the host's life. These individual differences are challenging to characterize. Certain studies have outlined that genetic factors can account for approximately 10% of the variation between human gut microbiome compositions, and the environment can account for approximately 20% (Scepanovic et al., 2019). While compositions can vary significantly from person to person, the required functionality remains consistent (Tian et al., 2020). Table 1.1 shows the main functionalities assigned to dominant phyla

Table 1.1.

Table 1.1. Documented role of prominent gut microbial phyla. Information shown for Actinobacteria, Firmicutes, Bacteroidetes and Proteobacteria. A simplified description of their roles can be found along with the appropriate citation. The roles described represent a sample and are a non-exhaustive list.

Phylum	Role	Citation
Actinobacteria	Protection from enteropathogenic infections	Binda et al., 2018
	Gut mucosal barrier function	Hardy et al., 2013
	Carbohydrate Metabolism	Macfarlane & Englyst, 1986
	T regulatory cell modulation	Binda et al., 2018
Firmicutes	Gut mucosal barrier integrity	Vaiserman et al., 2020
	Carbohydrate metabolism	Ottman et al., 2012
Bacteroidetes	Energy production	Ottman et al., 2012
	Amino acid metabolism	
	Carbohydrate metabolism	
Proteobacteria	Establish an environment for strict anaerobe colonization	Shin et al., 2015
	Pro-inflammatory responses	Shahi et al., 2017

Some of the mechanisms for the functionalities assigned to these taxa are extremely well-characterized, while others are only beginning to be explored. For example, carbohydrate metabolism is a well-established and well-known function of the gut microbiota (Macfarlane & Englyst, 1986; Ottman et al., 2012). Humans ingest many carbohydrates that we do not have the enzymes to ferment properly. Luckily, the bacteria can perform the necessary fermentation processes on these undigestible carbohydrates to produce human-usable short-chain fatty acids (SCFA). Depending on

the diet, these SCFA can account for approximately 10% of the required human caloric intake (den Besten et al., 2013).

While not novel, the exploration of the interplay between the gut microbiota and the immune system is still relatively new compared to the established roles in metabolism (Binda et al., 2018; Hardy et al., 2013; Shahi et al., 2017; Vaiserman et al., 2020). Several mechanisms to describe the interaction between SCFA and the immune system exist. For example, intestinal butyrate can be transported by the monocarboxylate transporter 1 across the gut epithelia and undergo beta-oxidation, which is the mechanism of breaking down fatty acids into energy, an essential process for maintaining the health of the gut epithelial cells. Butyrate could also interact with G protein-coupled receptors and initiate regulatory pathways such as mTOR, which control processes such as autophagy and immune function (Noureldein & Eid, 2018; Parada Venegas et al., 2019).

1.2. Initial Colonization and Early Life Events Associated with Shifts in the Infant the Gut Microbiota.

The initial gut microbiome composition of infants can vary greatly compared to adults. There is some evidence that some microbial colonization can take place *in utero*. However, most initial microbial colonization occurs vertically during the birthing process (Nguyen et al., 2016). An infant born vaginally will have its primary gut colonizers be very similar to the mother's vaginal microbiota, whereas a caesarean-born infant will have an initial composition more similar to the mother's skin. For this reason, researchers must also consider maternal microbiotas when anticipating infant trajectories. The maternal prenatal microbiota (both gut and vaginal) can influence the health of the mother and can predispose them to delivery complications which can, in turn, affect the initial infant gut colonization (Hiltunen et al., 2021).

However, an infant’s microbial gut colonizers change dynamically over the first few months, eventually stabilizing when the child has conformed to a more adult diet (Walker, 2013). The major bacterial phyla in the infant gut microbiome are Firmicutes, Actinobacteria and Bacteroidetes. Actinobacteria and Firmicutes play a significant role in short-chain fatty acid production pre and post-weaning, respectively (Odamaki et al., 2016; Rinninella et al., 2019). We see shifts like these because certain species, such as

the Actinobacterial genus *Bifidobacteria*, have specific pathways that require human milk oligosaccharides, so the transition to solid food would induce a phylogenetic shift. While the composition might change, the functional redundancy found in a healthy gut microbiota population aims to ensure consistent functionality regardless of age (Tian et al., 2020).

Evidence shows that certain environmental factors significantly impact gut microbiome composition in early life. Recent interest is the impact of caesarian versus vaginal births, as this can dramatically affect the initial colonizers populating the infant's gut. A study by Schmidt et al. showed that babies born via caesarian section showed almost a complete lack of *Bacteroides* in the gut microbiome for nearly a month after birth (Schmidt et al., 2018). Other environmental factors influencing gut microbiome composition include early microbe exposure, formula use and antibiotic use in early life (Stout et al., 2017). These dynamic changes result from different environmental exposures at critical points of development. Figure 1.3 shows a subset of environmental factors influencing gut microbial composition during infancyFigure 1.3.

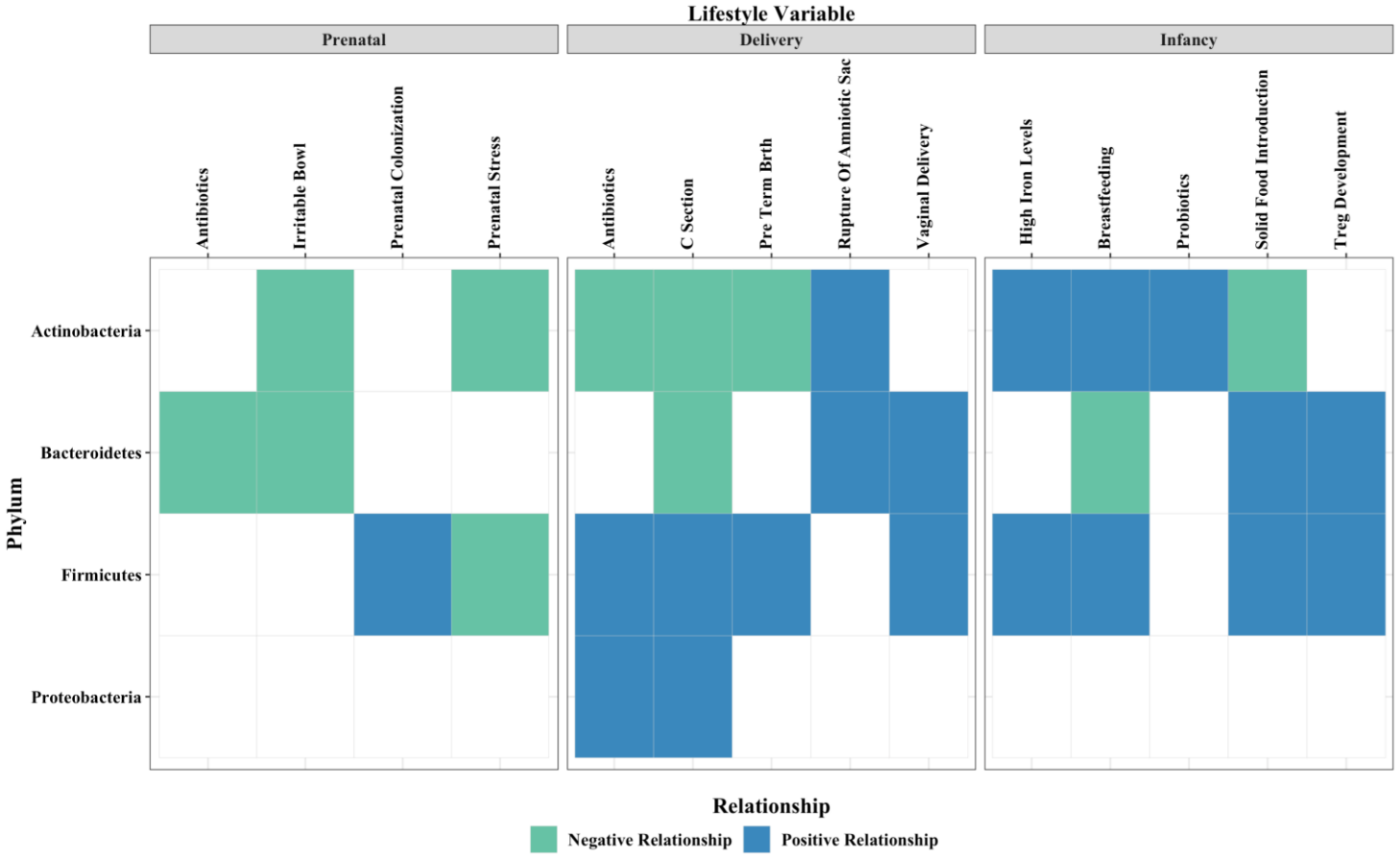

Figure 1.3. **Sample of known microbial and lifestyle interactions identified from the literature review. Known positive relationships are represented with a dark blue square, while negative relationships are identified with a green square. A positive relationship would indicate that the occurrence of the lifestyle variable increases the amount of microbe present in the gut microbiome. Associations are organized into three-time points, prenatal, at delivery and early infancy. Data summarized from Depner et al., 2020; Francino, 2014; Matamoros et al., 2013; Milani et al., 2017; Moore & Townsend, 2019; Nguyen et al., 2016; Tanaka & Nakayama, 2017; Torres et al., 2020**

1.3. Diseases Associated with The Gut Microbiota and Microbial Dysbiosis

Many medical conditions have been associated with microbial shifts. These conditions are not limited to inflammatory diseases but also metabolic and neuropsychiatric disorders. A subset of conditions with established microbial associations and their taxonomic associations can be seen in Table 1.2

Table 1.2. Documented associations of medical conditions and gut microbiota associations found in the literature. The table is adapted from Kho & Lal, 2018 and is a non-exhaustive list.

Disease	Association	Citation
Inflammatory Bowel Disease	Increase in Enterobacteriaceae, Bacteroides and Ruminococcus	Machiels et al., 2014; Png et al., 2010; Sokol et al., 2008; Willing et al., 2010
Atopic Disease	Increase in Clostridium difficile, increased C. difficile/Bifidobacteria ratio.	Kalliomäki et al., 2001
Obesity	Increased Firmicutes and Actinobacter	Koliada et al., 2017; Turnbaugh et al., 2009
Hypertension	Increase in Firmicutes-to-Bacteroidetes ratio	Yang et al., 2015
Depression	Increase in Eggerthella, Holdemania and others.	Kelly et al., 2016
Colorectal Cancer	Increase in Bacteroides decrease in butyrate-producing Faecalibacterium	Wang et al., 2012; Wu et al., 2018

In many cases, the biological hypothesis supporting these associations lies within the interaction of these microbial taxa to the gut immune environment. In the case of inflammatory bowel disease, the outlined microbial shifts would indicate a decrease in butyrate-producing bacteria. Colonocytes, the epithelial cells found in the colon, use butyrate as a fuel source. Without their fuel source, they cannot properly consume oxygen, and this can lead to an unfavourable environment for the commensal bacteria found in the gut (Litvak et al., 2018). Furthermore, proper gut barrier function and anti-inflammatory responses require butyrate (Donohoe et al., 2011). The biological hypothesis for colorectal cancer is also based on butyrate production from the microbiota. Without sufficient butyrate, there is less protection against oxidative stress and a diminished capacity to induce apoptosis in DNA-damaged cells (Wang et al.,

2012). When reviewing the literature, disease associations are established predominantly in adult cohorts. Research outlining potential interactions between gut microbial taxa and diseases in infants and children is limited. Furthermore, the potential longitudinal impact of early gut microbial dysbiosis remains unknown. To fill this knowledge gap, more longitudinal cohort studies are needed to collect microbiome and health data at various time points throughout a subject's life.

The manifestation of a disease is often a hallmark of microbial dysbiosis. In contrast to a healthy gut microbiome, microbial dysbiosis is a state that is often defined as an imbalance in the gut microbiome – in particular, an imbalance that inhibits an individual's ability to thrive in their environment, usually due to some disease (Messer & Chang, 2018). Notably, the identified shift differs from expected development, particularly in children where rapid changes are already expected. There are many ways researchers have tried to quantify and characterize microbial dysbiosis. One of those ways is using ratios of the different taxon concerning a specific disease state. Using ratios is a common practice when the sequencing data is already available due to its computational simplicity (Wei et al., 2021). Common measures of microbial dysbiosis include the Enterobacteriaceae-to-Bacteroidaceae (EB) ratio for a family level comparison or the Firmicutes-to-Bacteroidetes (FB) for a phylum level comparison (Ley et al., 2006). The FB ratio has been positively correlated with obesity, in both human adult cohorts and mouse models (Magne et al., 2020). The EB ratio has been positively correlated with atopic sensitization in infants (Azad et al., 2015). Identifying indicators of microbial dysbiosis is vital as they can be used to help predict and prevent adverse health outcomes. These ratios have been established in adult cohorts, with little data available on whether these trends would also be found in infant gut microbiomes, particularly for the FB ratio as there is current work in the CHILD Cohort Study that has looked at EB ratios (Azad et al., 2015, Mange et al., 2020). The validity of these ratios is still being established as there have been conflicting results when comparing these correlations in different age groups and geographic regions (Houtman et al., 2022).

1.4. The Gut Microbiota and Medication Use

Many established relationships between medication use and the gut microbiome have focused on adult cohorts. Most of the commonly used medications have been investigated in some capacity. For example, adult ibuprofen and celecoxib users have

been found to have increased Acidaminococcaceae and Enterobacteriaceae in their gut microbiota (Rogers & Aronoff, 2016). Several cohorts have also identified changes in gut microbial abundance associated with using proton pump inhibitors (Imhann et al., 2017). Gut microbiota changes related to medication use are not insignificant and may have long-lasting impacts on the health trajectories of adults, despite their more well-established gut microbiota populations.

Even in infants, there is some evidence that medication practices have long-lasting impacts. A study by Patrick et al., as part of the CHILD Cohort Study, is helping to address the need for more longitudinal studies to evaluate the impact of lifestyle on the development of the gut microbiota. This study found that increased antibiotic use in infancy was associated with the development of asthma, with the diversity of the gut microbiome being a potential modulator for this change (Patrick et al., 2020). An increase in non-antibiotic medications, such as vitamin D, has also been associated with an increase in gut Bifidobacterium and gut Lachnospiraceae in both infants and mothers (Kassem et al., 2020; Talsness et al., 2017).

1.5. Medication Use in Early Life

There are few regulations and recommendations in Canada surrounding specific medication use in infants under one year. Notable exceptions to this are the use of acetaminophen and vitamin D. For analgesics, it is recommended that acetaminophen be used as opposed to ibuprofen because acetaminophen use in infants has reported fewer adverse side effects. However, even though acetaminophen is the preferred option, there are still official guidelines for proper usage according to the infants' weight (Government of Canada, 2016). For Vitamin D, it is recommended that all infants get 400 IU per day of vitamin D to ensure proper bone development (Canadian Pediatrics Society, 2007). As an entire class, the use of antibiotics for all age groups has seen massive changes in regulations to mitigate risks of antimicrobial resistance (Lee et al., 2014; Marra et al., 2007).

However, antibiotics are still an incredibly prevalent class of medications in the infant population. A US study found amoxicillin to be the most prevalent prescription drug used in infants aged 0-1y (Olson & Mandl, 2012). However, there is evidence that the overall prevalence of amoxicillin and antibiotic use, in general, is decreasing.

Medications to treat respiratory issues, such as inhaled corticosteroids or albuterol, have also decreased (Hales et al., 2018). It is much easier to track the usage of prescription medications in these populations as there are records that exist. The same cannot be said for medications available over the counter (OTC). This is a problem that the CHILD Cohort Study is well placed to address, as they have collected information surrounding OTC and prescription health products. Some of the work being done to address this problem is outlined in Chapter 2.

1.6. The CHILD Cohort Study

The CHILD Cohort Study is a multidisciplinary, population-based longitudinal cohort study of healthy infants, with study centres in Vancouver, Edmonton, Winnipeg, and Toronto. The initial goal of the CHILD Cohort Study was to identify early life microbiome-based biomarkers and environmental exposures that predict asthma and inflammatory disease, but it has since expanded to study a more diverse array of infant and child health outcomes. Between 2009 and 2012, the CHILD Cohort Study recruited 3624 pregnant mothers in their second or third trimester. Ultimately, only 3542 infants were included in the study at one year because families became ineligible after recruitment. The subject would be ineligible for the CHILD Cohort Study if one or more of the following conditions were met: pregnancy via in vitro fertilization, miscarriage, premature birth (<35 weeks), multiple births, fetal death, or any other exception complications or abnormalities. A breakdown of CHILD Cohort Study demographics, which are approximately equal to the national averages, can be seen in Table 1.3 (Subbarao et al., 2015).

Table 1.3. Demographics of CHILD cohort participants. Demographics are reported as n (%) unless otherwise specified. The total sample size for the demographic collection was 3542 families. Families were recruited from 2008-2012, any families that became ineligible after recruitment are not reported in this summary.

Family Demographics		
Mean Gestational age in weeks at enrolment (SD)	26.7(6.3)	
Older Siblings in the home (%)	48.3	
Parental Demographics	**Maternal**	**Paternal**
Mean age in years (SD)	32.3(4.7)	33.8(5.6)
Reported ethnicity		
White Caucasian	2532(72.9)	2536 (73.7)
South East Asian	428 (12.3)	343 (10.0)
South Asian	91 (2.6)	106 (3.1)
First Nations	177 (5.1)	145 (4.2)
Black	77 (2.2)	110 (3.2)
Other	161 (4.6)	174 (5.1)
Unknown	6 (0.2)	29 (0.8)
Education		
High School or Less	295 (8.7)	502 (14.7)
College or university	2443 (72.2)	2151 (63.1)
Postgraduate education	647 (19.1)	756 (22.2)

Researchers have collected >47 million data points from 3263 children at nine time points from birth to 8 years (Subbarao et al., 2015). The data collected by the CHILD Cohort Study can be broadly classified into questionnaires, tests and samples. A breakdown of these, along with the time points at which they were collected, can be seen in Figure 1.4. The method of collection can differ depending on the information in question. Most questionnaires were self-completed, while a trained research assistant conducted the environmental assessments in person when the child was 3-4. Information about the delivery of the infant was collected via hospital records. For samples, in particular those relating to the microbiome, muconium and stool samples were collected to assess the gut microbiome of the infants, as stool collection is a common method for gut microbiome studies. While some participants have been lost over the years, the CHILD Cohort Study reports a retention rate of over 90% (Subbarao et al., 2015).

CHILD Study Schedule from Prenatal Recruitment to Age 5																
	Visit	18 Week	36 Week	Birth	3 Month home visit	3 Month clinic visit	6 Month	1 Year clinic visit	1.5 Years	1.5 Year clinic visit	2 Years	2.5 Years	3 Year clinic visit	4 Years	5 Year clinic visit
Questionnaires	Mother Profile/Residence	√			√			√			√		√	√	√
	Mother Health	√						√							√
	Mother Nutrition	√			√			√							√
	Mother Vitamins and Supplements	√			√			√							√
	Mother Medications	√		√	√		√	√							
	Mother Stress	√					√		√		√	√	√	√	√
	Mother Psychosocial (multiple Qs)		√					√					√	√	√
	Mother Life Stress Interview		√					√							
	Parenting Stress							√			√		√	√	√
	Father Health	√													√
	Socio-economic status (SES)	√						√					√	√	√
	Child Delivery Chart Extraction			√											
	Child Health				√		√	√	√	√	√	√	√	√	√
	Child Nutrition and Diet				√		√	√	√	√	√	√	√	√	√
	Child Medications			√	√		√	√	√		√	√	√	√	√
	Child Clinical Assessment							√					√		√
	Home Environment Questionnaire	√			√		√	√	√		√	√	√	√	√
	Home Assessment done by RA				√										
	Food packaging and prep							√					√	√	√
Tests	Mother Skin Prick Test							√							
	Mother Spirometry							√							
	Father Skin Prick Test	√													
	Father Spirometry	√													
	Child Skin Prick Test							√					√		√
	Child eNO			√		√		√		√					√
	Child Pulmonary Function Tests					√		√		√			√		√
Samples	Cord Blood			√											
	Mother Breast Milk				√										
	Mother Venous Blood	√						√							
	Mother/Father Buccal Swab*							√							
	Father Venous Blood	√													
	Child Venous Blood							√					√		√
	Child Buccal Swab *							√							
	Child Nasal Swab				√			√							
	Child Urine				√	√		√		√			√		√
	Child Meconium/Stool			√	√			√							
	Home Dust Collection				√										
TORONTO ONLY		VANCOUVER ONLY							WINNIPEG ONLY						
* Buccal swabs collected for child/parents if not able to collect venous blood															

Figure 1.4. Breakdown of CHILD study data collected from 18 Weeks Prenatal to age 5. Questionnaires are indicated along the rows in pink, tests in green and samples in orange. A checkmark in the box indicates that the specified type of information was collected at the timepoint displayed along the top. Any box coloured in green indicates the information was only collected in Vancouver. If a box is coloured in orange, the information was only collected in Toronto, and if the box is coloured in blue, then the information was only collected in Winnipeg.

To access data from the CHILD Cohort Study, users can register and explore data through CHILDdb. Through this portal, users can explore and select certain variables of interest and request that the complete data be made available for research purposes. Requests for data access are approved by the CHILD Cohort Study

administration after reviewing research, ethics and funding statements submitted via the portal. This database ensures that a diverse array of researchers can access and analyze this high-quality data (*CHILDdb*, n.d.).

1.7. Thesis Aims

My goal is to examine the impacts of medication practices on infant microbial dysbiosis and associated child outcomes. The CHILD Cohort Study has collected data on the medications used and self-reported reasons for use. I hypothesize that by integrating these data and the diverse data integrated into CHILDdb, I can identify medication use patterns associated with infant dysbiosis. However, much of these data, particularly the free-text descriptions of reasons for medication use, is unstandardized, making a systematic, accurate analysis difficult. Also, visualization of these complex data is non-trivial as traditional approaches do not scale to the diversity and number of variables we have. To address these issues and investigate my hypothesis, the following aims were completed and are described in this thesis.

The first aim described in Chapter 2 is to integrate and apply standardized vocabulary for medication use in CHILDdb. The second aim, described in Chapter 3, is to develop software for globe-based visualization analysis of correlation datasets. The outputs from Aims 1 and 2 are utilized in Aim 3 to correlate medication use profiles with patterns of microbiota development. Aim 3 is described in Chapters 4 and 5. Chapter 4 covers the creation of machine learning models for determining variables associated with microbial dysbiosis. Chapter 5 is a more in-depth investigation into variables of interest identified through the machine learning models described in Chapter 4. A discussion surrounding my conclusions and the future direction of this work is given in Chapter 6.

Chapter 2.

Applying Standardized Vocabulary for Medication Use Data in CHILDdb

This chapter aims to review the addition of standardized terms for reasons for medication use in the CHILD cohort and how this addition allows for more accurate and sophisticated analyses. Free text reasons for medication use were collected by other members of the CHILD Cohort Study. I lead the curation effort for the standardized terms using ontologies, the process of which is outlined in Section 2.3. Following the curation, I performed several analyses to show the utility of the standardized reasons for medication use data in the context of the CHILD Cohort Study. Section 2.4 outlines these analyses. The manuscript "From prescription drugs to natural health products: Medication use in Canadian infants", by Bédard et al., describing medication use in the first year of life has been submitted to Children. The manuscript "The use of prescription medications and non- prescription health products by breastfeeding mothers in a prospective cohort study", by Soliman et al., has been submitted to Breastfeeding Medicine. I am a co-author on both manuscripts.

I completed all work presented in this chapter with the following exceptions. The data collection was done from 2009-2017 by research assistants in the CHILD Cohort Study. I lead the manual curation effort. In addition to my manual curation, I coordinated additional curation by members of the Brinkman Lab, particularly Natalie Kumpula and Thorne Matthews.

2.1. Abstract

Medication habits in Canadian infants remain poorly characterized. The CHILD Cohort Study has collected data describing medication use and reasons for medication use for prescription, OTC and natural health products. However, the reasons for medication use data requires standardization before it can be appropriately analyzed. I hypothesized that by developing and applying an approach for assigning ontologies to reasons for medication use, I could perform analyses not previously possible that could provide new insights into medication use. This standardization can be completed using

curated ontologies. I led the ontology curation effort of 64817 entries of reasons for medication use. Employing automated and manual approaches, 99.4% of terms in the dataset were standardized. Using these standardized terms, I found evidence of off-label medication use in mothers using domperidone. I also identified teething and general discomfort as the top reasons for OTC medication use in infants. This landmark dataset will be available to a larger pool of researchers via the CHILD Cohort Study database. With this data, researchers can conduct streamlined analyses to improve the understanding of infant medication use in Canada. I also further apply this dataset in Chapters 4 and 5.

2.2. Introduction

There is limited existing literature describing the use of medications in children, particularly children under three years of age. Furthermore, the investigation into children's medication use tends to focus on prescription medication instead of OTC medication (Servais et al., 2021). This is likely due to the inherent trait that OTC medications are available without prescription and can be obtained by parents at any time, making it challenging to track consumption. Collecting information from parents about how often and why they choose to give medications to their children is essential for a better understanding of medication practices in Canada. Collecting data on reasons for use would help identify differences in health knowledge, literacy and perceived risk differences in groups of parents and how that could impact their children (Anderson et al., 2013).

The CHILD Cohort Study has collected information describing reasons for medication use in OTC, prescription medications, and natural health products. However, the information was collected as free-text responses where parents could write in whatever answer they felt was appropriate. Free text responses have several issues that make subsequent analyses difficult. Free text responses can be full of typos and synonyms, making summarization difficult. Depending on the context, respondents with domain knowledge might respond to a question using specific jargon instead of lay terms.

A solution is to map the free-text terms to an ontology. An ontology covers a specific domain of information and contains entities with unique identifiers and

descriptions. Furthermore, an ontology contains formal definitions that can be read by specialized computer programs to facilitate analyses (Hoehndorf et al., 2015). Ontologies exist in an encoded, hierarchical structure. An example of this structure can be seen in Figure 2.1 (Schriml et al., 2019). This hierarchical structure can help downstream researchers do searches and summaries at different levels of specificity depending on their research question. Integrating ontologies on a large scale can also facilitate large, cross-cohort analyses that would not be possible otherwise (Kourou et al., 2019).

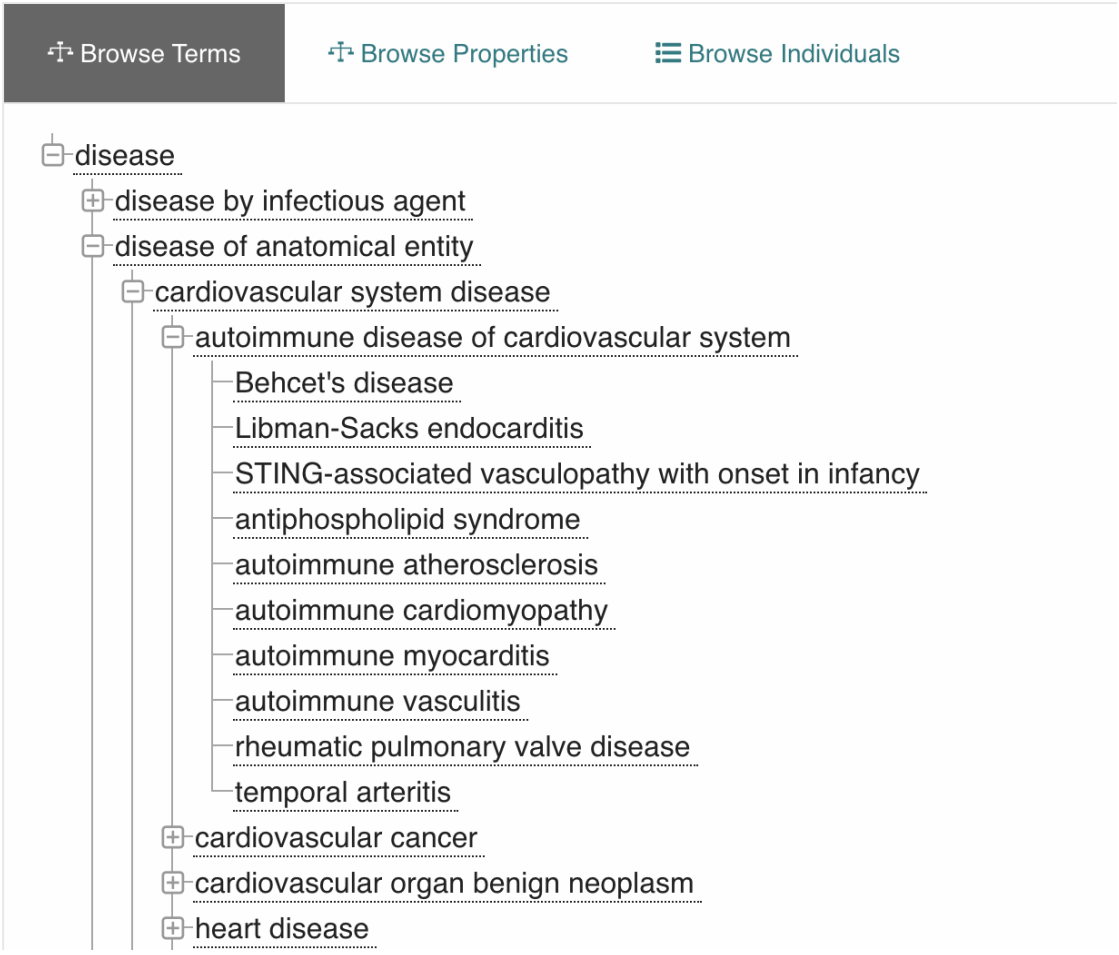

Figure 2.1. Example of Hierarchical structure of Ontology. Example is from the Human Disease Ontology (DOID) when looking at the structure on the Ontology Lookup Service repository maintained by the European Bioinformatics Institute. In this image, the progression from disease to disease of anatomical entity to cardiovascular system to autoimmune disease of cardiovascular system is shown. The Schmril Lab maintains DOID at the University of Maryland.

2.3. Methods

2.3.1. Data collection

CHILD Cohort participants were asked to indicate medications taken at the following time points: 18 weeks prenatal, Birth, 3 Months, 6 Months, 1 Year, 18 Months, 2 Years, 2.5 Years, 3 Years, 4 Years, and 5 Years. Both the mother and infants' information were collected from the 18 weeks prenatal to 1 year. After 1 year, information was only collected for the child. The questions did not ask for total medication at each time point, but the medications taken since the last survey. For example, the 6-month questionnaire would ask for all medications taken in between the three-month and 6-month questionnaire. This would reduce the chance of duplicate entries for the same medication for the same subject. For the birth and prenatal questionnaires, data was also collected for the mother's medication use. Maternal and infant entries are marked with the same subject ID. However, a field is available to filter by mother or infant for downstream analyses. After supplying the name of the medication taken, participants were also given the option to explain the reason for medication use. This answer was given as a free text response, meaning that the participant could write the answer they felt was appropriate. There were no drop-down menus or checkboxes with pre-set options.

2.3.2. Selection of Ontologies

A total of 64 817 total entries were supplied for the reasons for medication use questions taken from the previously specified timepoints. When summarizing for unique answers, this left 4424 distinct free-text responses. A major problem in this set is the number of misspellings. For example, 21 of the 4424 unique terms are different spellings of the word "hypothyroidism". Another issue is the level of specificity used by the respondents. For example, some participants would supply the answer "eczema," whereas others would supply the answer "eczema on face". In the dataset, there are 23 different references to eczema (with various spellings) at different levels of specificity.

There are many different ontologies covering different scopes of information. When looking to standardize data, one must decide on the most suitable. It is best to use a limited number of ontologies to take advantage of the hierarchical structure encoded in

the ontology. If one were to select only specific terms from multiple different ontologies, this might lose the relational nature of the terms if the different ontologies have not been mapped to each other by the developers themselves. In this context, mapping refers to association of terms in one ontology as equivalent to another. The ontologies chosen for this project were the Human Disease Ontology (DOID), Symptom Ontology (SYMP), Human Phenotype Ontology (HPO), Adverse Event Ontology(OAE) and National Cancer Institute thesaurus ontology (NCIT) (He et al., 2014; Köhler et al., 2017; Schriml et al., 2019; Sioutos et al., 2007). Several ontologies had to be used in this case to account for the wide variety of medication uses. Human Disease ontology would be used when a subject medicated for a specific disease, such as the common cold. This contrasts with medicating for something like swelling, which would find its term in the Symptom Ontology. Various allergies such as pet dander or shellfish would be found in the Human Phenotype ontology. Events such as bone breaks or muscle sprains would fall into the Adverse Event Ontology. Very general reasons for use (i.e. just indicating "bladder" or "head") would be covered by the National Cancer Institute Thesaurus.

2.3.3. Pre-Processing and Curation

Before starting the process of curating ontologies, all 4424 unique terms were put through a basic spell check and clustering algorithm using OpenRefine. This would remove small spelling errors ("hypothyroidims" would be corrected to "hypothyroidism") and cluster similar phrases together ("neck ache" and "neck aches" would become one term, "neck ache"). An example of the whole process can be seen in Table 2.2. This process resulted in 3551 unique terms. These terms would go through the ontology curation process and then be mapped back to their original free text entries.

Automatic curation of the ontologies was performed using the Python Module, Ontoma, which contains a wrapper for the Ontology Lookup Service tool, Zooma (*ZOOMA*, 2013/2022). Each corrected term would be searched against all 5 ontologies outlined in Section 2.3.1. As the output, I received the suggested term from each ontology, its unique ID and definition. In addition, the terms from each ontology were ranked from 1 to 5, with 1 being the most likely/best match according to the search algorithm and five being the least likely. Since I asked for multiple output options for each input term, this resulted in approximately 18,300 terms that required manual validation. To reduce the load on the curators, if any of the terms were an exact string

match, it was recorded as a match and would not require further validation. For example, both the input term and the curated terms were “asthma”. After this, automatic curation was completed. Curators were asked to assign one of 4 classifications to each of the terms. Classifications and examples can be seen in Table 2.1. After the first pass with this classification system, the terms were reshuffled and given to new curators, and the procedure was repeated. This was done to ensure that multiple people assessed the curated terms for validity and reduce the chances of introducing human error.

Table 2.1 Explanation of classification system used for evaluation of automatic ontology curation done by the Python Module Ontoma. Classifications were manually assigned to terms by curators. The specified classification is shown, along with a definition and example where the classification would be appropriate.

Rank	Definition	Example	
		Free Text	Ontology
Match	the assigned term was an appropriate term for the free text entry	Yeast infection	Candidiasis
No match	the assigned term was a not an appropriate term for the free text entry	Antibiotics in pregnancy	Oligohydramnios
Child	free-text term was more specific than the assigned ontology	Shellfish allergy	Allergy
Parent	free-text term was less specific than the ontology term	Pain	Leg pain

Any terms classified as a match after the second round of classification were removed from the curation process, as they were declared suitable standardized terms. Within an ontology structure, there are "parent" and "child" terms that are either above or below a term, respectively. For example, in Figure 2.1 “autoimmune cardiomyopathy” would be considered a child term of the parent term “autoimmune diseases of the cardiovascular system”. For any terms classified as "parent" or "child" by the curators, another automatic curation was run with Ontoma. This time telling the program to either be more specific with any terms classified as "child" or telling the program to be more general with any terms classified as "parent". If no suitable term was found, those terms would be manually curated using the Ontology Lookup Service and all the terms labelled

"no match". The breakdown of matches, no matches, parent and child terms can be seen in Figure 2.2.

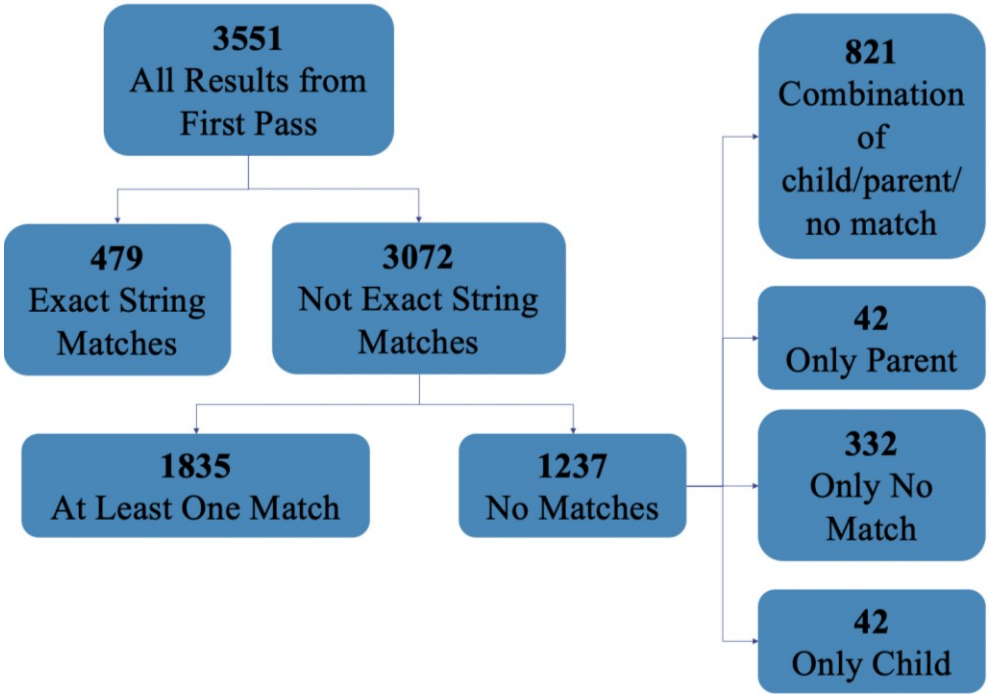

Figure 2.2. Breakdown of results from automatic curation after the first round of manual validation. As multiple options were provided for each input term, some terms had a combination results. Following this procedure, the 1237 terms with no match would go on to manual curation.

There were cases where multiple suitable terms were found for the free text entries. For example, the term "nausea" is present in the NCIT, SYMP and HP ontology. In these cases, one term would be chosen, and the term and unique ID added to the dataset. However, all appropriate records were stored in another field for mapping applications later. After a check to ensure as many terms as possible had an appropriate term, and all were in the format of "term; ID," the new standardized terms were added to the dataset. After curation, the 4424 unique, uncleaned terms were reduced to 2412 standardized terms.

Table 2.2. Example of Curation Process on Synthetic Dataset. Starting on the left, the original entry can be seen in the first column, then the data split by delimiters can be seen in the second column. The cluster results from OpenRefine can be seen in the third column. The fourth column contains the result assigned from Ontoma. The fifth column, or right-hand column, contains the final standardized term. The final column indicates whether this term resulted from automatic (headache, infection, pneumonia) or manual curation (teething, rash, candidiasis).

Free Text Entry	Reason for Use Split	Reason for Use Cluster	Ontology Term After Automatic Search	Ontology Term after Validation	Curation Type
Teething and headache	Teething	teething	No Match	Teething Syndrome; NCIT_C35063	Manual
	headache	headache	Headache;HP_0002315	Headache; HP_0002315	Automatic
Infection	infection	infection	Infection;NCIT_C128320	Infection; NCIT_C128320	Automatic
Rash, yeast infection	Rash	rash	Exanthem;DOID_0050486	Rash; SYMP_0000487	Manual
	Yeast infection	Yeast infection	Infection;NCIT_C128320	candidiasis; DOID_1508	Manual
Pnumonia	Pnumonia	pneumonia	Pneumonia;DOID_552	Pneumonia; DOID_552	Automatic
Head ache	Head ache	headache	Headache;HP_0002315	Headache; HP_0002315	Automatic

2.4. Results and Discussion

2.4.1. Comparison of Top Reasons for Use with and Without Ontology

The primary goal of applying ontologies to the reasons for medication use dataset was to create a clean dataset that would help facilitate more streamlined analyses. Out of the 64 817 original entries of reasons for medication use, 99.4% of the terms were successfully standardized. A basic comparison of the top 25 reasons for medication use with and without ontology can be seen in Table 2.3. While many indications remain consistent between the two groups, there were changes in the sum totals of each reported indication, which changed their respective rankings for most common reasons for medication use when comparing before versus after standardization.

It is not surprising that fever, the common cold, and teething are the most common reasons for medication use in this dataset. All indications are common and, in the case of teething, a natural part of development. As noted, the components of these top 25 lists remain fairly consistent. However, recorded occurrences have increased for the same or synonymous terms. This can be seen to a smaller degree when comparing mother's contraception use with (905 occurrences) and without (884 occurrences). To a larger degree, this is seen when comparing yeast infection (both infant and mother's) without ontology (513 occurrences) to the term candidiasis (1205 occurrences) with ontology. These standardized terms will make it easier for downstream researchers to investigate and determine more accurate sample sizes for future projects.

Table 2.3. Comparison of top 25 reasons for medication use for subjects in the CHILD Cohort Study from 18 Weeks prenatal to 5 years with and without the addition of standardized ontology terms. In the first column, the rank in the dataset can be seen, in the second and third column, the reason for use and the number of occurrences when using ontology is shown. The fourth and fifth columns show the reason for use and the number of occurrences without using ontology. Summaries are for 64,817 total entries for 3243 subjects.

With Ontology			Without Ontology		
Rank	**Reason**	**Occurrences**	**Rank**	**Reason**	**Occurrences**
1	Fever	9621	1	fever	9491
2	common cold	5572	2	cold symptoms	5448
3	Teething Syndrome	4811	3	discomfort from teething	4756
4	Discomfort	3569	4	ear infection	2874
5	Ear Infection	2952	5	discomfort	2282
6	Headache	1527	6	eczema	1290
7	Pain	1390	7	vitamin supplement	1282
8	Asthma	1377	8	headaches	1280
9	Eczema	1346	9	asthma	1230
10	Dietary Supplement	1286	10	pain	1141
11	Candidiasis	1205	11	discomfort from vaccinations	950
12	Allergic Reaction	1117	12	allergy symptoms	898
13	Skin Rash	1091	13	contraception	884
14	Contraception	905	14	diaper rash	858
15	Diaper Dermatitis	882	15	heartburn	792
16	Heartburn	827	16	nausea	764
17	Nausea	771	17	hypothyroidism	641
18	Hypothyroidism	711	18	skin rash	641
19	Infection	687	19	depression	574
20	Cough	595	20	nausea and vomiting	538
21	Depression	591	21	acid reflux	520
22	Nausea and Vomiting	554	22	yeast infection	513
23	Have Acid Reflux	525	23	infection	495
24	Wheezing	516	24	trouble breathing	475
25	Urinary Tract Infection	491	25	flu symptoms	448

Finally, there are some free text medication entries that did not have any suitable standardized terms. This occurred when there was insufficient information available in

the free text entry to be confident in an appropriate standard term. For example, the reason "48 hours only to rule out sepsis" does not indicate whether the subject had anything wrong with them. The terms "Preventative Intervention" and "Sepsis" may apply here, but in this process, borderline cases were left blank to avoid assuming information not in evidence. However, there were only 0.6% of the unique free text entries that were without a standardized term at the end of this process, so the missing data should be noted but will not have major impacts on the broader applications of this data, including the analyses to be outlined in Chapters 4 and 5.

2.4.2. Medication use in Breastfeeding Mothers

The breastfeeding habits of Canadian mothers are being studied by Dr. Lauren Kelly and her research group as part of the CHILD Cohort Study. When analyzing the medication usage in breastfeeding and non-breastfeeding mothers in the first year of life, I was able to provide reasons for medication use for each group for the top 10 medications taken by breastfeeding mothers at three months, six months and 12 months. The breakdown for the top medications at each time point can be seen in Figure 2.3, Figure 2.4 and, Figure 2.5 and the reasons for medication use can be seen in Table 2.4, Table 2.5 and, Table 2.6.

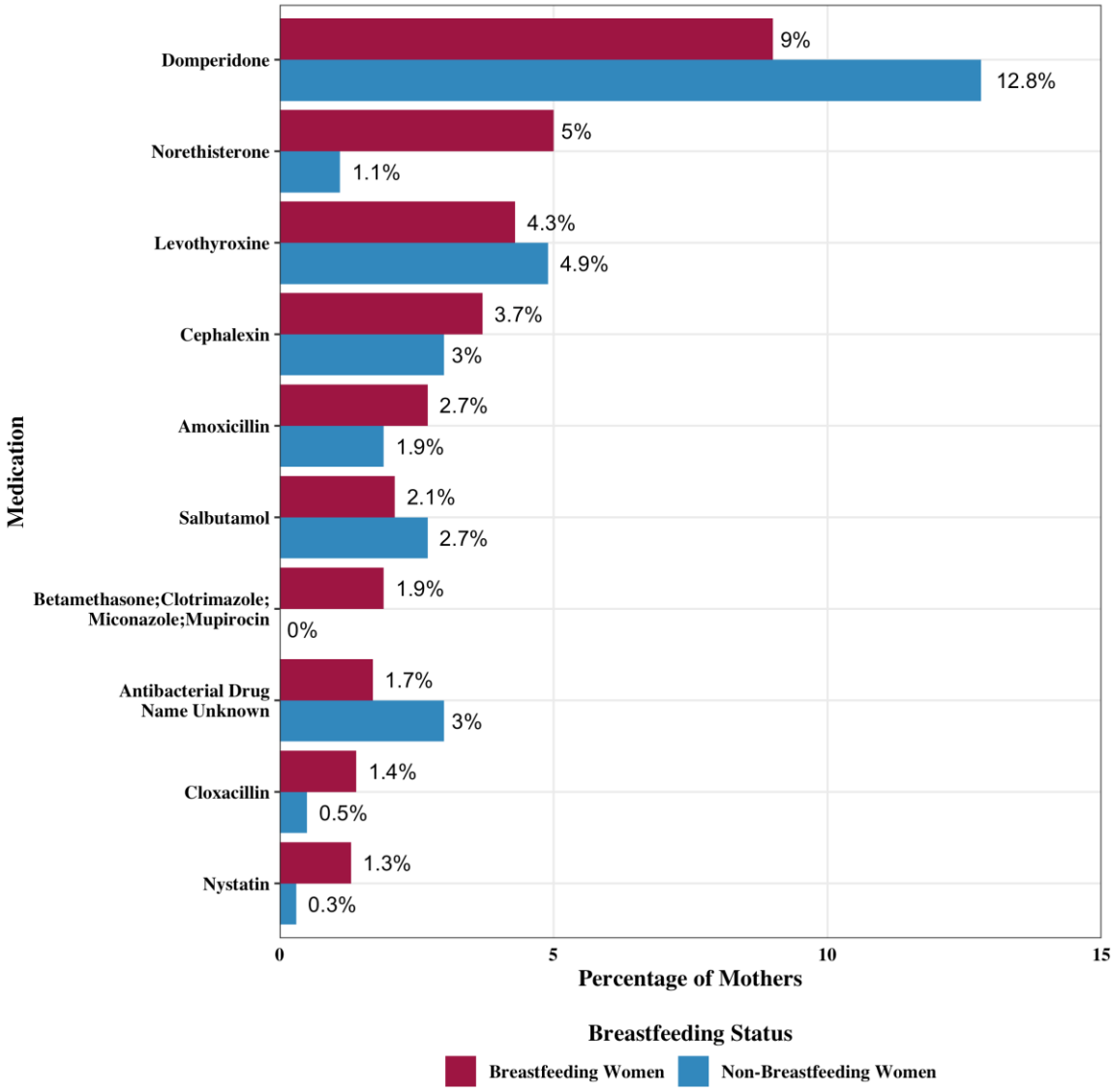

Figure 2.3. **The top 10 most frequently used prescription medications by mothers while breastfeeding at three months (n=2540) compared to non-breastfeeding mothers at three months (n=366). The percentage of mothers that reported taking medication is shown on the x-axis, and the medication is shown on the y. Breastfeeding Women are shown in red, while non-breastfeeding women are shown in blue. The percentage for each medication is shown to the right of the bars. Summaries completed by Soliman et al. (Submitted to *Breastfeeding Medicine*).**

Table 2.4. Ten most common prescription medications used by mothers while breastfeeding at three months (n=2540) compared to non-breastfeeding mothers (n=366) with the most common reasons for medication use in the CHILD Cohort Study. The status of breastfeeding vs non-breastfeeding was determined as outlined in Soliman et al. (Submitted to *Breastfeeding Medicine*).

	Breastfeeding Mothers n=2540		**Non-Breastfeeding Mothers n=366**	
Prescription Medication	**Reported Reason for Use***	**Mothers that used medication for specified reason.**	**Reported Reason for Use***	**Mothers that used medication for specified reason.**
		%		%
domperidone	Lactation Disorder	97.25	Lactation Disorder	100
	Gastroesophageal Reflux	1.38		
norethisterone	Contraception	99.22	Contraception	100
levothyroxine	Hypothyroidism	97.22	Hypothyroidism	88.89
cephalexin	Reason Not Given	71.91	Reason Not Given	100
	Mastitis	14.61		
	Infection	4.49		
amoxicillin	Infection	16.92	Reason Not Given	100
	Urinary Tract Infection	10.77		
	Bladder Infection	10.77		
	Mastitis	9.23		
salbutamol	Asthma	79.25	Asthma	70
	Chest Infection	3.77	Bronchitis	20
betamethasone;clotrimazole; miconazole;mupirocin	Reason Not Given	95.92	Reason Not Given	100
antibacterial drug name unknown	Reason Not Given	100		
cloxacillin	Mastitis	52.94	Reason Not Given	100
	Infection	23.53		
	Infection at Incision Site	5.88		
nystatin	Candidiasis Prevention	42.42	Reason Not Given	100
	Candidiasis Infection	18.18		
	Fungal Infection	15.15		
	Nipple Abnormality	9.09		

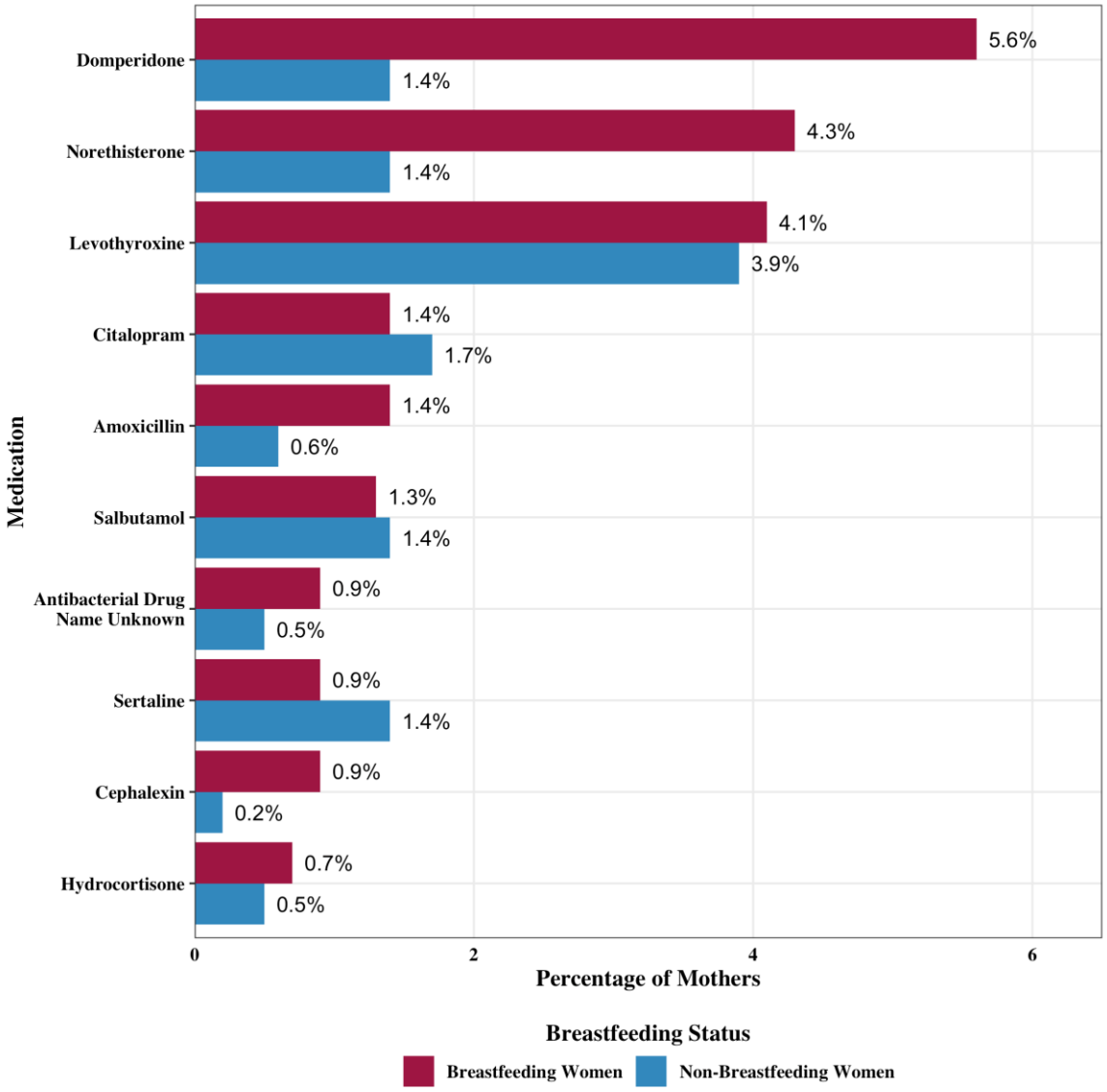

Figure 2.4. **The top 10 most frequently used prescription medications by mothers while breastfeeding at six months (n=1948) compared to non-breastfeeding mothers at six months (n=639). The percentage of mothers that reported taking medication is shown on the x-axis, and the medication is shown on the y. Breastfeeding Women are shown in red, while non-breastfeeding women are shown in blue. The percentage for each medication is shown to the right of the bars. Summaries completed as outlined in Soliman et al. (Submitted to *Breastfeeding Medicine*).**

Table 2.5. Ten most common prescription medications used by mothers while breastfeeding at six months (n=1948) compared to non-breastfeeding mothers (n= 639) with the most common reasons for medication use in the CHILD Cohort Study. The status of breastfeeding vs non-breastfeeding was determined as outlined in Soliman et al. (Submitted to *Breastfeeding Medicine*).

	Breastfeeding Mothers n=1948		Non-Breastfeeding Mothers n=639	
Prescription Medication	**Reported Reason for Use***	**Mothers that used medication for specified reason**	**Reported Reason for Use***	**Mothers that used medication for specified reason**
		%		**%**
Domperidone	Lactation Disorder	98.15	Lactation Disorder	88.89
Norethisterone	Contraception	100	Contraception	100
Levothyroxine	Hypothyroidism	97.44	Hypothyroidism	88
Amoxicillin	Sinusitis	21.43	Reason Not Given	100
	Urinary Tract Infection	10.71		
	Pharyngitis	10.71		
	Respiratory Tract Infection	7.14		
	Ear Infection	7.14		
Citalopram	Depression	77.78	Depression	81.82
	Anxiety	37.04	Anxiety	36.36
Salbutamol	Asthma	80.77	Asthma	88.89
Cephalexin	Reason Not Given	88.89	Reason Not Given	100
Antibacterial Drug Name Unknown	Reason Not Given	100	Reason Not Given	100
Sertraline	Depression	58.82	Depression	88.89
	Anxiety	41.18	Anxiety	33.33
	Postpartum Depression	11.76		
Hydrocortisone	Eczema	38.46	Reason Not Given	100
	Dermatitis	15.38		

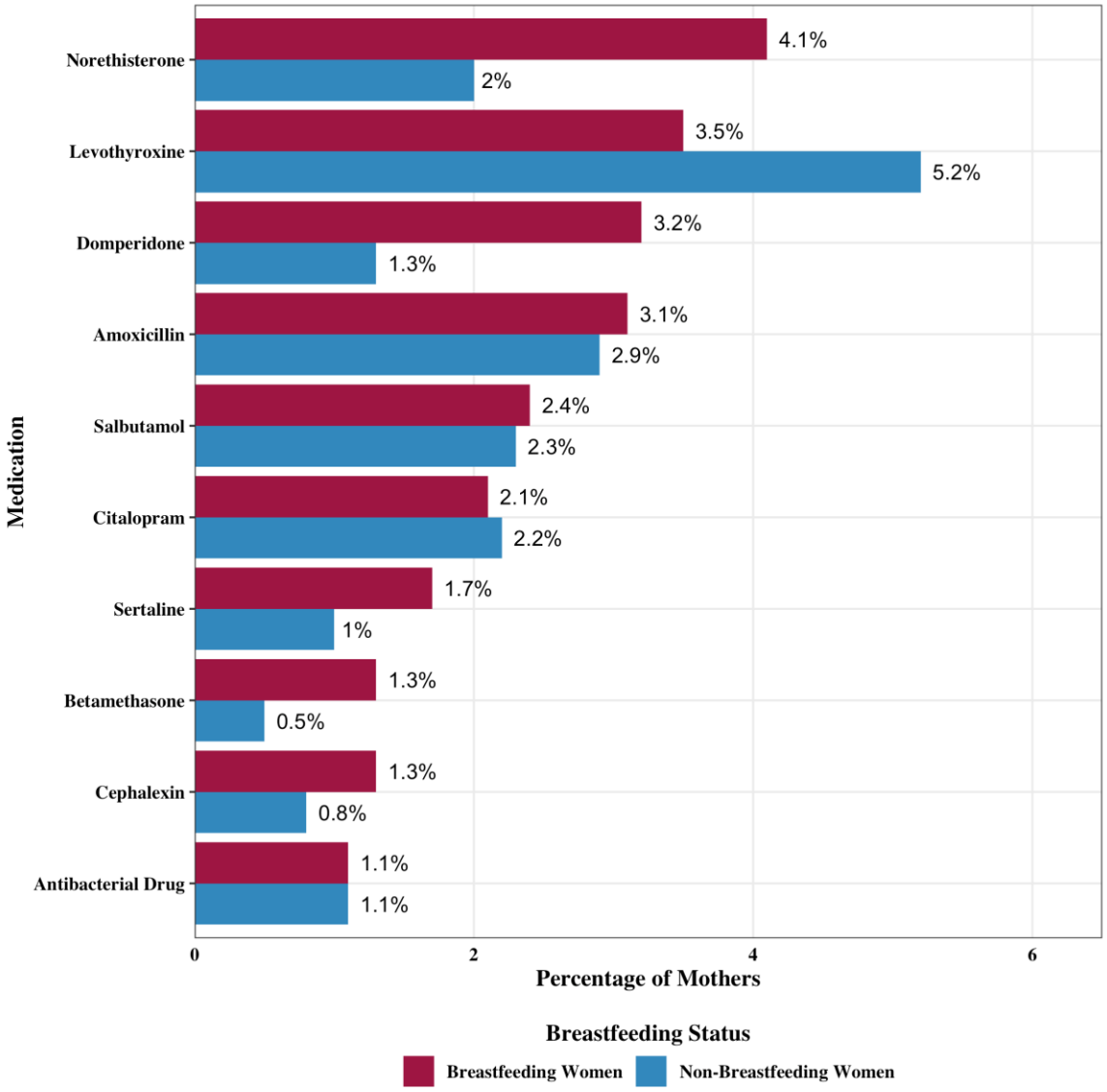

Figure 2.5. **The top 10 most frequently used prescription medications by mothers while breastfeeding at 12 months (n=1180) compared to non-breastfeeding mothers at 12 months (n=1413). The percentage of mothers that reported taking medication is shown on the x-axis, and the medication is shown on the y. Breastfeeding Women are shown in red, while non-breastfeeding women are shown in blue. The percentage for each medication is shown to the right of the bars. Summaries completed as outlined in Soliman et al. (Submitted to *Breastfeeding Medicine*).**

Table 2.6. Ten most common prescription medications used by mothers while breastfeeding at 12 months (n=1180) compared to non-breastfeeding mothers (n=1413) with the most common reasons for medication use in the CHILD Cohort Study. The status of breastfeeding vs non-breastfeeding was determined as outlined in Soliman et al. (Submitted to *Breastfeeding Medicine*).

	Breastfeeding Mothers n=1180		**Non-Breastfeeding Mothers n=1413**	
Prescription Medication	**Reported Reason for Use***	**Mothers that used medication for specified reason**	**Reported Reason for Use***	**Mothers that used medication for specified reason**
		%		%
Norethisterone	Contraception	100	Contraception	100
Levothyroxine	Hypothyroidism	97.37	Hypothyroidism	94.29
Domperidone	Lactation Disorder	86.84	Lactation Disorder	83.33
	Nausea	5.26		
Amoxicillin	Infection	18.92	Infection	18.92
	Ear Infection	16.22	Sinusitis	16.22
	Sinusitis	10.81	Pharyngitis	13.51
	Respiratory Tract Infection	5.41	Ear Infection	8.11
Salbutamol	Asthma	89.29	Asthma	90.63
	Common Cold	10.71		
Citalopram	Depression	69.57	Depression	90
	Anxiety	43.48	Anxiety	26.67
Sertraline	Depression	58.82	Depression	76.92
	Anxiety	23.53	Anxiety	15.38
Betamethasone	Eczema	64.29	Skin Rash	28.57
	Atopic Dermatitis	14.29	Reason Not Given	28.57
	Reason Not Given	14.29		
Antibacterial Drug	Pharyngitis	15.38	Urinary Tract Infection	30.77
Cephalexin	Reason Not Given	100	Reason Not Given	100

The integration of the standardized terms for medication use seen in Table 2.4, Table 2.5 and Table 2.6 facilitates some novel conclusions, specifically around using domperidone. Domperidone accounts for a large proportion of the medication use in

both breastfeeding and non-breastfeeding mothers. Domperidone is usually prescribed as an antiemetic used to treat gastrointestinal upsets. However, it can also be used off-label to help with lactation disorders. Table 2.7 shows how, by using the new curated reasons for medication use, provides evidence that mothers in the CHILD Cohort Study are using the drug outside the approved recommendations to treat lactation disorders.

Comparing both groups, we see that both breastfeeding and non-breastfeeding mothers use domperidone to combat lactation disorders. This suggests that the mothers classified as non-breastfeeding at a given time point are using medication to improve their chances of breastfeeding success. This provides novel insight into the motivations and decisions that go into the choice to breastfeed. Seeing a medication to improve lactation being frequently used in a non-breastfeeding group would indicate that the group is motivated to breastfeed but is inhibited by ability rather than making the active choice to not breastfeed their infant.

Table 2.7. Medication Reasons For Use for Breastfeeding (N=218) and Non-Breastfeeding (n=44) Mothers that Report using domperidone. Columns on the left are the summary results when using the standardized terms, columns on the left are the summary results when using the uncorrected free-text data. The status of breastfeeding vs non-breastfeeding was determined by Soliman et al. (Submitted to *Breastfeeding Medicine*).

With Ontology				Without Ontology			
Breastfeeding Mothers		Non-Breastfeeding Mothers		Breastfeeding Mothers		Non-Breastfeeding Mothers	
Reported Reason for Use	Mothers that used medication for specified reason	Reported Reason for Use	Mothers that used medication for specified reason	Reported reason for use	Mothers that used medication for specified reason	Reported Reason for Use	Mothers that used medication for specified reason
	%		%		%		%
Lactation Disorder	97.25	Lactation Disorder	100	increase breast milk	73.39	increase breast milk	70.45
				increase breastmilk	16.97		
				increase milk production	2.29		
				increase breast milk supply	0.92		
				increase breat milk	0.46	increase breastmilk	22.73
				increse breastmilk	0.46		
				increase breastfiding	0.46		
				milk production	0.46		
Gastroesophogeal Reflux	1.38			increase brest milk	0.46	increased breast milk	4.55
				increased breastmilk	0.46		
				lactation aid	0.46		
				increase breast milk production	0.46		
				acid reflux	1.38	increase breask milk	2.27

2.4.3. Medication Use in First Year of Life

Medication use in the first year of life was the major focus of a CHILD Cohort Study project led by Pascale Bédard. When looking at the top medications taken by the infant in the first year of life, I was able to supply the reason for medication use for these top medications. An excerpt of these results can be seen in Table 2.8. The table shows that the top medications are vitamin D, acetaminophen, ibuprofen, hydrocortisone, and amoxicillin. While no novel associations between medication use and reason for use were noted here, by integrating the reasons for medication use for OTC medications, we now have a dataset that can help establish trends in health literacy by comparing parents' justifications to recommended guidelines. The addition of the standardized reason for medication use data allowed for a comparison of medication practices for Canadian infants with infant cohorts from other geographic areas. For example, this study found that teething symptoms were more often treated with analgesic use than natural health products, which aligns with established practices in North America (Thompson & Huntington, 2019).

Table 2.8. Prevalence of use of most commonly used pharmaceutical products by age. Products are grouped into their respective 1st level ATC code: Alimentary tract and metabolism (A), Dermatologicals (D), Anti-infectives for systemic use (J), Nervous system (N) and Respiratory system (R), or into natural health products (NHP) or over-the-counter (OTC) drugs. Medication summaries determined by Bedard et al. Reported reasons for use were collected from standardized reasons for medication use data.

Product	Type	Prevalence of use (%)				Most common reported reasons for use
		0-3m	4-6m	7-12m	0-12m	
		(n=2930)	(n=2003)	(n=2218)	(n=3050)	
Vitamin D	ATC	66.2	56.9	50.1	78.7	Supplement
Acetaminophen	OTC	33.9	49.7	58.7	67.3	Discomfort, teething
Ibuprofen	OTC	2.6	8	23.5	20.4	Discomfort, teething
Topical	OTC	5.6	5.2	6.9	11.3	Eczema, rash
Amoxicillin	ATC	1.2	2.2	11	9.7	Ear infection
Nystatin	ATC	8.2	1.8	2.2	9.7	Candidiasis
Simethicone	OTC	7.6	2.7	0.6	8.4	Gas
Gripe water	NHP	6.3	2.7	0.9	7.3	Gas
Topical zinc oxide	NHP	4.6	1.6	2.2	6.3	Rash, diaper
Topical clotrimazole	OTC	3.2	0.6	1.9	4.6	Rash, diaper
Homeopathic	NHP	0.7	2.9	3.5	4.5	Discomfort, teething
Ranitidine	OTC	3.7	2.1	0.9	4.2	Gastroesophageal
Salbutamol	ATC	0.9	1.3	3.2	3.5	Wheezing
Sodium chloride	OTC	1.9	1	1.4	3.2	Nasal congestion
Homeopathic cold	NHP	1	1.1	1.9	2.7	Common cold
Diphenhydramine	OTC	0.1	0.6	2.6	2.3	Allergy
Topical benzocaine	OTC	0.3	1.8	1.3	2.1	Discomfort, teething
Topical fusidic acid	ATC	1	0.5	0.9	2	Rash
Topical erythromycin	ATC	1.1	0.3	0.5	1.6	Eye infection
Lansoprazole	ATC	1	0.9	0.7	1.5	Gastroesophageal
Dexamethasone	ATC	0.2	0.2	1.7	1.5	Croup
Topical polymyxin B	OTC	0.8	0.6	0.6	1.5	Eye infection
Clarithromycin	ATC	0.1	0.2	1.7	1.4	Bronchitis
Homeopathic	NHP	0.2	1	0.9	1.4	Teething
Probiotics	NHP	1	0.6	0.3	1.4	Intestinal flora
Fluticasone	ATC	0.2	0.5	1.4	1.3	Wheezing
Gentian violet	OTC	1.4	0.2	0	1.3	Candidiasis
Moisturizing creams	OTC	0.5	0.8	0.4	1.2	Dry skin
Cephalexin	ATC	0.3	0.3	0.8	1.1	Urinary tract infection,
Azithromycin	ATC	0.1	0.2	1.2	1	Ear infection, chest
Topical	ATC	0.3	0.3	0.7	1	Eczema, rash
Homeopathic colic	NHP	0.9	0.3	0.1	1	Gas, colic

2.4.4. Reasons for Hospital Visits

Another field of information deemed suitable for standardization was the free text field "reason for hospital/doctor visit". This free text field would have the same issues outlined previously with the reasons for medication use information. In short, there were misspellings, synonym use and differences in the level of detail that would make an analysis of the raw information difficult. The reasons for medication use data were very useful in expediting the process of standardizing these data.

While very similar, medication use data and reason for hospital visit data have some distinct differences in scope. Hospital visit data tend to be more focused on adverse events and well-defined illnesses, whereas reasons for medication use are more symptom-based and preventative. Therefore, there was some overlap between the datasets, and I quickly standardized 533 out of 3161 terms from existing standardizations in the reason for medication use dataset. Going through the free text entries for reasons for hospital visits and applying standardizations provided some novel insights. For example, without standardization, researchers would be unable to capture all the incidents of foreign bodies in the noses of infants and toddlers because each one is referred to very differently in the free text data. This also provides support for keeping both the free text and standardized entry. The standardized term helps summarize and search, but the free text entry is useful for gaining more insight into the events that may be too specific to have an established ontology term in the database. To demonstrate this, a table describing all incidents of objects stuck in the noses of children from birth to 5 years can be seen in Table 2.9.

Table 2.9. Example application of the standardization of the reason for hospital visit dataset. Summary of objects found in noses of subjects when looking for the term "Foreign Body" with the anatomical entity of the nose in subjects from birth to 5 years.

Object	Free text	Reason For Visit	Anatomy
Bead	put a bead up her nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004
	a small bead stuck in her nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004
Cheerio	cheerio stuck in nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004
fluff	fluff in her nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004
Lego	lego stuck in nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004
	lego up her nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004
Nut and m&m	nut and m&m shoved up his nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004
Paper	stuck paper in his nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004
Pea	pea stuck up nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004
Popcorn	popcorn kernel stuck up her nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004
Raisin	put raisin up nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004
	raisin in nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004
	rock in nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004
Rock	rock stuck in nostril	Foreign Body;NCIT_C34620	nose;UBERON_0000004
	rock up her nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004
	same rock stuck in nostril	Foreign Body;NCIT_C34620	nose;UBERON_0000004
Trail mix	trail mix stuck up nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004
Unknown	foreign body nasal	Foreign Body;NCIT_C34620	nose;UBERON_0000004
	foreign body removal from nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004
	foreign object in nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004
	object in nostril	Foreign Body;NCIT_C34620	nose;UBERON_0000004
Vitamin	stuck vitamin up his nose	Foreign Body;NCIT_C34620	nose;UBERON_0000004

The reasons for hospital use data were not incorporated into the downstream machine learning models outlined later in this thesis since this procedure was carried out after the construction of the models and was also considered too granular to be used as each separate reason would have to be one-hot-encoded to be used for the model. However, these data are being used in other CHILD Cohort projects to study the health trajectories of children with the goal of better defining what makes a healthy child, which are beyond the scope of this thesis.

2.5. Concluding Remarks

This dataset consisting of 64,817 total entries for 3243 subjects describing their reasons for medication use is a landmark dataset. Consisting of prescription, OTC and natural health products, this dataset will reveal associations between medications and associated reasons for use that were not able to be previously appreciated. Taking on the task of standardizing these data has removed the burden from future researchers interested in these data and will help them conduct more streamlined and efficient analyses.

Having access to standardized cohort data will also make it easier to harmonize data between cohort studies. Harmonized cohort studies are incredibly important as they allow for more robust conclusions to be drawn that may not be possible within a single cohort. The Common Infrastructure for National Cohorts In Europe, Canada and Africa, or CINECA project, is one such group that I have worked with that is focused on making data transferable across different cohorts. Their third work package, "Cohort Level metadata representation," is focused on the implementation of ontologies to facilitate better data sharing on a global scale (Lawson et al., 2021).

The integration of standardized terms for reasons for medication use will have far-reaching applications outside the scope of this thesis project. The CHILD Cohort has been working on CHILDdb, a database through which researchers can request CHILD Cohort data to investigate their specific research topics. The curated ontology terms will be integrated into CHILDdb, and researchers from all over the country will have access to these data.

There are some caveats that must be addressed with these data. The first is when discussing the percentage of cohort participants that use medication for a certain symptom, we must qualify the statement by saying "X% of participants who provided a reason for use" because cohort participants were not required to submit a reason for medication use for their reported medications. Therefore, there is a possibility that not all reasons for medication use are accounted for due to missing responses. Similarly, if a researcher were to use the reasons for medication use data as an indicator of health outcomes in a cohort, they would have to specify that these are "medicated health outcomes" because if a participant had some adverse health outcome and did not

medicate for it, it would not show up in this dataset. However, these caveats do not completely negate the utility of this data or undermine its inclusion in the analyses outlined in Chapters 4 and 5.

Chapter 3.

Globe-Based Visualization of Correlation Datasets

This chapter explores the creation of a dynamic, interactive tool to aid in the visualization of correlation datasets. This visualization tool can be found at globecorr.ca. For this tool, I handled the creation and organization of static content (i.e. tutorial and FAQ pages, example dataset curation, navigation text). Once the tool was live, I led testing out trial versions of software and organizing new user tests. I would then collect and direct feedback to the appropriate developers, Mariam Arab or Nolan Woods, who wrote the code to generate the visualizations. I led the writing of the manuscript describing this tool, with the participation of the other authors, and this manuscript is in preparation for submission to Bioinformatics. The functionality of GlobeCorr is outlined in Section 3.3, and applications and a comparison of this method to existing correlation methods are summarized in Section 3.4. This tool was also used to further analyze data in subsequent chapters.

3.1. Abstract

In many fields, increasingly complex correlation datasets are being generated. This includes omics datasets with associated, diverse categories of metadata. Correlations between variables help examine potential confounding factors. Static, globe-based visualizations of pairwise correlations have been used to aid statistical analyses, but to capitalize on the benefits of dynamic visualizations, GlobeCorr was created. GlobeCorr.ca is a web-based application that provides interactive visualization and analysis of correlation datasets. Users upload tabular data, and GlobeCorr will create a dynamic visualization using ribbons to represent positive and negative correlations, optionally grouped by domain/category if required. GlobeCorr provides a simple method for users to visualize and identify potential confounders and quickly visually summarize complex datasets. This tool applies to a wide range of disciplines and domains of interest.

3.2. Introduction

Correlation analysis is a common method used in many different fields of research. Visualizing correlations can help identify confounding variables. As more complex datasets are being generated, there is a need to expand upon the visualization tools available for correlation analysis that will scale well with the large amount of data being created, especially in the omics field. The goal of creating GlobeCorr was to complement existing strategies, not entirely replace well-established and accepted methods.

Heat maps are a well-established visualization tool for identifying correlated variables in large datasets. While they can be highly customizable for users with coding experience, the large datasets found in omics studies are laborious and time-consuming to analyze using heat maps. Krzywinski et al. created circos plots to organize and identify relationships in rich datasets for genomics.

However, the need for new visualization tools is not limited to genomics. For other domains, Patel and Manrai introduced the concept of exposome globes for mapping environment-wide associations and visualizing them simply. These static globes can be created using the R codes provided by Patel and Manrai. In this case, creating correlation globes requires some relatively advanced experience and knowledge of R. Even when using the provided code, it can be difficult to understand the choices made in the visualization. For example, the domains are arranged to minimize crossover in the middle of the globe. This is not immediately evident without reading the background information and can lead to confusion. GlobeCorr was created to be an accessible platform for researchers to create static and dynamic visualizations for their complex datasets without requiring any complex coding experience.

3.3. Methods

3.3.1. Website Specifications

GlobeCorr is a visualization software. It does not perform any statistical analyses or store data on its servers and is easily run off-line. This makes it a suitable tool for users worried about data security. The web platform for GlobeCorr is written using the

Vue framework, and the visualizations are created using the amCharts 4 JavaScript library. A diagram of the visualization portal can be seen in Figure 3.1. The major elements highlighted in the figure are the website navigation menu, upload box, visualization window, and globe options menu. The website navigation menu is used to switch between pages such as the FAQ and tutorial. The upload box is where users can drag the appropriate csv or click the box to open a dialogue box to select their data. The visualization window is where the dynamic globe will appear, and the Globe Options menu is where users can personalize the globe and find the export function to create a static image.

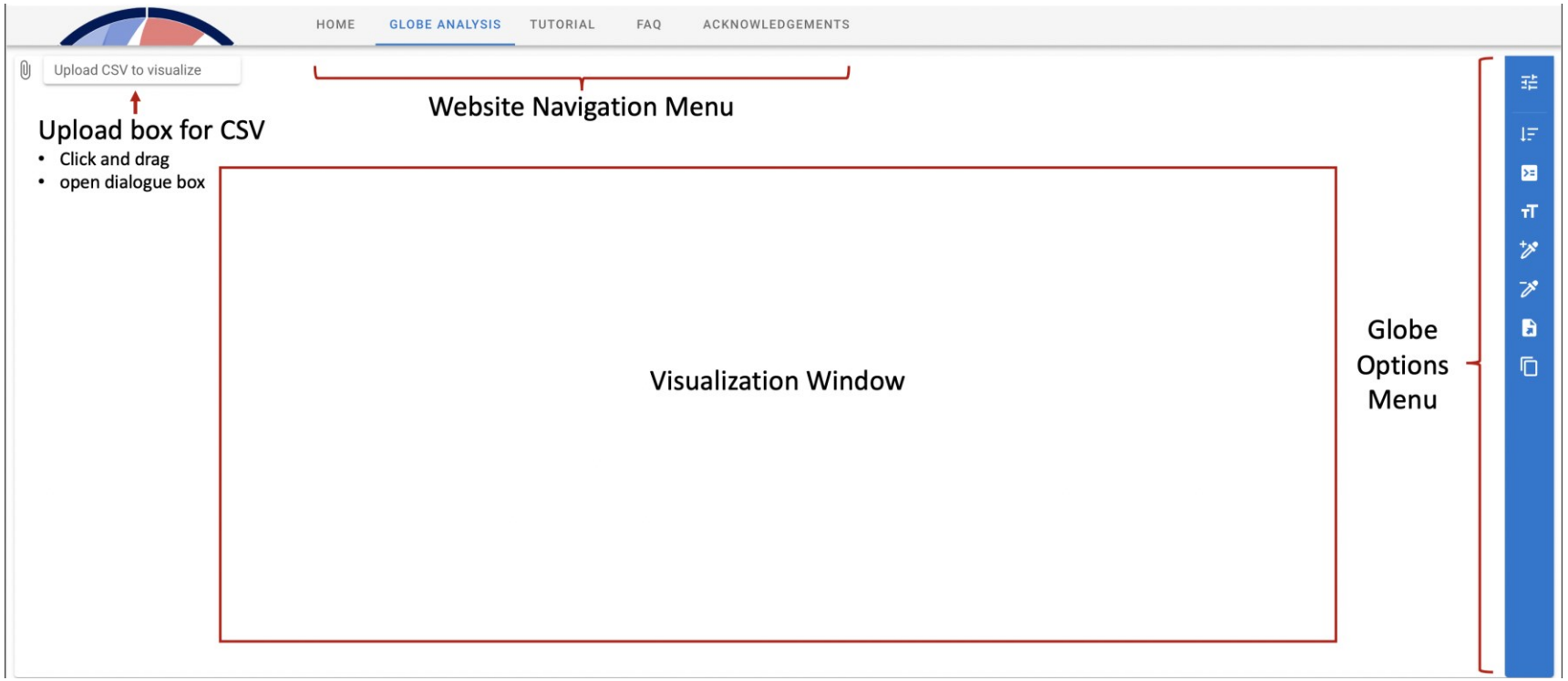

Figure 3.1. Diagram of GlobeCorr visualization portal as viewed through the Google Chrome browser on a MacBook Pro. Important features indicated are the upload box for a users data, the website navigation menu, the visualization window where the globe will appear and the Globe Options menu which contains the tools needed to customize a visualization.

This window can be seen at globecorr.ca/globe

3.3.2. Visualization Creation

To generate visualizations, GlobeCorr requires a comma-delimited file with the columns containing the following information (in order): variable1, variable1 domain, variable2, variable2 domain, correlation coefficient, with the following column headings "variable1, var1_domain, variable2, var2_domain, coef". An example of the format is provided in the tutorial at globecorr.ca/tutorial and can also be seen in Table 3.1. GlobeCorr uses the PapaPare.js library to read these files.

Table 3.1. Example of csv format required for GlobeCorr.ca visualizations. The first column indicated the name of the first variable, the second column indicates the domain of the first variable, the third column indicates the name of the second variable, the fourth column shows the domain of the second variable and the fifth column is the correlation coefficient associated with the two variables.

variable1	var1_domain	variable2	var2_domain	coef
varU	domain1	varI	domain2	-0.5041308
varZ	domain3	varR	domain1	0.5093516
varM	domain4	varN	domain2	-0.7682316
varZ	domain5	varE	domain6	0.40976778
varG	domain2	varB	domain7	0.50423833
varP	domain2	varT	domain3	0.41072329
varA	domain2	varX	domain2	-0.420769

A full version of this table can be found at globecorr.ca/tutorial

Examples of GlobeCorr visualizations can be seen in Figure 3.2. The arcs around the circle's circumference will each have a unique colour and correspond to one domain specified in the input file. The default arrangement of the domains around the circumference is by size, with the largest domain being first when plotting in a clockwise direction. However, there is also the option to organize by the order found in the file. A comparison of these two options can be seen in Figure 3.2. Domains describe groups of variables all relating to the same type of information. For example, if one was looking at the correlations in dietary patterns, all variables relating to vegetable consumption could be visualized under the domain "Vegetables".

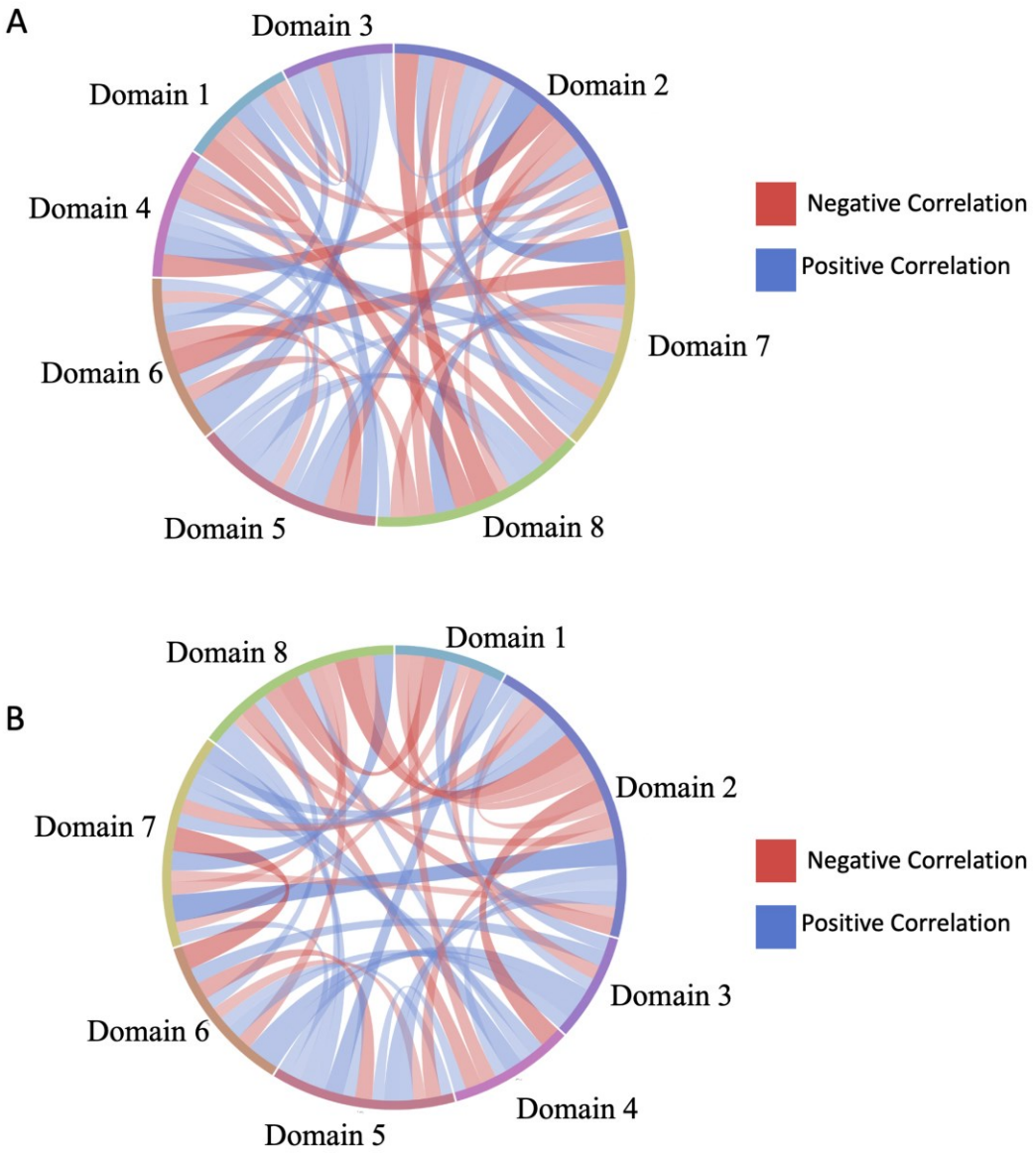

Figure 3.2. Example of GlobeCorr diagrams. A sample globe arranging by domain size can be seen in panel A, and an example globe arranging by order in input file can be seen in panel B. The data being vizualized is the example dataset found on GlobeCorr.ca

Domains can also be moved manually by clicking on the domain of interest and dragging it to another point on the edge of the circle. Domain and associated correlations can be removed entirely from the visualization by clicking the domain. An example of this can be seen in Figure 3.3.

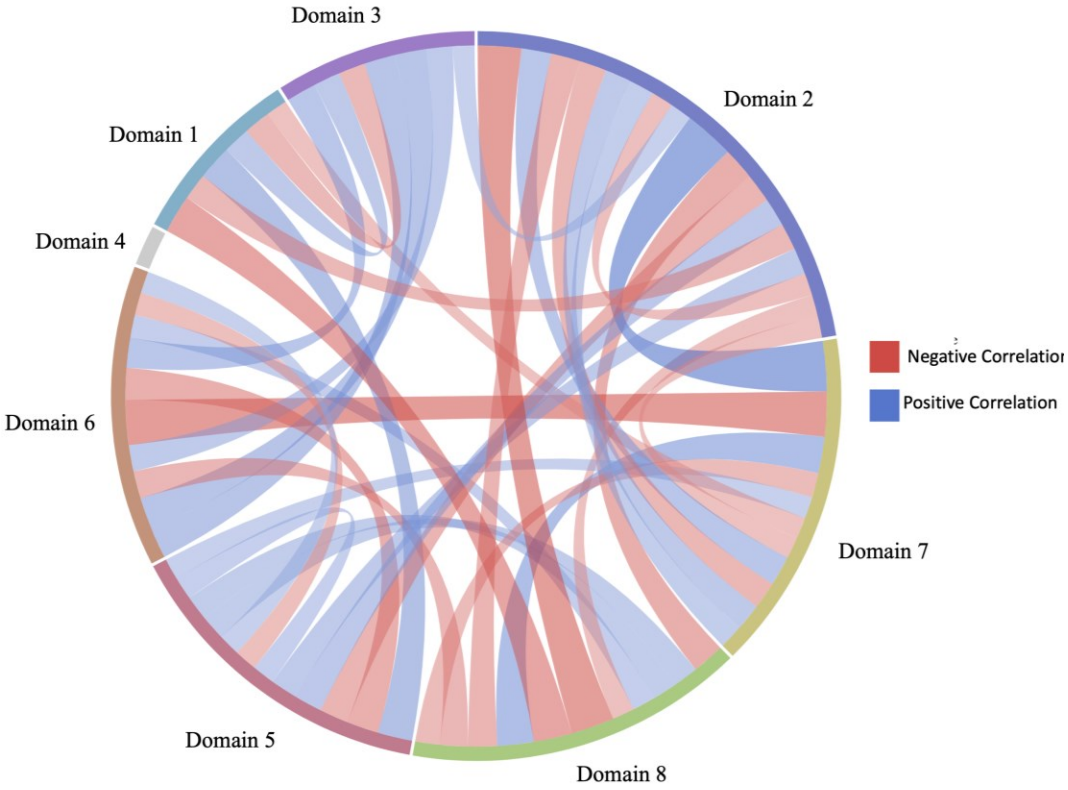

Figure 3.3. Example of GlobeCorr image from Figure 3.2A with domain 4 removed from the visualization by clicking on the corresponding arc. Data used for visualization is the example dataset found on GlobeCorr.ca

The bands going across the circle are coloured according to whether the correlation coefficient is positive or negative, and the width of the band corresponds to the magnitude. When looking at the dynamic visualization, hovering over the band of interest with your mouse will display the associated variables and their coefficients. An example of this can be seen in Figure 3.4. To reduce the number of correlation bands in the globe, users can specify a correlation threshold using the options in the Globe Options menu on the right-hand side of the screen. Users can select a threshold using the slider or by typing in a specific value.

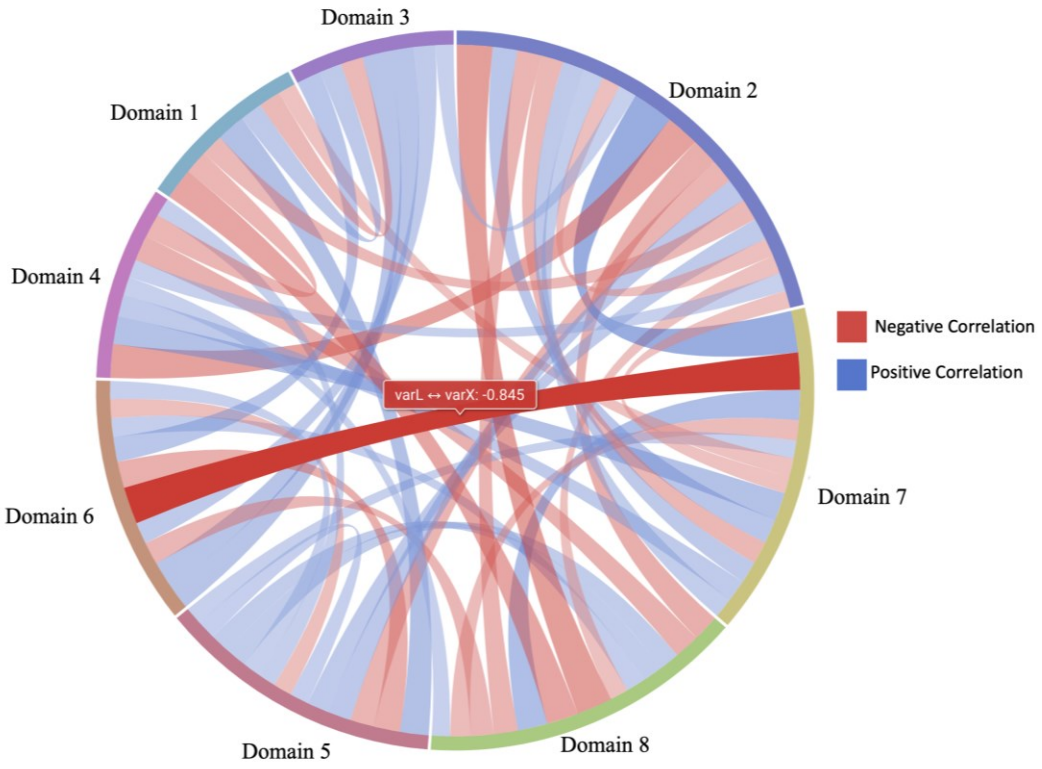

Figure 3.4. Example of GlobeCorr image highlighting a correlation ribbon to shown variable labels and correlation coefficient. Data used for visualization is the example dataset found on GlobeCorr.ca

For aesthetics, the user can change the font size of the domain labels in the Globe Options menu. For colours, the user can only change the colour of the correlation bands, not the domains. The positive and negative bands can have their colours changed independently to improve visibility and accessibility or match a preferred colour scheme for export. Users can also find the options for exporting under the Globe Options menu. File formats available for static files are png, svg, jpg, or pdf. Users can save the visualization settings (i.e. the domain plot order, correlation threshold and band colours) via the copy and paste of a text field that maintains JSON encoded values of the settings found in the GlobeCorr Globe options menu.

3.4. Results and Discussion

While GlobeCorr was created to help researchers visualize and interact with their correlation datasets, it is not meant to replace existing methods for visualization. Instead,

the aim is to complement existing methods such as heat maps. To show this comparison, Figure 3.5 and Figure 3.6 show the same dataset visualized in a heat map and a correlation globe, respectively.

Unlike heatmaps that will label every variable, GlobeCorr only shows the variable labels when hovering over a specific correlation ribbon in the dynamic view. Someone looking at a static GlobeCorr plot would require more background information on the variables contained within the globe compared to someone looking at a heatmap, where an unfamiliar user could just read along the axis to get a better idea of the specific variables involved in the analysis.

GlobeCorr diagrams are best suited to correlation analyses where the correlations are between variables of the same “type.” For example it means that a user is doing a study looking at how different microbial taxa correlate with each other instead of seeing how microbial taxa correlate with environmental contaminants. An example of these two situations, comparing the heat map vs. Globe, can be seen in Figure 3.5 and Figure 3.6, respectively. The layout of a circle works well for information of the same type, but when there are two “types,” i.e. microbial taxa and contaminant, the division of those two using the two axes of a heat map is preferable for clarity.

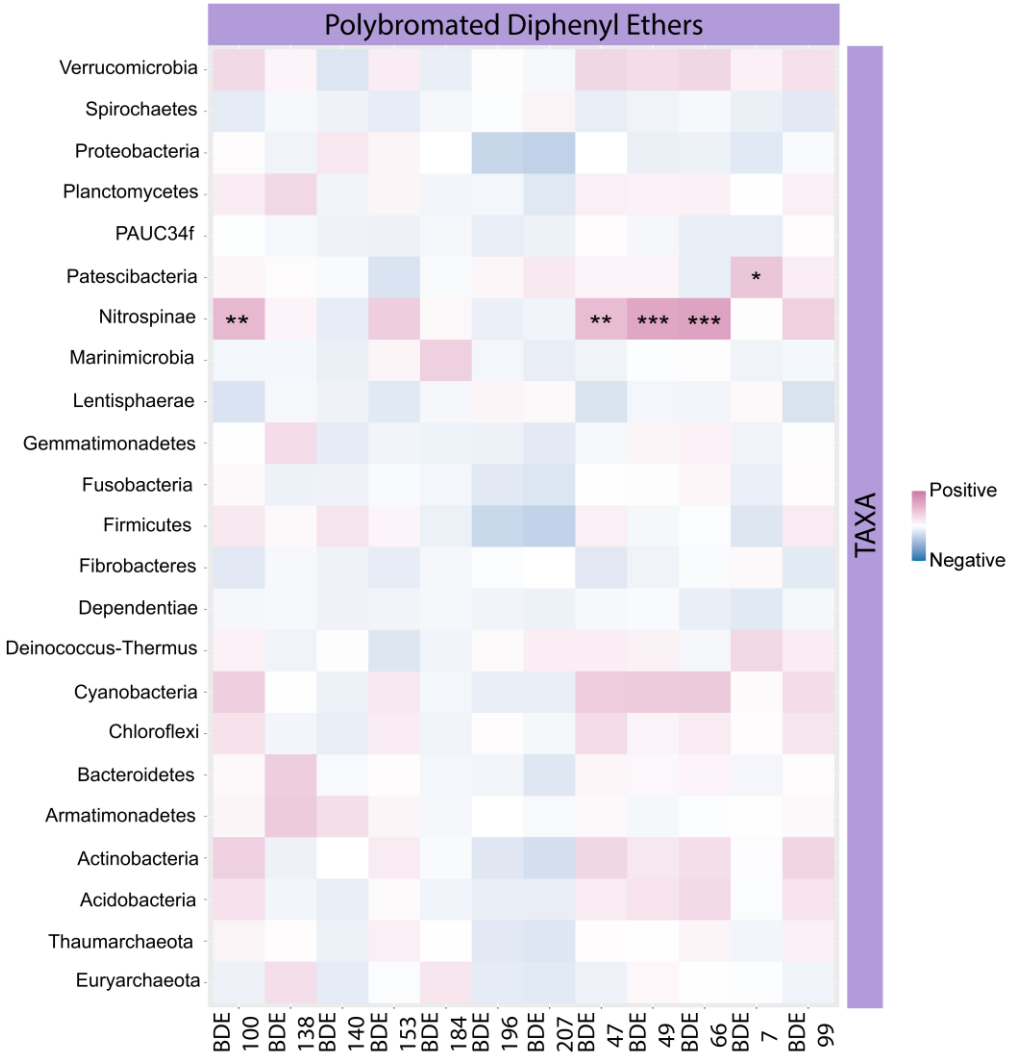

Figure 3.5. Comparison of Heatmap visualization for Beluga skin microbiome and environmental toxin analysis. Heat map depicting the relationship between Beluga whale (Delphinapterus leucas) microbiome (shown as domain; phylum) and blubber contaminants. Blubber contaminants are arranged along the horizontal while microbiome components are shown along the vertical at the phylum level. A diverging colour scale is used to show the strength of the correlations, with dark pink being positive and blue being negative correlations. The significance for correlation is indicated by:*: p<0.001, **: p<0.01, *p<0.05.**

Heatmap courtesy of Justin Jia (Jia et al., 2022, Submitted to *Frontiers in Environmental Science*)

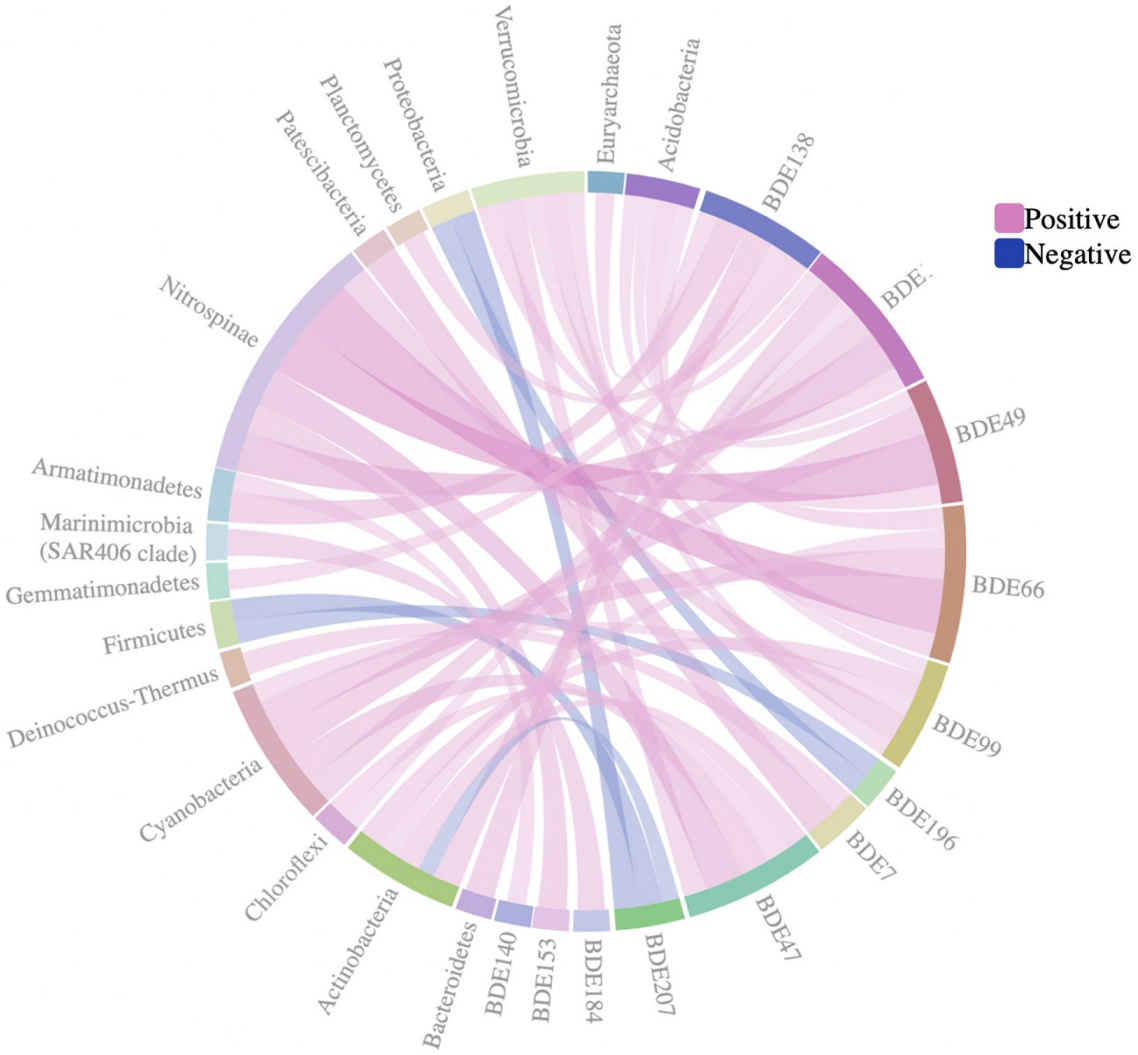

Figure 3.6. Correlation Globe visualization for Beluga microbiome and environmental toxin analysis. Sample GlobeCorr plot depicting the relationship between Beluga whale (*Delphinapterus leucas*) microbiome and blubber contaminants with a correlation cut-off of |0.2|. Blubber contaminants are located on the right and side of the image and microbiome components are depicted on the left at the phylum level.

3.5. Concluding Remarks

GlobeCorr has limitations and best use cases like any other method. Regarding the number of correlations that GlobeCorr can handle, the dynamic aspect of GlobeCorr tends to struggle with larger csvs (approximately anything over 500 correlations), and

the responsiveness of the website lags, making the desired customizations challenging to achieve. With larger globes, the static visualization can be lacking in necessary detail.

GlobeCorr is an excellent tool for dynamic data exploration through a browser. This could be on an individual or a larger scale if a user is willing to share their browser during a presentation. The use of GlobeCorr facilitates engagement with the visualization that is impossible with a static image. It can be challenging to transfer the dynamic images generated in GlobeCorr to a compelling static image. Long domain labels can be cut off, and the legend is very small. The individual ribbons also have no labels, making it difficult for users to know what is going on in the static globe unless they have an accompanying dynamic version. A presenter can go through individual slides with highlighted ribbons and include each globe as a separate image. However, this can be time-consuming and make formatting difficult. When using GlobeCorr, users must reflect on their needs and the properties of their data and decide if GlobeCorr is the most appropriate tool for them.

By understanding the limitations of the GlobeCorr software, users can customize figures to their specifications with limited coding experience and use these results in harmony with other visualization tools to explore their data to generate hypotheses and identify confounding variables. The creation of GlobeCorr addresses the need for an accessible and dynamic visualization to examine correlations within complex datasets.

Chapter 4.

Machine Learning-Based Analysis of Medication Use and Lifestyle Factors Associated with Microbial Dysbiosis

This chapter outlines the creation of Random Forest and Gradient Boosting models to predict microbial dysbiosis as measured by the Enterobacteriaceae-to-Bacteroidaceae (EB) ratio or Firmicutes-to-Bacteroidetes (FB) ratio. This task was undertaken with the hypothesis that there are patterns of medication usage predictive of microbial dysbiosis that can be identified through machine learning. Section 4.3 outlines the methods used. I curated a dataset of 1547 variables for 564 participants from the CHILD Cohort Study and used multiple imputation by chained equations to handle any missing data in the dataset. With a complete dataset, I created Random Forest regression and Gradient Boosting regression models, tuning model parameters where necessary to optimize their error scores. Upon completion, I evaluated the importance scores of the variables included in the machine learning models to determine candidates for more in-depth analyses to explore the associations between medications and microbial dysbiosis. This evaluation procedure is outlined in Section 4.4.

I completed all work presented in this chapter with the following exceptions: The data collection was done from 2009-2012 by various research assistants in the CHILD Cohort Study. The microbiome data were generated by the Turvey lab at UBC and made available through the CHILD Cohort Study. All software was run using R v4.1.0 on Cedar, supported by The Digital Research Alliance of Canada.

4.1. Abstract

The interaction between medication usage and microbial dysbiosis is likely complex and multifactorial. While some medication use, such as antimicrobial use, has been associated with dysbiosis, the impact of most medications on gut microbiomes is not known, especially in infants. To initiate the exploration of these complex relationships, machine learning was used to identify patterns of medication usage that may be predictive of microbial dysbiosis. Using a curated dataset of 1547 variables for

564 participants, Random Forest and Gradient Boosting regression models were built to predict infant microbial dysbiosis at three months and one year, using Firmicutes-to Bacteroidetes ratio and the Enterobacteriaceae-to-Bacteroidaceae ratios. Using the variable importance scores from both models, antimicrobials, analgesics, and vitamin D were identified as medications with strong associations to the microbial dysbiosis outcomes. Evaluation of these models provides a preliminary prioritization strategy for medication and microbial dysbiosis associations requiring further investigation and the investigation of more causative factors associated with microbial dysbiosis. The ranking of variables in the machine learning models may suggest that other lifestyle factors are more causative of microbial dysbiosis versus medication use. This work illustrates the benefit of more holistic integration of lifestyle and medication use data. It provides a base for further exploration of the complex relationship between medication use, lifestyles, and microbial dysbiosis.

4.2. Introduction

There are many decision tree-based methods that are commonly used for the analysis of large multidimensional datasets in biomedical and life sciences (Banerjee et al., 2019). One of the most common decision tree-based methods is the Random Forest. Decision trees can be used for regression and classification. Decision trees generally start with a singular node. This node represents a variable that can effectively split the data into distinct groups. From there, one can trace down to intermediate nodes, which contain other variables that can be used to split the data. Ultimately, by tracing down these nodes, one will arrive at a leaf node, which will give you the final value or classification for your data point. An example of this structure can be found in Figure 4.1.

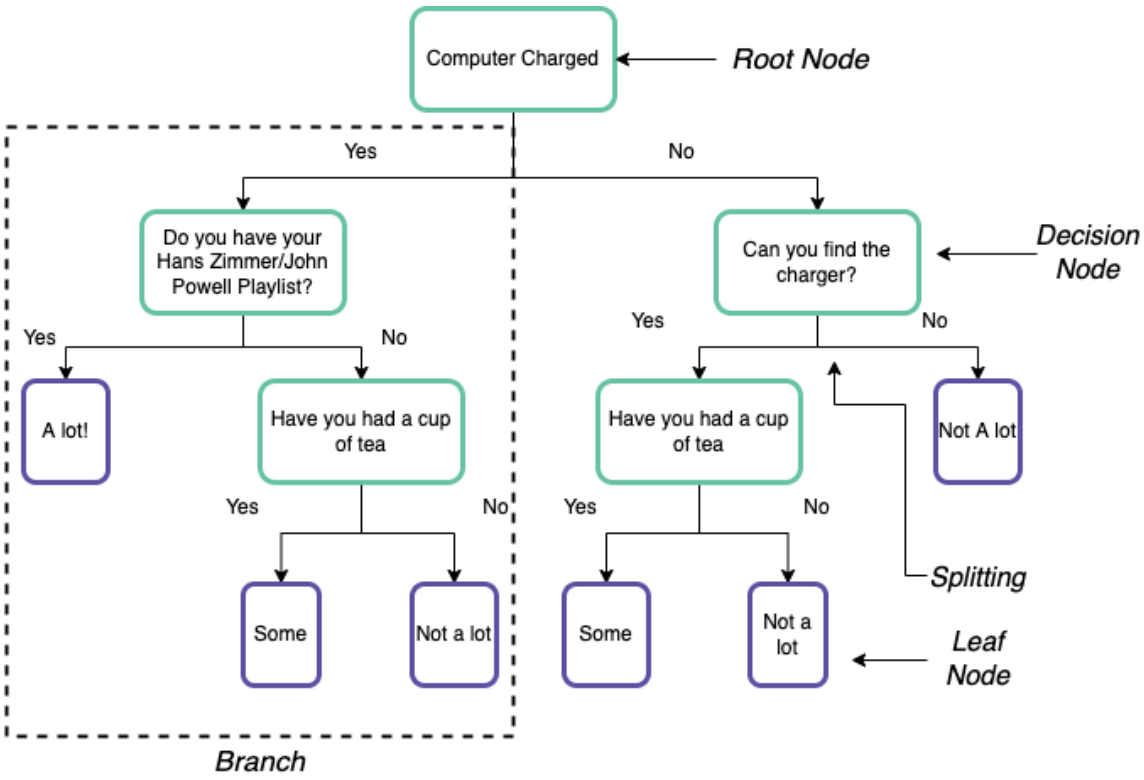

Figure 4.1. Example of Decision Tree Anatomy. This decision tree was created to predict the amount of work a bioinformatics master's student will get done given the conditions of her working environment. Elements shown are a root node, branch, decision node, splitters and leaf nodes.

As the name might suggest, Random Forest models involve a collection of decision trees. For these models, a certain number of trees are made, each having been constructed slightly differently. This group of *n* unique trees will all sort the same data according to their structure and come up with an answer, and each tree's answer is compiled before giving the result. Multiple distinct trees are used in this instance to avoid overfitting and reduce the error rates of the predictions (Liaw & Wiener, 2002).

When working with machine learning models, there is always a concern for optimization. There are several metrics in which to evaluate optimization. A popular method for regression trees is using the Root Mean Squared Error (RMSE). When dealing with a regression model, a regression line is generated, and the RMSE describes how far away the actual data points are from their value when following the regression line. The smaller the RMSE values, the better. The RMSE is interpreted in the same units as the outcome of interest. Therefore, the value of a "suitable" RMSE is

subject to the range of the outcome variable. In the case of Random Forest models, several parameters can be tuned to decrease the RMSE and optimize the model. The mtry parameter refers to the number of variables included in a tree at a time, as the construction of the Random Forest tree relies on a subset of the given variables for each tree created. Another parameter, referred to as ntree, describes the number of trees built in the model. These two parameters were altered in different permutations to find the models with the lowest RMSE values for a given outcome.

Decision trees are also a hallmark of the Gradient Boosting method. However, the construction of the model is significantly different. While Random Forest models create independent trees, Gradient Boosted models build trees sequentially, building off information from the preceding tree. Like Random Forest, Gradient Boosting models can also be run as regression or classification models.

When building a Gradient Boosting regression model, the tree starts as a singular node that is the average value that we are trying to predict. This is used as the predicted value for all entries in the dataset. Then, pseudo-residuals are calculated based on the difference between the observed and predicted values. For this first tree, all predicted values are the same. Once the pseudo-residuals have been established, a new tree is built to predict the pseudo-residuals, not the outcome of interest. The predicted pseudo-residuals are then added to the average value of the outcome variable, and a second, new pseudo residual is calculated based on the observed value and the average value plus the pseudo residual predicted from the first tree. However, the pseudo residual predicted from the first tree is weighted before being added to the average value. This weight is referred to as the learning rate. The algorithm attempts to drive the model through small, sequential improvements by employing the learning rate. This iterative process will continue until a certain number of trees have been completed or the model is no longer significantly improving. This iterative learning process is often the justification for selecting a Gradient Boosting approach. By tuning the trees each time, users generally end up with a smaller error rate for their predicted values compared to other methods (Natekin & Knoll, 2013).

To evaluate the performance of the models, Gradient Boosting also uses the RMSE. This value can be interpreted in the same as one would for a Random Forest, with units identical to the outcome variable and a "good" score dependent on the

magnitude of the score themselves. As with Random Forest, some parameters can be changed to optimize the RMSE values of Gradient Boosting models: learning rate, interaction depth, minimum nodes and the bag fraction. The learning rate is the weight applied to the newly predicted residuals at each iteration of the gradient boost descent. The interaction depth is the number of splits on a node. The minimum nodes is the number of observations allowed in a terminal node. Should there be multiple observations in a terminal node, will have the average taken and then that average value will be used to determine the pseudo residuals at that stage. The bag fraction determines the percentage of the observations used for each tree.

4.3. Methods

4.3.1. Data collection

All data was collected via the CHILD Cohort Study. As this project aims to investigate early life associations of the gut microbiome, the first criteria to be met for inclusion in this analysis was whether there was an available stool microbiota sample for that participant. A preliminary data quality assessment indicated that I should use the dataset from Turvey et al. This dataset had 564 participants with microbiome samples at three months and one year, with taxa assigned at the phylum and family level. Following this, I assessed additional numeric, binary, ordinal and continuous data for inclusion in this model. For any character values, such as the response "fever" to the question "reason for medication use," these values were one-hot encoded. An example of one hot encoding can be seen in Figure 4.2.

Subject	Medication
Subject 1	Acetaminophen
Subject 2	Crystal Violet
Subject 3	Salbutamol
Subject 4	Cephalosporin

Label Encoded-Character String

Subject	Acetaminophen	Cephalosporin	Crystal Violet	Salbutamol
Subject 1	1	0	0	0
Subject 2	0	0	1	0
Subject 3	0	1	0	0
Subject 4	0	0	0	1

One Hot Encoded-Binary Variable

Figure 4.2. Example of label encoding and conversion to one-hot encoding.

The additional data assessed for these models described can be assigned to the following categories or "domains": infant's healthcare, mother's healthcare, infant nutrition, mother's nutrition, infant medications, mother's medications, home

environment, external environment, lifestyle, and demographics. The data was collected at a range of time points from the prenatal questionnaire to 1 year. One exception is that metrics for height and weight were also obtained from the 5-year time point to assess any potential association with the early-life microbial imbalance and BMI at five years.

The microbiome data were available as raw counts from the CHILD Cohort Study. However, to account for the compositional nature of the dataset, everything was expressed as a ratio to other taxa (Gloor et al., 2017). For the microbiota data at the phylum level, there were 22 distinct taxa. For the microbiota data available at the family level, there were 165 distinct taxa. It is important to note that while this solves the compositional problem, using ratios introduces two mathematical problems. Having a zero in the numerator will give a result of zero, and having zero in the denominator will give a result of undefined. Biologically, this indicates that relationships with low abundance taxa would not be captured by this method. For this project, I was interested in examining the influence of these variables on microbial dysbiosis. As a dysbiosis metric, I used the Firmicutes-to-Bacteroidetes Ratio (FB) and the Enterobacteriaceae-to-Bacteroidaceae (EB) (Ley et al., 2006; Vu et al., 2021). Using ratios is a common practice when the sequencing data is already available due to its computational simplicity (Wei, 2021). The FB ratio has been positively correlated with obesity, in both human adult cohorts and mouse models (Mange et al., 2020). The EB ratio has been positively correlated with atopic sensitization in infants (Azad et al., 2015). . With these ratios, it is important to note that gut microbial composition changes with age, and especially during infancy as new foods are introduced to the infant's diet. However, microbial dysbiosis is still a concern as dramatic changes still occur even in this context and warrant more research (Underwood et al., 2020).

Microbiome data was available at both the three-month and one year timepoint, so these ratios were calculated at both points. Giving a total of four outcome variables: Firmicutes-to-Bacteroidetes at three months (FB_3m), Firmicutes-to-Bacteroidetes at one year (FB_1y), the Enterobacteriaceae-to-Bacteroidaceae (EB_3m) and the Enterobacteriaceae-to-Bacteroidaceae at one year (EB_1y). Summary figures of the FB and EB ratios can be seen in Figures 4.3 and 4.4, respectively. The values for cohort subjects are shown in black, with a reference dot in red to represent documented ratios for those microbes for cohorts of similar ages. In the machine learning models, all outcomes were treated as a continuous variables.

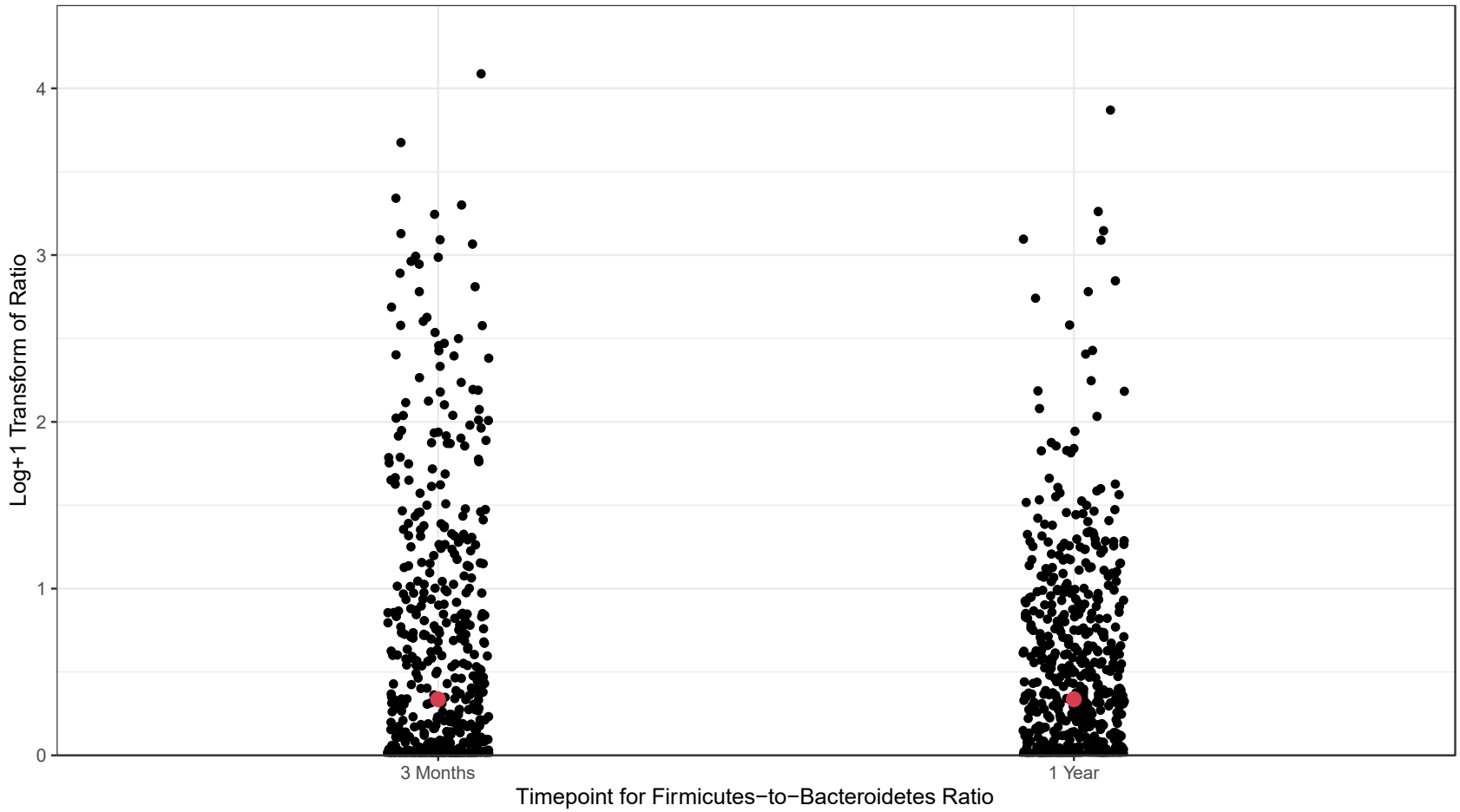

Figure 4.3. **Violin plot of Firmicutes-to-Bacteroidetes ratios for infants in the CHILD Cohort Study. The x axis shows the two timepoints where data was collected (3 months and 1 year), and the Y axis shows the ratios with a log+1 transform to improve readability. The spread of the data is indicated by the green violins, with the individual measures represented by black points. The larger pink datapoint at both timepoints represents the documented Firmicutes-to-Bacteroidetes ratio in infants of 0.4 (Mariat et al., 2009).**

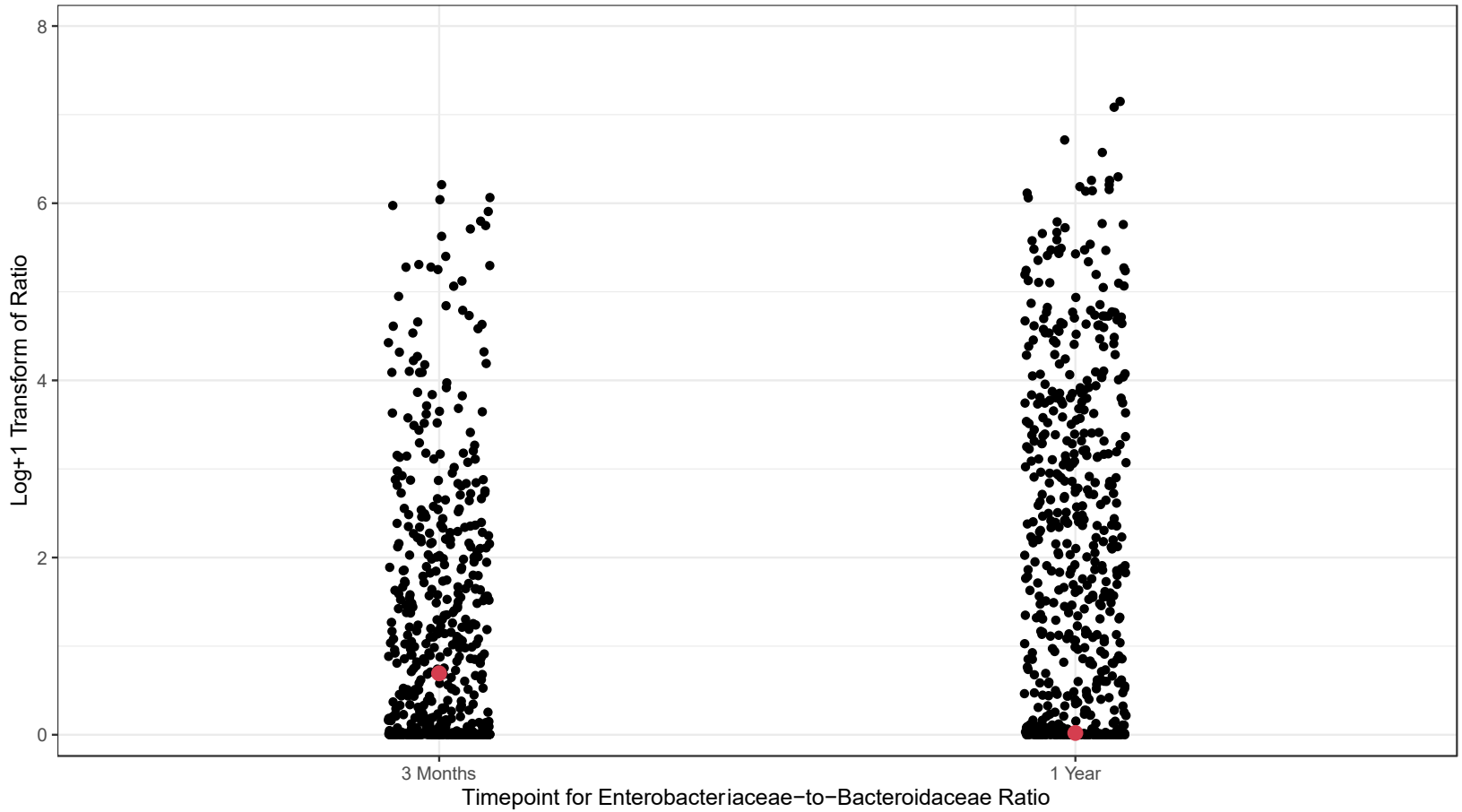

Figure 4.4. **Violin plot of the Enterobacteriaceae-to-Bacteroidaceae ratios for infants in the CHILD Cohort Study. The x axis shows the two timepoints where data was collected (3 months and 1 year), and the Y axis shows the ratios with a log+1 transform to improve readability. The spread of the data is indicated by the green violins, with the individual measures represented by black points. The larger pink datapoint at both timepoints represents the documented Enterobacteriaceae-to-Bacteroidaceae ratio in infants of 1.0 for 3 months and 0.02 for 1 year (Azad et al., 2015)**

4.3.2. Multiple Imputation by Chained Equations

When creating the dataset for the model, the only requirement for data completeness was for all subjects to have complete gut microbiome data at three months and one year. From this, 564 participants were identified. These 564 participants were not required to have completed all the other surveys and tests curated for the rest of the dataset. This resulted in many instances of missing data. For example, breast serum metabolomes and urine phthalate levels were also tested for the CHILD Cohort Study. However, these datasets did not overlap and had anywhere from 40-89% missing data since the subjects that had valid microbiome data at 3 months and 1 year did not have the serum metabolomes or urine phthalate levels available. Overall, the variables with the highest percentage of null responses are from other diagnostic tests (such as the metabolome, phthalate levels and household dust composition) as opposed to a lack of survey responses from the participants.

When dealing with missing data, there are two main approaches. The first is a complete case analysis, wherein the researchers will only use the participants that have completed all questions. This was not a feasible option for this project, as this left me with insufficient sample size. The other option is imputation, which was the strategy employed for this project.

Multiple imputation by chained equations (MICE) was used to account for the wide array of data. I implemented this method using the MICE package in R (Buuren & Groothuis-Oudshoorn, 2011). This method uses the values found in the rest of the dataset to predict the missing values in a specified column. To specify which columns would be used for imputation, I created a predictor matrix in R, which would be applied during the imputation process to ensure that columns such as subject id were not being used to impute values that would be later used in the machine learning models.

Different imputation methods for different variable types can be specified when using MICE. MICE classifies variables as continuous, ordinal and binary, which will be automatically applied as the procedure takes place. Table 4.1 shows the methods used for each type of variable when using MICE in this project.

Table 4.1. Methods used for different variable types in Multiple Imputation by Chained Equations. Methods implemented using MICE v. 3.14.0 in R v 4.1.1.

Variable Type	Method	Rationale
Continuous	Classification and Regression Trees	Preferred when dealing with non-linear relationships and skewed distributions
Binary	Logistic Regression	Mice default method for binary
Ordinal	Proportional odds model	Only option for ordinal data

Another hallmark of MICE is the creation of multiple datasets that may have slight variations in the imputed values, which helps account for some of the unknown variability when trying to predict missing values. After creating multiple datasets, the same analyses would be run on all n imputed datasets and the results pooled. For this project, an n of 5 was selected when running MICE.

There was a wide range of data completeness for the questions asked in the CHILD Cohort Study for the 564 participants selected for this project. To look at the impacts of including variables that were more vs less complete (therefore requiring less or more imputation), different datasets with different data completeness thresholds were created. The baseline dataset where variables containing any missing percentage were used, as well as datasets requiring 70%, 80%, 85% and 90% complete variables. This gave a total of 5 different datasets with the different missing data allowances, and each would be imputed five times per the MICE protocol. This resulted in 25 different datasets for the machine learning models.

There is some debate with imputation and machine learning about whether the data should be split into training and testing datasets before imputation. The argument is that should all data be imputed at once, the testing dataset will influence the values in the training datasets. This would be an indication of data leakage, which is ideally avoided. However, I chose to impute all the data beforehand and then split it into testing and training datasets. This was done for practical reasons and justified in this project's goal. Practically, this would have meant I was working and organizing 50 datasets instead of 25. In terms of the project goal, it is not the intention to create a tool that can accurately predict microbial dysbiosis in infants. The machine learning model is being

used to identify potential variables of importance that may require further research into their impact on infant microbiota.

4.3.3. Machine Learning Model Construction

Random Forest

For this project, all models will be treated as regression models. This is because I am treating the outcome of interest, one of the four microbial dysbiosis and time combinations, as a continuous as opposed to a categorical variable. In this project, I will not be classifying subjects as having "poor" or "favourable" ratios. These models will be used to identify variables associated with shifts in common measures of microbial dysbiosis.

To evaluate the Random Forest regression models in this project, I have chosen to evaluate the models based on RMSE. The units for RMSE are identical to the units of the outcome of interest. However, in this case, as we are predicting a ratio, the RMSE is unitless. Furthermore, as this model is being built with only the intention of identifying potentially influential variables and not as a method for microbial dysbiosis prediction, RMSE values were not treated as definitive markers for the success of these models. The actual construction and running of the Random Forest models were completed in R using the randomForest and Caret packages (Kuhn, 2008; Liaw & Wiener, 2002). In R, the parameters for mtry and ntree were altered using the grid-search functionality to find the models with the lowest RMSE values for a given outcome.

There were five different data completeness thresholds for each of the four outcomes, and each data completeness threshold had five different imputed datasets. By the nature of multiple imputation, a Random Forest would be run on each of the 25 datasets for each outcome (5 data completeness * 5 imputations) before pooling the results for each level of data completeness (ie. 70%, 80%, 85% and 90% complete variables), giving five models to evaluate for each outcome. Expanding this, this is 100 Random Forest models (4 outcomes *5 data completeness * 5 imputations), pooled down to 20 (4 outcomes* 5 pooled for data completeness). Table B.1 in Appendix B shows the organization of these Random Forest models as described in the text above.

Once the results are pooled, variables can be described by their Mean Decrease Accuracy Score, or %IncMSE. This describes how much the model improves upon including a particular variable. In broad terms, a positive value indicates that the variable does improve model performance, and a negative value indicates it hinders model performance. The magnitude indicates the strength of the impact in whatever direction would be implied by the sign of the value. These values are calculated by running the model with and without the specified variable and then evaluating the change in Mean Squared Error due to the exclusion of the variable in question. In this project, the %IncMSE values were used to determine variables of importance requiring further investigation.

However, upon what I will now refer to as the first iteration of the machine learning models, the top %IncMSE values were overwhelmed with variables containing one of the taxa of interest. For example, the ratio of Firmicutes-to-Bacteroidetes was used as a metric to describe a potential state of microbial dysbiosis. Other ratios like the Bacteroidetes-to-Actinobacteria would also be incredibly influential for the overall outcome. This is expected given that the outcome value was calculated based on one of the taxa included in the Bacteroidetes-to-Actinobacteria ratio. Therefore, it is comforting to see these values come up as high scoring because the model is performing as one would expect, given the known relationships between the variables. However, it is not good practice to include these types of variables due to the collinearity of the results.

To remedy this, a smaller set of Random Forest models were run again. Based on the RMSE scores of the models, the models that required 90% and 85% data completion were performing the best. Therefore, only the 90% and 85% models were re-run on the second iteration, where any microbiome ratios containing the outcome taxa were removed. To expand on the example noted above, in that case the Bacteroidetes-to-Actinobacteria ratio would be removed from the dataset for the model trying to identify predictive factors for the Firmicutes-to-Bacteroidetes ratio. This second iteration followed the same procedure as the first, with the %IncMSE scores from this second iteration used to determine the variables of importance requiring further discussion, which is outlined in Chapter 5.

Gradient Boosting

To complement the Random Forest models, Gradient Boosting regression was also used in this project. The RMSE is used as described above to evaluate the Gradient Boosting models' performance. However, when looking at the values during the procedure, this is referred to as the cv. error. This value can be interpreted as indicated in the previous section, with units identical to the outcome variable and a "good" score being dependent on the magnitude of the score themselves. To optimize the error score, the following parameters were altered: learning rate, interaction depth, minimum nodes and the bag fraction.

There were five different data completeness thresholds for each of the four outcomes, and each data completeness threshold had five different imputed datasets. By the nature of multiple imputation, a Random Forest would be run on each of the 25 datasets for each outcome (5 data completeness * 5 imputations) before pooling the results for each level of data completeness, giving five models to evaluate for each outcome. Expanding this, this is 100 Random Forest models (4 outcomes *5 data completeness * 5 imputations), pooled down to 20 (4 outcomes* 5 pooled for data completeness). Table B.2 in Appendix B shows the organization of these Gradient Boosting models as described in the text above.

The Random Forest models use %IncMSE to describe the importance of specific variables. The relative influence score or RelInf is the equivalent for Gradient Boosting. For each split, the change in MSE for the model is determined by the inclusion and exclusion of the variable in question. The most influential variables are those with the highest relative influence.

As with the Random Forest procedure, a second iteration of the Gradient Boosting models were run on a reduced dataset, omitting the variables that contain the values of the taxa of interest and using only the 90% and 85% data completeness thresholds. The RelInf scores from this second iteration were used to determine the important variables discussed in Chapter 5.

4.4. Results and Discussion

4.4.1. Random Forest Model Analysis

The Random Forest models using the datasets requiring 85% and 90% response rates before imputation were used to determine variables of importance. Both models were tuned, altering the number of variables included in the model (mtry) and the number of trees created (ntree). All models were trained independently, so it is possible that different models will have different parameters that resulted in the lowest RMSE value. While RMSE was minimized, that does not mean that all models are perfect. Table 4.2 and Table 4.3 show the summary statistics for the 85 and 95% models, respectively. For all outputs, the RMSE values are larger than the mean of the outcome variable. This indicates the models would be unsuitable if one were planning to rely on these models for the prediction of Firmicutes-to-Bacteroidetes or Enterobacteriaceae-to-Bacteroidaceae ratios. However, the goal of this project is not to create a perfect predictive model but to narrow down the most important factors associated with microbial dysbiosis.

Table 4.2. Summary Statistics for Random Forest model for 85 percent complete data. 20 models are shown (5 imputations of each 4 outcomes), the means of the outcome variables are shown, along with the model parameters for the number of variables included in the model (mtry) and the number of trees used in the model (ntree). RMSE is shown to describe overall quality of the model.

Imputation	Outcome	Mean	mtry	Ntree	RMSE
1	EB_1y	44.33	2200	500	113.86
2	EB_1y	44.33	2200	1000	109.05
3	EB_1y	44.33	2200	300	114.96
4	EB_1y	44.33	2200	400	107.53
5	EB_1y	44.33	2200	300	108.92
1	EB_3m	15.44	2200	300	61.25
2	EB_3m	15.44	2200	1000	61.85
3	EB_3m	15.44	2200	400	62.52
4	EB_3m	15.44	2200	300	60.28
5	EB_3m	15.44	2200	600	58.36
1	FB_1y	1.25	2200	800	1.95
2	FB_1y	1.25	2200	800	1.94
3	FB_1y	1.25	2200	400	1.89
4	FB_1y	1.25	2200	1000	1.92
5	FB_1y	1.25	2200	1000	1.96
1	FB_3m	1.84	2200	1000	4.1
2	FB_3m	1.84	2200	1000	4.02
3	FB_3m	1.84	2200	800	4.16
4	FB_3m	1.84	2200	800	3.76
5	FB_3m	1.84	2200	800	3.93

Table 4.3. Summary Statistics for Random Forest model for 90 percent complete data. 20 models are shown (5 imputations of each 4 outputs), the means of the outcome variables are shown, along with the model parameters for the number of variables included in the model (mtry) and the number of trees used in the model (ntree). RMSE is shown to describe overall quality of the model.

Imputation	Outcome	Mean	mtry_list	ntree_list	rmse
1	EB_1y	44.33	2200	500	116.71
2	EB_1y	44.33	2200	800	106.34
3	EB_1y	44.33	2200	1000	111.03
4	EB_1y	44.33	2200	500	113.56
5	EB_1y	44.33	2200	1000	115.33
1	EB_3m	15.44	2200	1000	62.61
2	EB_3m	15.44	2200	1000	59.12
3	EB_3m	15.44	2200	300	63.91
4	EB_3m	15.44	2200	600	59.13
5	EB_3m	15.44	2200	800	58.63
1	FB_1y	1.25	2200	300	2.03
2	FB_1y	1.25	2200	1000	1.99
3	FB_1y	1.25	2200	400	1.85
4	FB_1y	1.25	2200	300	1.97
5	FB_1y	1.25	2200	300	1.97
1	FB_3m	1.84	2200	400	4.08
2	FB_3m	1.84	2200	300	4.05
3	FB_3m	1.84	2200	600	4.21
4	FB_3m	1.84	2200	300	3.96
5	FB_3m	1.84	2200	600	4.06

Figure 4.5 is an example figure that the log of true vs predicted values for the FB ratio at 1 year given by random forest models run on each imputed dataset that required 90% complete variables. Log scales used to improve visibility of data across a wide range of values. The figures for the other outcomes can be found in Appendix A. In these figures, points that are closer to the best fit line indicate predicted values that are closer to their true value. From these plots, we see that there is a range of values for which the model appears to perform better when the random forest models predict the FB ratio at 1 year. In all panels, this cluster appears just above the 0 value on the x axis. These values appear to be around the document average FB ratio for infants which is 0.4 (Mariat et al., 2009). This may indicate that the model does not perform well with

values at the extremes, which is a significant drawback for predicting microbial dysbiosis. However, this may also be a case where overfitting in the model has been avoided.

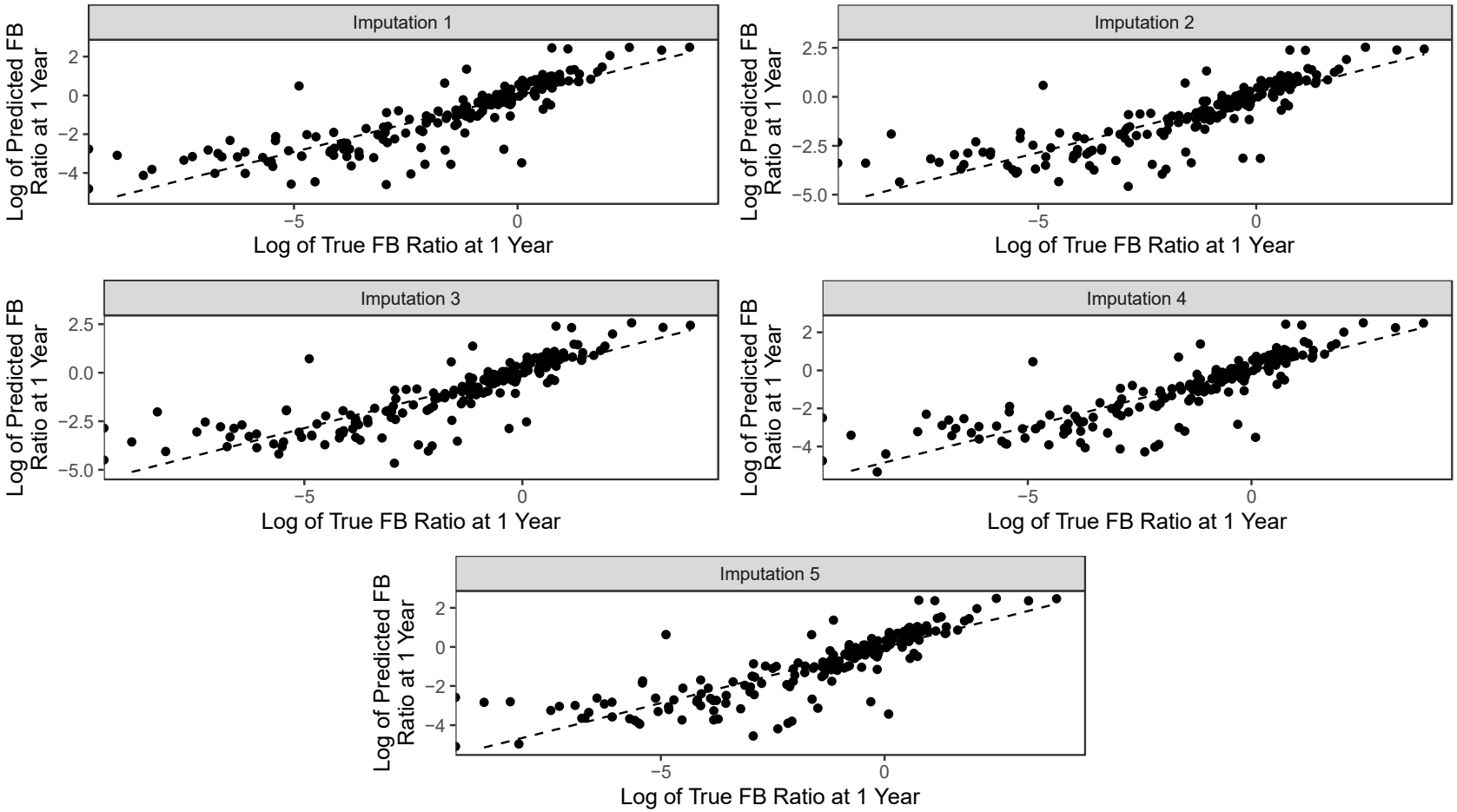

Figure 4.5. Example of Log of Predicted Firmicutes-to-Bacteroidetes ratios vs Log of True Firmicutes-to-Bacteroidetes ratois at 1 year using the Random Forest model with variables at least 90% complete. Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibility of data across a wide range of values.

For all five imputed versions of the dataset, the variable importance score, which describes how influential each variable is for predicting a given ratio, was pooled, giving four different lists of variable importance. The overall composition of the top variables for the Random Forest models can be seen in Figure 4.6. Microbiome data (not including ratios with the outcome taxa, as described in section 4.3.3) accounts for many high scoring variables. The inclusion of this microbiome data may pose some limitations as to the interpretability of these variable importance scores. It is possible that the inclusions of these variables, which would be highly correlated given the relational composition of a gut microbial community, could influence the model and skew the variable importance scores. With this in mind, we can also note that home environment, healthcare and medications comprise notable portions of the high scoring variables.

As this project focuses on medications, subsets of these lists can be found in Table 4.4, Table 4.5, Table 4.6, and Table 4.7. The entire list, consisting of all information domains and raw importance scores, is available in the supplemental materials described in Appendix C The subsets show the top 25 medication variables for the outcome. For each table, the variable is shown, ranking in the overall list (with all domains of information included) and then ranking within the medication domain. The rankings for the 85% complete information and 90% complete information are shown. For both the 85 and 90% cases, the same variables are in the top 25; however, their rankings are not identical.

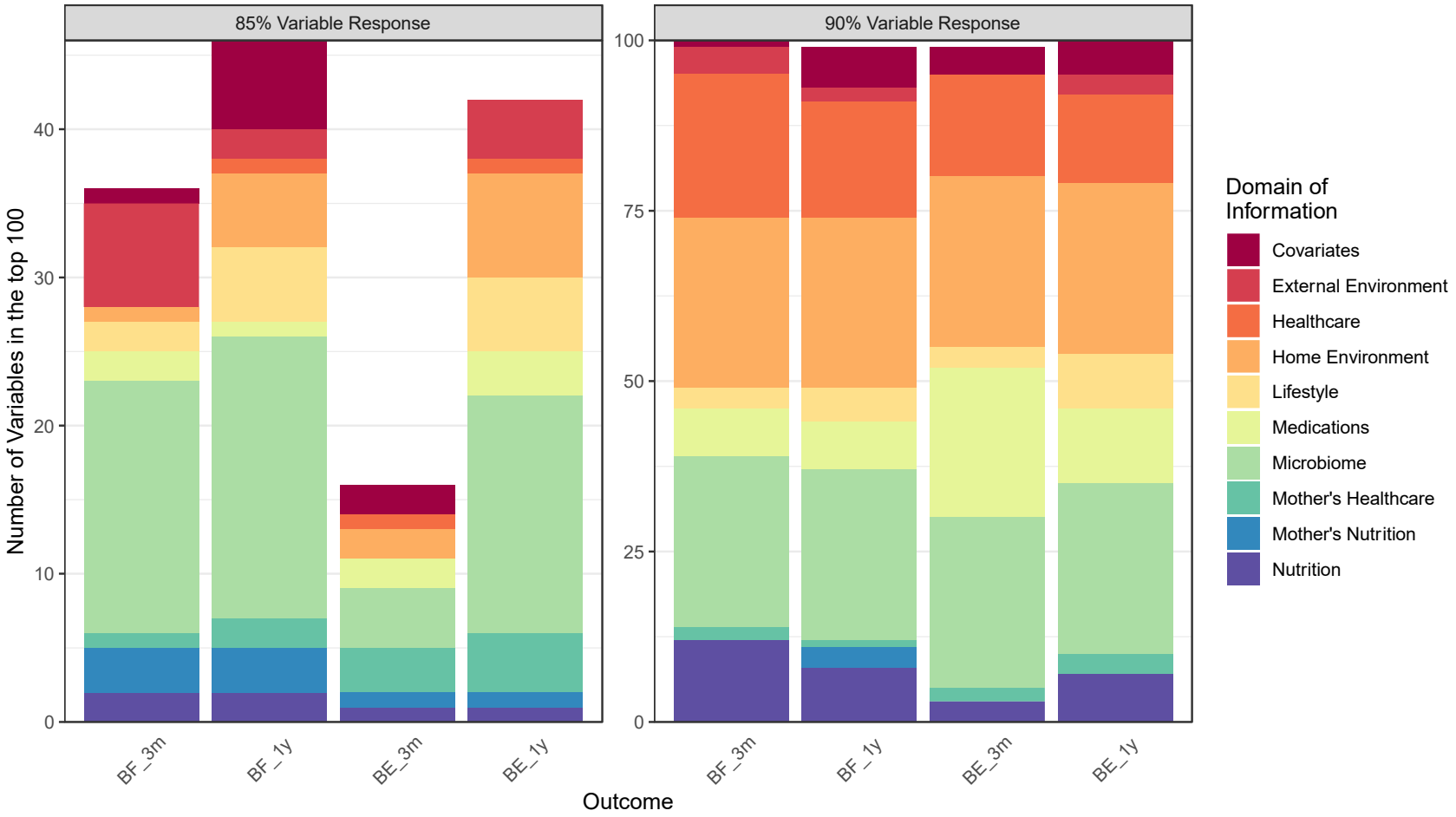

Figure 4.6. **Domain representation in the 100 highest scoring variables for the Random Forest models The panel on the left is for the models run requiring 85% variable response, the panel on the right is for 90% variabel response. The x axis shows the 4 different outcomes predicted, and the y axis shows the variable counts for each domain of information. Colours of the bars correspond to the domains of information. Unless otherwise stated, the domain of information pertains to the infant.**

Table 4.4. Top 25 Medication Variables from Random Forest Models for the outcome the Enterobacteriaceae-to-Bacteroidaceae ratio at 3 months. Results shown for the 90% and 85% variable response models. Within those, the overall rank when considering all variable domains in shown, along with the ranking within the medication domain itself. Table Organized by medication rank for 90%.

	90 Percent Variable Response		85 Percent Variable Response	
Medication	**Overall Rank**	**Medication Rank**	**Overall Rank**	**Medication Rank**
number of antifungals at 6 months	75	1	310	6
rash medications at 1y	81	2	161	5
respiratory therapy at 6 months	88	3	318	7
Estrogenic substances at 6 months	89	4	319	8
medications for third cold at 1 year	90	5	320	9
clobestasone butyrate at 1 year	91	6	321	10
desloratadine at 1 year	95	7	322	11
drugs at 6 months	101	8	129	2
Salbutamol at 1 year	106	9	330	12
Cefprozil at 1 year	122	10	341	13
took ibuprofen at 6 months	129	11	348	14
Homeopathy Therapy at 6 months	130	12	349	15
Diphenhydramine hydrochlorideat 1 year	132	13	351	16
Cetalkonium chloride at 1 year	136	14	148	4
Anitfungal drug at 1 year	135	15	353	17
Diphenhydramine hydrochloride at 6 months	140	16	130	3
Diflucortolone valerate at 1 year	138	17	355	18
Omeprazole at 1 year	139	18	356	19
respiratory therapy at 1 year	141	19	357	20
anti asthmatic drug at 1 year	142	20	358	21
lactulose at 1 year	143	21	359	22
fluticasone at 6 months	144	22	128	1
cefuroxime axetil at 1 year	146	23	361	23
steroid at 6 months	147	24	362	24
azithromycin at 1 year	148	25	363	25

Table 4.5. Top 25 Medication Variables from Random Forest Models for the outcome the Enterobacteriaceae-to-Bacteroidaceae ratio at 1 year. Results shown for the 90% and 85% variable response models. Within those, the overall rank when considering all variable domains in shown, along with the ranking within the medication domain itself. Table Organized by medication rank for 90%.

	90 Percent Variable Response		85 Percent Variable Response	
Medication	**Overall Rank**	**Medication Rank**	**Overall Rank**	**Medication Rank**
cough medication at 1 year	27	1	277	3
epinephrine at 1 year	73	2	309	4
amoxicillin at 1 year	82	3	315	5
prednisone at 1 year	88	4	318	6
number of acetaminophen at 3 months	89	5	319	7
took antiallergic medication at 6 months	95	6	323	8
glycerin at 1 year	98	7	324	9
vitamin D at 1 year	111	8	336	10
Diphenhydramine hydrochloride at 6 months	112	9	337	11
antibacterial drug at 1 year	117	10	342	12
number of analgesic	123	11	140	2
took ibuprofen at 6 months	124	12	346	13
fluticasone propionate at 1 year	127	13	348	14
fluticasone at 6 months	130	14	86	1
clavulanic acid at 1 year	133	15	351	15
number of antihelminthic at 6 months	135	16	353	16
fusidic acid at 6 months	136	17	354	17
erythromycin at 1 year	137	18	355	18
ambroxol hydrochloride at 1 year	138	19	356	19
cefaclor at 1 year	139	20	357	20
mupirocin at 1 year	140	21	358	21
phenobarbital at 6 months	141	22	359	22
furosemide at 6 months	142	23	360	23
fluocinolone acetonide at 1 year	143	24	361	24
moxifloxacin hydrochloride at 6 months	144	25	362	25

Table 4.6. Top 25 Medication Variables from Random Forest Models for the outcome Firmicutes-to-Bacteroidetes ratio at 3 months. Results shown for the 90% and 85% variable response models. Within those, the overall rank when considering all variable domains in shown, along with the ranking within the medication domain itself. Table Organized by medication rank for 90%.

	90 Percent Variable Response		85 Percent Variable Response	
Medication	**Overall Rank**	**Medication Rank**	**Overall Rank**	**Medication Rank**
number of ibuprofen at 1 year	29	1	272	4
benzocaine at 6 months	80	2	300	5
number of antifungal at 3 months	81	3	301	6
vitamin D at 1 year	82	4	111	3
Iron Supplement at 1 year	91	5	308	7
number of antibacterial at 6 months	92	6	309	8
took antiallergic medication at 3 months	99	7	315	9
number of antiallergic at 3 months	104	8	318	10
number of antipyretic	105	9	319	11
albuterol at 1 year	110	10	321	12
medications for first cold at 1 year	111	11	322	13
other vitamins and supplements at 1 year	117	12	102	2
took antiallergic medication at 6 months	120	13	329	14
number of anti asthmatic medications	121	14	330	15
fluticasone propionate at 1 year	122	15	331	16
moxifloxacin hydrochloride at 6 months	123	16	332	17
fluocinolone acetonide at 1 year	139	17	348	18
azithromycin at 6 months	146	18	351	19
number of antiallergic at 1 year	155	19	359	20
took ibuprofen at 6 months	159	20	360	21
Vitamin D Supplement at 1 year	162	21	92	1
number of antifungal at 1 year	164	22	362	22
antibacterial drug at 1 year	165	23	363	23
cefixime at 6 months	166	24	364	24
cefprozil at 6 months	169	25	367	25

Table 4.7. Top 25 Medication Variables from Random Forest Models for the outcome Firmicutes-to-Bacteroidetes ratio at 1 year. Results shown for the 90% and 85% variable response models. Within those, the overall rank when considering all variable domains in shown, along with the ranking within the medication domain itself. . Table Organized by medication rank for 90%.

	90 Percent Variable Response		85 Percent Variable Response	
Medication	**Overall Rank**	**Medication Rank**	**Overall Rank**	**Medication Rank**
acetaminophen at 1 year	31	1	292	7
oral thrush medications at 1 year	34	2	295	8
albuterol at 1 year	60	3	316	9
ear infection medications at 1 year	62	4	119	3
took antifungal at 3 months	71	5	322	10
took ibuprofen at 6 months	72	6	323	11
fluconazole at 6 months	75	7	326	12
benzocaine at 1 year	88	8	338	13
laxative at 6 months	92	9	341	14
clotrimazole at 1 year	98	10	79	1
cephalexin at 1 year	99	11	97	2
hydrocortisone at 6 months	97	12	344	15
number of antiallergic at 3 months	119	13	360	16
fluticasone at 1 year	120	14	361	17
beclometasome dipropionate at 6 months	127	15	365	18
ratitidine at 6 months	129	16	151	6
took ibuprofen at 6 months	133	17	142	5
homeopathy therapy at 1 year	134	18	369	19
steroid at 1 year	136	19	371	20
number of antibacterial at 3 months	138	20	372	21
nystatin at 6 months	140	21	374	22
number of antipyretic	141	22	375	23
belladonna at 1 year	145	23	126	4
vitamins and supplements	150	24	381	24
Vitamin D at 6 months	153	25	383	25

4.4.2. Gradient Boosting Model Analysis

The Gradient Boosting models using the datasets requiring 85% and 90% response rates before imputation were used to determine variables of importance. Both models were tuned using the learning rate, interaction depth, minimum nodes (ntree) and the bag fraction. The learning rate is the weight applied to the newly predicted residuals at each iteration of the gradient boost descent. The interaction depth is the number of splits on a node. The bag fraction determines the percentage of the observations used for each tree. To evaluate the performance of the models, Gradient Boosting also uses the RMSE as described above. However, when looking at the values during the procedure, this is referred to as the cv.error.

Table 4.8and Table 4.9 show the summary statistics for the 85% and 95% models, respectively. For all outputs, the cv.error values are approximately equal to the standard deviation of the outcome variable in question, and all cv values are larger than the mean of the outcome variable. These results would have the same implications as described with the Random Forest models in Section 4.2.1.

Table 4.8. Summary Statistics for Gradient Boosting Machine for 85 percent complete data. 20 models are shown (5 imputations of each 4 outputs), the means of the outcome variables are shown, along with the model parameters for the number of trees (ntree), learning rate, interaction depth and bag fraction. The cv.error is shown to describe overall quality of the model.

Imputation	Outcome	Mean	Ntree	Learning Rate	Interaction Depth	Bag Fraction	cv. Error
1	EB_1y	44.33	43	0.3	5	0.65	114.65
2	EB_1y	44.33	148	0.1	3	0.8	111.48
3	EB_1y	44.33	49	0.3	3	1	109.96
4	EB_1y	44.33	180	0.3	1	0.8	113.26
5	EB_1y	44.33	12	0.3	5	1	112.14
1	EB_3m	15.44	8	0.1	3	0.8	58.02
2	EB_3m	15.44	18	0.3	1	0.8	58.36
3	EB_3m	15.44	6	0.1	3	0.8	57.67
4	EB_3m	15.44	18	0.3	1	0.8	58.05
5	EB_3m	15.44	29	0.3	5	1	58.63
1	FB_1y	1.25	102	0.1	3	0.8	2
2	FB_1y	1.25	74	0.1	3	0.65	1.99
3	FB_1y	1.25	27	0.3	1	0.65	2.1
4	FB_1y	1.25	7	0.3	5	0.65	2.07
5	FB_1y	1.25	9	0.3	5	0.65	1.98
1	FB_3m	1.84	33	0.3	5	0.65	4.42
2	FB_3m	1.84	158	0.3	3	0.8	4.24
3	FB_3m	1.84	688	0.1	5	0.8	4.28
4	FB_3m	1.84	41	0.3	5	0.65	4.36
5	FB_3m	1.84	180	0.3	5	0.8	4.2

Table 4.9. Summary Statistics for Gradient Boosting Machines for 90% complete data. 20 models are shown (5 imputations of each 4 outputs) along with the model parameters for the number of trees (ntree), learning rate, interaction depth and bag fraction. The cv.error is shown to describe overall quality of the model.

Imputation	Outcome	Mean	Ntree	Learning Rate	Interaction Depth	Bag Fraction	cv. Error
1	EB_1y	44.33	32	0.3	3	0.8	118.4
2	EB_1y	44.33	87	0.3	5	0.65	114.06
3	EB_1y	44.33	29	0.3	5	0.65	116.37
4	EB_1y	44.33	4993	0.01	5	1	112.79
5	EB_1y	44.33	460	0.1	3	0.8	113.29
1	EB_3m	15.44	51	0.3	5	1	58.51
2	EB_3m	15.44	6	0.1	3	0.8	57.83
3	EB_3m	15.44	24	0.3	1	1	57.96
4	EB_3m	15.44	6	0.1	3	0.8	58.26
5	EB_3m	15.44	24	0.3	1	1	58.24
1	FB_1y	1.25	4	0.3	3	0.65	2.06
2	FB_1y	1.25	9	0.3	3	0.65	2.08
3	FB_1y	1.25	4	0.3	3	0.65	2.09
4	FB_1y	1.25	6	0.3	3	0.65	2.12
5	FB_1y	1.25	4097	0.01	3	1	1.93
1	FB_3m	1.84	73	0.3	5	0.65	4.41
2	FB_3m	1.84	90	0.3	3	0.65	4.24
3	FB_3m	1.84	124	0.3	3	0.8	4.29
4	FB_3m	1.84	476	0.1	5	0.8	4.15
5	FB_3m	1.84	361	0.1	5	0.8	4.17

Figure 4.7 is an example figure of the log of true vs predicted values for the FB ratio at 3 months given by gradient boosting models run on each imputed dataset that required 90% complete variables. Log scales used to improve visibility of data across a wide range of values. The figures for the other outcomes can be found in Appendix A. Unlike the Random Forest plots in Figure 4.5, there does not appear to be a range of values for which the model appears to perform better. Furthermore, the plots in Figure 4.5 appear more uniform when comparing each imputation version. In Figure 4.7, Imputation 1 and Imputation 2 have similar shapes, while the other 3 versions share a different spread. These differences may be due to the fact that Random Forest has been

known to handle noise introduced by large datasets better than Gradient Boost (Hastie et al., 2009).

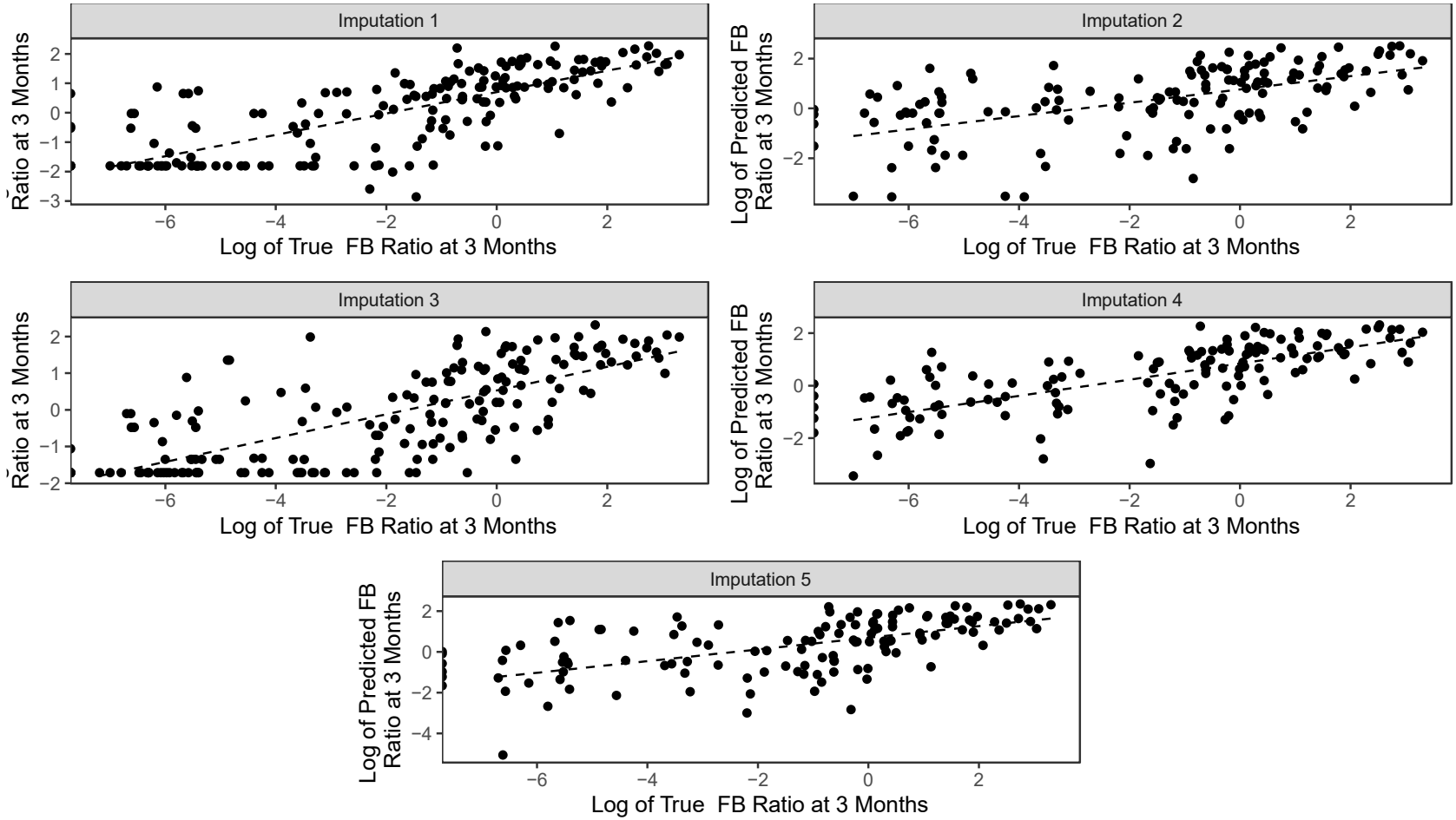

Figure 4.7. Example of Log of Predicted Firmicutes-to-Bacteroidetes ratios vs Log of True Firmicutes-to-Bacteroidetes ratois at 3 months using the Gradient Boosting model with variables at least 85% complete. Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibility of data across a wide range of values.

For all five imputed versions of the dataset, the variable importance for each outcome in Gradient Boosting machines was pooled, giving four lists of variable importance. The overall composition of the top variables for the Gradient Boosting models can be seen in Figure 4.8. Microbiome data (not including ratios with the outcome taxa, as described in section 4.3.3) accounts for many high scoring variables. The inclusion of this microbiome data may pose some limitations as to the interpretability of these variable importance scores. It is possible that the inclusions of these variables, which would be highly correlated given the relational composition of a gut microbial community, could influence the model and skew the variable importance scores. With this in mind, we can also note that home environment, external environment and mother's nutrition make up for notable portions of the high scoring variables.

As this project focuses on medications, subsets of these lists can be found in Table 4.10, Table 4.11, Table 4.12 and Table 4.13. The entire list, consisting of all information domains and raw importance scores, is available in the supplemental materials described in Appendix C. The subsets show the top 25 medication variables for the outcome. For each table, the variable is shown, ranking in the overall list (with all domains of information included) and then ranking within the medication domain. The rankings for the 85% complete information and 90% complete information are shown. While the components of the 85% and 90% lists are consistent, which was also the case with the variable importance lists from the Random Forest models, the rankings of the variables within the medication domain are much more consistent between groups when using the Gradient Boosting algorithm.

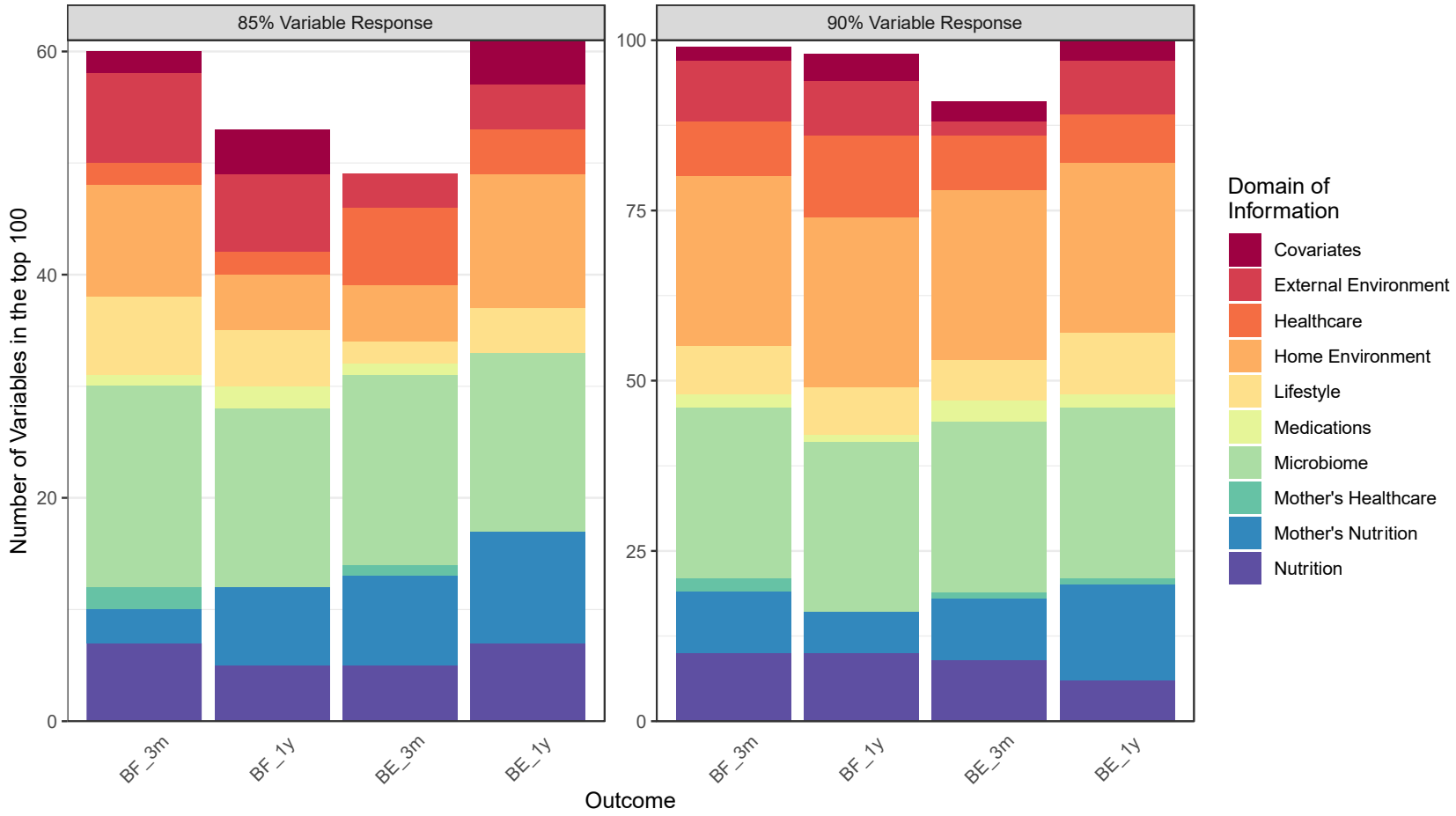

Figure 4.8. Domain representation in the 100 highest scoring variables for the Gradient Boosting models The panel on the left is for the models run requiring 85% variable response, the panel on the right is for 90% variabel response. The x axis shows the 4 different outcomes predicted, and the y axis shows the variable counts for each domain of information. Colours of the bars correspond to the domains of information. Unless otherwise stated, the domain of information pertains to the infant.

Table 4.10. Top 25 Medication Variables from Gradient Boosting Machine for the outcome the Enterobacteriaceae-to-Bacteroidaceae ratio at 3 months. Results shown for the 90% and 85% variable response models. Within those, the overall rank when considering all variable domains in shown, along with the ranking within the medication domain itself. Table Organized by medication rank for 90%.

	90 Percent Variable Response		85 Percent Variable Response	
Medication	**Overall Rank**	**Medication Rank**	**Overall Rank**	**Medication Rank**
hydrocortisone valerate at 1 year	61	1	98	2
number of antifungal at 3 months	69	2	67	1
nystatin at 1 year	75	3	106	3
took acetaminophen at 3 months	127	4	157	4
number of acetaminophen at 3 months	128	5	158	5
took ibuprofen at 3 months	129	6	159	6
took antibacterial at 3 months	130	7	160	7
number of antibacterial at 3 months	131	8	161	8
took antifungal at 3 months	132	9	162	9
took antihelminthic at 3 months	133	10	163	10
number of antihelminthic at 3 months	134	11	164	11
took antiallergic at 3 months	135	12	165	12
number of antiallergic at 3 months	136	13	166	13
took acetaminophen at 6 months	137	14	167	14
number of acetaminophen at 6 months	138	15	168	15
took ibuprofen at 6 months	139	16	169	16
number of ibuprofen at 6 months	140	17	170	17
took antibacterial at 6 months	141	18	171	18
number of antibacterial at 6 months	142	19	172	19
took antifungal at 6 months	143	20	173	20
number of antifungals at 6 months	144	21	174	21
took antiallergic at 6 months	145	22	175	22
number of antiallergic at 6 months	146	23	176	23
took acetaminophen at 1 year	147	24	177	24
number of acetaminophen at 1 year	148	25	178	25

Table 4.11. Top 25 Medication Variables from Gradient Boosting Machine for the outcome the Enterobacteriaceae-to-Bacteroidaceae ratio at 1 Year. Results shown for the 90% and 85% variable response models. Within those, the overall rank when considering all variable domains in shown, along with the ranking within the medication domain itself. Table Organized by medication rank for 90%.

	90 Percent Variable Response		85 Percent Variable Response	
Medication	**Overall Rank**	**Medication Rank**	**Overall Rank**	**Medication Rank**
benzocaine at 1 year	83	1	156	2
number of acetaminophen at 6 months	98	2	162	3
took antbacterial at 3 months	107	3	168	4
number of ibuprofen at 6 months	170	4	202	5
zinc oxide at 1 year	171	5	203	6
took acetaminophen at 6 months	176	6	207	7
number of antifungals at 6 months	179	7	209	8
took acetaminophen at 1 year	182	8	116	1
number of antibacterial at 1 year	183	9	210	9
took antibacterial at 1 year	184	10	211	10
acetaminophen at 1 year	185	11	212	11
vitamins and supplements at 1 year	188	12	214	12
sulfamethoxazole at 1 year	189	13	215	13
hydrocortisone at 6 months	190	14	216	14
herbal remedy supplement at 1 year	191	15	217	15
sodium chloride at 1 year	192	16	218	16
herbal remedy supplement at 6 months	193	17	219	17
rash medications at 1 year	196	18	222	18
number of antifungal at 1 year	198	19	224	19
medications for second cold at 1 year	200	20	226	20
other vitamins and supplements at 1 year	201	21	227	21
albuterol at 6 months	203	22	229	22
clarithromycin at 1 year	204	23	230	23
drug cream at 6 months	205	24	231	24
amoxicillin at 6 months	206	25	232	25

Table 4.12. Top 25 Medication Variables from Gradient Boosting Machine for the outcome Firmicutes-to-Bacteroidetes ratio at 3 months. Results shown for the 90% and 85% variable response models. Within those, the overall rank when considering all variable domains in shown, along with the ranking within the medication domain itself. Table Organized by medication rank for 90%.

	90 Percent Variable Response		85 Percent Variable Response	
Medication	**Overall Rank**	**Medication Rank**	**Overall Rank**	**Medication Rank**
hydrocortisone valerate at 1 year	61	1	98	2
number of antifungal at 3 months	69	2	67	1
nystatin at 1 year	75	3	106	3
took acetaminophen at 3 months	127	4	157	4
number of acetaminophen at 3 months	128	5	158	5
took ibuprofen at 3 months	129	6	159	6
took antbacterial at 3 months	130	7	160	7
number of antibacterial at 3 months	131	8	161	8
took antifungal at 3 months	132	9	162	9
took antihelminthic at 3 months	133	10	163	10
number of antihelminthic at 3 months	134	11	164	11
took antiallergic at 3 months	135	12	165	12
number of antiallergic at 3 months	136	13	166	13
took acetaminophen at 6 months	137	14	167	14
number of acetaminophen at 6 months	138	15	168	15
took ibuprofen at 6 months	139	16	169	16
number of ibuprofen at 6 months	140	17	170	17
took antibacterial at 6 months	141	18	171	18
number of antibacterial at 6 months	142	19	172	19
took antifungal at 6 months	143	20	173	20
number of antifungals at 6 months	144	21	174	21
took antiallergic at 6 months	145	22	175	22
number of antiallergic at 6 months	146	23	176	23
took acetaminophen at 1 year	147	24	177	24
number of acetaminophen at 1 year	148	25	178	25

Table 4.13. Top 25 Medication Variables from Gradient Boosting Machine for the outcome Firmicutes-to-Bacteroidetes ratio at 1 Year. Results shown for the 90% and 85% variable response models. Within those, the overall rank when considering all variable domains in shown, along with the ranking within the medication domain itself. Table Organized by medication rank for 90%.

	90 Percent Variable Response		85 Percent Variable Response	
Medication	**Overall Rank**	**Medication Rank**	**Overall Rank**	**Medication Rank**
drug solution at 6 months	43	1	43	2
vitamins and supplements at 1 year	122	2	140	3
number of antifungal at 1 year	132	3	149	4
number of acetaminophen at 3 months	137	4	154	5
took antifungal at 1 year	139	5	156	6
vitamin D at 1 year	140	6	157	7
number of anti inflammatory	150	7	167	8
took acetaminophen at 3 months	153	8	170	9
took antibacterial at 1 year	156	9	172	10
number of antibacterial at 1 year	158	10	174	11
number of analgesic	217	11	20	1
took ibuprofen at 3 months	190	12	205	12
took antibacterial at 3 months	191	13	206	13
number of antibacterial at 3 months	192	14	207	14
took antifungal at 3 months	193	15	208	15
number of antifungal at 3 months	194	16	209	16
took antihelminthic at 3 months	195	17	210	17
number of antihelminthic at 3 months	196	18	211	18
took antiallergic at 3 months	197	19	212	19
number of antiallergic at 3 months	198	20	213	20
took acetaminophen at 6 months	199	21	214	21
number of acetaminophen at 6 months	200	22	215	22
took ibuprofen at 6 months	201	23	216	23
number of ibuprofen at 6 months	202	24	217	24
took antibacterial at 6 months	203	25	218	25

4.5. Evaluation of Variables of Importance

There are several caveats that must be considered for these results. The first being that while microbial ratios including the outcome taxa of interest were removed

from the dataset, there were still other microbial ratios included in the dataset. It is possible that the inclusions of these variables, which would be highly correlated given the relational composition of a gut microbial community, could have influenced the model and skewed the variable importance scores such that the highly ranked features are not interpretable. In future research, it would be good to compare results when removing all microbial data from the dataset.

With that caveat in mind, the variables relating to medication did not score as high as variables relating to the lifestyle and home environment in the machine learning models, suggesting such latter variables may play a. more significant role. Scores for all variables included in the machine learning models can be found in Appendix C. Using a large, diverse dataset for this project is a benefit. Without the incorporation of diverse data domains, the fact that other domains of information may play a more significant role would not have been apparent. However, the goal of this project was not to create a perfect predictive model but rather to narrow down some of the most important potential factors associated with microbial dysbiosis, which then may warrant further study in more detail.

The outcome variables for the machine learning models are for the 3 month and 1-year timepoints. Looking at the variable importance scores, one sees that there are medication use variables from birth, 3 months, 6 months and 1 year. The variables used for the predictions of the 3-month ratios were not limited to the variables collected before that timepoint, variables from all available timepoints were used. The situation was the same for the 1 year timepoint. This was done to avoid mistakenly assuming the directionality of any of the relationships, as it is still unclear whether it is illness or medication use impacting the microbiota or vice versa. Such directionality warrants further investigation through use of other methods that the Brinkman lab aims to explore in the future (See Future Directions).

However, since this project focuses on the potential associations with medication, I examined the high-scoring variables within the medication domain. For all four microbial dysbiosis outcomes, vitamin D, antimicrobial, and analgesic use came up as important variables in the machine learning models. In this case “important” means that the inclusion of the variables would lower the RMSE of the machine learning

models, indicating better model performance when the variables in question were included to predict microbial dysbiosis ratios.

Antimicrobials in the first year of life were not unexpected. It is a logical assumption that medications created to remove microbes would impact the gut microbiome. Looking at Tables 4.4 to 4.7 and Tables 4.10 to 4.13, antimicrobials come up between 4 to11 times out of 25 variables. However, it is notable that there were a large number of antimicrobial-related variables included in the model. This includes both variables describing specific antimicrobials (ex. Azithromycin use) and variables describing the use of the overarching medication class (ex. Antibacterial use at 6 months).

Analgesics were also notable, but not a surprise, as they are a very common medication class in the first year of life. Variables relating to analgesic use occur between 1 and 11 times as important variables shown in Tables 4.4 to 4.7 and Tables 4.10 to 4.13. Note that, like antimicrobials, specific drugs were incorporated, such as acetaminophen, and the overarching drug class of analgesics. Both specific analgesics and the entire drug class came up in the variables of importance.

Vitamin D showed up fewer times, only between 1-3 times. However, compared to the drug classes of antimicrobials and analgesics, there were not as many vitamin D-related variables included in the model. Vitamin D impacts in the context of breastfeeding have been previously investigated in the CHILD Cohort Study, and an association with microbial shifts was identified for the taxa Lachnospiraceae and Rikenellaceae (Drall et al., 2020).

When looking at the high scoring medication variables, one would see that these medications are very common medications for infants to be taking. For example, acetaminophen and vitamin D. The CHILD Cohort study has more variables relating to these medications, simply due to their abundance in infant populations. Mathematically, this could result in a sampling bias given that less common medications, such as salbutamol, do not have as many variables present in the dataset that describe their use. However, given the prevalence of antimicrobial, analgesic and vitamin D use in the infant population, these medications are still worth follow up analyses to investigate potential associations with the gut microbiome. A more detailed investigation into the three

specified medication classes/types (antimicrobials, analgesics, and vitamin D) is described in Chapter 5 of this thesis.

4.6. Concluding Remarks

From a technical perspective, the RMSE values of both the Random Forest and Gradient Boosting models indicate that the model as it currently stands should not be used as a predictive model. However, that was not the goal of this project, the models constructed here are meant to act as a baseline for identifying variables of potential importance that can be examined further. Future projects within the CHILD Cohort Study can use these values as baseline comparisons to evaluate model performance improvement as new information about weight and influence is added to the models. Regardless of the RMSE scores, these models could produce variable importance scores that could guide a more in-depth investigation into the relationship between medications in the first year of life and microbial dysbiosis measurements at three months and one year. It should be emphasized though that the inclusion of the microbiome data in these models may impact the interpretability of these results. The inclusion of microbiome data should therefore be avoided in future models to evaluate the impact of this inclusion.

Antimicrobials, analgesics, and vitamin D were identified as medications with strong associations to the microbial dysbiosis outcomes based on the variable importance scores from both models. Evaluation of these models provides a preliminary prioritization strategy for medication and microbial dysbiosis associations requiring further investigation and the investigation of more causative factors associated with microbial dysbiosis. When evaluating the lists of variable importance, the results predicting the FB ratio are more consistent than those predicting the EB ratio. This may indicate that the FB ratio may have stronger associations with lifestyle events than the EB ratio. Going forward, the investigation will only focus on contextualizing the results within the FB ratio.

When looking at the results of these models, what isn’t present is as important as what is present. From these results, we see no clear associations medications that

drastically shift an infant's microbiome toward FB or EB ratios that have previously been found to have negative associations with health in adult cohorts. This is reassuring as these results demonstrate no strong and obvious harmful associations between early life FB and EB ratios and medications.

Chapter 5.

Exploration of the Association of Antimicrobial Use, Analgesic Use, And Vitamin D, With Microbial Dysbiosis.

This chapter aims to further examine select high-scoring variables from the machine learning models outlined in Chapter 4, hypothesizing that the approach will identify specific medication use patterns associated with infant microbial dysbiosis. The methods used are summarized in Section 5.3. For all medications, I look at the differential abundance in the gut microbiota at the family and phylum levels. I also use the reasons for medication use dataset (Chapter 2) to compare medication use patterns. Finally, I use GlobeCorr (Chapter 3) and more traditional statistical techniques to explore correlations that a given medication use has with other information domains present in the dataset. Based on the initial analysis, I focused on analyses for three medications (Section 5.4): antimicrobials, analgesics, and vitamin D and their potential associations with microbial dysbiosis and differential abundance of gut microbial taxa.

I completed all work outlined in this chapter.

5.1. Abstract

Characterizing patterns in medication use and microbial dysbiosis requires an in-depth investigation of differential taxonomic abundance, medication habits and lifestyle correlations. From the results presented in Chapter 4, I selected three medication groups for further investigation based on their ability to predict microbial dysbiosis ratios in machine learning models antimicrobials, analgesics, and vitamin D. For each medication group, differential microbial abundance was investigated using ANCOM-BC. Medication habits were analyzed using the curated dataset of reasons for medication use from the CHILD Cohort Study (Chapter 2). Finally, pairwise correlations were calculated using a diverse dataset of 1546 variables for 564 participants from the CHILD cohort study to investigate broader lifestyle associations with medication use. From this analysis, I identified associations with taxa abundance such as Lachnospiraceae and Rikenellaceae with the use of antimicrobials and vitamin D. The analysis of medication

habits and other correlated variables suggests that the use of any of these specified medications is a proxy for a lifestyle that tends to be high in medication use, which may make one more prone to microbial dysbiosis. Overall, these analyses will help direct research into more microbial functional hypotheses and provides a base for further investigations of the causative factors associated with microbial dysbiosis.

5.2. Introduction

A healthy microbiota is generally high in species diversity, maintains homeostasis and is compatible with the host's diet, allowing an individual to thrive in their unique environment (McBurney et al., 2019.; Yatsunenko et al., 2012). While the composition and abundance of species vary from person to person and can change over the host's life, overall functionality remains relatively consistent (Tian et al., 2020). Carbohydrate metabolism is a well-explored and characterized functionality found in gut microbiota phyla such as Actinobacteria and Bacteroidetes (Macfarlane & Englyst, 1986; Ottman et al., 2012b). There are many carbohydrates ingested by humans that we do not have the enzymes to properly degrade ourselves. The fermentation processes performed by bacteria on these undigestible carbohydrates produce short-chain fatty acids (SCFA) that account for approximately 10% of the human caloric intake (den Besten et al., 2013).

While an infant's microbial colonizers change dynamically over the first few months, their microbiota is still composed of the same major taxa in different proportions until they conform to a more adult diet (Walker, 2013). The major bacterial phyla in the infant microbiome are Firmicutes, Actinobacteria and Bacteroidetes. For healthy gut microbiotas, Actinobacteria and Firmicutes play a major role in short-chain fatty acid production pre and post-weaning, respectively (Odamaki et al., 2016; Rinninella et al., 2019).

In contrast to a healthy gut microbiome, microbial dysbiosis is a state that is often defined as an imbalance or change in the gut microbiome that differs from expected development. In particular, microbial dysbiosis is an imbalance that inhibits an individual's ability to thrive in their environment, usually due to the manifestation of some type of disease (Messer & Chang, 2018). There are many ways researchers have tried to quantify and characterize microbial dysbiosis. One of those ways is using ratios of the

different taxon in relation to a specific disease state. A common measure of microbial dysbiosis at the phylum level is the Firmicutes-to-Bacteroidetes FB ratio (Ley et al., 2006).

A low FB ratio has been associated with irritable bowel syndrome, while a high FB ratio has been associated with obesity (Stojanov et al., 2020). It is difficult to define “low” and “high” with exact values, as it can vary greatly depending on diet, geography, and age. However, in a European cohort, a FB ratio below 10.9 would be considered below average (Mariat et al., 2009). In the case of inflammatory bowel disease, the outlined microbial shifts would indicate a decrease in butyrate-producing bacteria, which includes members of the Firmicutes phylum (Vaiserman et al., 2020). Butyrate is used as a fuel source for colonocytes, the epithelial cells found in the colon. Without their fuel source, colonocytes cannot properly consume oxygen, and this can lead to an unfavourable environment for the commensal bacteria found in the gut (Litvak et al., 2018). Without the proper commensal gut microbes, humans may be more susceptible to autoimmune disorders and other medication conditions (Wu and Wu., 2012) Furthermore, butyrate is required for proper gut barrier function and anti-inflammatory responses (Donohoe et al., 2011).

Regarding obesity, Firmicutes can digest many carbohydrates that humans do not have the appropriate enzymes for (den Besten et al., 2013). The bacteria ferment carbohydrates and produce SCFA. However, the literature suggests the obesity phenotype is due to excess production of SCFA, leading to increased energy storage on the part of the host (Khan et al., 2016).

Many established relationships between medication use and the microbiome have focused on adult cohorts. When considering the infant population, the most focus is on antibiotics and the infant gut microbiota. Previously, The CHILD Cohort Study found that increased antibiotic use in infancy was associated with the development of asthma, with the diversity of the gut microbiome being a potential modulator for this change. Mechanisitically, it is suggested that increased antibiotic usage decreases the presence of bacteria such as *Faecalibacterium prausnitzii,* which are important for the production of anti-inflammatory metabolites.(Patrick et al., 2020). The study by Patrick et al. is helping to address the need for more longitudinal studies to evaluate the impact of lifestyle on the development of the gut microbiota. The CHILD Cohort Study is well

placed to fill this knowledge gap because of its diverse dataset. Using this dataset has allowed for an exploration into the links between microbial dysbiosis and medication use and medication use and other lifestyle factors. Researchers can use these complex networks to identify relationships to help direct research towards more microbial functional hypotheses.

5.3. Methods

5.3.1. Differential Abundance

To evaluate differential microbial abundance for subjects that have vs have not taken a specific drug, ANCOM-BC was used. ANCOM-BC is a tool to detect differential abundance in microbiome samples where the groups in question are not of equal size. For example, when there are more participants who have taken azithromycin as opposed to those who haven't. ANCOM-BC is an improvement upon ANCOM-II. While both detect differential abundance, ANCOM-BC is based on a linear regression framework and can include confounders in the analysis, as well as output test statistics, p-values, corrected p-values and standard errors in addition to a true/false matrix for whether or not a taxon can be considered differentially abundant between 2 or more groups. Microbiome samples at family and phylum levels are suitable for the ANCOM-BC analysis.

When deciding on the variables that would be investigated for the ANCOM-BC analysis, any variables related to the medication in question were used. For example, if the question "number of analgesics" had been flagged from the machine learning analysis, that would not be the only variable included in the ANCOM-BC analysis for the analgesics, variables such as "acetaminophen taken" and "taken analgesics" would also be used. The total number of variables investigated for each ANCOM-BC analysis will be specified in the appropriate section. Multiple variables in each analysis made a p-value correction necessary to avoid the type I error inflation. For this, all analyses used a Bonferroni correction to correct the p-values. In all figures presented for the ANCOM-BC analysis, all taxa found to be significantly differentially abundant (corrected $p < 0.05$) are shown in the plots. If there is a taxonomic group absent from the plots at the family or phylum level, it was not found to be differentially abundant.

For each analysis, regardless of medication type, all models were corrected for the mode of delivery (vaginal vs cesarean), visit (3 months vs one year), and sex (male vs female). Default values were used for all other parameters in the model. In practice, integrating covariates into the regression model creates comparisons for the impact of the medication use for each level of the covariates. The use of regression accounts for the small groups that may arise when the data is stratified by multiple covariates.

5.3.2. Correlation Globes

The machine learning analysis used 1547 variables, as outlined in Chapter 4. As part of the follow-up analyses, I wanted to look at the relationships between these variables by looking at correlation patterns. To do this, I first calculated the pairwise correlations, using different methods as appropriate depending on the data. The methods used can be seen in Table 5.1.

Table 5.1. Correlation Methods are used according to the variable's data type found in the curated dataset.

	Continuous	**Ordinal**	**Binary**
Continuous	Spearman Rank	Tau B or Spearman Rank	Point Biserial
Ordinal	Tau B or Spearman Rank	Kendall's Coefficient	Point Biserial
Binary	Point Biserial	Point Biserial	Matthews Correlation Coefficient

It would be difficult to visualize all correlations from the all-to-all correlations calculated from the 1547 variables. GlobeCorr starts to lag, and its dynamic features suffer. With large files, images become difficult to interact with when there are over 1000 bands around the globe. Therefore, a random subset of 500 correlations was extracted and visualized using the GlobeCorr software to get an idea of overall trends. This was completed based on the assumption that a random subsample can capture prominent trends in the dataset as a whole. This correlation globe can be viewed in Figure 5.1. Many of the correlations in this subset are interdomain correlations. However, there is a notable crossover between the home environment, health domains, and medication and health domains. Correlations between medication use and healthcare are not surprising, as health issues are often treated with medication. Seeing correlations between

healthcare and home environment is also not unexpected as it has been well established that there can be significant correlations between home environment and health trajectories. Seeing results consistent with established trends is encouraging as it indicates that a robust cross-cohort comparison may be feasible due to baseline similarities.

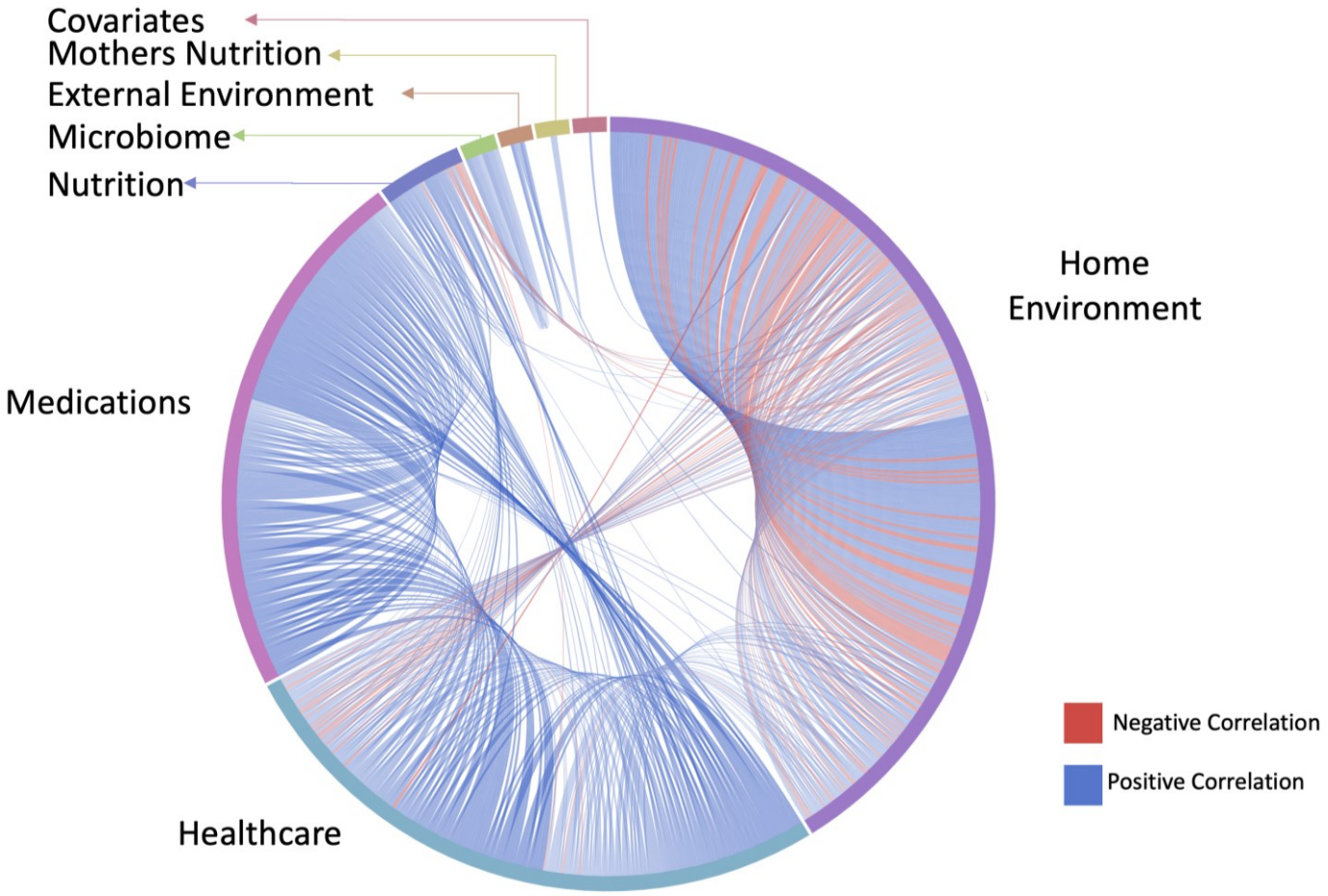

Figure 5.1. **Random Subsample of 500 correlations with a correlation coefficient with an absolute value above 0.4. from the all against all correlation analysis using the 1547 variables used in the machine learning analysis. Positive correlations are shown with a blue ribbon. Negative correlations are shown in red. For smaller domains, the label is offset with an arrow indicating the proper label.**

For further analyses into the medications of interest (antimicrobials, analgesics and vitamin D), GlobeCorr was used to investigate variables correlated with a particular medication. Correlations would be shown in the globes as long as one of the variables in the correlation was one of the variables relating to medication use, as described in Section 5.3.1. Once all the correlations were identified, a minimum threshold for the correlation coefficient was applied in R to reduce the dataset. GlobeCorr is also able to filter visualizations based on correlation thresholds. Still, in this context, it was easier to reduce the dataset beforehand to avoid any lags on the GlobeCorr website that have been associated with larger datasets. The chosen thresholds were specific to each medication and will be declared in the following sections.

Each variable relating to medication use was treated as its own domain. So, in the input csv, variable 1 would have the same value as var1_domain. This allowed for each medication variable to be represented by its own distinct arc around the circumference of the correlation globe. For other variables, they were organized into their domains as specified in Section 4.3.2. In the GlobeCorr portal, the domains were arranged so that all the medication variables were located on one part of the globe, and the non-medication variables were clustered on another. There is no functionality to add a "super-domain" in use cases like this, so another arc was added after the fact to indicate the "super-domain" of the medication variables. The domain labels were also adjusted outside of GlobeCorr to improve the clarity of the static visualizations.

5.3.3. Medication Use Patterns

I used the reasons for medication use data described in Chapter 2 to look at medication use patterns for the 564 subjects included in this analysis. Including the records for both children and mothers, we have reasons for medication use for 712 different medications from prenatal to 5 years. For this analysis, I was only interested in acetaminophen, ibuprofen, hydrocortisone, amoxicillin, and vitamin D use in the first year. These medications were chosen as they are the most common ones taken in the first year of life, as identified in the paper by Bédard et al.

For each of the three medications of interest from the machine learning models, which were antimicrobials, analgesics, and vitamin D, the subjects were divided into groups based on whether they had ever taken the medication. For example, one group

had taken vitamin D at any point in the first year, and one group had never taken vitamin D at any point in the first year. From there, the reasons for medication use for the top 5 medications identified by Bédard et al. will be summarized for the two groups. After calculating the numbers for the two distinct groups, I calculated the difference between the two by subtracting the number calculated from the non-medication group from the number calculated for the medicated group. This would mean that if a reason for medication use had a higher number, it was a more common reason for use in the group that did take the medication of interest (ex., Vitamin D). If the number was negative, it was more common in the group that did not take the medication of interest. If the difference were zero, then we would see no difference between the two groups as it pertains to that reason for medication use. A visual description of this process can be seen in Figure 5.2.

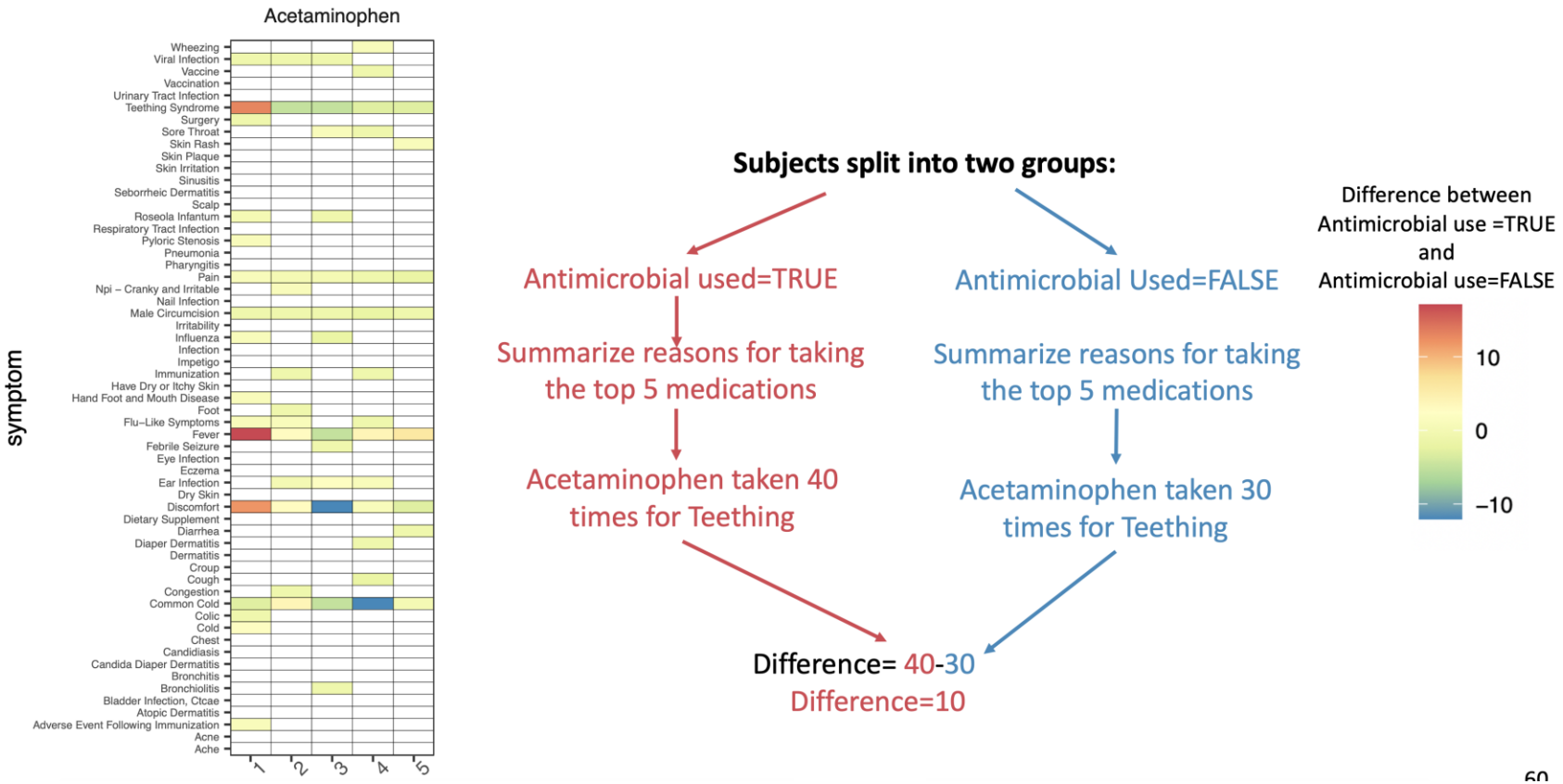

Figure 5.2. **Example of calculation used for examing reasons for medication use. Example shows subjects being split into two groups based on the presence or abense of anitmicorbial useage in the first year. A summary of the gorups usage of acetaminophen for teething is shown and difference calculated. Numbers calculated based on data available for reasons for medication use. Numbers on the bottom axis represent quintiles of bacterial ratios (either FB or EB) at the specified timepoint (3 months or 1 year). Quintile 1 would have subjects with the lowest values, with quintile 5 containing subjects with the highest ratios.**

5.4. Results and Discussion

5.4.1. Antimicrobials

Antimicrobial Use and Differential Abundance in Gut Microbiota Taxa

Differential abundance, compared between groups of CHILD Cohort Study participants that did versus did not take any antimicrobial, was calculated using ANCOM-BC as described in Section 5.3.1. To capture all reports of antimicrobial use in the first year of life for cohort participants, 41 variables were used. The full list of variables can be seen in Table B.9. Figures 5.3-5.5 show the results of the ANCOM-BC analysis. Figure 5.3 shows all taxa at the family level, organized by phylum that were found to be differentially abundant between groups that did and did not take the medication specified by the bar's colour. Declaration of differential abundance associated with the specified medication variable was only stated if the adjusted p-value was below 0.05 to account for multiple testing. However, for all variables shown for the antimicrobial comparisons, the n of the groups is barely within the threshold recommended by ANCOM-BC for comparison (ranging from n= 12-22). In Figure 5.3 the family level comparison is shown, and the phylum level is shown in Figure 5. 4 .Antibacterial use at one year, antibacterial use at three months, cephalosporin use at one year and crystal violet at three months were all identified as antimicrobials with associations to changes in the infant gut microbiome. Figure 5.3 is a summary figure, and its components are subsequently broken down into individual figures. Figure 5.5 shows differential abundance associated with crystal violet. Figure 5.6 shows all associations for the family Rikenellaceae. The other individual comparisons can be found as supplemental figures in the Appendix.

In all figures presented for the ANCOM-BC analysis, all taxa found to be significantly differentially abundant (corrected $p< 0.05$) are shown in the plots. If there is a taxonomic group absent from the plots at the family or phylum level, it was not found to be differentially abundant.

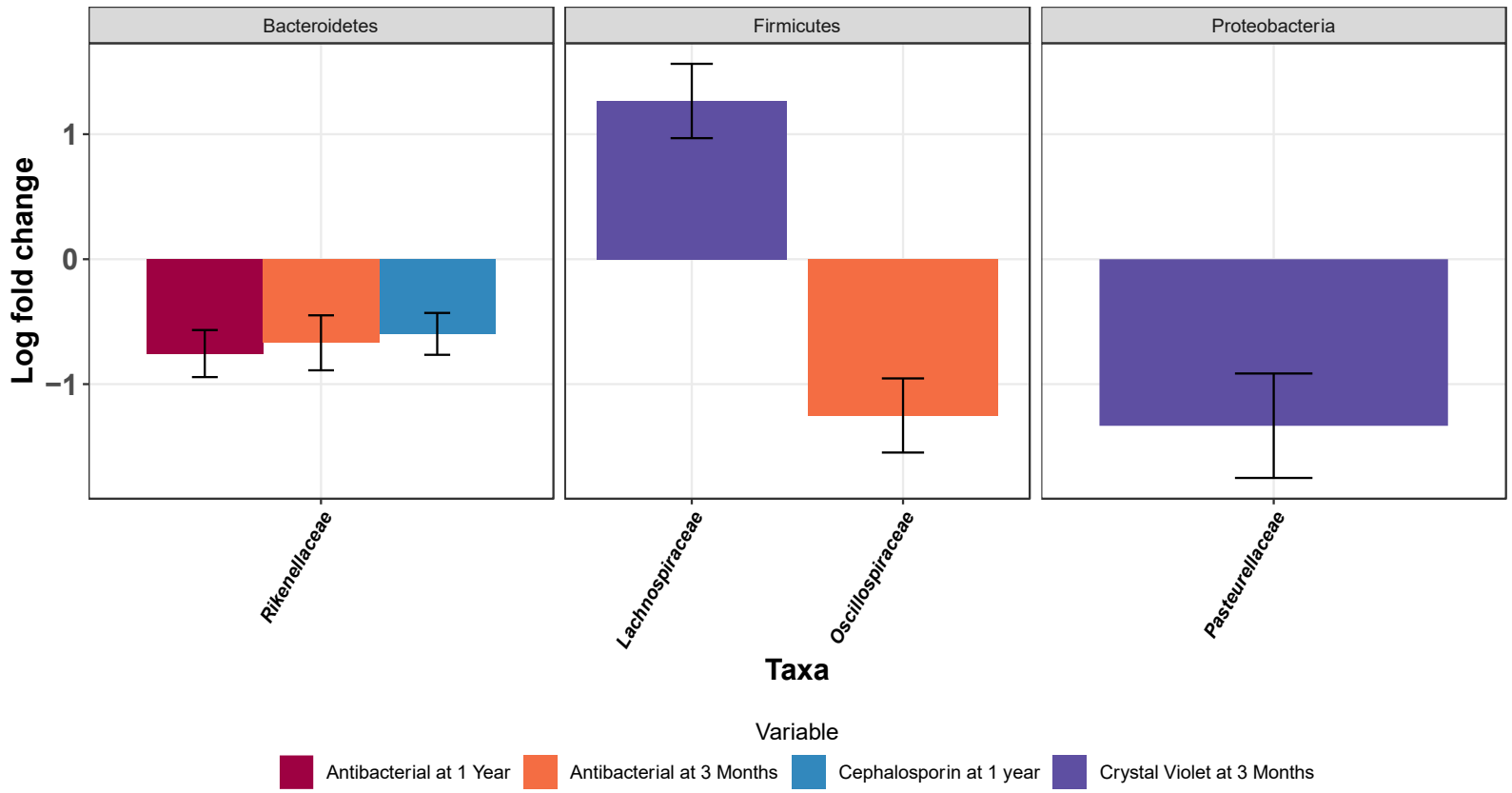

Figure 5.3. **ANCOM-BC Family Level Analysis Summary for Antimicrobial Medications. Analysis was performed using ANCOM-BC while correcting for gender, delivery mode and visit. The Y axis represents the differential abundance between groups that did and did not take the medication indicated in the legend with a log 2 fold change. The X axis shows the assigned taxonomic group. The analysis was run on information available at the family level, but has been organized into phylum groups for easier comparison. Red denotes the comparison for taking an antimicrobial at 1 year, orange the comparison for antimicrobials at 3 months. Blue is the comparison for cephalosporin users at 1 year and the purple is the comparision of crystal violet users at 3 months. All taxa found to be significantly differentially abundant (Bonferroni corrected p< 0.05) are shown in the plots.**

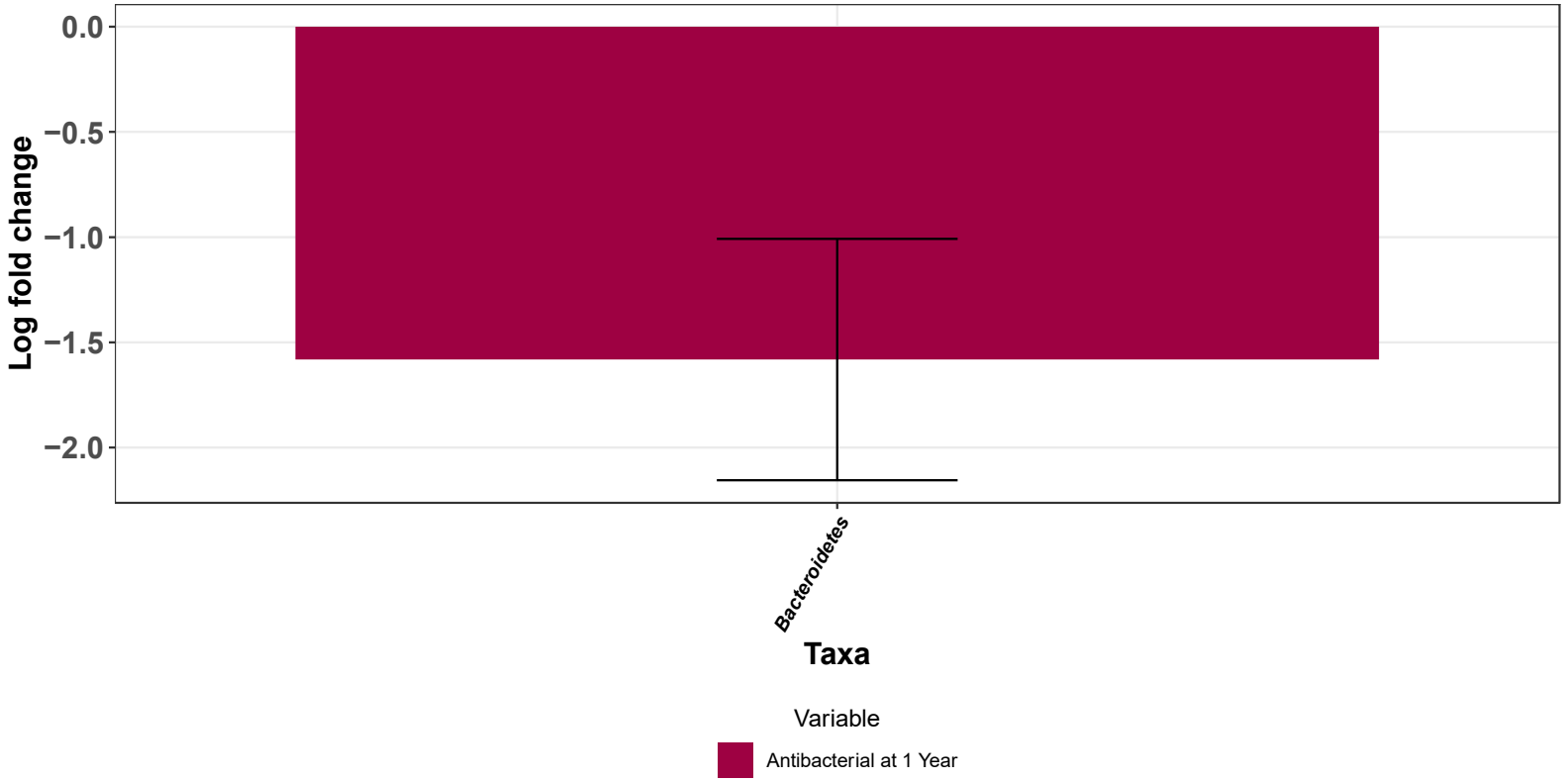

Figure 5.4 **ANCOM-BC Phylum Level Analysis Summary for Antimicrobial Medications. Analysis was performed using ANCOM-BC while correcting for gender, delivery mode and visit. The Y axis represents the differential abundance between groups that did and did not take the medication indicated in the legend with a log 2 fold change. The X axis shows the assigned taxonomic group. The analysis was run on information available at the phylum levelRed denotes the comparison for taking an antimicrobial at 1 year, orange the comparison for antimicrobials at 3 months. All taxa found to be significantly differentially abundant (Bonferroni corrected p< 0.05) are shown in the plots.**

As illustrated in Figure 5.4, there is a notable shift in the Bacteroidetes phylum, likely associated with the shift seen at the family level with the taxa Rikenellaceae, which is a member of the Bacteroidetes phylum as noted in Figure 5.3. When considering the known associations with FB ratio, and if the proportion of Firmicutes remains constant, a the decrease in Bacteroidetes shown in Figure 5.4 could cause an increase in the FB ratio. Higher FB ratios have been previously linked to higher rates of obesity in both adult and child cohorts (Stojanov et al., 2020). In adults, it has been suggested that antibiotics modulate this link (Del Fiol et al., 2018). A discussion in the literature on whether this association would also be found in infants is lacking, however the results from this analysis suggest more research is needed to establish whether this link is present in infant cohorts and adults. I found no statistically significant associations between obesity and any of the four microbial dysbiosis ratios in this project. The number of antimicrobials taken in the first year of life also had no statistically significant correlation with metrics of obesity at 3 months, 1 year or 5 years.

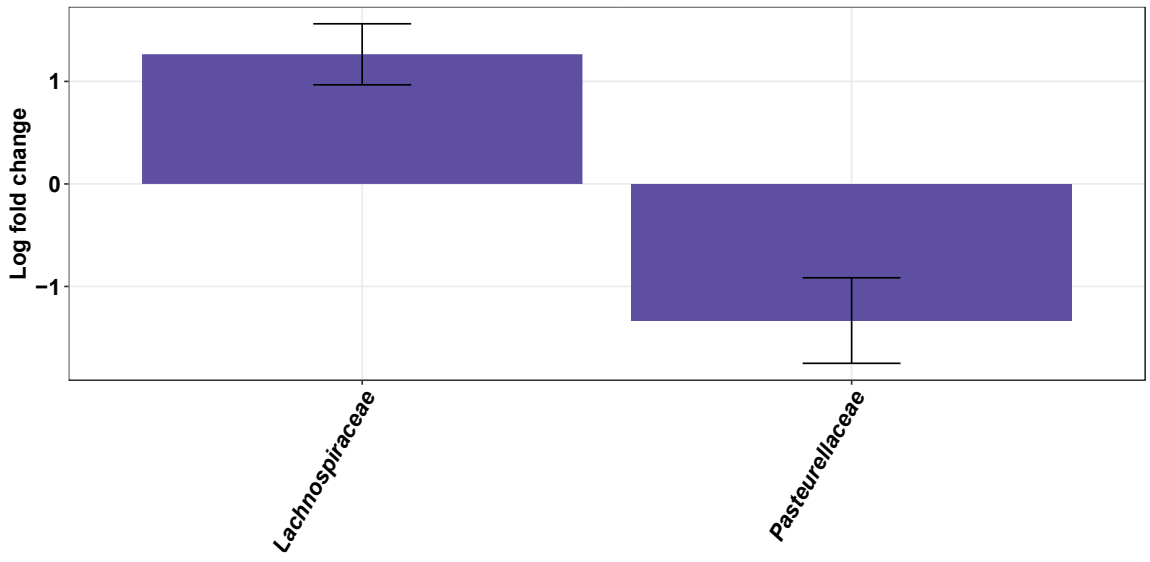

Figure 5.5. Differential Abundance for gut microbial taxa between crystal violet users and non users at 1 year of life. Analysis was performed using ANCOM-BC while correcting for gender, delivery mode and visit. The log 2 fold change is shown along the Y axis, with standard error bars and the taxa (at the family level) with differential abundance (adj p <0.05) are shown along the X axis. All taxa found to be significantly differentially abundant (Bonferroni corrected p< 0.05) are shown in the plots.

To contextualize the discussion regarding crystal violet, note that the infant would have been reaching three months between 2009-2010, at which point crystal violet was recommended as a topical treatment for oral yeast infections in infants. Crystal violet is not strictly antibacterial and has antifungal and anti-helminthic properties. Looking at the reason for medication use data curated as described in Chapter 2, we see that oral yeast infections were the primary indication for crystal violet use in the CHILD Cohort. However, in 2019, crystal violet was removed from Canadian pharmacies due to its carcinogenic behaviour. So, any discussion about implementation or changes in crystal violet usage resulting from this analysis is purely academic as this medication is no longer in use.

In Figure 5.3 and 5.5, we see that Lachnospiraceae, a member of the Firmicutes phylum, has an increased abundance associated with using crystal violet at three months. Previous literature has associated an increase in Lachnospiraceae with an increase in obesity. In this project, obesity was represented using BMI, which is calculated by dividing a subject's mass in kg by their height in meters squared ($BMI=kg/m^2$). However, BMI is not considered a valid marker of obesity in subjects under two years old (CDC, 2015). Regardless, an analysis was run to see if there was any correlation between Lachnospiraceae and the metric of BMI in infants. No statistically significant association between BMI and Lachnospiraceae levels was found in this project. It is worth noting that an increase in Lachnospiraceae has also been associated with an increase in anti-inflammatory responses due to their role in SCFA production. Based on the information available in the literature, we cannot say whether the increase in Lachnospiraceae associated with the use of crystal violet is strictly favourable or unfavourable. Previous research has associated antibacterial use in infants with decreased Lachnospiraceae (Ramirez et al., 2020). The other taxa with differential abundance associated with crystal violet usage at three months are Pasteurellaceae. A decrease in Pasteurellaceae has previously been associated with a decrease in gut inflammation in an adult cohort. Given only this association, this would be the preferable shift for this taxon. The varying associations with Lachnospiraceae abundance, combined with the favourable decrease in Pasteurellaceae, make it difficult to say definitively whether or not the use of crystal violet at three months is associated with favourable or unfavourable changes in the infant gut microbiome.

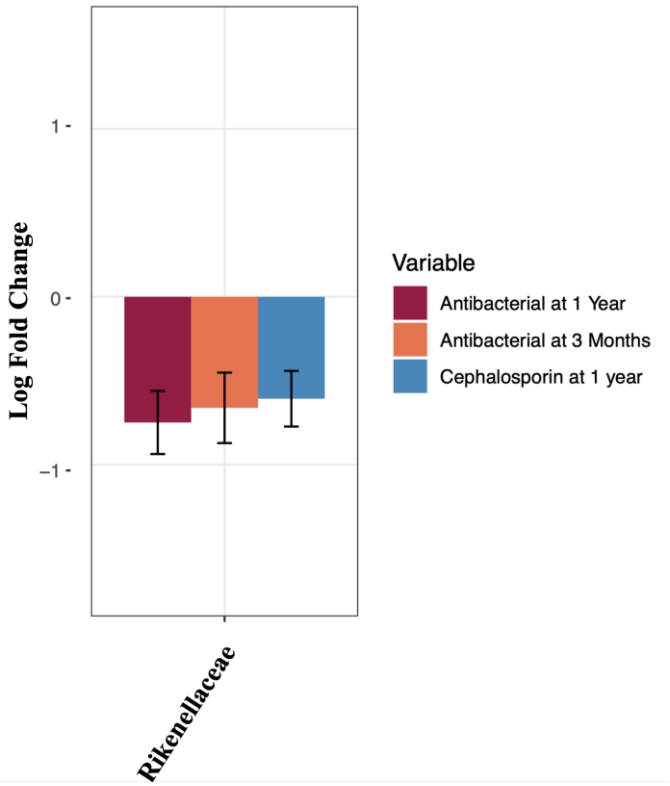

Figure 5.6. Differential abundance in the Rikenellaceae family associated with antimicrobial usage. Analysis was performed using ANCOM-BC while correcting for gender, delivery mode and visit. The log 2 fold change is shown along the Y axis, with standard error bars. and the taxa with differential abundance (adj p <0.05) are shown along the X axis. Red denotes the comparison for taking an antimicrobial at 1 year, orange the comparison for antimicrobials at 3 months. Blue is the comparison for cephalosporin users at 1 year.

The family Rikenellaceae was found to have associations with three different variables relating to the use of antimicrobials in the first year of life,as seen in Figure 5.6. As previously stated, this decrease in a member of the Bacteroidetes family could indicate a shift towards a more obese phenotype, assuming the Firmicutes abundances remain unchanged. More specifically, the taxa Rikenellaceae has been associated with increased obesity in adult cohorts (Stojanov et al., 2020; Vallianou et al., 2021). However, no significant association between Rikenellaceae and BMI was established in this project. The medications associated with this shift in Rikenellaceae abundance are not unexpected. Cephalosporin is known to decrease the abundance of gram-negative bacteria such as Rikenellaceae. However, the association of cephalosporin with microbial abundance shifts linked to obesity phenotypes is novel as literature has focused on the association between cephalosporin and how it increases the chances for

Clostridium difficile infections and the development of vancomycin-resistant *Enterococcus* (Bhalodi et al., 2019).

Overall, through the lens of differential abundance in the gut microbiome, this study did not find a significant correlation between the Firmicutes-to-Bacteroidetes ratios at 3m or 1y with their corresponding BMI measures, even though it has been previously established in other adult cohorts. However, the directionality in the shifts of taxonomic abundance associated with antimicrobial use in infants is consistent with the expected changes associated with obesity in adults. This study provides a novel insight into the potential linkage between the antimicrobial use in infants and the outcome of obesity by associating it with changes in the early life gut microbiota. This helps build upon known connections between FB and obesity, providing early life medication use as a possible explanation for the difference in microbial abundance. However, the use of an antimicrobial itself was not found to have a statistically significant impact on the FB ratio at any time point under study. When investigating the impact of antimicrobials, we see that trends that have been well established in adult microbiomes are also occurring in infant microbiomes, such as the impact of cephalosporin on the Rikenellaceae family. Regardless of the outcome, this study also addresses the need for more region-specific longitudinal studies to characterize the gut microbiome throughout one's lifespan.

Reasons for Medication Use in Antimicrobial and Non-Antimicrobial Users

The comparison in medication use patterns for CHILD cohort participants that never used an antimicrobial vs used an antimicrobial at any point during the first year can be seen in Figure 5.7. For most of the figure, we see that the cells are yellow, indicating that the differences between the two groups are very close to zero. Contextually, this would mean that regardless of their choice to use an antimicrobial, their habits regarding medicating for other adverse events are similar. Furthermore, there appear to be few instances where there is a higher proportion of one group at the extremes of the Firmicutes-to-Bacteroidetes ratio (indicated at the bottom of each panel from 1-5). Figure 5.7 shows the associations with the Firmicutes-to-Bacteroidetes ratio; the other ratios' figures can be found in the supplementary figures in Appendix A.

There are some notable differences when looking at lower bins for the Firmicutes-to-Bacteroidetes ratios and acetaminophen use. We see that for teething, fevers and discomfort, the group that took antimicrobials has more members that are

also giving acetaminophen for the aforementioned symptoms. It is possible that this can be explained by an overall increase in medication use linked to antimicrobials.

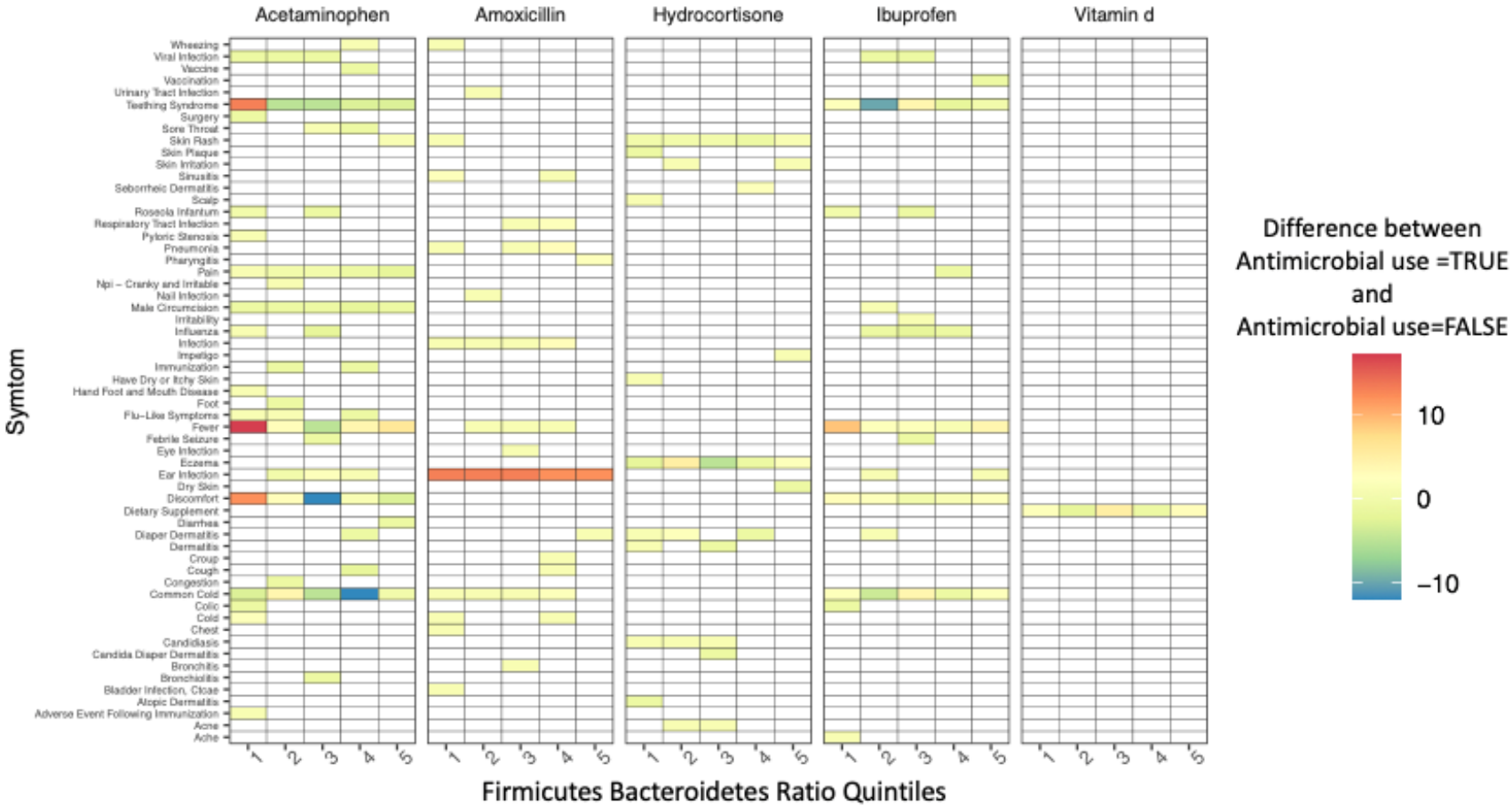

Figure 5.7. **Quintiles for FB_1y and medication patterns for top 5 most common medications Each facet represents a separate medication, and the rows potential reasons for usage. Within the facets, rows represent bins of subjects (n=564) based on their EB ratio at 1y. Subjects in cluster 1 have the lowest ratios, 5 the highest. Colour scale indicates which group tends to use a medication for a specific reason. Red indicates more common in the antimicrobial usage group, blue is the non antimicrobial usage group**

Interdomain Correlations associated with Antimicrobial Use

Looking at the correlation globe in Figure 5.8, we see an overwhelming number of positive correlations, especially with other medications. Based on this observation, I performed a Wilcoxon test to determine whether there was any statistically significant association between taking an antimicrobial at any point in the first year of life and the total number of medications taken over that same period. I was able to confirm a statistically significant relationship ($p<0.05$) between having taken an antimicrobial and the total number of medications taken in the first year. However, the total number of medications did not correlate with any of the measures of microbial dysbiosis or the microbial dysbiosis quintiles depicted in Figure 5.7.

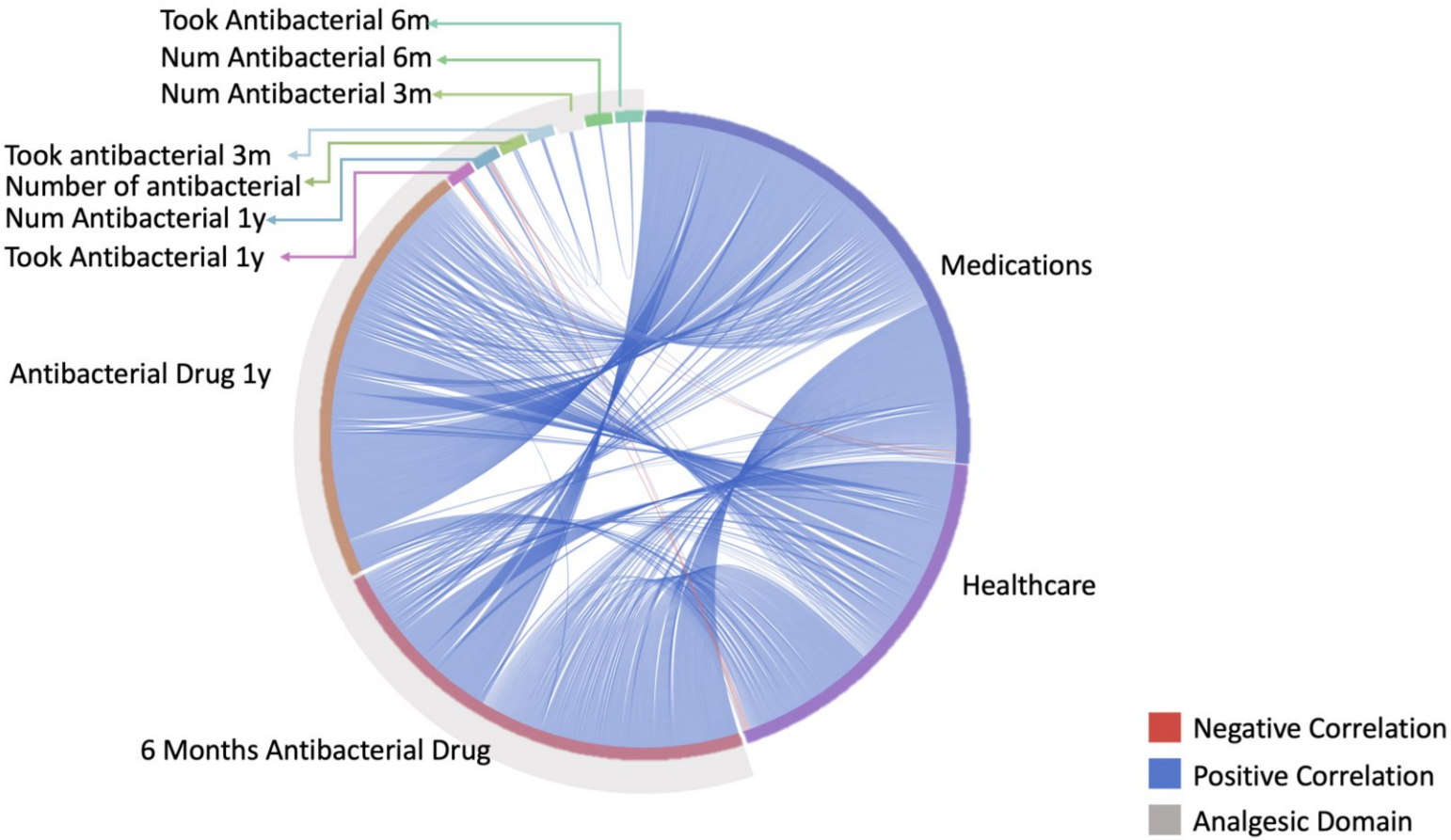

Figure 5.8. GlobeCorr Diagram for variables correlating with antimicrobial usage. All correlations are shown in the correlation coefficient was above 0.4. Positive correlations are indicated with blue ribbons, negative correlations are shown with red ribbons,

5.4.2. Analgesics

Analgesic Use and Differential Abundance in gut microbiota taxa

Due to the n of the groups being below what ANCOM-BC recommends for effective normalization for differing sample sizes, this study found no associations between analgesic use and differentially abundant taxa. Very few associations have been identified to date when looking at the literature, so this is not an uncommon or unexpected outcome. However, the research that does exist on the impact of analgesics on the microbiome once again focuses on the adult microbiome as opposed to the infant microbiome. Some known associations with analgesic use in adults include associations between Eysipelotrichaceae, Verrucomicrobia and, *Butyvibrio* (Yun et al., 2017). More specific relationships have been identified between ibuprofen and Acidaminococcaceae*,* Enterobacteriaceae*,* Propionibacteriaceae*,* Pseudomonaceae and*,* Rikenellaceae in adult gut microbiomes (Rogers & Aronoff, 2016).

To follow up on the lack of statistically valid associations between taxa and specific analgesics, the total number of analgesics taken in the first year of life was compared to the four microbial dysbiosis ratios. No statistically significant associations were identified. This may be partly due to the prevalence of analgesic use in infants. Analgesics were the most common class of medication taken by infants in the first year of life, which could be why it is difficult to get a large enough control group to compare the differential abundance between children that did and did not take specific analgesics in the first year of life. The lack of statistically significant follow-up analyses does not necessarily invalidate its inclusion in the variables of importance from the machine learning models. Analgesic use shows up in the machine learning results, likely due to its association with overall medication use.

Reasons for Medication Use in Analgesic and Non-Analgesic Users

Looking at trends in medication use for analgesic and non-analgesic users in Figure 5.9, we see that acetaminophen appears to be more frequently used than ibuprofen in the first year of life. From a healthcare and policy perspective, this is a positive result, as current guidelines recommend using acetaminophen over ibuprofen for infants (Government of Canada, 2016). This is a helpful observation when looking to

characterize the CHILD Cohort, as it indicates the parents in our group are (intentionally or unintentionally) following recommended guidelines regarding their children.

There appears to be a diverse array of symptoms that parents have chosen to use acetaminophen for their children. It is impossible to state from this data whether this qualifies as “overuse” of acetaminophen by the group. However, it does justify further examining the ways parents decide to use acetaminophen with their infants. However, no statistically significant associations existed between the number of analgesics and any microbial dysbiosis ratios used. Furthermore, there was no statistically significant association between the FB quintiles and the total number of medications at three months or one year.

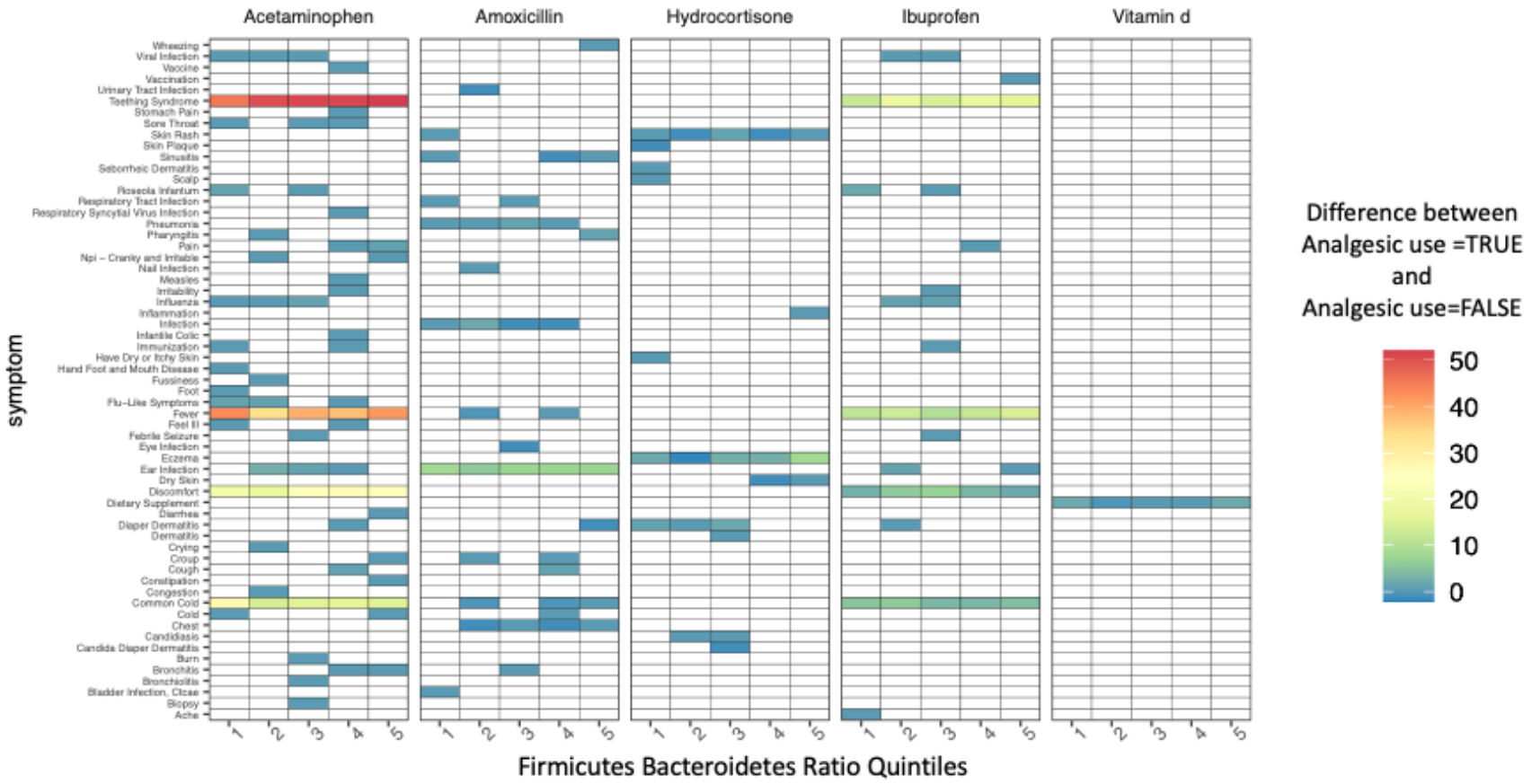

Figure 5.9. **Quintiles for FB_1y and medication patterns for top 5 most common medications Each facet represents a separate medication, and the rows potential reasons for usage. Within the facets, rows represent bins of subjects (n=564) based on their EB ratio at 1y. Subjects in cluster 1 have the lowest ratios, 5 the highest. Colour scale indicates which group tends to use a medication for a specific reason. Red indicates more common in the analgesic usage group, blue is the non-analgesic usage group.**

Interdomain Correlations associated with Analgesic Use

When looking at the interdomain correlations associated with analgesic use in Figure 5.10, there are many negative correlations with variables in the healthcare domain. This is surprising as these variables include information such as the number of fevers, generic infections, headaches etc. This healthcare information is all from the reasons for medication use dataset outlined in Chapter 2, so it was expected that the use of analgesics would positively correlate with the reports of medicated health issues. However, the negative correlation would imply that as the number of analgesics taken goes up, the health incidents decrease. This may imply that the high number of analgesics taken is linked to using these medications for maintenance and preventative purposes rather than for treatment of manifested symptoms. As with antimicrobials, a statistically significant correlation (adj $p < 0.05$) was identified between having ever taken an analgesic in the first year of life and the total number of medications taken in that same period.

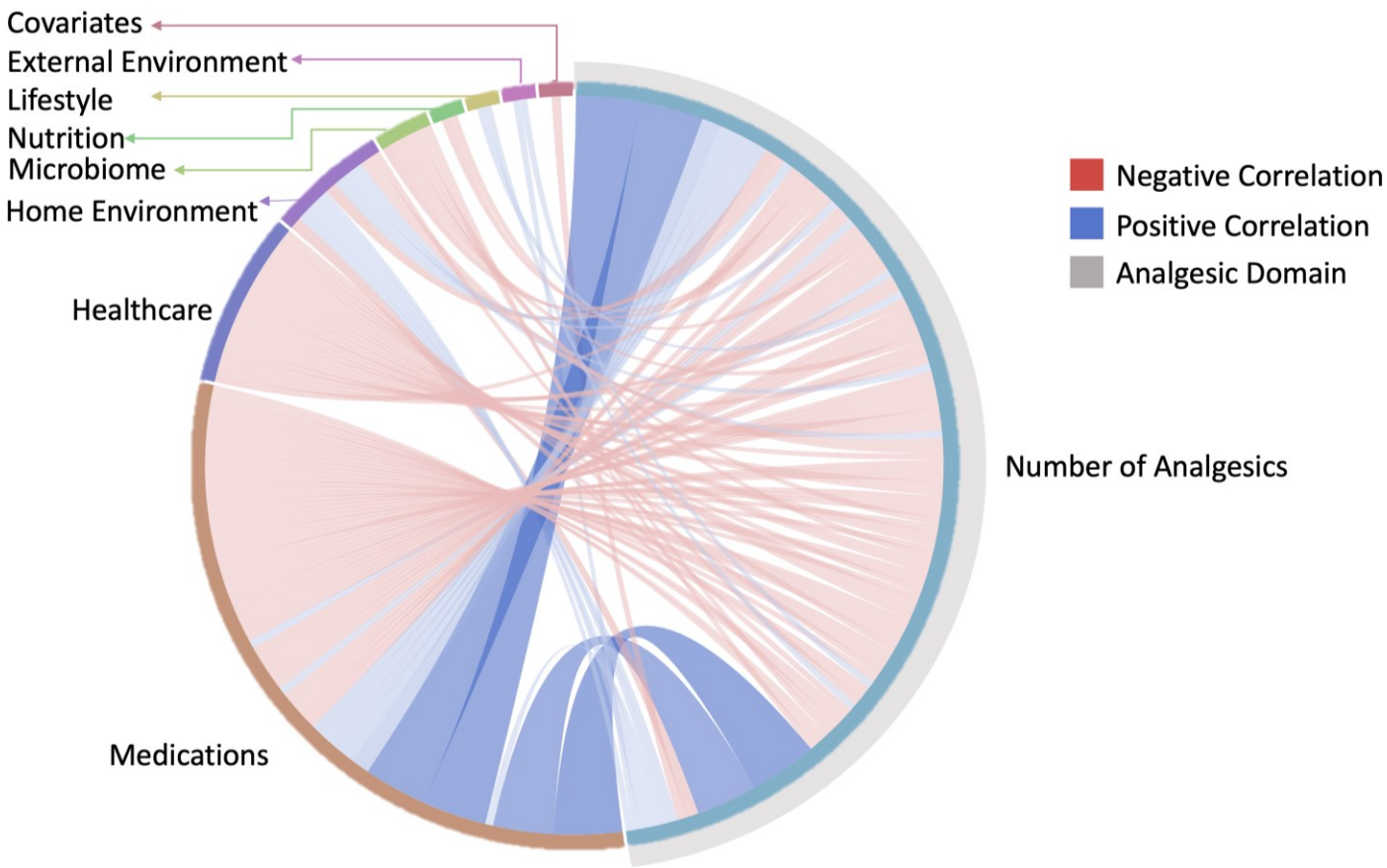

Figure 5.10. GlobeCorr Diagram for variables correlating with analgesic usage. All correlations are shown in the correlation coefficient was above 0.4. Positive correlations are indicated with blue ribbons, negative correlations are shown with red ribbons.

5.4.3. Vitamin D

Vitamin D Use and Differential Abundance in gut microbiota taxa

Differential abundance compared between groups of CHILD Cohort Study participants that did vs did not take any vitamin D was calculated using ANCOM-BC as described in Section 5.3.1. To capture all reports of vitamin D use in the first year of life for cohort participants, eight variables were used. The full list of variables can be seen in Table B.12. Figures 5.11-5.14 show the results of the ANCOM-BC analysis. Figure 5.11 shows all taxa that were found to be differentially abundant between groups that did and did not take the medication specified by the bar's colour at the family level, organized by phylum. The phylum level comparison can be seen in Figure 5.12. Declaration of differential abundance associated with the specified medication variable was only stated if the adjusted p-value was below 0.05 to account for multiple testing. Taking vitamin D at three months and having vitamin D supplementation at one year were associated with changes in the infant gut microbiome. Figures 5.11 and 5.12 are summary figurse, and their components are subsequently broken down into individual figures. Figure 5.13 shows differential abundance associated with taking vitamin D at three months, and Figure 5.14 shows the associations for vitamin D supplementation by one year.

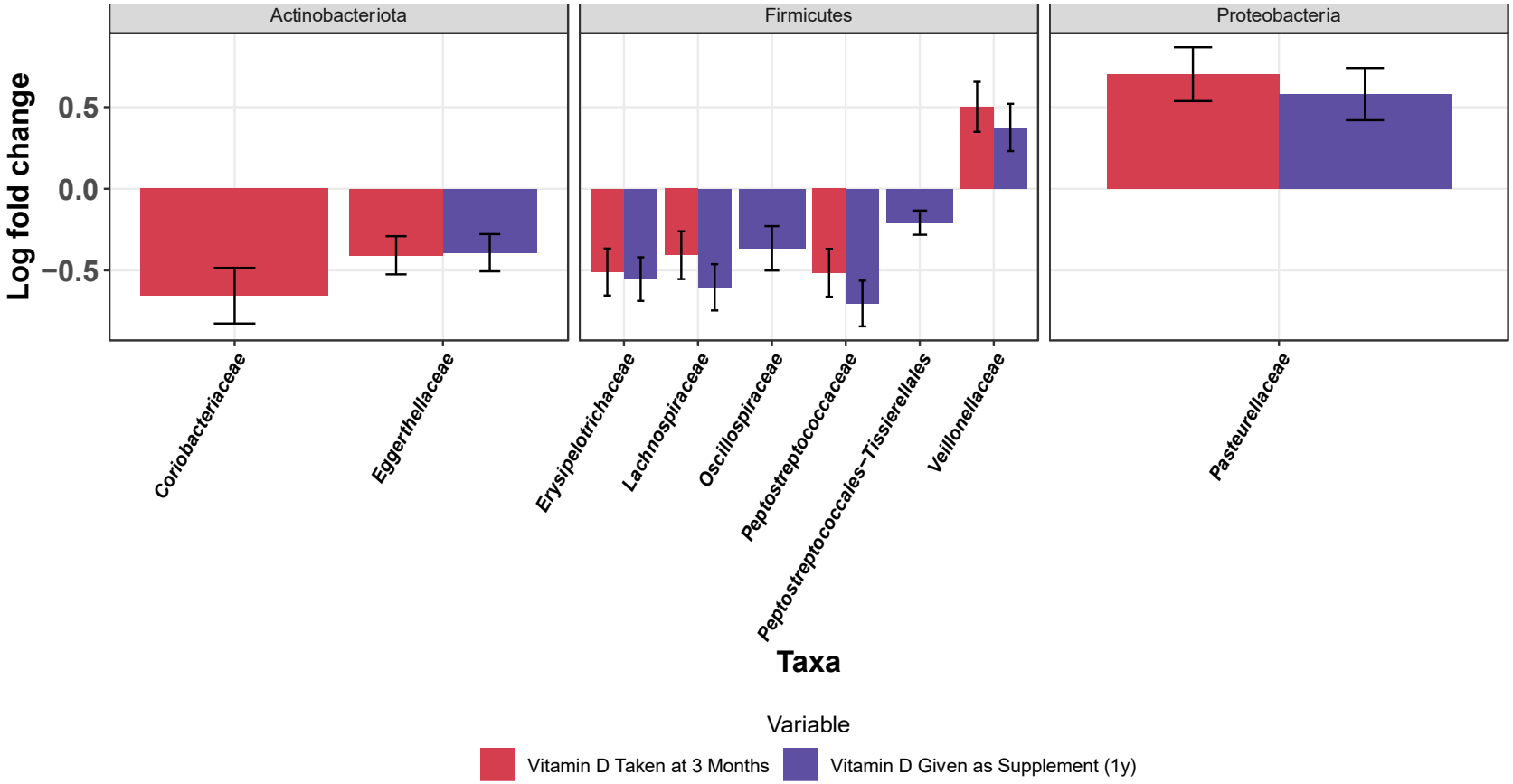

Figure 5.11. ANCOM-BC analysis at the family level for Vitamin D. Analysis was performed using ANCOM-BC while correcting for gender, delivery mode and visit. The Y axis represents the differential abundance between groups that did and did not take the medication indicated in the legend. The X axis shows the assigned taxonomic group. The analysis was run on information available at the family level, but has been organized into phylum groups for easier comparison. Red is for Vitamin D taken at 3 months, and Purple is for vitamin D given as a supplement at 1 year. All taxa found to be significantly differentially abundant (Bonferroni corrected p< 0.05) are shown in the plots.

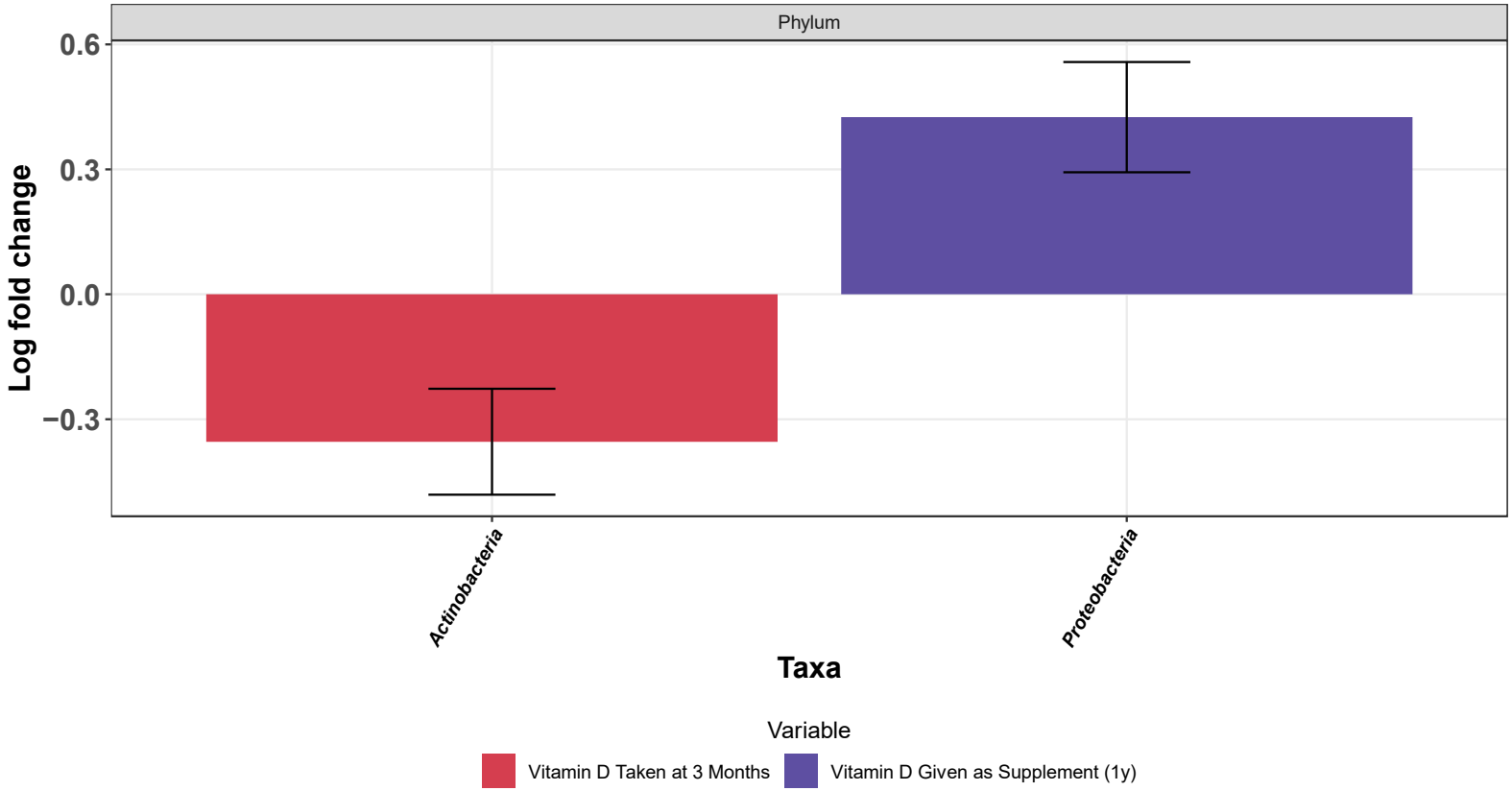

Figure 5.12. ANCOM-BC analysis at the Phylum level for Vitamin D. Analysis was performed using ANCOM-BC while correcting for gender, delivery mode and visit. The Y axis represents the differential abundance between groups that did and did not take the medication indicated in the legend. The X axis shows the assigned taxonomic group. The analysis was run on information available at the phylum level. Red is for Vitamin D taken at 3 months, and Purple is for vitamin D given as a supplement at 1 year. All taxa found to be significantly differentially abundant (Bonferroni corrected p< 0.05) are shown in the plots.

In Figure 5.11 6 of the identified taxa are members of the Firmicutes phylum, Erysipelotrichaceae*,* Lachnospiraceae, Peptostreptococcaceae, Peptostreptococcales-Tissierellales, and Veillonellaceae. Of these, 5 are shown to decrease in abundance with the associated medication variables. Under the assumption that members of the Bacteroidetes phylum stay consistent, this would decrease the Firmicutes-to-Bacteroidetes ratio. A decreased FB ratio has been negatively associated with obesity in previous literature, and the usage of vitamin D has also been associated with lower rates of obesity (Filgueiras et al., 2020). However, no statistically significant correlation was found between the differentially abundant taxa and BMI. While a decreased FB ratio has been negatively associated with obesity, it has been positively associated with a pro-inflammatory phenotype. Of particular interest in the family Lachnospiraceae. As short-chain fatty acid producers, this group is essential for modulating inflammatory responses and ensuring that those infants do not develop a pro-inflammatory phenotype.

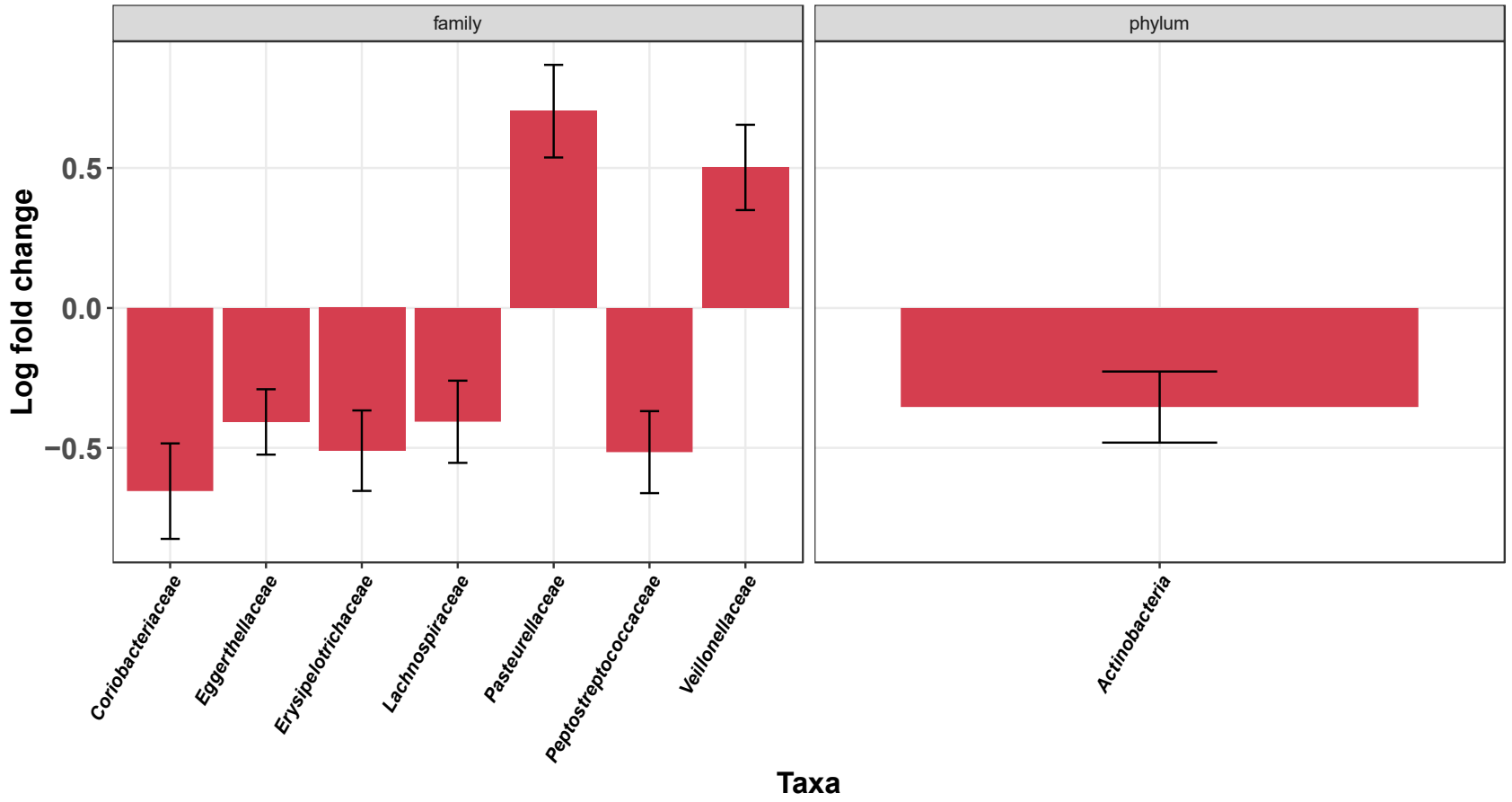

Figure 5.13. Differential Abundance for gut microbial taxa between Vitamin D users (n=706) and non users (n=422) at 3 months of life. Analysis was performed using ANCOM-BC while correcting for gender, delivery mode and visit. The log 2 fold change is shown along the Y axis, with standard error bars and the taxa (at the family level) with differential abundance (adj p <0.05) are shown along the X axis. All taxa found to be significantly differentially abundant (Bonferroni corrected p< 0.05) are shown in the plots.

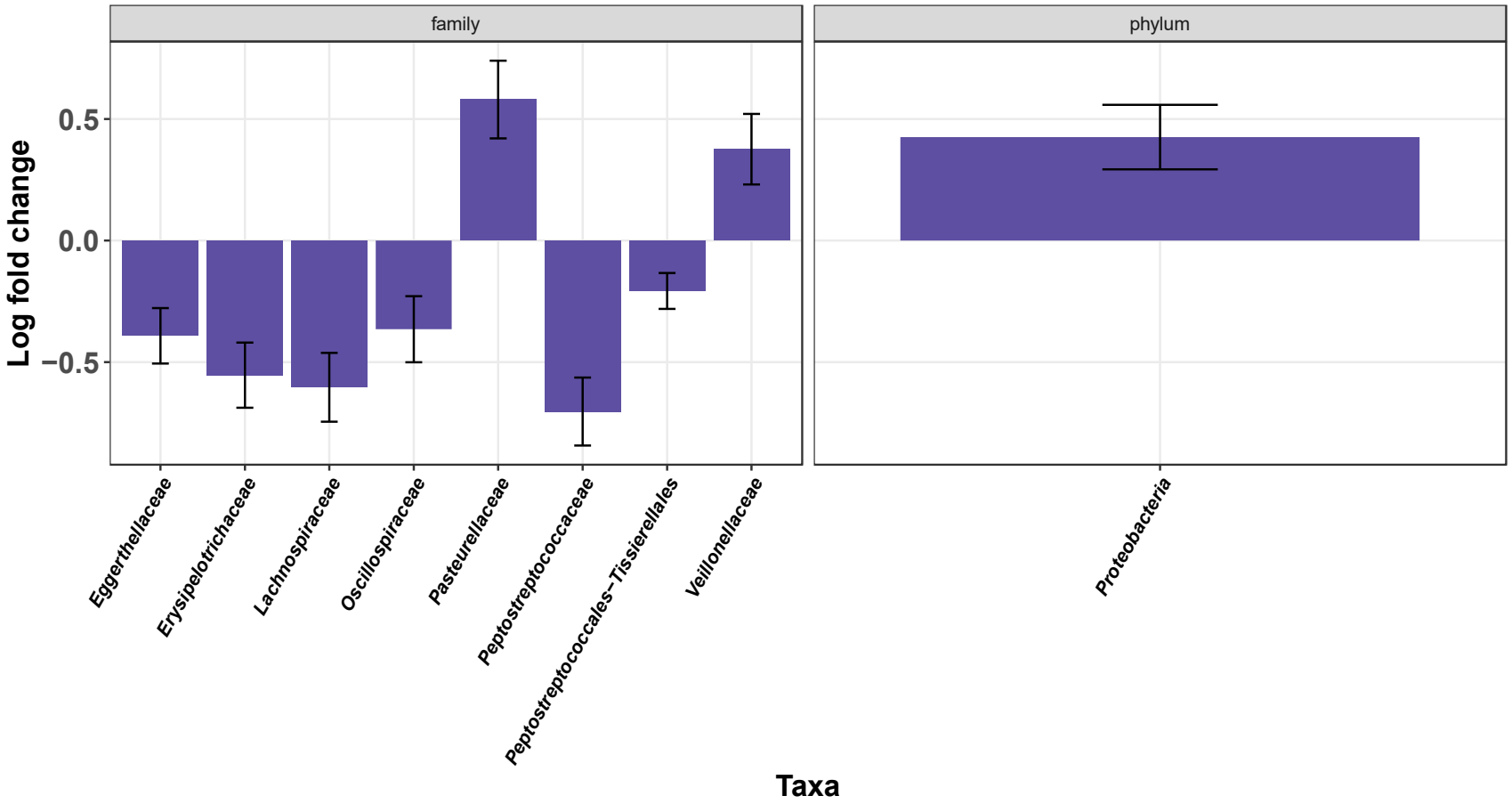

Figure 5.14. **Differential Abundance for gut microbial taxa between vitamin D supplement users (n=538) and non users (n=590) at 1 year of life. Analysis was performed using ANCOM-BC while correcting for gender, delivery mode and visit. The log 2 fold change is shown along the Y axis, with standard error bars and the taxa (at the family level) with differential abundance (adj p <0.05) are shown along the X axis. All taxa found to be significantly differentially abundant (Bonferroni corrected p< 0.05) are shown in the plots.**

It is important to contextualize these results with other studies looking at the relationship between vitamin D at different points of development and in different metabolic forms. Vitamin D can be obtained in 3 ways. Vitamin D can be found in food, and in Canada, many dairy products are supplemented with it. Vitamin D supplements are also available in liquid dropper form for infants. Vitamin D can also be metabolically produced after the skin comes in contact with the sun. However, vitamin D is not automatically in a metabolically active form. It must first go to the liver, where it is converted to 25-hydroxyvitamin D, and to the kidney, where it becomes 1,25-hydroxyvitamin D and can then go on and perform its various roles.

In this study, taking a vitamin D supplement was found to have associations with microbiome changes. Figures 5.11, 5.13, and 5.14 show that a decrease in the family Lachnospiraceae is associated with vitamin D usage and supplementation. A previous project in the CHILD Cohort Study also found that maternal prenatal vitamin D supplementation was associated with a decrease in Lachnospiraceae in exclusively breastfed infants at three months (Drall et al., 2020). There have been other studies that have looked at vitamin D at different stages that have contrasting results. For example, a US cohort study found that prenatal 25-hydroxyvitamin D (the form vitamin D takes after being metabolized in the liver) is associated with an increase in the abundance of Lachnospiraceae in infants (Kassem et al., 2020). This finding and comparing it to the findings in this project and the other study in the CHILD cohort project bring up an important consideration for microbiome and lifestyle studies as there are many possible explanations for these differences that must be considered.

In the prenatal US study, it could be that high levels of 25-hydroxyvitamin D impact the overall health status of the mother, which in turn alters the development of the infant's microbiome. In this study, the microbial associations were made with the variables asking whether or not the infant and any kind of vitamin D supplementation. While we know that taking vitamin D could increase the amount of vitamin D in the infant's system, the decision to use vitamin D may also be correlated to other lifestyle factors that interact with the infant's developing microbiome. For example, in Figure 5.12 and Figure 5.13, we see a decrease in the abundance of the phylum actinobacteria associated with vitamin D taken at three months. Bifidobacteria make up a large component of the actinobacteria phylum found in infants and a decrease in Bifidobacteria has also been found in formula-fed infants (Ottman et al., 2012). Another

possibility put forward by Yamamoto and Jørgensen is that vitamin D doesn't impact the microbiota but different immunoregulatory properties, which contributes to microbiome shifts (Yamamoto & Jørgensen, 2020). Regardless, these results suggest that more research is needed to better characterize the most effective vitamin D supplementation strategies for microbiome trajectories.

This study analyzed vitamin D because this medication was flagged during the machine learning analyses. Vitamin D usage was associated with several taxonomic shifts in the infant gut microbiome. However, comparing these results to other studies looking at vitamin D usage and metabolism pre and postnatally, there is a wide range of impacts. More research is required to fully describe the associations between vitamin d usage, vitamin D metabolism, lifestyle, and the implications for the infant microbiome.

Reasons for Medication Use in Vitamin D and Non-Vitamin D Users

The comparison in medication use patterns for CHILD cohort participants that never used vitamin D vs used vitamin D at any point during the first year can be seen in Figure 5.15. It is recommended that infants receive 400 IU of vitamin D per day to ensure proper bone growth and avoid diseases such as Ricketts (Canadian Pediatrics Society, n.d.). Another well-established recommendation for infants is that acetaminophen should be used as opposed to ibuprofen when giving analgesics. Similar to the outlined observations for analgesics in Section 5.4.2. When looking at patterns in acetaminophen and vitamin D supplementation, it appears that if parents follow recommendations for one of the medications, they follow it for the other.

The difference between vitamin D users and non-users is highest when looking at the use of acetaminophen for teething symptoms. The use of a medication to treat teething symptoms is an example of "maintenance" medication use, where there is not actually anything wrong with the infant that requires fixing via medication. Teething is a natural part of the developmental process. Likewise, the supplementation of vitamin D is a maintenance/preventative medication usage. This suggests that there may be a group of parents more liberal in their medication use than others. This medication lifestyle could manifest for various reasons, the full characterization of which would be outside the scope of this master's thesis.

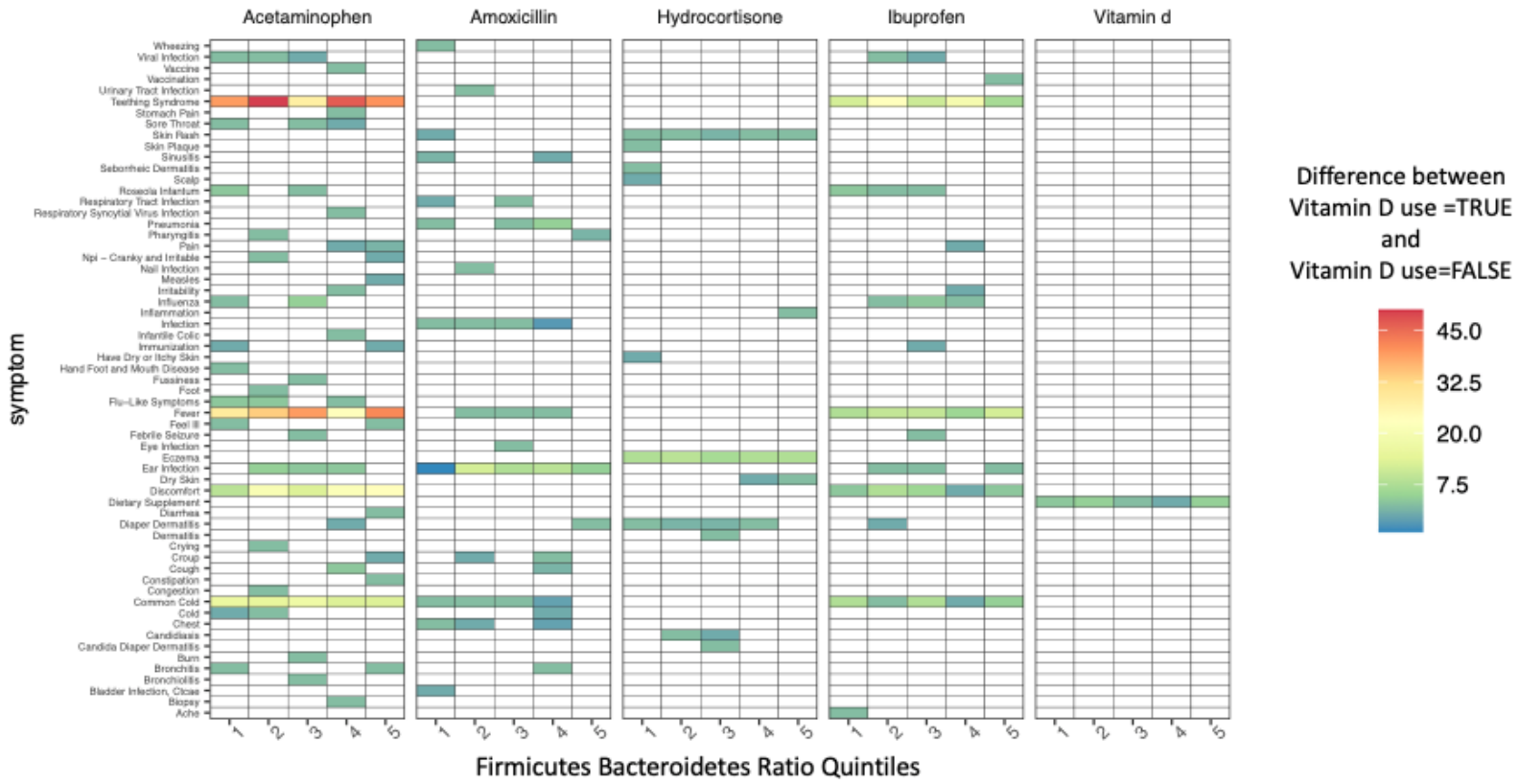

Figure 5.15. Quintiles for FB_1y and medication patterns for top 5 most common medications Each facet represents a separate medication, and the rows potential reasons for usage. Within the facets, rows represent bins of subjects (n=564) based on their FB ratio at 1y. Subjects in cluster 1 have the lowest ratios, 5 the highest. Colour scale indicates which group tends to use a medication for a specific reason. Red indicates more common in the Vitamin D usage group, blue is the non Vitamin D usage group.

Interdomain Correlations associated with Vitamin D

Correlations with vitamin D variables can be seen in Figure 5.16. Unlike with antimicrobials and analgesics, taking vitamin D at any point in the first year did not significantly correlate with the total number of medications taken in the first year. Considering the observations made from the previous sections, it is interesting that most of the strong correlations are with other medications and healthcare variables such as number of hospital visits and incidents of common colds. There are very few links between variables relating to the infant and the mother. We also see this trend in Figure 5.1, where a random subset of correlations is shown. This is another incident the importance of what we don't see. No correlations in this globe immediately help clarify the complex and multifactorial relationship between infant supplementation of vitamin D and the infant microbiome. However, the abundant positive correlations with other medications may help to further characterize the liberal medication lifestyle that was hypothesizes based on the results in Section 5.4.2.

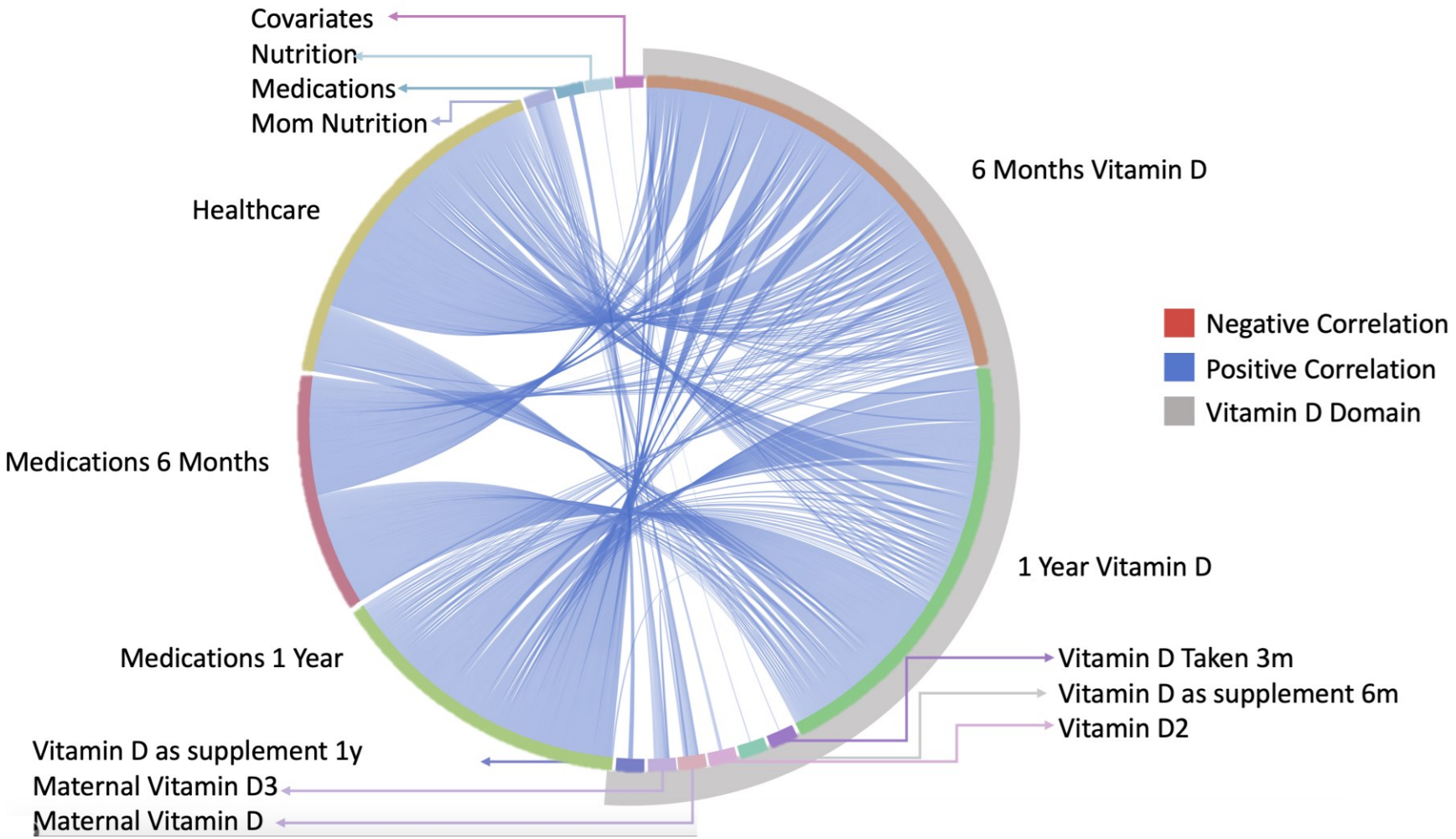

Figure 5.16. GlobeCorr Diagram for variables correlating with Vitamin D usage. All correlations are shown in the correlation coefficient was above 0.4. Positive correlations are indicated with blue ribbons, negative correlations are shown with red ribbons.

5.5. Concluding Remarks

In recent literature reviewing the current state of understanding for microbiome development, there is a call for more region-specific longitudinal studies. The CHILD Cohort Study is well situated to help fill this knowledge gap. The results obtained from this project helped lay the foundation for more targeted analyses surrounding the relationship between medication use and microbiome development in infants. When shifting from an investigation into the differential abundance to more general medication use habits in this cohort, there are some notable observations. Having taken an antimicrobial or an analgesic at any point in the first year had a statistically significant (adj $p < 0.05$) correlation between the medication class ever being taken and the total number of medications taken. Given that the use of these medication classes was flagged by the machine learning models but failed to show any statistically significant correlations to microbial dysbiosis or obesity outcomes, it is possible that taking an analgesic or antimicrobial is a proxy for a lifestyle that tends to be high in medication use overall. However, due to its multifactorial nature, further research is required.

Furthermore, when examining how and what medications are used, it appears that the parents in the CHILD Cohort Study are acting in accordance with recommended guidelines. Whether this is intentional or coincidental is unknown. There are more reports of acetaminophen usage than ibuprofen, which is the preferred trend based on healthcare recommendations. Vitamin D supplementation is also recommended for infants, and we see that the group supplementing with vitamin D is also using more acetaminophen than the non-vitamin D using counterparts. If these habits are not a coincidence and result from proper education around the current best practices for medication use in infancy, then this indicates that we are currently employing effective strategies for parental education and that these strategies can be implemented in the future to help disseminate results from the CHILD Cohort Study.

However, it does not avoid the common pitfalls of microbiota and disease association studies. It is very difficult to claim causation in many of these relationships. At its most simple, these situations are a three-way relationship, we have the illness or adverse event requiring medication, we have the act of taking the medication, and we have the microbiota shifts. In some cases, it may be that there is a direct impact of the drug on the microbiome, and this shift in the microbiome induces a disease. However,

there are cases where the disease creates a shift in the microbiome, and the drug taken to treat the condition has no impact. In the case of infant colonization, it is possible that mothers taking supplements have healthier pregnancies and are less likely to deliver prematurely. In this case, it may be that the fact they are able to carry the baby to term impacts the microbiome and not the vitamin supplements. With the methods employed in this project, it is impossible to identify any causal or directional relationships. However, the establishment of potential associations is a helpful first step for future work to evaluate causal relationships.

Chapter 6.

Conclusion

6.1. Summary

This project aimed to examine medication practices' impacts on infant microbial dysbiosis and associated child outcomes. This project helps address the need for region-specific longitudinal studies of the gut microbiome throughout one's lifespan. To do this, ontologies were applied to a dataset describing reasons for medication use. This data had been collected as user-supplied free text but having access to properly cleaned and organized data describing reasons for medication use had several advantages both for this project and for collaborative projects within the CHILD Cohort Study. These collaborative projects include evaluating medication usage for infants in the first year of life and the medication habits of mothers and supplying the justifications provided for the medication usage. In future, this data can be made available through CHILDdb to streamline analyses for other researchers across Canada. This study provides variables that can be used to incorporate health outcomes in machine learning models built to identify influencers of microbial dysbiosis.

A wide variety of variables were included in the machine learning models. To better visualize the relationships between these variables, the visualization tool GlobeCorr was created. This tool helps to create interactive, dynamic visualization of large correlation datasets. This tool was made available at GlobeCorr.ca, and in addition to curating examples, I contributed to developing static content and learning resources on the site. Conclusions drawn from these visualizations helped to complement the results from the machine learning models.

The machine learning models were built to try and identify influencers of the FB ratio and EB ratio at three months and one year, as these ratios have been previously associated with negative health outcomes. The goal when building these models was not to create a tool that would predict microbial dysbiosis ratios but to help identify potential medication practices that may influence these outcomes. The machine learning models identified antimicrobials, analgesics and vitamin D as medications with potential associations to microbial dysbiosis.

After investigating these medications further, the results from this research suggest that trends present in adult microbiomes are also found in infant microbiomes. This is significant as it provides insight into where actionable windows of change might be present at different life stages. There were no direct statistically significant associations between medication classes and microbial dysbiosis. This may suggest that medication use, particularly antimicrobial and analgesic, can be used as an overall proxy for a lifestyle more prone to microbial dysbiosis. However, more research is needed to characterize the links between medication use fully, metabolism, and lifestyle to identify the most effective intervention points for preventing microbial dysbiosis.

6.2. Future Directions

The work completed in this thesis project lays a foundation for further investigation into the links between medication use and microbiome development in the CHILD cohort. Improvements can be made to the machine learning models to try and improve predictive power. Potentially, more research can be done into establishing normal and abnormal ranges of microbial dysbiosis ratios to bin participants into "preferable" and "non -preferable" taxonomic ratios. Setting up the data as such would allow for the creation of Random Forest or machine learning classifiers instead of regression models. The accuracy metric could then be evaluated, and it is possible that by binning the subjects into more biologically relevant groups, the most important variables are easier to spot and identify with statistically significant correlations.

The results of this project can contribute to ongoing CHILD Cohort Study Projects. The standardization of reasons for medication uses and its application to reasons for hospital visits is being made available to other CHILD cohort researchers to create a definition of a "healthy child." A complimentary study looks to characterize overall lifestyles associated with microbial dysbiosis. The information collected regarding medication use practices for antimicrobials, analgesics and vitamin D can be used to add dimension and more robustly characterize different groups of cohort participants.

This project was mainly exploratory, which is a necessary step when attempting to characterize relationships, as techniques such as Structural Equation Modelling (SEM) can only test relationships the researcher has previously identified. SEM is a multivariate statistical technique. Running these models will give distinct p values for

each relationship, allowing the researcher to determine if the association is significant and, if so, the magnitude of its impact. The CHILD Cohort Study has previously employed SEMs to establish relationships between microbial dysbiosis and asthma (Patrick et al., 2020). Going forward, Geoff Winsor of the Brinkman Lab will be focusing on employing SEM to answer questions surrounding early life exposures, gut microbiome and child health. This project's results can help guide the formation of these SEM.

An option to further characterized this data would be to employ the ARISTOTLE method developed by Mansouri et al. to evaluate the associations identified in this project (Mansouri et al., 2022). While the SEM outlined above requires some a priori decision-making about what relationships to include, a hallmark of ARISTOTLE is that it can take large, multi-omics datasets and reduce them down to more manageable clusters, assign weights to the variables, filter based on variable weight, and then evaluate probable causes based on groups of heavily weighted features. Overall, a combination of the SEM and ARISTOTLE methods would be a suitable direction to complement the results of this project.

References

Acknowledging the Alliance | Digital Research Alliance of Canada. (n.d.). Retrieved June 14, 2022, from https://alliancecan.ca/en/services/advanced-research-computing/research-portal/acknowledging-alliance

Anderson, C., Rolfe, P., & Brennan-Hunter, A. (2013). Administration of Over-the-Counter Medication to Children at Home—A Survey of Parents from Community Health Centers. *Journal of Community Health Nursing*, *30*(3), 143–154. https://doi.org/10.1080/07370016.2013.806716

Azad, M. B., Konya, T., Guttman, D. S., Field, C. J., Sears, M. R., HayGlass, K. T., Mandhane, P. J., Turvey, S. E., Subbarao, P., Becker, A. B., Scott, J. A., Kozyrskyj, A. L., & Investigators, the C. S. (2015). Infant gut microbiota and food sensitization: Associations in the first year of life. *Clinical & Experimental Allergy*, *45*(3), 632–643. https://doi.org/10.1111/cea.12487

Banerjee, M., Reynolds, E., Andersson, H. B., & Nallamothu, B. (2019). Tree-Based Analysis: A Practical Approach to Create Clinical Decision Making Tools. *Circulation. Cardiovascular Quality and Outcomes*, *12*(5), e004879. https://doi.org/10.1161/CIRCOUTCOMES.118.004879

Binda, C., Lopetuso, L. R., Rizzatti, G., Gibiino, G., Cennamo, V., & Gasbarrini, A. (2018). Actinobacteria: A relevant minority for the maintenance of gut homeostasis. *Digestive and Liver Disease*, *50*(5), 421–428. https://doi.org/10.1016/j.dld.2018.02.012

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*, 1–67. https://doi.org/10.18637/jss.v045.i03

Canadian Pediatrics Society. (2007). Vitamin D supplementation: Recommendations for Canadian mothers and infants | Canadian Paediatric Society. *2007*. https://cps.ca/en/documents//position//vitamin-d/

CHILDdb. (n.d.). Retrieved May 2, 2022, from https://childdb.ccm.sickkids.ca/

den Besten, G., van Eunen, K., Groen, A. K., Venema, K., Reijngoud, D.-J., & Bakker, B. M. (2013). The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *Journal of Lipid Research*, *54*(9), 2325–2340. https://doi.org/10.1194/jlr.R036012

Depner, M., Taft, D. H., Kirjavainen, P. V., Kalanetra, K. M., Karvonen, A. M., Peschel, S., Schmausser-Hechfellner, E., Roduit, C., Frei, R., Lauener, R., Divaret-Chauveau, A., Dalphin, J.-C., Riedler, J., Roponen, M., Kabesch, M., Renz, H., Pekkanen, J., Farquharson, F. M., Louis, P., … Ege, M. J. (2020). Maturation of the gut microbiome during the first year of life contributes to the protective farm effect on childhood asthma. *Nature Medicine*, *26*(11), 1766–1775. https://doi.org/10.1038/s41591-020-1095-x

Donohoe, D. R., Garge, N., Zhang, X., Sun, W., O'Connell, T. M., Bunger, M. K., & Bultman, S. J. (2011). The Microbiome and Butyrate Regulate Energy Metabolism and Autophagy in the Mammalian Colon. *Cell Metabolism*, *13*(5), 517–526. https://doi.org/10.1016/j.cmet.2011.02.018

Drall, K. M., Field, C. J., Haqq, A. M., de Souza, R. J., Tun, H. M., Morales-Lizcano, N. P., Konya, T. B., Guttman, D. S., Azad, M. B., Becker, A. B., Lefebvre, D. L., Mandhane, P. J., Moraes, T. J., Sears, M. R., Turvey, S. E., Subbarao, P., Scott, J. A., & Kozyrskyj, A. L. (2020). Vitamin D supplementation in pregnancy and early infancy in relation to gut microbiota composition and C. difficile colonization: Implications for viral respiratory infections. *Gut Microbes*, *12*(1), 1799734. https://doi.org/10.1080/19490976.2020.1799734

Francino, M. P. (2014). Early Development of the Gut Microbiota and Immune Health. *Pathogens*, *3*(3), 769–790. https://doi.org/10.3390/pathogens3030769

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, *8*. https://www.frontiersin.org/article/10.3389/fmicb.2017.02224

Government of Canada. (2016, September 22). *Acetaminophen and children* [Education and awareness]. https://www.canada.ca/en/health-canada/services/drugs-medical-devices/acetaminophen-and-children.html

Hales, C. M., Kit, B. K., Gu, Q., & Ogden, C. L. (2018). Trends in Prescription Medication Use Among Children and Adolescents—United States, 1999-2014. *JAMA*, *319*(19), 2009–2020. https://doi.org/10.1001/jama.2018.5690

Hardy, H., Harris, J., Lyon, E., Beal, J., & Foey, A. D. (2013). Probiotics, Prebiotics and Immunomodulation of Gut Mucosal Defences: Homeostasis and Immunopathology. *Nutrients*, *5*(6), 1869–1912. https://doi.org/10.3390/nu5061869

He, Y., Sarntivijai, S., Lin, Y., Xiang, Z., Guo, A., Zhang, S., Jagannathan, D., Toldo, L., Tao, C., & Smith, B. (2014). OAE: The Ontology of Adverse Events. *Journal of Biomedical Semantics*, *5*(1), 29. https://doi.org/10.1186/2041-1480-5-29

Hiltunen, H., Collado, M. C., Ollila, H., Kolari, T., Tölkkö, S., Isolauri, E., Salminen, S., & Rautava, S. (2021). Spontaneous preterm delivery is reflected in both early neonatal and maternal gut microbiota. *Pediatric Research*, 1–8. https://doi.org/10.1038/s41390-021-01663-8

Hoehndorf, R., Schofield, P. N., & Gkoutos, G. V. (2015). The role of ontologies in biological and biomedical research: A functional perspective. *Briefings in Bioinformatics*, *16*(6), 1069–1080. https://doi.org/10.1093/bib/bbv011

Houtman, T. A., Eckermann, H. A., Smidt, H., & de Weerth, C. (2022). Gut microbiota and BMI throughout childhood: The role of firmicutes, bacteroidetes, and short-chain fatty acid producers. *Scientific Reports*, *12*(1), 3140. https://doi.org/10.1038/s41598-022-07176-6

Imhann, F., Vich Vila, A., Bonder, M. J., Lopez Manosalva, A. G., Koonen, D. P. Y., Fu, J., Wijmenga, C., Zhernakova, A., & Weersma, R. K. (2017). The influence of proton pump inhibitors and other commonly used medication on the gut microbiota. *Gut Microbes*, *8*(4), 351–358. https://doi.org/10.1080/19490976.2017.1284732

Kalliomäki, M., Kirjavainen, P., Eerola, E., Kero, P., Salminen, S., & Isolauri, E. (2001). Distinct patterns of neonatal gut microflora in infants in whom atopy was and was not developing. *Journal of Allergy and Clinical Immunology*, *107*(1), 129–134. https://doi.org/10.1067/mai.2001.111237

Kassem, Z., Sitarik, A., Levin, A. M., Lynch, S. V., Havstad, S., Fujimura, K., Kozyrskyj, A., Ownby, D. R., Johnson, C. C., Yong, G. J. M., Wegienka, G., & Cassidy-Bushrow, A. E. (2020). Maternal and cord blood vitamin D level and the infant gut microbiota in a birth cohort study. *Maternal Health, Neonatology and Perinatology*, *6*(1), 1–10. https://doi.org/10.1186/s40748-020-00119-x

Kelly, J. R., Borre, Y., O' Brien, C., Patterson, E., El Aidy, S., Deane, J., Kennedy, P. J., Beers, S., Scott, K., Moloney, G., Hoban, A. E., Scott, L., Fitzgerald, P., Ross, P., Stanton, C., Clarke, G., Cryan, J. F., & Dinan, T. G. (2016). Transferring the blues: Depression-associated gut microbiota induces neurobehavioural changes in the rat. *Journal of Psychiatric Research*, *82*, 109–118. https://doi.org/10.1016/j.jpsychires.2016.07.019

Khan, M. J., Gerasimidis, K., Edwards, C. A., & Shaikh, M. G. (2016). Role of Gut Microbiota in the Aetiology of Obesity: Proposed Mechanisms and Review of the Literature. *Journal of Obesity*, *2016*, e7353642. https://doi.org/10.1155/2016/7353642

Kho, Z. Y., & Lal, S. K. (2018). The Human Gut Microbiome – A Potential Controller of Wellness and Disease. *Frontiers in Microbiology*, *9*. https://www.frontiersin.org/article/10.3389/fmicb.2018.01835

King, C. H., Desai, H., Sylvetsky, A. C., LoTempio, J., Ayanyan, S., Carrie, J., Crandall, K. A., Fochtman, B. C., Gasparyan, L., Gulzar, N., Howell, P., Issa, N., Krampis, K., Mishra, L., Morizono, H., Pisegna, J. R., Rao, S., Ren, Y., Simonyan, V., … Mazumder, R. (2019). Baseline human gut microbiota profile in healthy people and standard reporting template. *PLOS ONE*, *14*(9), e0206484. https://doi.org/10.1371/journal.pone.0206484

Köhler, S., Vasilevsky, N. A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S. M., Boerkoel, C. F., Boycott, K. M., Brudno, M., Buske, O. J., Chinnery, P. F., Cipriani, V., Connell, L. E., Dawkins, H. J. S., DeMare, L. E., Devereau, A. D., de Vries, B. B. A., … Robinson, P. N. (2017). The Human Phenotype Ontology in 2017. *Nucleic Acids Research*, *45*(D1), D865–D876. https://doi.org/10.1093/nar/gkw1039

Koliada, A., Syzenko, G., Moseiko, V., Budovska, L., Puchkov, K., Perederiy, V., Gavalko, Y., Dorofeyev, A., Romanenko, M., Tkach, S., Sineok, L., Lushchak, O., & Vaiserman, A. (2017). Association between body mass index and Firmicutes/Bacteroidetes ratio in an adult Ukrainian population. *BMC Microbiology*, *17*(1), 120. https://doi.org/10.1186/s12866-017-1027-1

Kourou, K. D., Pezoulas, V. C., Georga, E. I., Exarchos, T. P., Tsanakas, P., Tsiknakis, M., Varvarigou, T., De Vita, S., Tzioufas, A., & Fotiadis, D. I. (2019). Cohort Harmonization and Integrative Analysis From a Biomedical Engineering Perspective. *IEEE Reviews in Biomedical Engineering*, *12*, 303–318. https://doi.org/10.1109/RBME.2018.2855055

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*, 1–26. https://doi.org/10.18637/jss.v028.i05

Lawson, J., Cabili, M. N., Kerry, G., Boughtwood, T., Thorogood, A., Alper, P., Bowers, S. R., Boyles, R. R., Brookes, A. J., Brush, M., Burdett, T., Clissold, H., Donnelly, S., Dyke, S. O. M., Freeberg, M. A., Haendel, M. A., Hata, C., Holub, P., Jeanson, F., … Courtot, M. (2021). The Data Use Ontology to streamline responsible access to human biomedical datasets. *Cell Genomics*, *1*(2), 100028. https://doi.org/10.1016/j.xgen.2021.100028

Lee, G. C., Reveles, K. R., Attridge, R. T., Lawson, K. A., Mansi, I. A., Lewis, J. S., & Frei, C. R. (2014). Outpatient antibiotic prescribing in the United States: 2000 to 2010. *BMC Medicine*, *12*, 96. https://doi.org/10.1186/1741-7015-12-96

Ley, R. E., Turnbaugh, P. J., Klein, S., & Gordon, J. I. (2006). Human gut microbes associated with obesity. *Nature*, *444*(7122), 1022–1023. https://doi.org/10.1038/4441022a

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Litvak, Y., Byndloss, M. X., & Bäumler, A. J. (2018). Colonocyte metabolism shapes the gut microbiota. *Science (New York, N.Y.)*, *362*(6418), eaat9076. https://doi.org/10.1126/science.aat9076

Macfarlane, G. T., & Englyst, H. N. (1986). Starch utilization by the human large intestinal microflora. *Journal of Applied Bacteriology*, *60*(3), 195–201. https://doi.org/10.1111/j.1365-2672.1986.tb01073.x

Machiels, K., Joossens, M., Sabino, J., Preter, V. D., Arijs, I., Eeckhaut, V., Ballet, V., Claes, K., Immerseel, F. V., Verbeke, K., Ferrante, M., Verhaegen, J., Rutgeerts, P., & Vermeire, S. (2014). A decrease of the butyrate-producing species Roseburia hominis and Faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis. *Gut*, *63*(8), 1275–1283. https://doi.org/10.1136/gutjnl-2013-304833

Magne, F., Gotteland, M., Gauthier, L., Zazueta, A., Pesoa, S., Navarrete, P., & Balamurugan, R. (2020). The Firmicutes/Bacteroidetes Ratio: A Relevant Marker of Gut Dysbiosis in Obese Patients? *Nutrients*, *12*(5), 1474. https://doi.org/10.3390/nu12051474

Mansouri, M., Khakabimamaghani, S., Chindelevitch, L., & Ester, M. (2022). Aristotle: Stratified causal discovery for omics data. *BMC Bioinformatics*, *23*(1), 42. https://doi.org/10.1186/s12859-021-04521-w

Marra, F., Monnet, D. L., Patrick, D. M., Chong, M., Brandt, C. T., Winters, M., Kaltoft, M. S., Tyrrell, G. J., Lovgren, M., & Bowie, W. R. (2007). A comparison of antibiotic use in children between Canada and Denmark. *The Annals of Pharmacotherapy*, *41*(4), 659–666. https://doi.org/10.1345/aph.1H293

Matamoros, S., Gras-Leguen, C., Le Vacon, F., Potel, G., & de La Cochetiere, M.-F. (2013). Development of intestinal microbiota in infants and its impact on health. *Trends in Microbiology*, *21*(4), 167–173. https://doi.org/10.1016/j.tim.2012.12.001

McBurney, M. I., Davis, C., Fraser, C. M., Schneeman, B. O., Huttenhower, C., Verbeke, K., Walter, J., & Latulippe, M. E. (n.d.). Establishing What Constitutes a Healthy Human Gut Microbiome: State of the Science, Regulatory Considerations, and Future Directions. *The Journal of Nutrition*. https://doi.org/10.1093/jn/nxz154

Messer, J. S., & Chang, E. B. (2018). Chapter 36—Microbial Physiology of the Digestive Tract and Its Role in Inflammatory Bowel Diseases. In H. M. Said (Ed.), *Physiology of the Gastrointestinal Tract (Sixth Edition)* (pp. 795–810). Academic Press. https://doi.org/10.1016/B978-0-12-809954-4.00036-0

Milani, C., Duranti, S., Bottacini, F., Casey, E., Turroni, F., Mahony, J., Belzer, C., Palacio, S. D., Montes, S. A., Mancabelli, L., Lugli, G. A., Rodriguez, J. M., Bode, L., Vos, W. de, Gueimonde, M., Margolles, A., Sinderen, D. van, & Ventura, M. (2017). The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. *Microbiology and Molecular Biology Reviews*, *81*(4). https://doi.org/10.1128/MMBR.00036-17

Moore, R. E., & Townsend, S. D. (2019). Temporal development of the infant gut microbiome. *Open Biology*, *9*(9), 190128. https://doi.org/10.1098/rsob.190128

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*. https://www.frontiersin.org/articles/10.3389/fnbot.2013.00021

Nguyen, Q. N., Himes, J. E., Martinez, D. R., & Permar, S. R. (2016). The Impact of the Gut Microbiota on Humoral Immunity to Pathogens and Vaccination in Early Infancy. *PLoS Pathogens*, *12*(12). https://doi.org/10.1371/journal.ppat.1005997

Noureldein, M. H., & Eid, A. A. (2018). Gut microbiota and mTOR signaling: Insight on a new pathophysiological interaction. *Microbial Pathogenesis*, *118*, 98–104. https://doi.org/10.1016/j.micpath.2018.03.021

Odamaki, T., Kato, K., Sugahara, H., Hashikura, N., Takahashi, S., Xiao, J., Abe, F., & Osawa, R. (2016). Age-related changes in gut microbiota composition from newborn to centenarian: A cross-sectional study. *BMC Microbiology*, *16*(1), 90. https://doi.org/10.1186/s12866-016-0708-5

Olson, K. L., & Mandl, K. D. (2012). Temporal Patterns of Medications Dispensed to Children and Adolescents in a National Insured Population. *PLOS ONE*, *7*(7), e40991. https://doi.org/10.1371/journal.pone.0040991

Ottman, N., Smidt, H., de Vos, W. M., & Belzer, C. (2012a). The function of our microbiota: Who is out there and what do they do? *Frontiers in Cellular and Infection Microbiology*, *2*, 104. https://doi.org/10.3389/fcimb.2012.00104

Ottman, N., Smidt, H., de Vos, W. M., & Belzer, C. (2012b). The function of our microbiota: Who is out there and what do they do? *Frontiers in Cellular and Infection Microbiology*, *2*, 104. https://doi.org/10.3389/fcimb.2012.00104

Parada Venegas, D., De la Fuente, M. K., Landskron, G., González, M. J., Quera, R., Dijkstra, G., Harmsen, H. J. M., Faber, K. N., & Hermoso, M. A. (2019). Short Chain Fatty Acids (SCFAs)-Mediated Gut Epithelial and Immune Regulation and Its Relevance for Inflammatory Bowel Diseases. *Frontiers in Immunology*, *10*. https://www.frontiersin.org/article/10.3389/fimmu.2019.00277

Patrick, D. M., Sbihi, H., Dai, D. L. Y., Al Mamun, A., Rasali, D., Rose, C., Marra, F., Boutin, R. C. T., Petersen, C., Stiemsma, L. T., Winsor, G. L., Brinkman, F. S. L., Kozyrskyj, A. L., Azad, M. B., Becker, A. B., Mandhane, P. J., Moraes, T. J., Sears, M. R., Subbarao, P., … Turvey, S. E. (2020). Decreasing antibiotic use, the gut microbiota, and asthma incidence in children: Evidence from population-based and prospective cohort studies. *The Lancet Respiratory Medicine*. https://doi.org/10.1016/S2213-2600(20)30052-7

Png, C. W., Lindén, S. K., Gilshenan, K. S., Zoetendal, E. G., McSweeney, C. S., Sly, L. I., McGuckin, M. A., & Florin, T. H. J. (2010). Mucolytic Bacteria With Increased Prevalence in IBD Mucosa AugmentIn VitroUtilization of Mucin by Other Bacteria. *Official Journal of the American College of Gastroenterology | ACG*, *105*(11), 2420–2428. https://doi.org/10.1038/ajg.2010.281

Rinninella, E., Raoul, P., Cintoni, M., Franceschi, F., Miggiano, G. A. D., Gasbarrini, A., & Mele, M. C. (2019). What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases. *Microorganisms*, *7*(1), 14. https://doi.org/10.3390/microorganisms7010014

Rogers, M. A. M., & Aronoff, D. M. (2016). The Influence of Nonsteroidal Anti-Inflammatory Drugs on the Gut Microbiome. *Clinical Microbiology and Infection : The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, *22*(2), 178.e1-178.e9. https://doi.org/10.1016/j.cmi.2015.10.003

Scepanovic, P., Hodel, F., Mondot, S., Partula, V., Byrd, A., Hammer, C., Alanio, C., Bergstedt, J., Patin, E., Touvier, M., Lantz, O., Albert, M. L., Duffy, D., Quintana-Murci, L., Fellay, J., Abel, L., Alcover, A., Aschard, H., Astrom, K., … The Milieu Intérieur Consortium. (2019). A comprehensive assessment of demographic, environmental, and host genetic associations with gut microbiome diversity in healthy individuals. *Microbiome*, *7*(1), 130. https://doi.org/10.1186/s40168-019-0747-x

Schmidt, T. S. B., Raes, J., & Bork, P. (2018). The Human Gut Microbiome: From Association to Modulation. *Cell*, *172*(6), 1198–1215. https://doi.org/10.1016/j.cell.2018.02.044

Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., Felix, V., Jeng, L., Bearer, C., Lichenstein, R., Bisordi, K., Campion, N., Hyman, B., Kurland, D., Oates, C. P., Kibbey, S., Sreekumar, P., Le, C., Giglio, M., & Greene, C. (2019). Human Disease Ontology 2018 update: Classification, content and workflow expansion. *Nucleic Acids Research*, *47*(D1), D955–D962. https://doi.org/10.1093/nar/gky1032

Servais, J., Ramage-Morin, P. L., Gal, J., & Hales, C. M. (2021). Prescription medication use among Canadian children and youth, 2012 to 2017. *Health Reports*, *32*(3), 1–16.

Shahi, S. K., Freedman, S. N., & Mangalam, A. K. (2017). Gut microbiome in multiple sclerosis: The players involved and the roles they play. *Gut Microbes*, *8*(6), 607–615. https://doi.org/10.1080/19490976.2017.1349041

Shin, N.-R., Whon, T. W., & Bae, J.-W. (2015). Proteobacteria: Microbial signature of dysbiosis in gut microbiota. *Trends in Biotechnology*, *33*(9), 496–503. https://doi.org/10.1016/j.tibtech.2015.06.011

Sioutos, N., Coronado, S. de, Haber, M. W., Hartel, F. W., Shaiu, W.-L., & Wright, L. W. (2007). NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, *40*(1), 30–43. https://doi.org/10.1016/j.jbi.2006.02.013

Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermúdez-Humarán, L. G., Gratadoux, J.-J., Blugeon, S., Bridonneau, C., Furet, J.-P., Corthier, G., Grangette, C., Vasquez, N., Pochart, P., Trugnan, G., Thomas, G., Blottière, H. M., Doré, J., Marteau, P., Seksik, P., & Langella, P. (2008). Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(43), 16731–16736. https://doi.org/10.1073/pnas.0804812105

Stout, M. J., Zhou, Y., Wylie, K. M., Tarr, P. I., Macones, G. A., & Tuuli, M. G. (2017). Early pregnancy vaginal microbiome trends and preterm birth. *American Journal of Obstetrics and Gynecology*, *217*(3), 356.e1-356.e18. https://doi.org/10.1016/j.ajog.2017.05.030

Subbarao, P., Anand, S. S., Becker, A. B., Befus, A. D., Brauer, M., Brook, J. R., Denburg, J. A., HayGlass, K. T., Kobor, M. S., Kollmann, T. R., Kozyrskyj, A. L., Lou, W. Y. W., Mandhane, P. J., Miller, G. E., Moraes, T. J., Pare, P. D., Scott, J. A., Takaro, T. K., Turvey, S. E., … Investigators, the C. S. (2015). The Canadian Healthy Infant Longitudinal Development (CHILD) Study: Examining developmental origins of allergy and asthma. *Thorax*, *70*(10), 998–1000. https://doi.org/10.1136/thoraxjnl-2015-207246

Talsness, C. E., Penders, J., Jansen, E. H. J. M., Damoiseaux, J., Thijs, C., & Mommers, M. (2017). Influence of vitamin D on key bacterial taxa in infant microbiota in the KOALA Birth Cohort Study. *PLOS ONE*, *12*(11), e0188011. https://doi.org/10.1371/journal.pone.0188011

Tanaka, M., & Nakayama, J. (2017). Development of the gut microbiota in infancy and its impact on health in later life. *Allergology International: Official Journal of the Japanese Society of Allergology*, *66*(4), 515–522. https://doi.org/10.1016/j.alit.2017.07.010

Thompson, K., & Huntington, M. K. (2019). Methods of Symptomatic Relief of Teething in Infants and Young Children Recommended by South Dakota Physicians. *South Dakota Medicine: The Journal of the South Dakota State Medical Association*, *72*(11), 509–512.

Thursby, E., & Juge, N. (2017). Introduction to the human gut microbiota. *Biochemical Journal*, *474*(11), 1823–1836. https://doi.org/10.1042/BCJ20160510

Tian, L., Wang, X.-W., Wu, A.-K., Fan, Y., Friedman, J., Dahlin, A., Waldor, M. K., Weinstock, G. M., Weiss, S. T., & Liu, Y.-Y. (2020). Deciphering functional redundancy in the human microbiome. *Nature Communications*, *11*(1), 6217. https://doi.org/10.1038/s41467-020-19940-1

Torres, J., Hu, J., Seki, A., Eisele, C., Nair, N., Huang, R., Tarassishin, L., Jharap, B., Cote-Daigneault, J., Mao, Q., Mogno, I., Britton, G. J., Uzzan, M., Chen, C.-L., Kornbluth, A., George, J., Legnani, P., Maser, E., Loudon, H., … Peter, I. (2020). Infants born to mothers with IBD present with altered gut microbiome that transfers abnormalities of the adaptive immune system to germ-free mice. *Gut*, *69*(1), 42–51. https://doi.org/10.1136/gutjnl-2018-317855

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., Egholm, M., Henrissat, B., Heath, A. C., Knight, R., & Gordon, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature*, *457*(7228), 480–484. https://doi.org/10.1038/nature07540

Underwood, M. A., Mukhopadhyay, S., Lakshminrusimha, S., & Bevins, C. L. (2020). Neonatal intestinal dysbiosis. *Journal of Perinatology*, *40*(11), 1597–1608. https://doi.org/10.1038/s41372-020-00829-2

Vaiserman, A., Romanenko, M., Piven, L., Moseiko, V., Lushchak, O., Kryzhanovska, N., Guryanov, V., & Koliada, A. (2020). Differences in the gut Firmicutes to Bacteroidetes ratio across age groups in healthy Ukrainian population. *BMC Microbiology*, *20*(1), 221. https://doi.org/10.1186/s12866-020-01903-7

Vu, K., Lou, W., Tun, H. M., Konya, T. B., Morales-Lizcano, N., Chari, R. S., Field, C. J., Guttman, D. S., Mandal, R., Wishart, D. S., Azad, M. B., Becker, A. B., Mandhane, P. J., Moraes, T. J., Lefebvre, D. L., Sears, M. R., Turvey, S. E., Subbarao, P., Scott, J. A., & Kozyrskyj, A. L. (2021). From Birth to Overweight and Atopic Disease: Multiple and Common Pathways of the Infant Gut Microbiome. *Gastroenterology*, *160*(1), 128-144.e10. https://doi.org/10.1053/j.gastro.2020.08.053

Walker, W. A. (2013). Initial Intestinal Colonization in the Human Infant and Immune Homeostasis. *Annals of Nutrition and Metabolism*, *63*(Suppl. 2), 8–15. https://doi.org/10.1159/000354907

Wang, T., Cai, G., Qiu, Y., Fei, N., Zhang, M., Pang, X., Jia, W., Cai, S., & Zhao, L. (2012). Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *The ISME Journal*, *6*(2), 320–329. https://doi.org/10.1038/ismej.2011.109

Wei, S., Bahl, M. I., Baunwall, S. M. D., Hvas, C. L., & Licht, T. R. (2021). Determining Gut Microbial Dysbiosis: A Review of Applied Indexes for Assessment of Intestinal Microbiota Imbalances. *Applied and Environmental Microbiology*, *87*(11), e00395-21. https://doi.org/10.1128/AEM.00395-21

Willing, B. P., Dicksved, J., Halfvarson, J., Andersson, A. F., Lucio, M., Zheng, Z., Järnerot, G., Tysk, C., Jansson, J. K., & Engstrand, L. (2010). A Pyrosequencing Study in Twins Shows That Gastrointestinal Microbial Profiles Vary With Inflammatory Bowel Disease Phenotypes. *Gastroenterology*, *139*(6), 1844-1854.e1. https://doi.org/10.1053/j.gastro.2010.08.049

Wu, Y., Wu, J., Chen, T., Li, Q., Peng, W., Li, H., Tang, X., & Fu, X. (2018). Fusobacterium nucleatum Potentiates Intestinal Tumorigenesis in Mice via a Toll-Like Receptor 4/p21-Activated Kinase 1 Cascade. *Digestive Diseases and Sciences*, *63*(5), 1210–1218. https://doi.org/10.1007/s10620-018-4999-2

Yang, T., Santisteban, M. M., Rodriguez, V., Li, E., Ahmari, N., Carvajal, J. M., Zadeh, M., Gong, M., Qi, Y., Zubcevic, J., Sahay, B., Pepine, C. J., Raizada, M. K., & Mohamadzadeh, M. (2015). Gut Dysbiosis Is Linked to Hypertension. *Hypertension*, *65*(6), 1331–1340. https://doi.org/10.1161/HYPERTENSIONAHA.115.05315

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. A., Lauber, C., Clemente, J. C., Knights, D., … Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, *486*(7402), 222–227. https://doi.org/10.1038/nature11053

ZOOMA. (2022). [Java]. EBISPOT. https://github.com/EBISPOT/zooma (Original work published 2013)

Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., Sun, H., Xia, Y., Liang, S., Dai, Y., Wan, D., Jiang, R., Su, L., Feng, Q., Jie, Z., Guo, T., Xia, Z., Liu, C., Yu, J., … Xiao, L. (2019). 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nature Biotechnology*, *37*(2), 179–185. https://doi.org/10.1038/s41587-018-0008-8

Appendix A.

Supplemental Figures

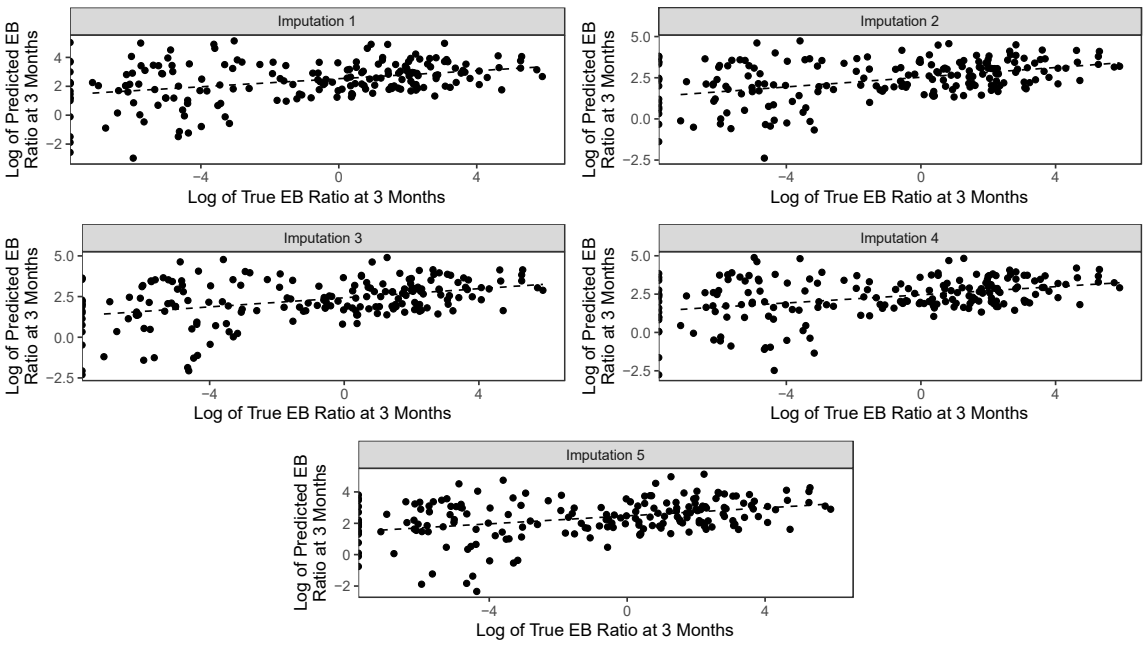

Figure A.1. Log of Predicted Enterobacteriaceae-to-Bacteroidaceae ratios vs Log of True Enterobacteriaceae-to-Bacteroidaceae ratios at 3 months using the random forest model with the variables at least 85% complete.Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibily of data across a wide range of values.

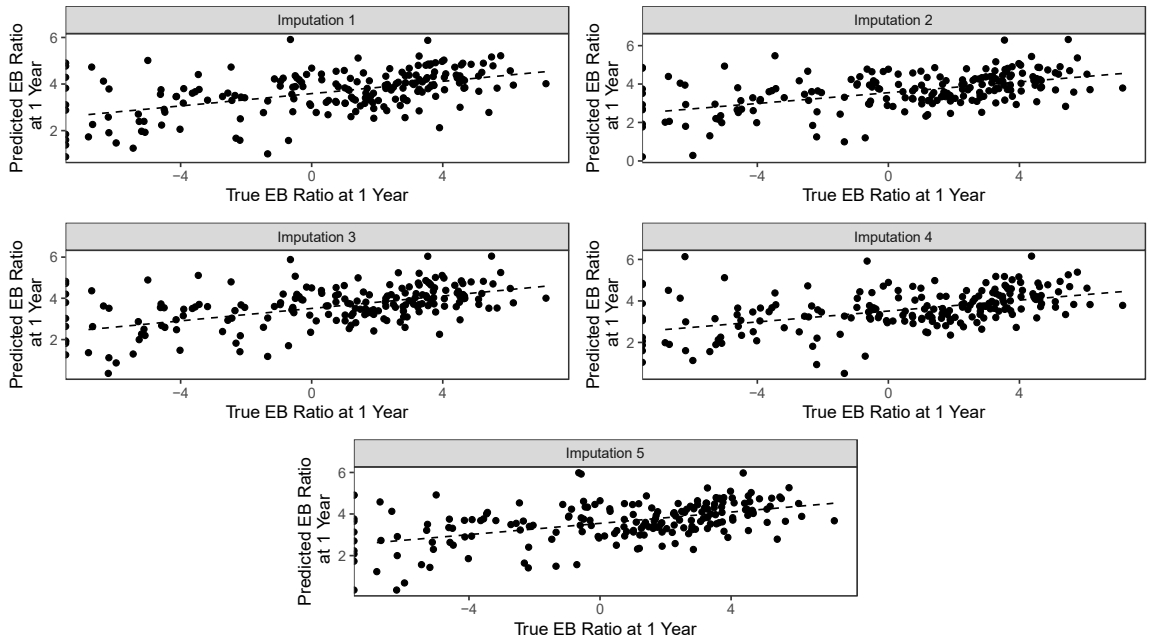

Figure A.2. Log of Predicted Enterobacteriaceae-to-Bacteroidaceae ratios vs Log of True Enterobacteriaceae-to-Bacteroidaceae ratios at 1 yearusing the random forest model with the variables at least 85% complete.Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibily of data across a wide range of values.

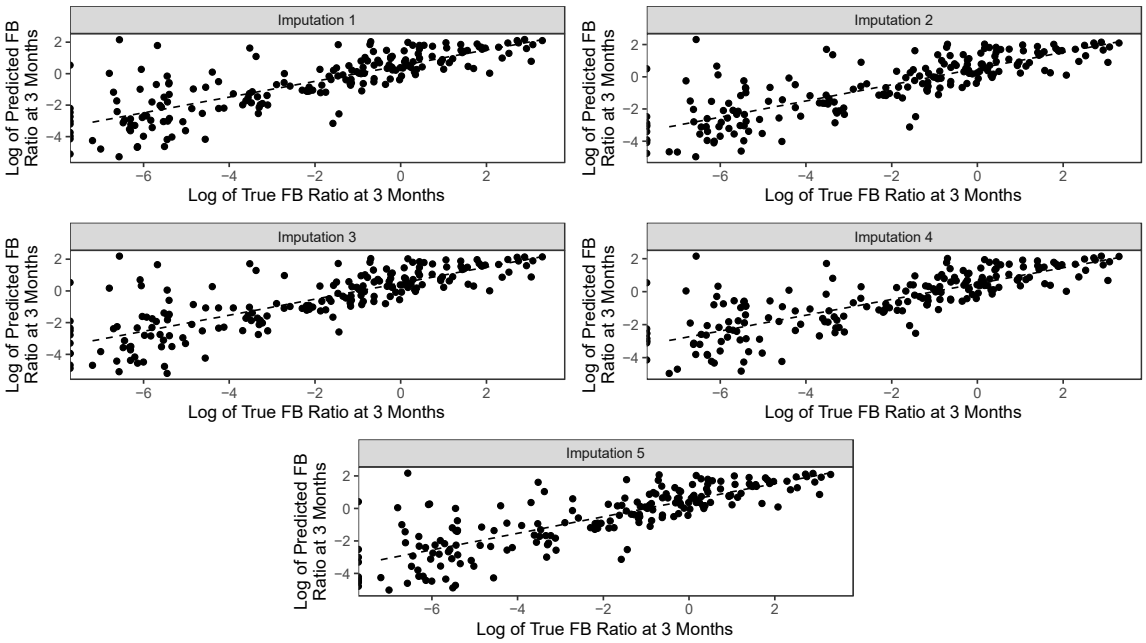

Figure A.3. Log of Predicted Firmicutes-to-Bacteroidetes ratios vs Log of True Firmicutes-to-Bacteroidetes ratios at 3 months using the random forest model with the variables at least 85% complete.Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibily of data across a wide range of values.

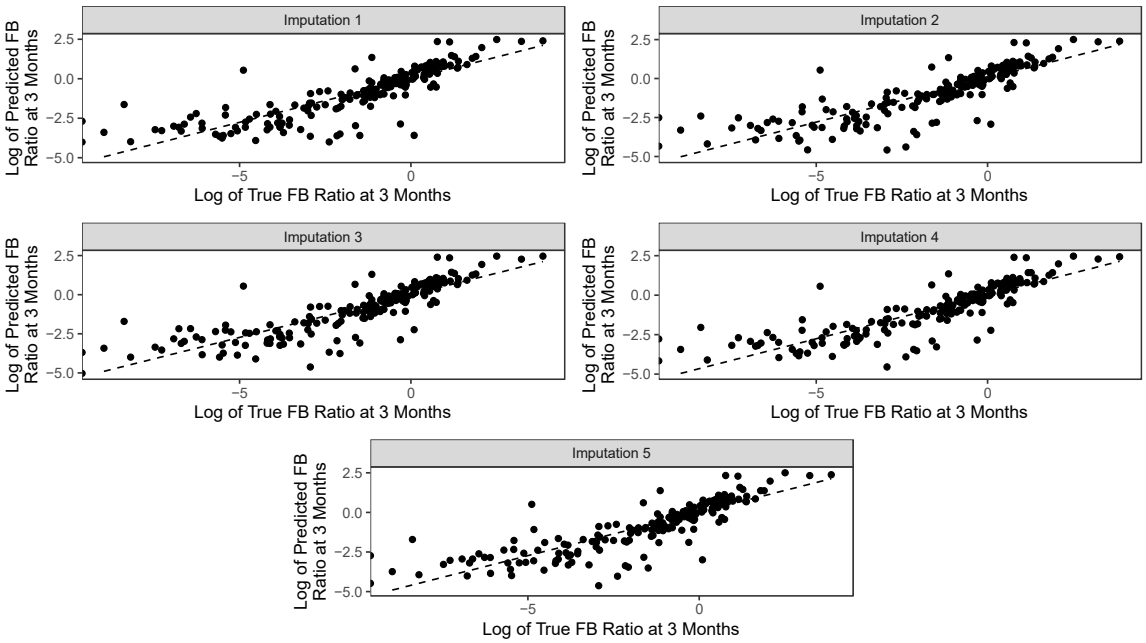

Figure A.4. Log of Predicted Firmicutes-to-Bacteroidetes ratios vs Log of True Firmicutes-to-Bacteroidetes ratios at 1 yearusing the random forest model with the variables at least 85% complete.Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibily of data across a wide range of values.

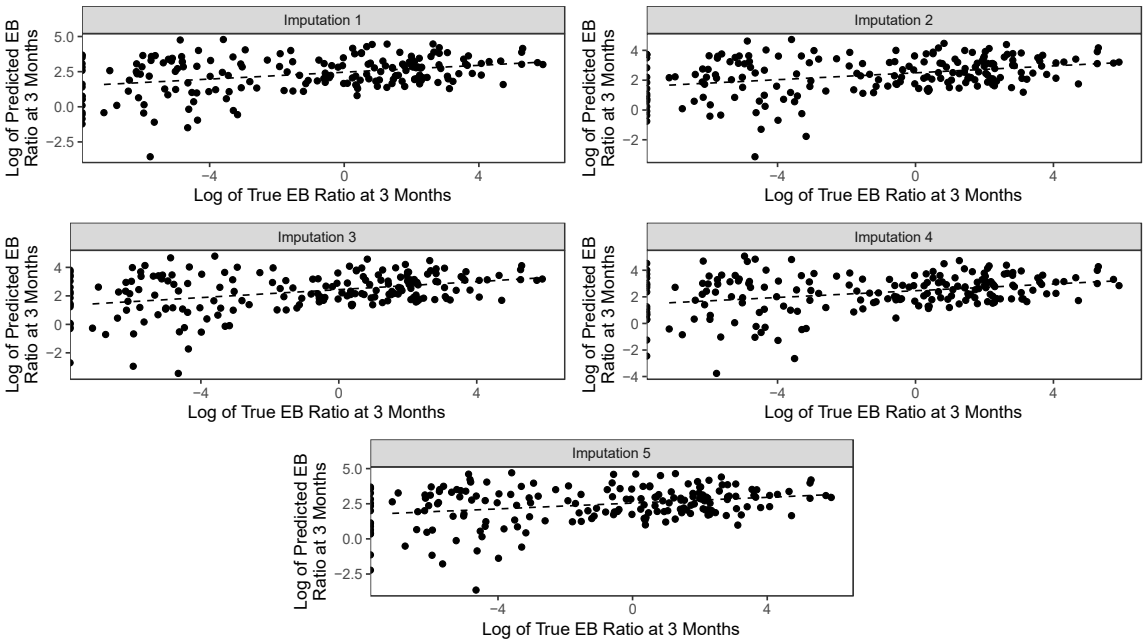

Figure A.5. Log of Predicted Enterobacteriaceae-to-Bacteroidaceae ratios vs Log of True Enterobacteriaceae-to-Bacteroidaceae ratios at 3 months using the random forest model with the variables at least 90% complete.Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibily of data across a wide range of values.

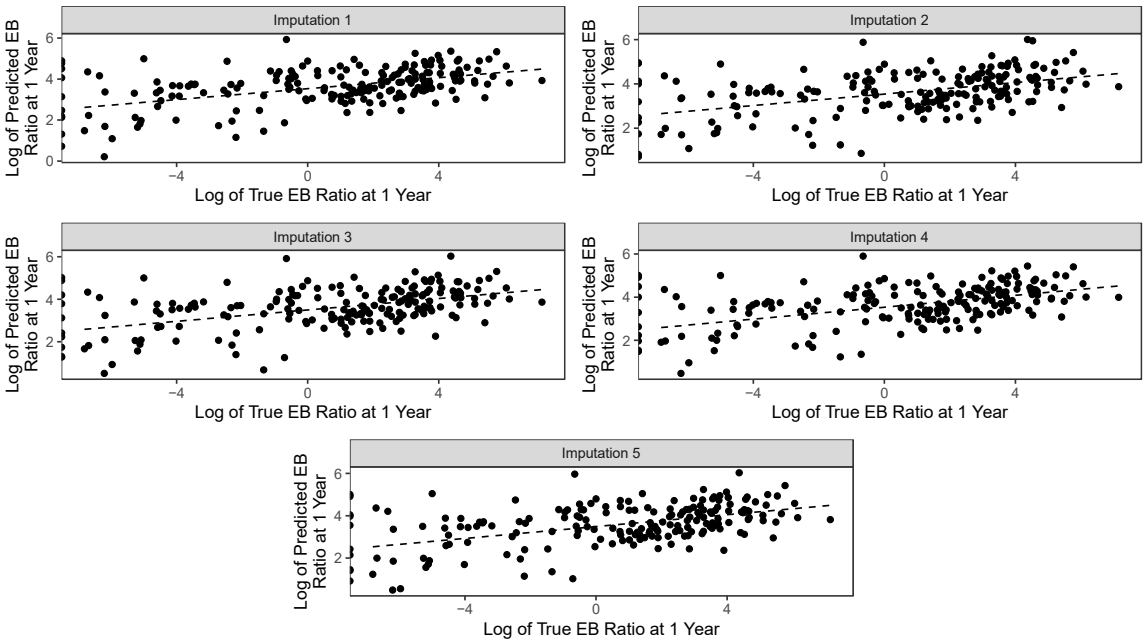

Figure A.6. Log of Predicted Enterobacteriaceae-to-Bacteroidaceae ratios vs Log of True Enterobacteriaceae-to-Bacteroidaceae ratios at 1 yearusing the random forest model with the variables at least 90% complete.Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibily of data across a wide range of values.

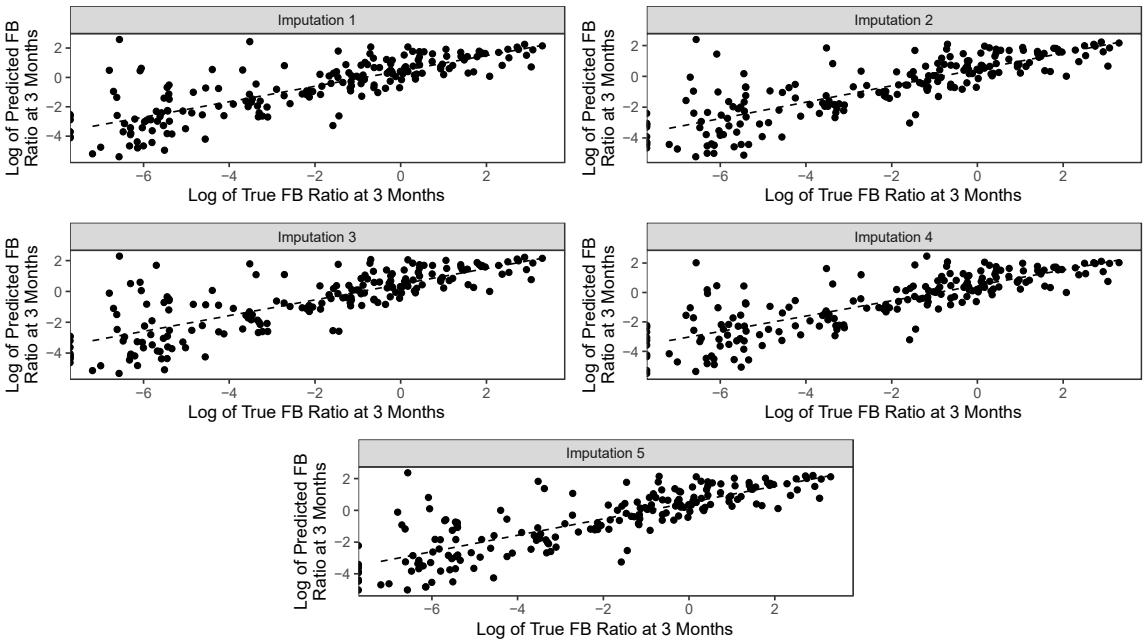

Figure A.7. Log of Predicted Firmicutes-to-Bacteroidetes ratios vs Log of True Firmicutes-to-Bacteroidetes ratios at 3 months using the random forest model with the variables at least 90% complete.Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibily of data across a wide range of values.

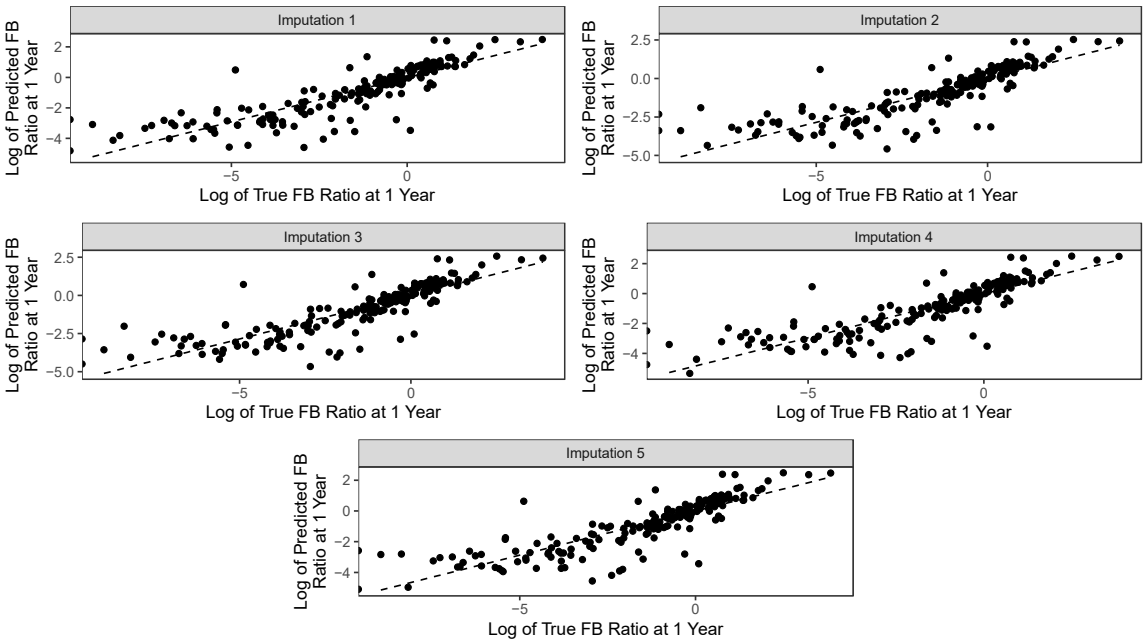

Figure A.8. Log of Predicted Firmicutes-to-Bacteroidetes ratios vs Log of True Firmicutes-to-Bacteroidetes ratios at 1 yearusing the random forest model with the variables at least 90% complete.Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibily of data across a wide range of values.

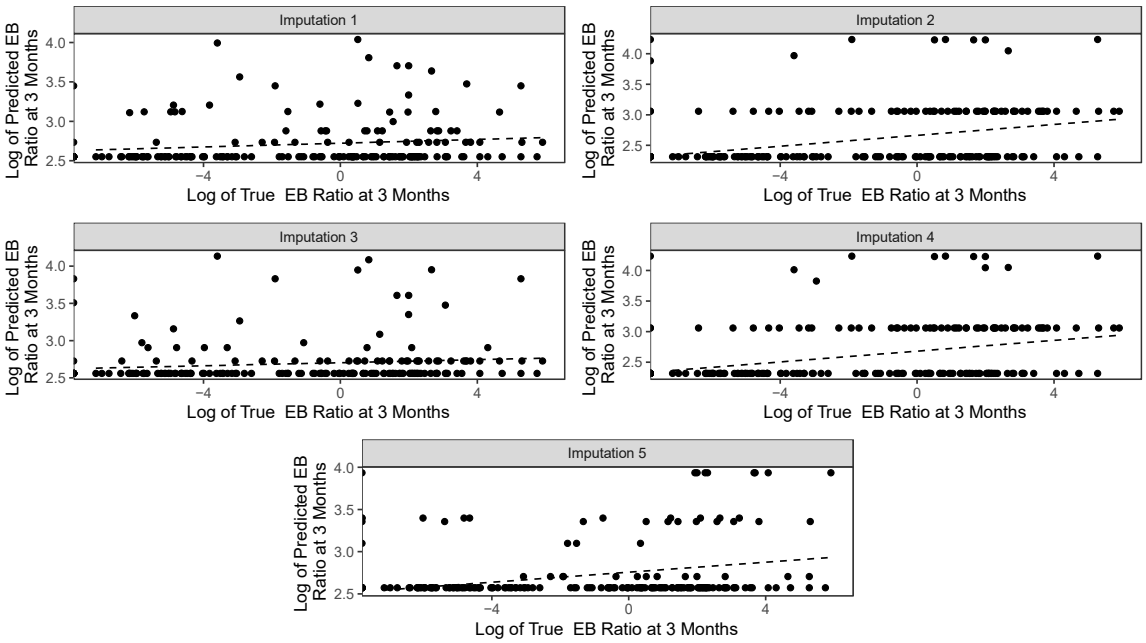

Figure A.9. Log of Predicted Enterobacteriaceae-to-Bacteroidaceae ratios vs Log of True Enterobacteriaceae-to-Bacteroidaceae ratios at 3 months using the gradient boosting model with the variables at least 85% complete.Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibily of data across a wide range of values.

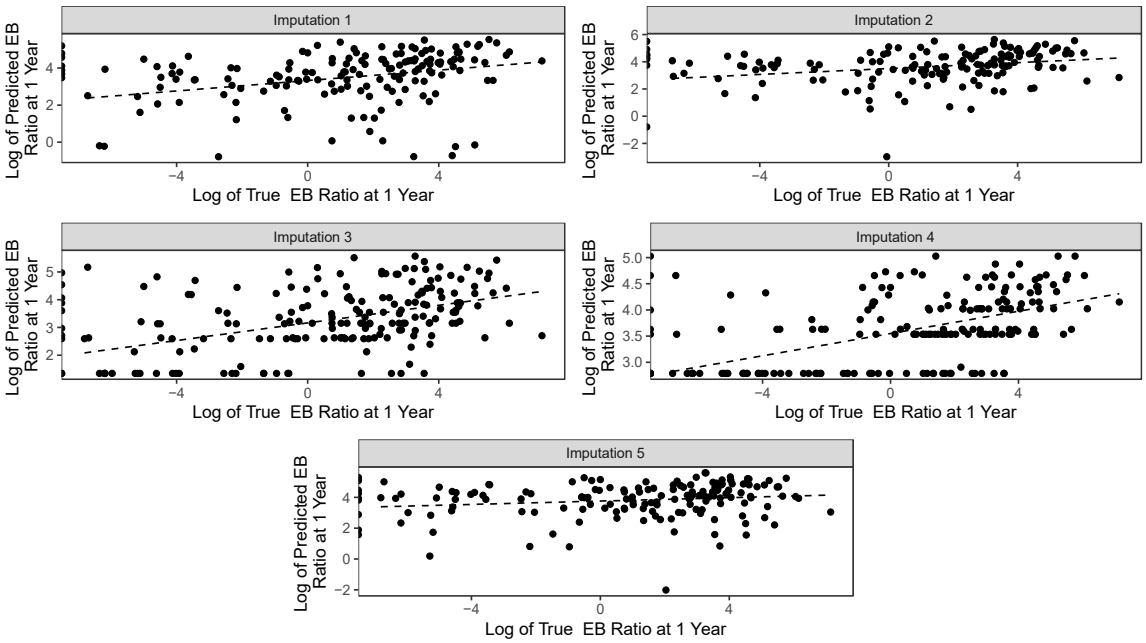

Figure A.10. Log of Predicted Enterobacteriaceae-to-Bacteroidaceae ratios vs Log of True Enterobacteriaceae-to-Bacteroidaceae ratios at 1 yearusing the gradient boosting model with the variables at least 85% complete.Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibily of data across a wide range of values.

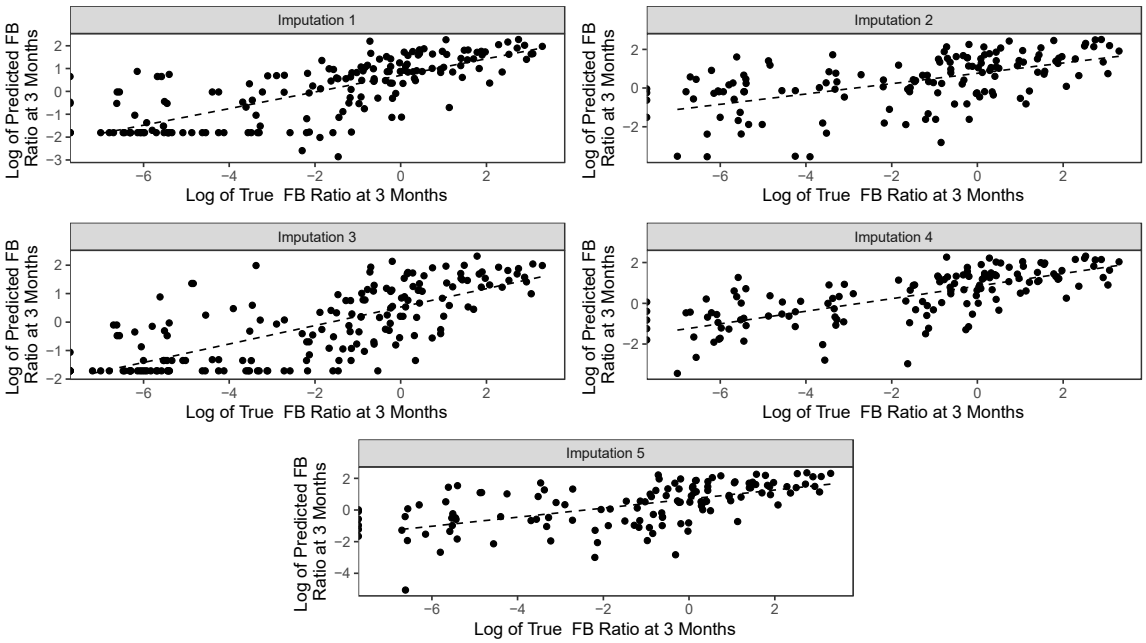

Figure A.11. Log of Predicted Firmicutes-to-Bacteroidetes ratios vs Log of True Firmicutes-to-Bacteroidetes ratios at 3 months using the gradient boosting model with the variables at least 85% complete.Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibily of data across a wide range of values.

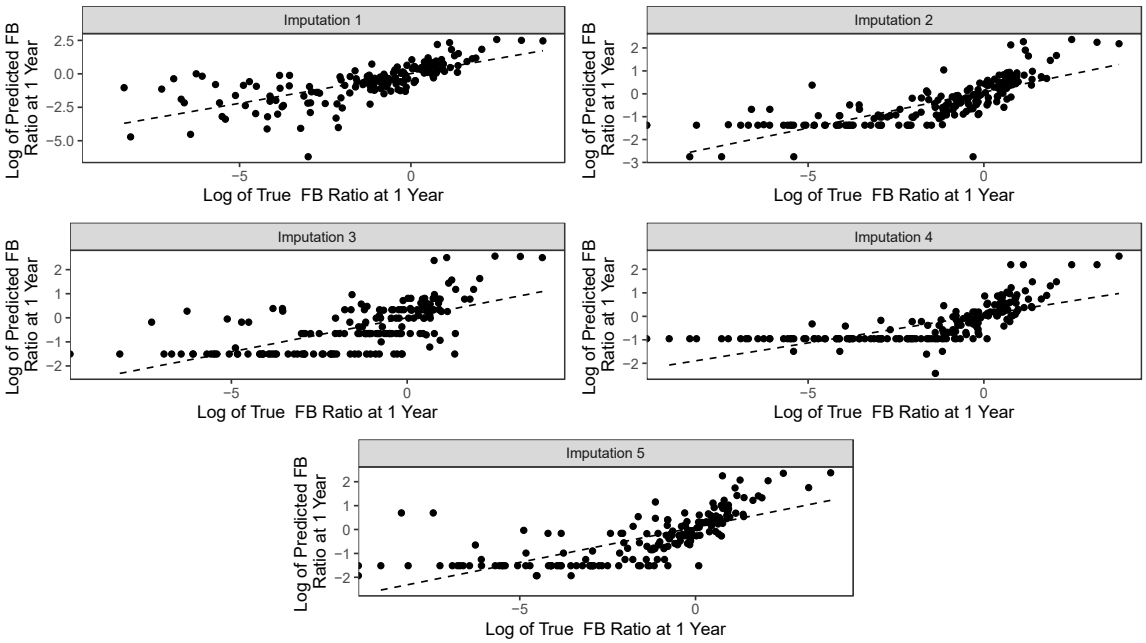

Figure A.12. Log of Predicted Firmicutes-to-Bacteroidetes ratios vs Log of True Firmicutes-to-Bacteroidetes ratios at 1 yearusing the gradient boosting model with the variables at least 85% complete.Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibily of data across a wide range of values.

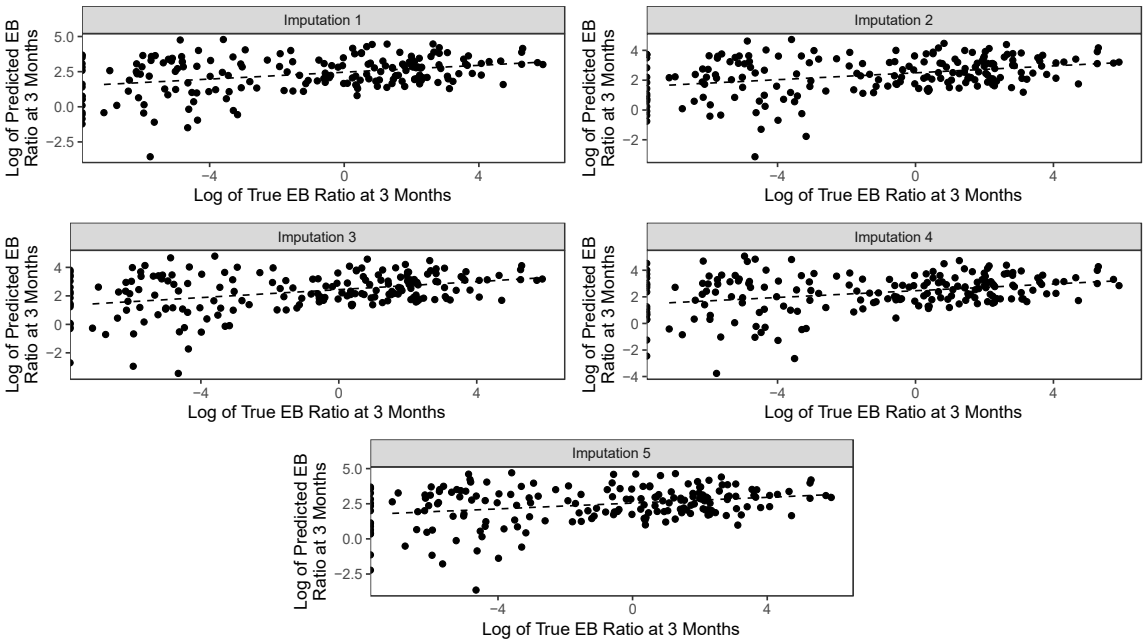

Figure A.13. Log of Predicted Enterobacteriaceae-to-Bacteroidaceae ratios vs Log of True Enterobacteriaceae-to-Bacteroidaceae ratios at 3 months using the gradient boosting model with the variables at least 90% complete.Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibily of data across a wide range of values.

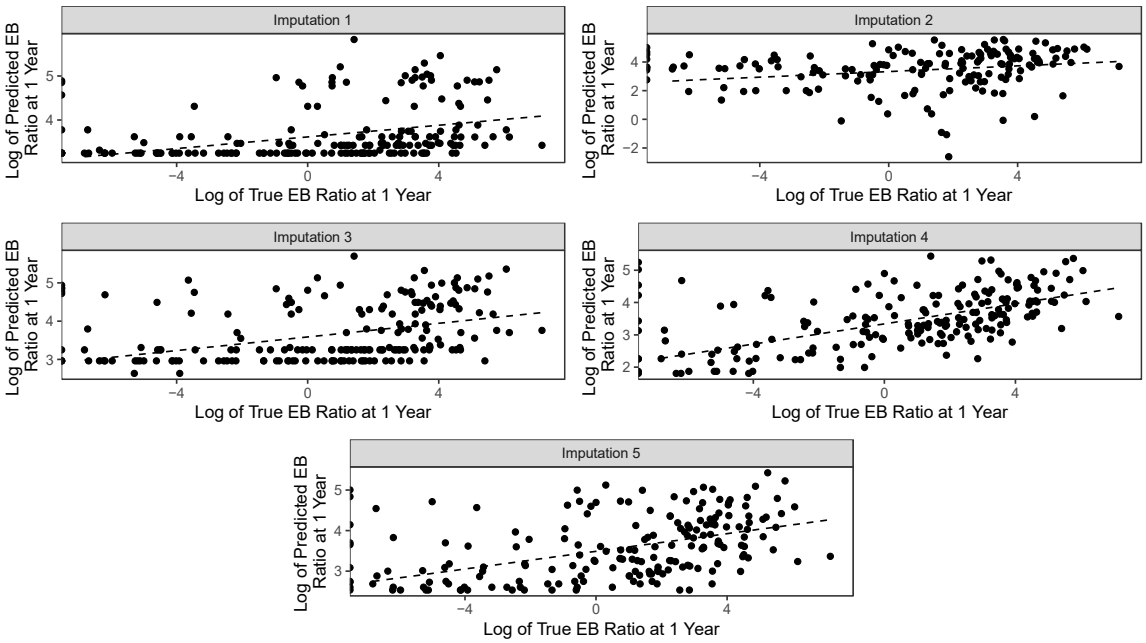

Figure A.14. Log of Predicted Enterobacteriaceae-to-Bacteroidaceae ratios vs Log of True Enterobacteriaceae-to-Bacteroidaceae ratios at 1 yearusing the gradient boosting model with the variables at least 90% complete.Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibily of data across a wide range of values.

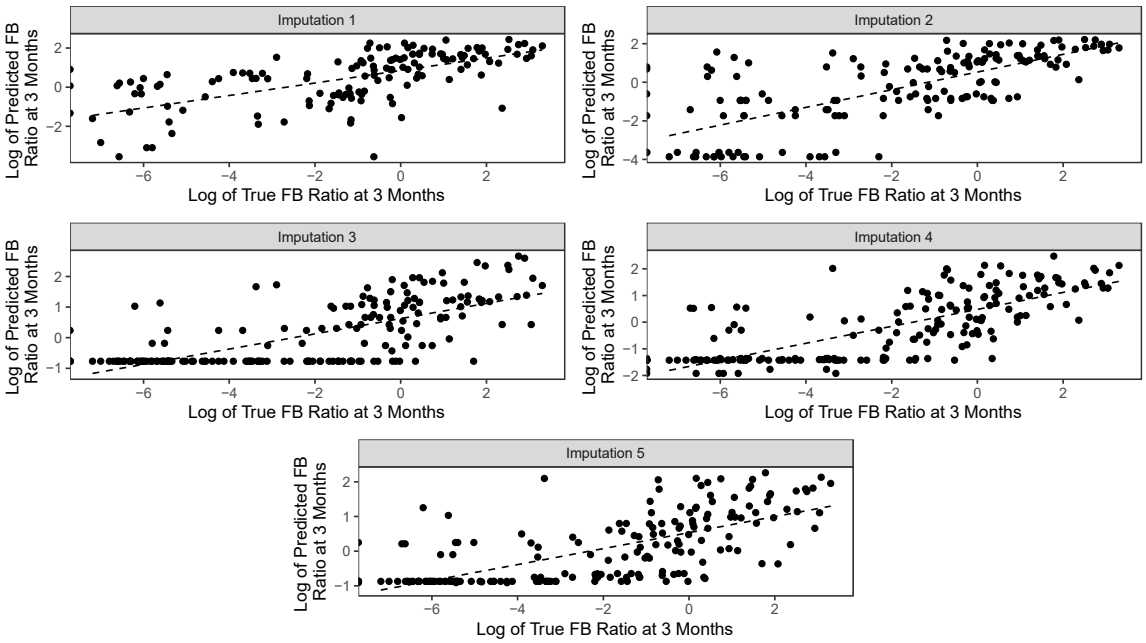

Figure A.15. Log of Predicted Firmicutes-to-Bacteroidetes ratios vs Log of True Firmicutes-to-Bacteroidetes ratios at 3 months using the gradient boosting model with the variables at least 90% complete.Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibily of data across a wide range of values.

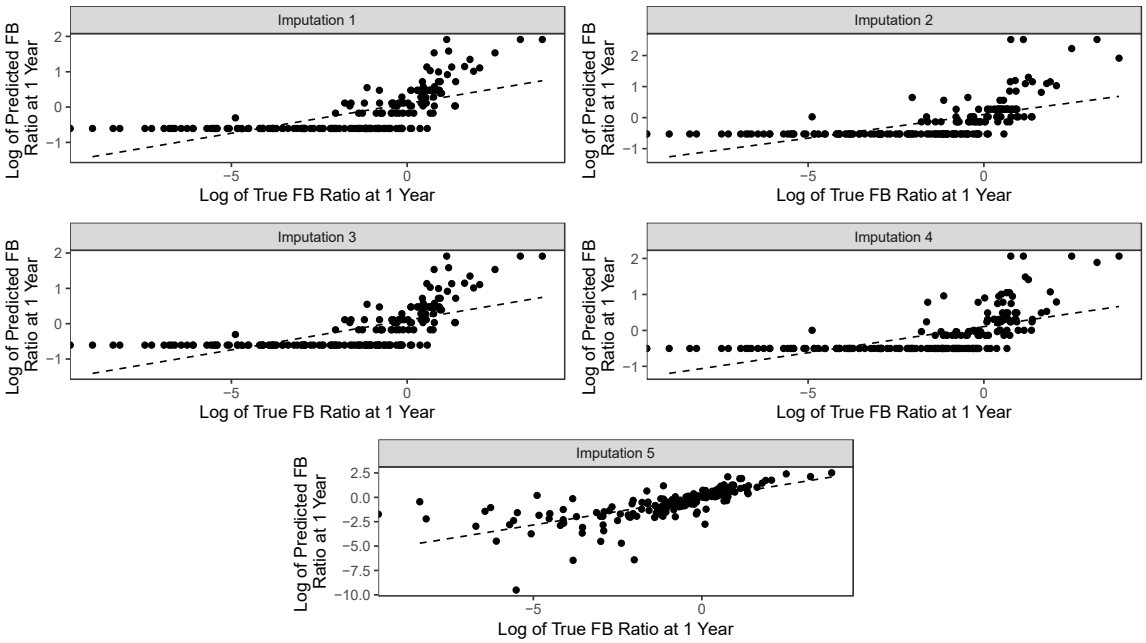

Figure A.16 **Firmicutes-to-Bacteroidetes ratios at 1 yearusing the gradient boosting model with the variables at least 90% complete.Each panel represents the predictions using one of the five imputed datasets, as indicated by the banner at the top of the panel. Log scales used to improve visibily of data across a wide range of values.**

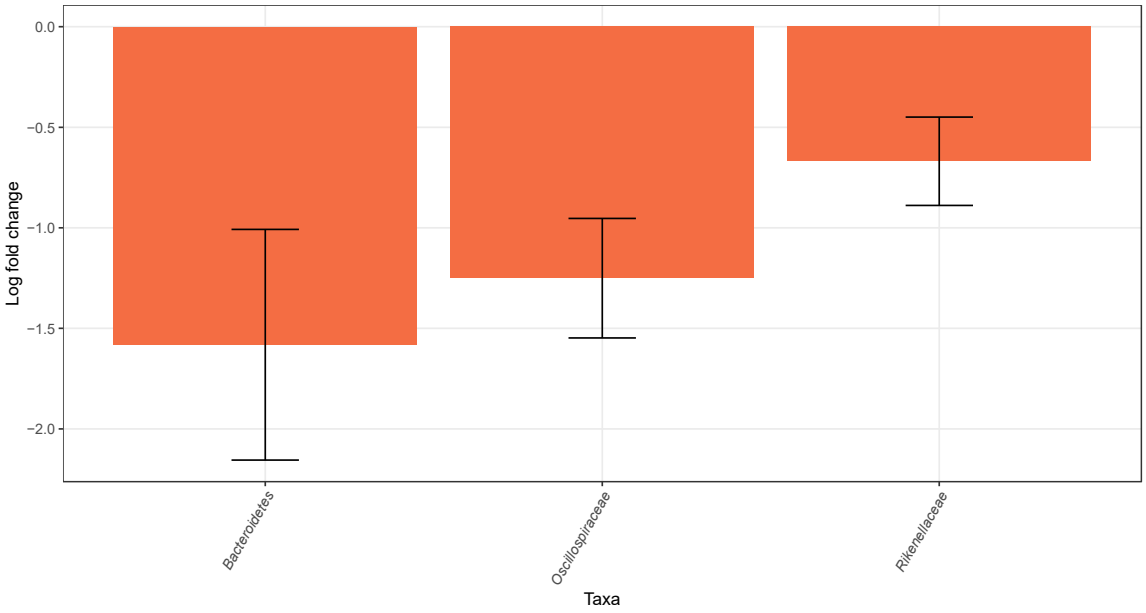

Figure A.17. **Differential Abundance for gut microbial taxa between Antimicrobial users and non users at 3 months Analysis was performed using ANCOM-BC while correcting for gender, delivery mode and visit. The log 2 fold change is shown along the Y axis, with standard error bars. and the taxa (at the family level) with differential abundance (adj p <0.05) are shown along the X axis.**

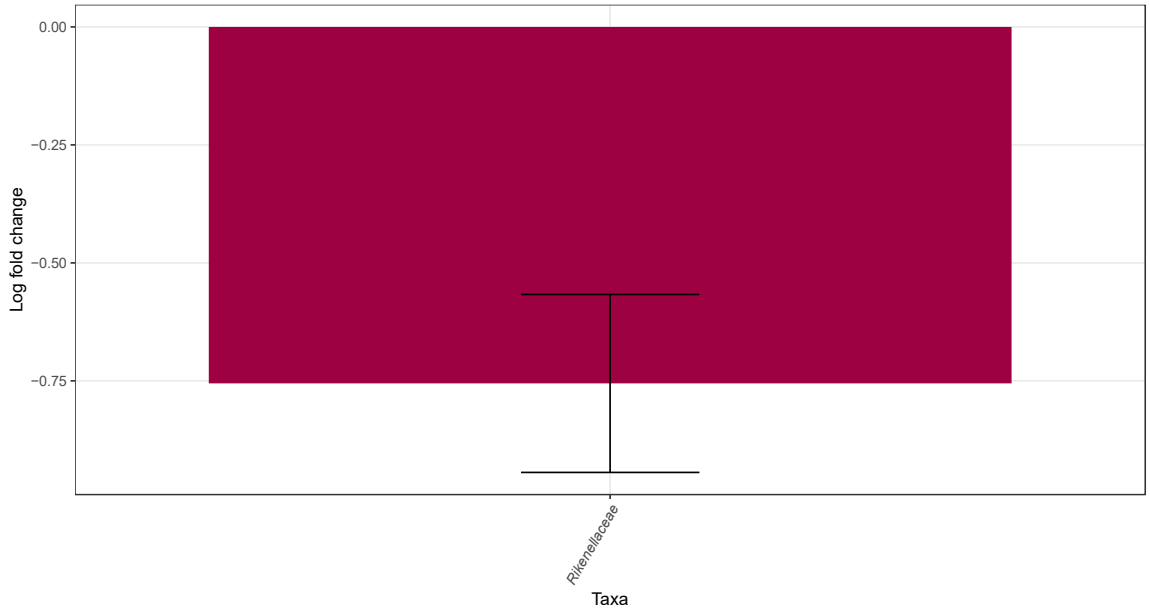

Figure A.18. Differential Abundance for gut microbial taxa between Antimicrobial users and non users at 1 year. Analysis was performed using ANCOM-BC while correcting for gender, delivery mode and visit. The log 2 fold change is shown along the Y axis, with standard error bars. and the taxa (at the family level) with differential abundance (adj p <0.05) are shown along the X axis

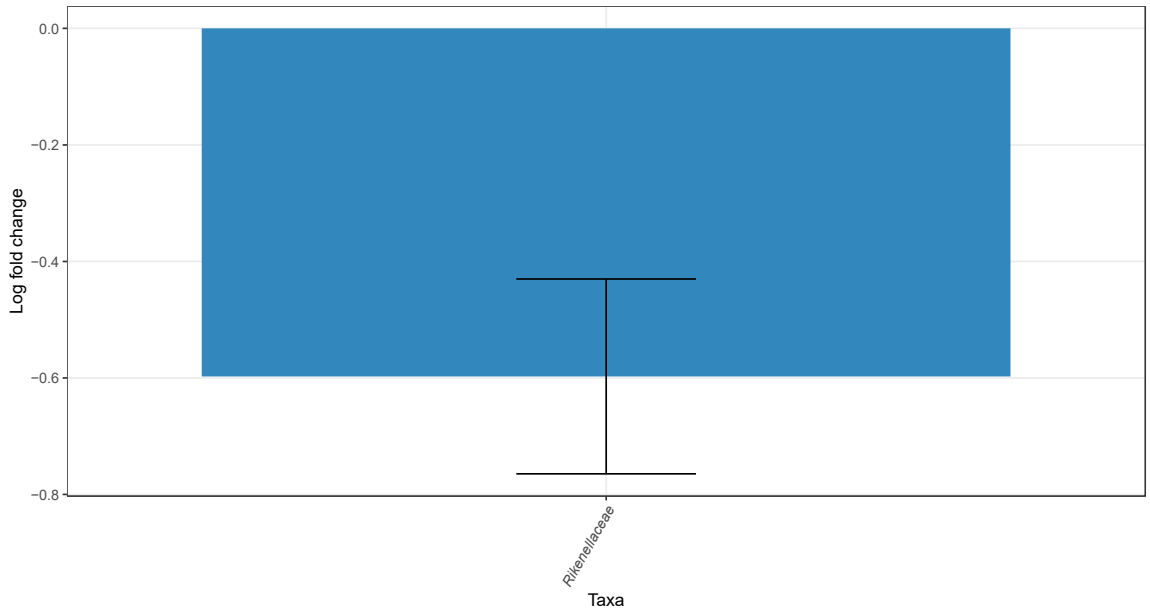

Figure A.19. Differential Abundance for gut microbial taxa between cephalosporin users and non users at 1 year. Analysis was performed using ANCOM-BC while correcting for gender, delivery mode and visit. The log 2 fold change is shown along the Y axis, with standard error bars. and the taxa (at the family level) with differential abundance (adj p <0.05) are shown along the X axis.

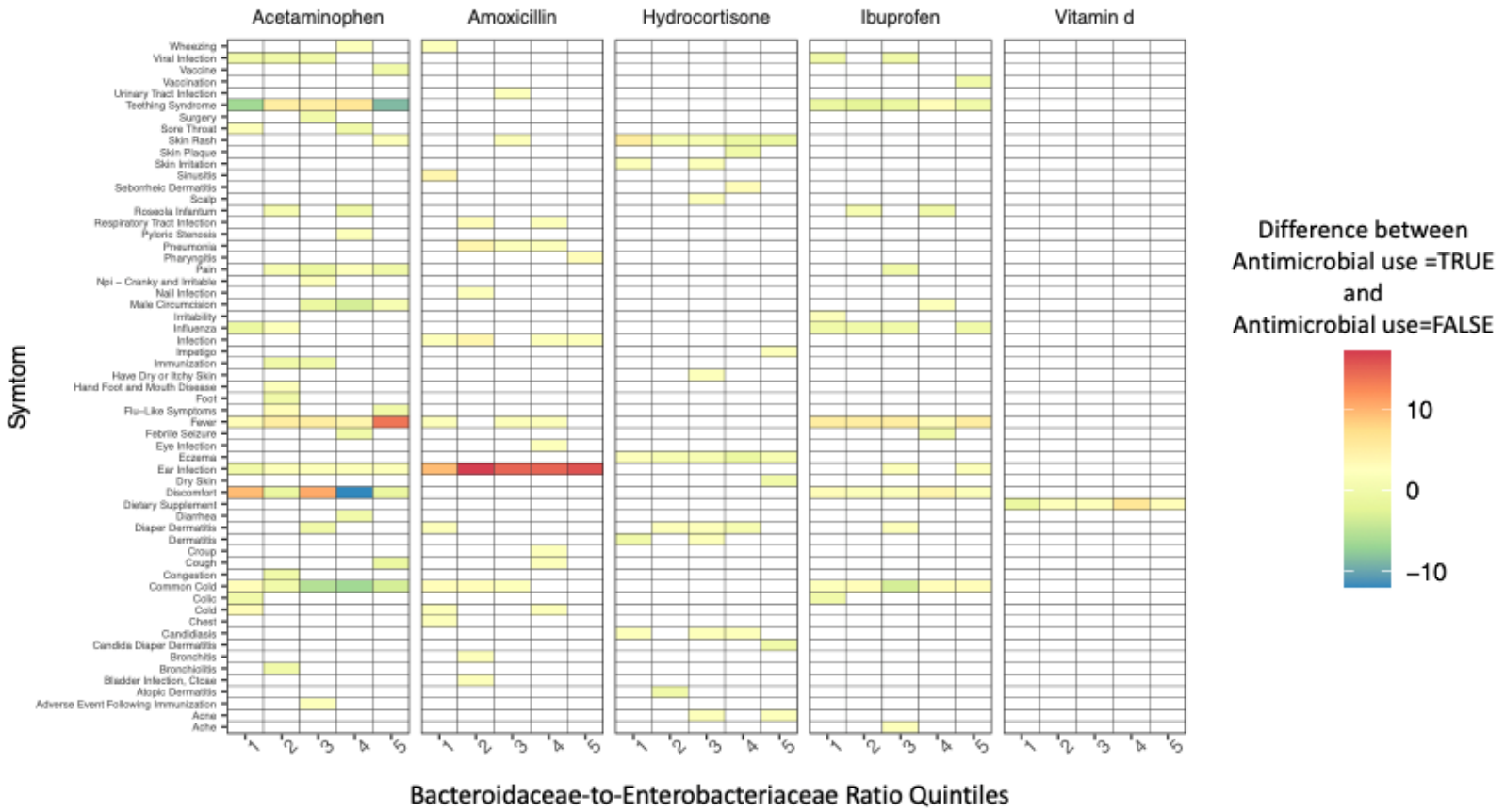

Figure A.20. Quintiles for EB_1y and medication patterns for top 5 most common medications Each facet represents a separate medication, and the rows potential reasons for usage. Within the facets, rows represent bins of subjects (n=564) based on their EB ratio at 1y. Subjects in cluster 1 have the lowest ratios, 5 the highest. Colour scale indicates which group tends to use a medication for a specificreason. Red indicates more common in the antimicrobial usage group, blue is the non antimicrobial usage group.

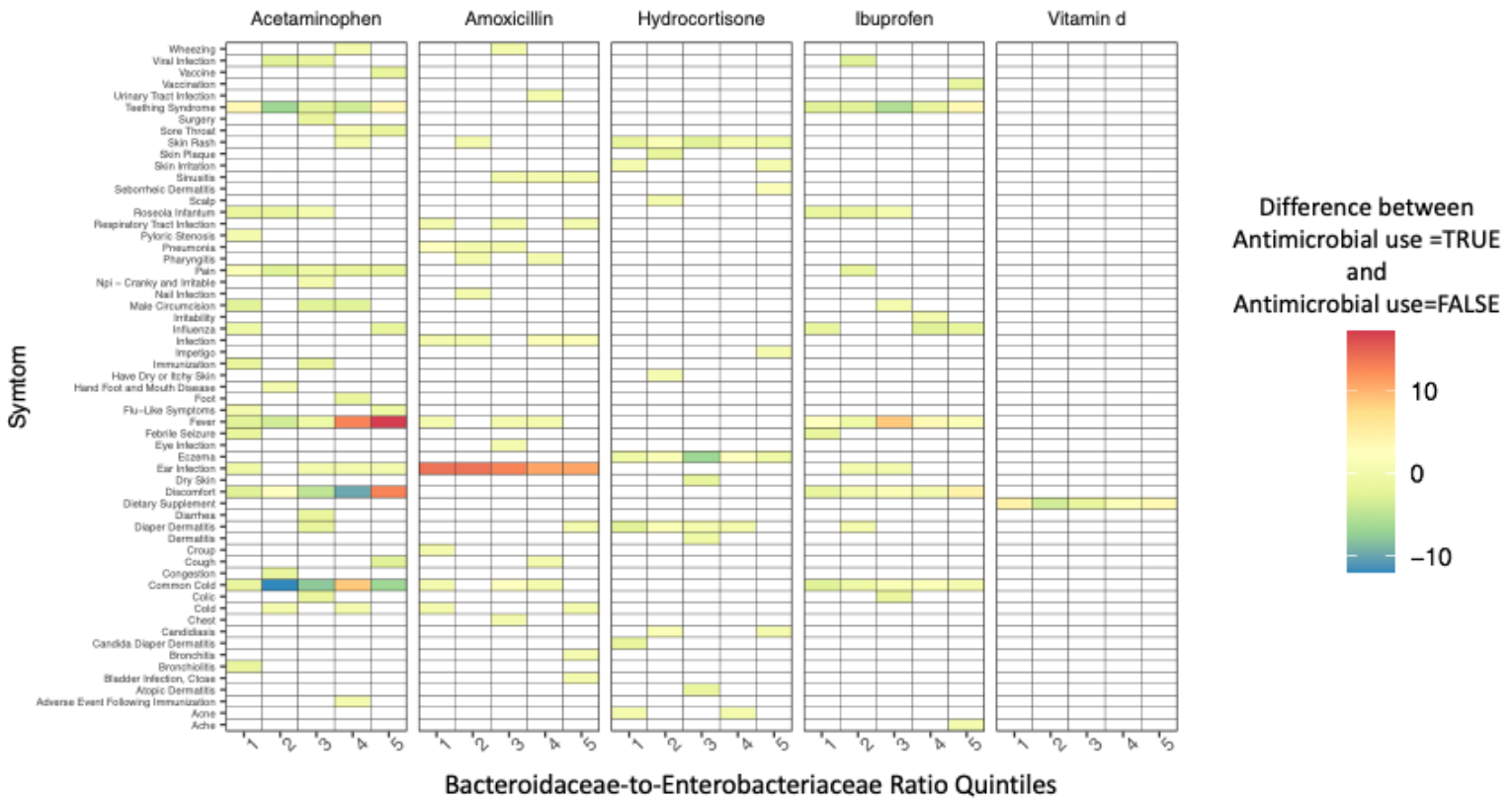

Figure A.21. Quintiles for EB_3m and medication patterns for top 5 most common medications Each facet represents a separate medication, and the rows potential reasons for usage. Within the facets, rows represent bins of subjects (n=564) based on their EB ratio at 3m. Subjects in cluster 1 have the lowest ratios, 5 the highest. Colour scale indicates which group tends to use a medication for a specificreason. Red indicates more common in the antimicrobial usage group, blue is the non antimicrobial usage group.

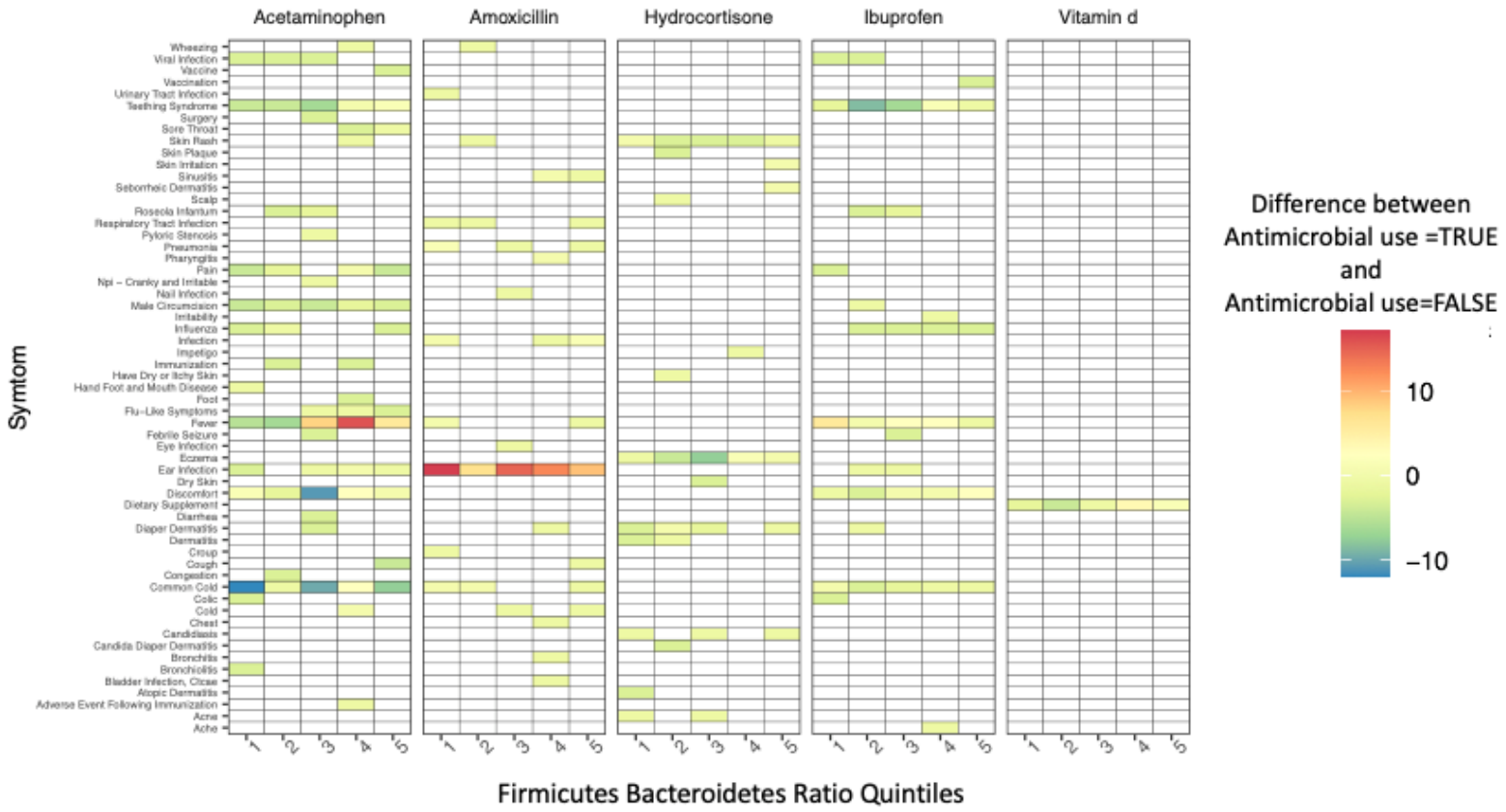

Figure A.22. Quintiles for FB_3m and medication patterns for top 5 most common medications. Each facet represents a separate medication, and the rows potential reasons for usage. Within the facets, rows represent bins of subjects (n=564) based on their EB ratio at 1y. Subjects in cluster 1 have the lowest ratios, 5 the highest. Colour scale indicates which group tends to use a medication for a specificreason. Red indicates more common in the antimicrobial usage group, blue is the non antimicrobial usage group.

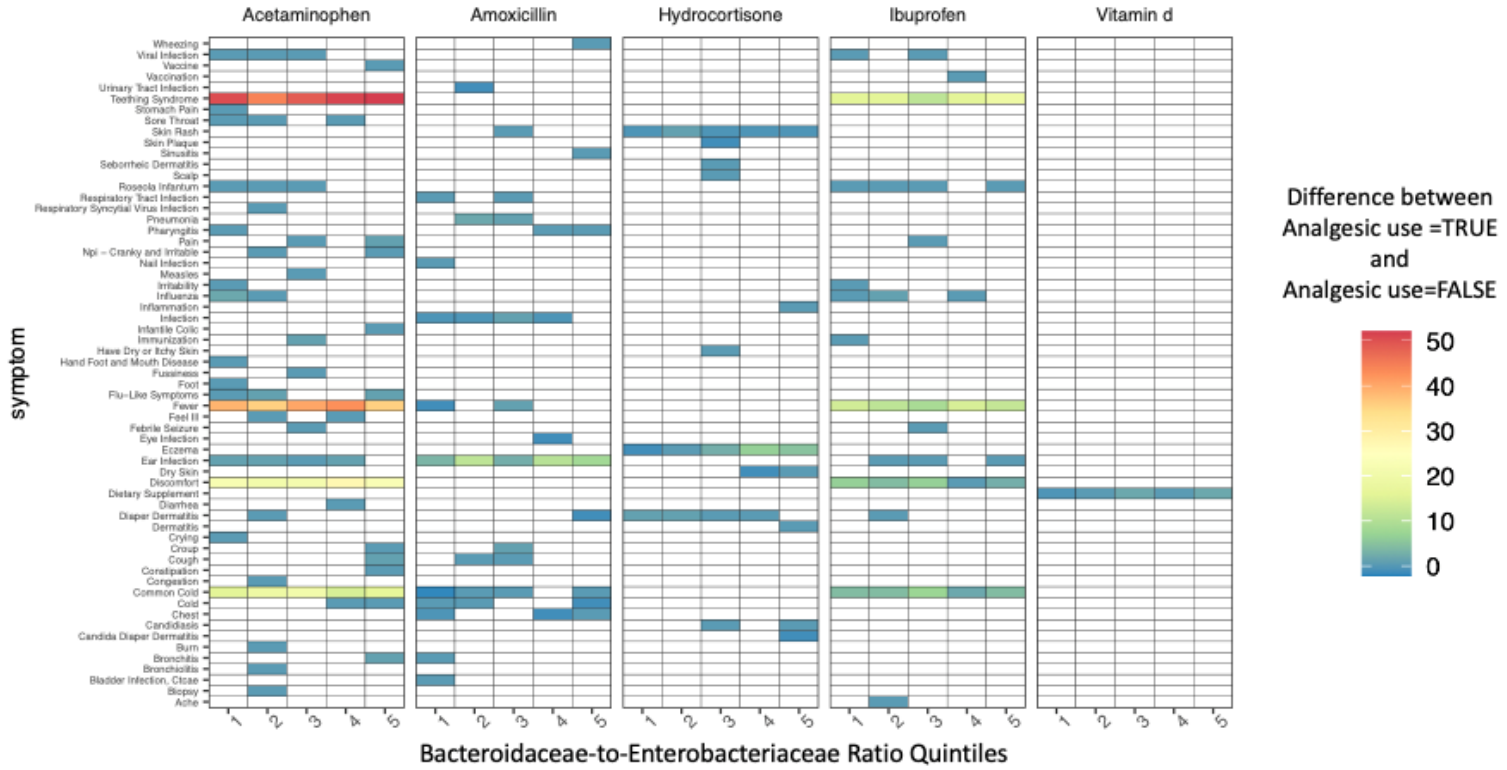

Figure A.23. Quintiles for EB_1y and medication patterns for top 5 most common medications Each facet represents a separate medication, and the rows potential reasons for usage. Within the facets, rows represent bins of subjects (n=564) based on their EB ratio at 1y. Subjects in cluster 1 have the lowest ratios, 5 the highest. Colour scale indicates which group tends to use a medication for a specificreason. Red indicates more common in the analgesic usage group, blue is the non analgesic usage group.

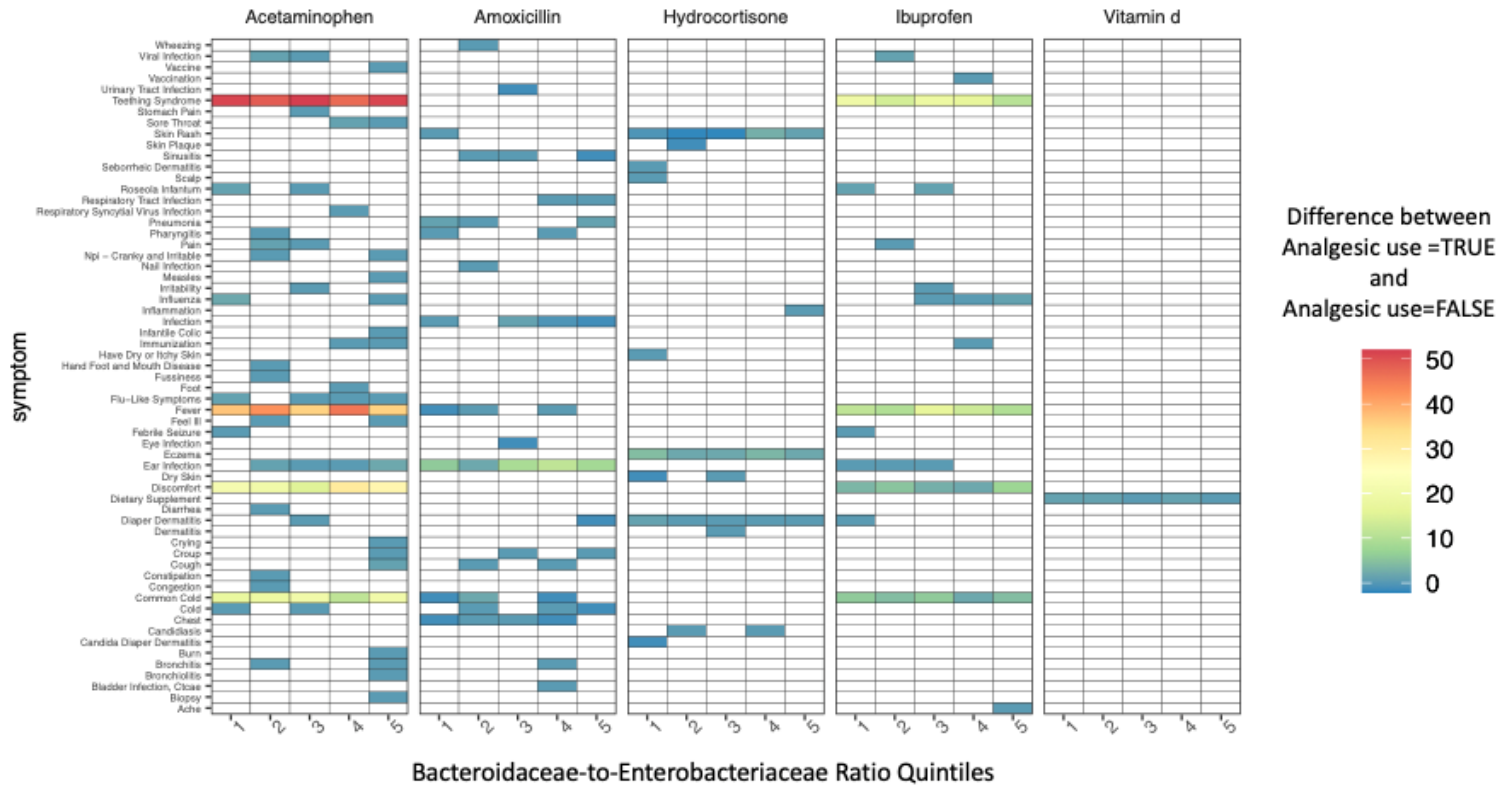

Figure A.24. Quintiles for EB_3m and medication patterns for top 5 most common medications Each facet represents a separate medication, and the rows potential reasons for usage. Within the facets, rows represent bins of subjects (n=564) based on their EB ratio at 1y. Subjects in cluster 1 have the lowest ratios, 5 the highest. Colour scale indicates which group tends to use a medication for a specificreason. Red indicates more common in the analgesic usage group, blue is the non analgesic usage group.

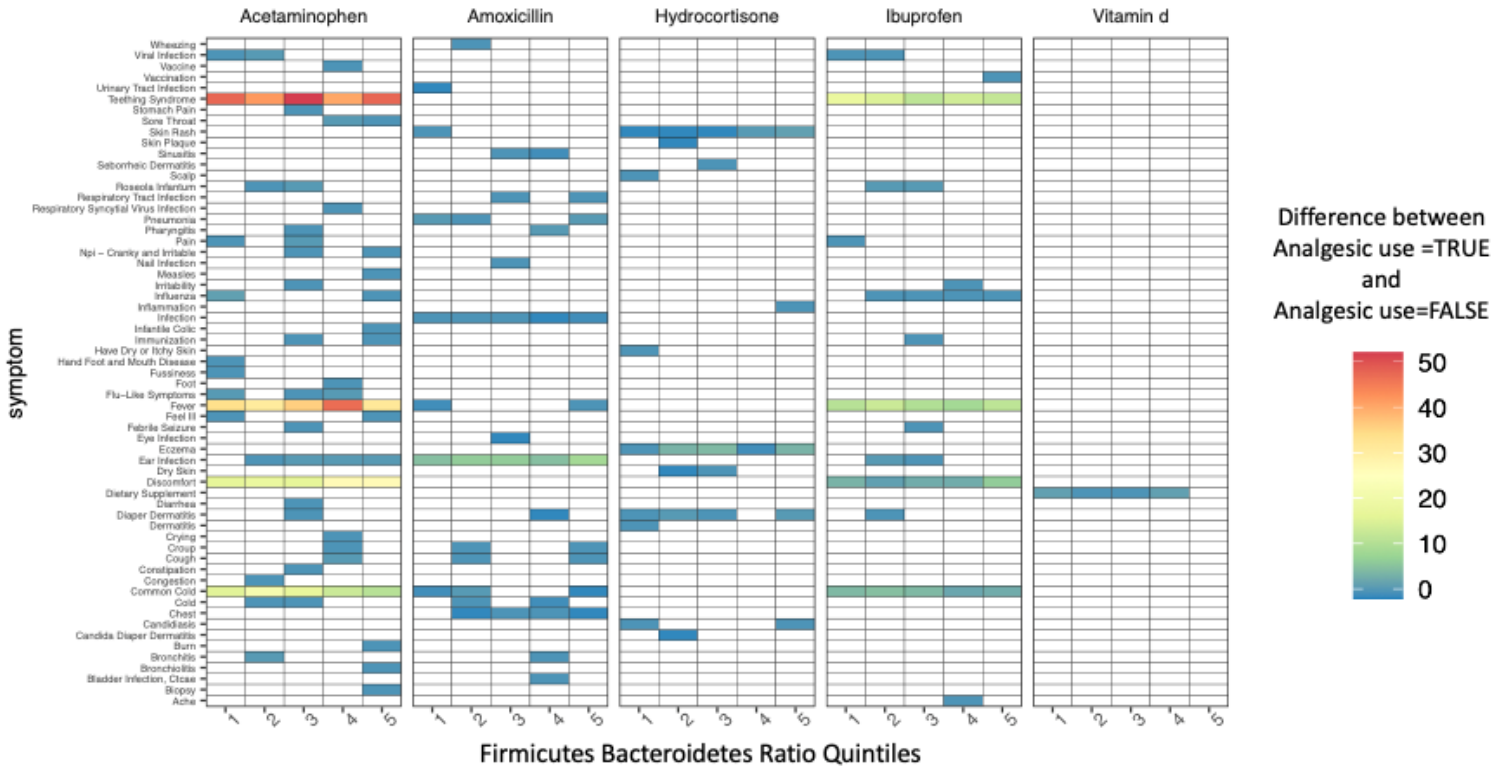

Figure A.25. Quintiles for FB_3m and medication patterns for top 5 most common medications Each facet represents a separate medication, and the rows potential reasons for usage. Within the facets, rows represent bins of subjects (n=564) based on their EB ratio at 1y. Subjects in cluster 1 have the lowest ratios, 5 the highest. Colour scale indicates which group tends to use a medication for a specificreason. Red indicates more common in the analgesic usage group, blue is the non analgesic usage group.

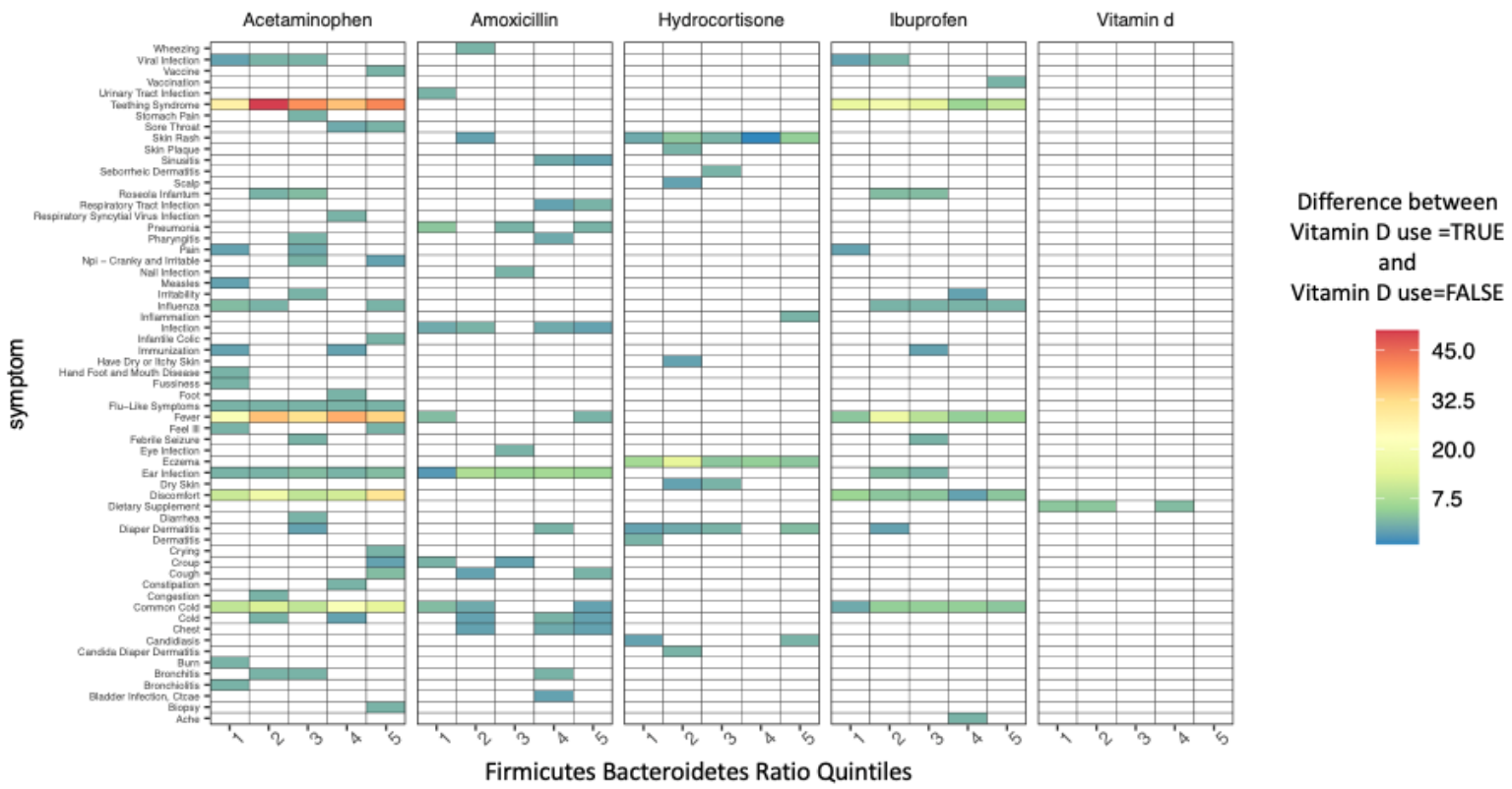

Figure A.26. Quintiles for FB_3m and medication patterns for top 5 most common medications Each facet represents a separate medication, and the rows potential reasons for usage. Within the facets, rows represent bins of subjects (n=564) based on their FB ratio at 3m. Subjects in cluster 1 have the lowest ratios, 5 the highest. Colour scale indicates which group tends to use a medication for a specificreason. Red indicates more common in the Vitamin D usage group, blue is the non Vitamin D usage group

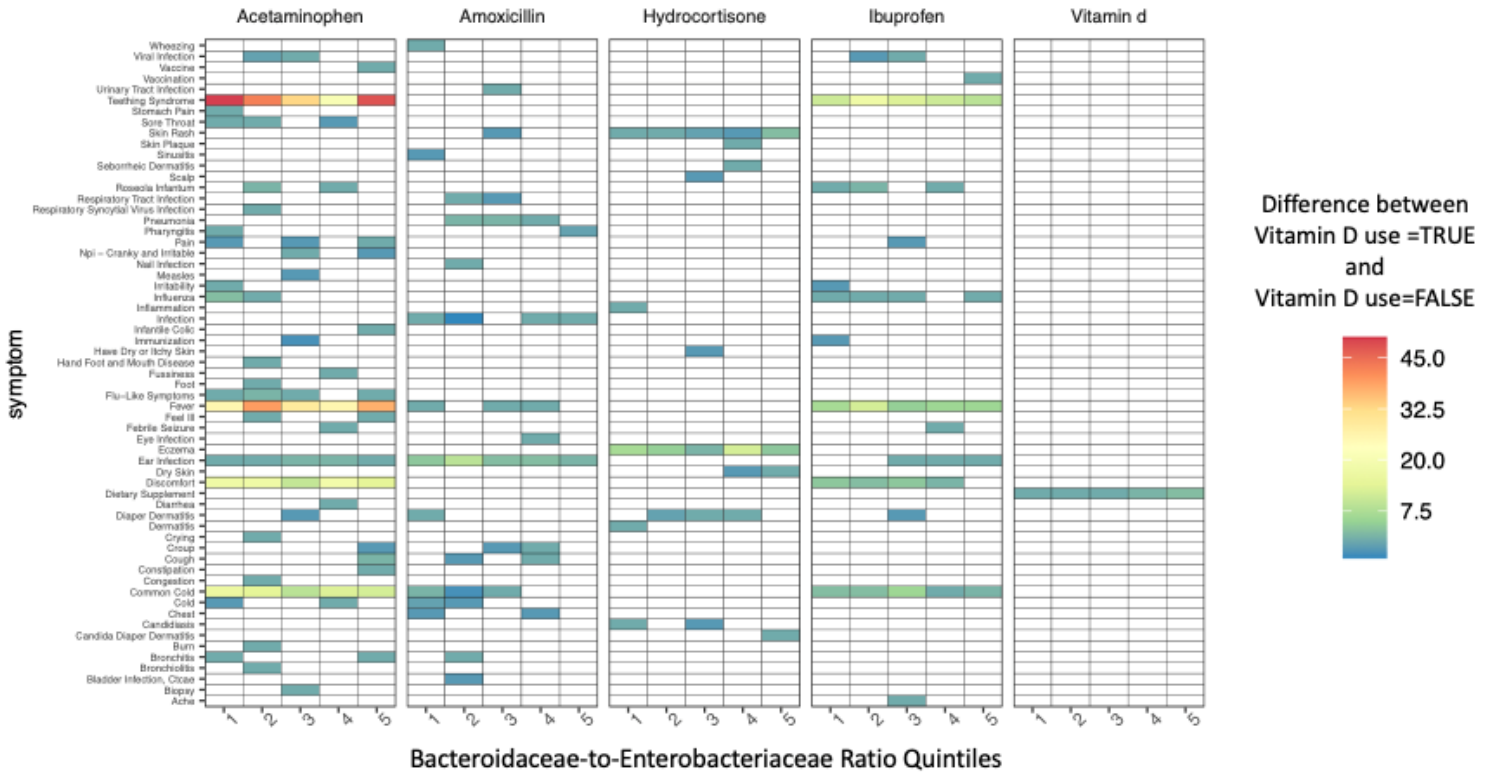

Figure A.27. Quintiles for EB_1y and medication patterns for top 5 most common medications Each facet represents a separate medication, and the rows potential reasons for usage. Within the facets, rows represent bins of subjects (n=564) based on their EB ratio at 1y. Subjects in cluster 1 have the lowest ratios, 5 the highest. Colour scale indicates which group tends to use a medication for a specificreason. Red indicates more common in the Vitamin D usage group, blue is the non Vitamin D usage group.

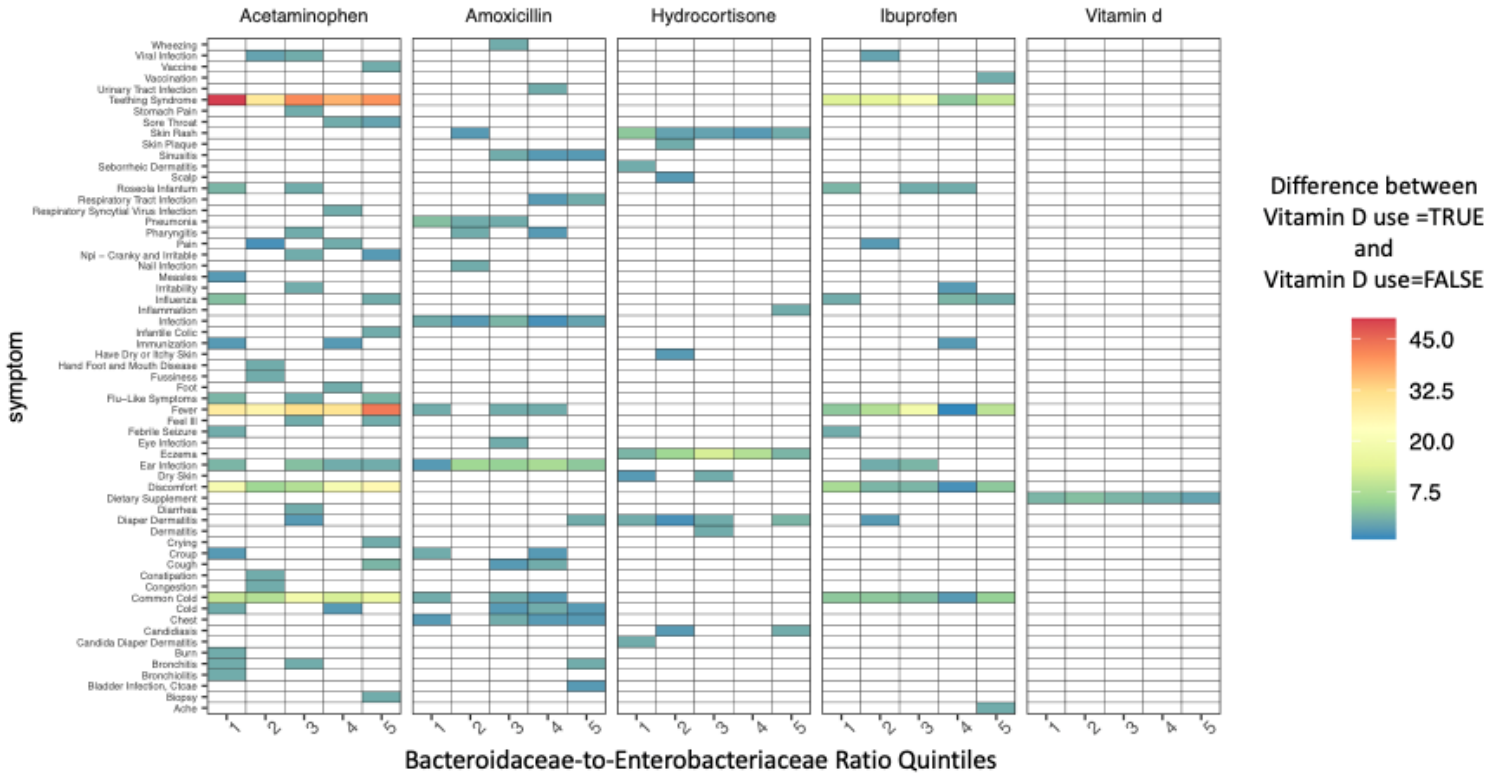

Figure A.28. Quintiles for EB_3m and medication patterns for top 5 most common medications Each facet represents a separate medication, and the rows potential reasons for usage. Within the facets, rows represent bins of subjects (n=564) based on their EB ratio at 3m. Subjects in cluster 1 have the lowest ratios, 5 the highest. Colour scale indicates which group tends to use a medication for a specificreason. Red indicates more common in the Vitamin D usage group, blue is the non Vitamin D usage group.

Appendix B.

Supplemental Tables

Table B.1. Organization of Random Forest Models given different outcomes, missing data thresholds and imputed versions of the dataset. The four outcomes are indicated in the first column, while different data allowances are shown across the columns. From MICE, five datasets were imputed for each combination are indicated by imp1:5.

	Threshold				
Outcome	**All**	**90**	**85**	**80**	**75**
EB_1y	Imp 1:5	Imp 1:5	Imp 1:5	Imp 1:5	Imp 1:5
EB_3m	Imp 1:5	Imp 1:5	Imp 1:5	Imp 1:5	Imp 1:5
FB_1y	Imp 1:5	Imp 1:5	Imp 1:5	Imp 1:5	Imp 1:5
FB_3m	Imp 1:5	Imp 1:5	Imp 1:5	Imp 1:5	Imp 1:5

Table B.2. Organization of Gradient Boosting Models given different outcomes, missing data thresholds and imputed versions of the dataset. The four outcomes are indicated in the first column, while different data allowances are shown across the columns. From MICE, five datasets were imputed for each combination are indicated by imp1:5.

	Threshold				
Outcome	**All**	**90**	**85**	**80**	**75**
EB_1y	Imp 1:5	Imp 1:5	Imp 1:5	Imp 1:5	Imp 1:5
EB_3m	Imp 1:5	Imp 1:5	Imp 1:5	Imp 1:5	Imp 1:5
FB_1y	Imp 1:5	Imp 1:5	Imp 1:5	Imp 1:5	Imp 1:5
FB_3m	Imp 1:5	Imp 1:5	Imp 1:5	Imp 1:5	Imp 1:5

Table B.3. Top 25 Medication Variables from Random Forest Models for the outcome the Enterobacteriaceae-to-Bacteroidaceae ratio at 3 months with variable and descriptions. Results shown for the 90% and 85% variable response models. Within those, the overall rank when considering all variable domains in shown, along with the ranking within the medication domain itself. Table Organized by medication rank for 90%.

Variable Name	Description	Overall Rank	Medication Rank	Overall Rank	Medication Rank
num_antifungal_6m	number of antifungals at 6 months	75	1	310	6
medications_rash_1_Year1	rash medications at 1y	81	2	161	5
X6_months_respiratory.therapy	respiratory therapy at 6 months	88	3	318	7
X6_months_Estrogenic.substances..conjugated.	NA	89	4	319	8
medications_third_cold_1_Year1	medications for third cold at 1 year	90	5	320	9
X1_year_clobetasone.butyrate	clobestasone butyrate at 1 year	91	6	321	10
X1_year_desloratadine	desloratadine at 1 year	95	7	322	11
X6_months_drug	drugs at 6 months	101	8	129	2
X1_year_salbutamol	Salbutamol at 1 year	106	9	330	12
X1_year_cefprozil	Cefprozil at 1 year	122	10	341	13
took_ibuprofen_6m1	took ibuprofen at 6 months	129	11	348	14
X6_months_Homeopathy.Therapy	Homeopathy Therapy at 6 months	130	12	349	15
X1_year_diphenhydramine.hydrochloride	Diphenhydramine hydrochlorideat 1 year	132	13	351	16
X1_year_cetalkonium.chloride	Cetalkonium chloride at 1 year	136	14	148	4
X1_year_antifungal.drug	Anitfungal drug at 1 year	135	15	353	17
X6_months_diphenhydramine.hydrochloride	Diphenhydramine hydrochloride at 6 months	140	16	130	3
X1_year_diflucortolone.valerate	Diflucortolone valerate at 1 year	138	17	355	18
X1_year_omeprazole	Omeprazole at 1 year	139	18	356	19
X1_year_respiratory.therapy	respiratory therapy at 1 year	141	19	357	20
X1_year_anti.asthmatic.drug	anti asthmatic drug at 1 year	142	20	358	21
X1_year_lactulose	lactulose at 1 year	143	21	359	22
X6_months_fluticasone	fluticasone at 6 months	144	22	128	1
X1_year_cefuroxime.axetil	cefuroxime axetil at 1 year	146	23	361	23
X6_months_steroid	steroid at 6 months	147	24	362	24
X1_year_azithromycin	azithromycin at 1 year	148	25	363	25

Table B.4. Top 25 Medication Variables from Random Forest Models for the outcome the Enterobacteriaceae-to-Bacteroidaceae ratio at 1 year with variable and descriptions. Results shown for the 90% and 85% variable response models. Within those, the overall rank when considering all variable domains in shown, along with the ranking within the medication domain itself. . Table Organized by medication rank for 90%.

Variable	Description	Overall Rank	Medication Rank	Overall Rank	Medication Rank
medications_for_coughs_1_Year1	cough medication at 1 year	27	1	277	3
X1_year_epinephrine	epinephrine at 1 year	73	2	309	4
X1_year_amoxicillin	amoxicillin at 1 year	82	3	315	5
X1_year_prednisone	prednisone at 1 year	88	4	318	6
num_acetaminophen_3m	number of acetaminophen at 3 months	89	5	319	7
took_antiallergic_6m1	took antiallergic medication at 6 months	95	6	323	8
X1_year_glycerin	glycerin at 1 year	98	7	324	9
X1_year_vitamin.d	vitamin D at 1 year	111	8	336	10
X6_months_diphenhydramine.hydrochloride	Diphenhydramine hydrochloride at 6 months	112	9	337	11
X1_year_antibacterial.drug	antibacterial drug at 1 year	117	10	342	12
num_analgesic	number of analgesic	123	11	140	2
took_ibuprofen_1y1	took ibuprofen at 6 months	124	12	346	13
X1_year_fluticasone.propionate	fluticasone propionate at 1 year	127	13	348	14
X6_months_fluticasone	fluticasone at 6 months	130	14	86	1
X1_year_clavulanic.acid	clavulanic acid at 1 year	133	15	351	15
num_anthelminthic_6m	number of antihelminthic at 6 months	135	16	353	16
X6_months_fusidic.acid	fusidic acid at 6 months	136	17	354	17
X1_year_erythromycin	erythromycin at 1 year	137	18	355	18
X1_year_ambroxol.hydrochloride	ambroxol hydrochloride at 1 year	138	19	356	19
X1_year_cefaclor	cefaclor at 1 year	139	20	357	20
X1_year_mupirocin	mupirocin at 1 year	140	21	358	21
X6_months_phenobarbital	phenobarbital at 6 months	141	22	359	22
X6_months_furosemide	furosemide at 6 months	142	23	360	23
X1_year_fluocinolone.acetonide	fluocinolone acetonide at 1 year	143	24	361	24
X6_months_moxifloxacin.hydrochloride	moxifloxacin hydrochloride at 6 months	144	25	362	25

Table B.5. Top 25 Medication Variables from Random Forest Models for the outcome Firmicutes-to-Bacteroidetes ratio at 3 months with variable and descriptions. Results shown for the 90% and 85% variable response models. Within those, the overall rank when considering all variable domains in shown, along with the ranking within the medication domain itself. . Table Organized by medication rank for 90%.

Variable	Description	Overall Rank	Medication Rank	Overall Rank	Medication Rank
num_ibuprofen_1y	number of ibuprofen at 1 year	29	1	272	4
X6_months_benzocaine	benzocaine at 6 months	80	2	300	5
num_antifungal_3m	number of antifungal at 3 months	81	3	301	6
X1_year_vitamin.d	vitamin D at 1 year	82	4	111	3
vitamins_supplements_Poly.Vi.Sol_iron_1_Year1	NA	91	5	308	7
num_antibacterial_6m	number of antibacterial at 6 months	92	6	309	8
took_antiallergic_3m1	took antiallergic medication at 3 months	99	7	315	9
num_antiallergic_3m	number of antiallergic at 3 months	104	8	318	10
num_antipyretic	number of antipyretic	105	9	319	11
X1_year_albuterol	albuterol at 1 year	110	10	321	12
medications_first_cold_1_Year1	medications for first cold at 1 year	111	11	322	13
vitamins_supplements_other_1_Year1	other vitamins and supplements at 1 year	117	12	102	2
took_antiallergic_6m1	took antiallergic medication at 6 months	120	13	329	14
num_anti_asthmatic	number of anti asthmatic medications	121	14	330	15
X1_year_fluticasone.propionate	fluticasone propionate at 1 year	122	15	331	16
X6_months_moxifloxacin.hydrochloride	moxifloxacin hydrochloride at 6 months	123	16	332	17
X1_year_fluocinolone.acetonide	fluocinolone acetonide at 1 year	139	17	348	18
X6_months_azithromycin	azithromycin at 6 months	146	18	351	19
num_antiallergic_1y	number of antiallergic at 1 year	155	19	359	20
took_ibuprofen_6m1	took ibuprofen at 6 months	159	20	360	21
vitamins_supplements_D.Vi.Sol_D.drops_vitamin_D_1_Year1	NA	162	21	92	1
num_antifungal_1y	number of antifungal at 1 year	164	22	362	22
X1_year_antibacterial.drug	antibacterial drug at 1 year	165	23	363	23
X6_months_cefixime	cefixime at 6 months	166	24	364	24
X6_months_cefprozil	cefprozil at 6 months	169	25	367	25

Table B.6. Top 25 Medication Variables from Random Forest Models for the outcome Firmicutes-to-Bacteroidetes ratio at 1 year with variable and descriptions.. Results shown for the 90% and 85% variable response models. Within those, the overall rank when considering all variable domains in shown, along with the ranking within the medication domain itself. . Table Organized by medication rank for 90%.

Variable	Description	Overall Rank	Med Rank	Overall Rank	Med Rank
X1_year_acetaminophen	acetaminophen at 1 year	31	1	292	7
medications_oral_thrush_1_Year1	oral thrush medications at 1 year	34	2	295	8
X1_year_albuterol	albuterol at 1 year	60	3	316	9
medications_ear_infection_1_Year1	ear infection medications at 1 year	62	4	119	3
took_antifungal_3m1	took antifungal at 3 months	71	5	322	10
took_ibuprofen_1y1	took ibuprofen at 6 months	72	6	323	11
X6_months_fluconazole	fluconazole at 6 months	75	7	326	12
X1_year_benzocaine	benzocaine at 1 year	88	8	338	13
X6_months_laxative	laxative at 6 months	92	9	341	14
X1_year_clotrimazole	clotrimazole at 1 year	98	10	79	1
X1_year_cephalexin	cephalexin at 1 year	99	11	97	2
X6_months_hydrocortisone	hydrocortisone at 6 months	97	12	344	15
num_antiallergic_3m	number of antiallergic at 3 months	119	13	360	16
X1_year_fluticasone	fluticasone at 1 year	120	14	361	17
X6_months_beclometasone.dipropionate	beclometasome dipropionate at 6 months	127	15	365	18
X6_months_ranitidine	ratitidine at 6 months	129	16	151	6
took_ibuprofen_6m1	took ibuprofen at 6 months	133	17	142	5
X1_year_Homeopathy.Therapy	homeopathy therapy at 1 year	134	18	369	19
X1_year_steroid	steroid at 1 year	136	19	371	20
num_antibacterial_3m	number of antibacterial at 3 months	138	20	372	21
X6_months_nystatin	nystatin at 6 months	140	21	374	22
num_antipyretic	number of antipyretic	141	22	375	23
X1_year_belladonna	belladonna at 1 year	145	23	126	4
vitamins_supplements_Poly.Vi.Sol_vitamins_A_C_D_B1_B2_B3_B6_1_Year1	vitamins and supplements	150	24	381	24
X6_months_vitamin.d	Vitamin D at 6 months	153	25	383	25

Table B.7. Top 25 Medication Variables from Gradient Boosting Machine for the outcome the Enterobacteriaceae-to-Bacteroidaceae ratio at 3 months. Results shown for the 90% and 85% variable response models. Within those, the overall rank when considering all variable domains in shown, along with the ranking within the medication domain itself. Table Organized by medication rank for 90%.

Variable	Description	Overall Rank	Medication Rank	Overall Rank	Medication Rank
X1_year_hydrocortisone.valerate	hydrocortisone valerate at 1 year	61	1	98	2
num_antifungal_3m	number of antifungal at 3 months	69	2	67	1
X1_year_nystatin	nystatin at 1 year	75	3	106	3
took_acetaminophen_3m	took acetaminophen at 3 months	127	4	157	4
num_acetaminophen_3m	number of acetaminophen at 3 months	128	5	158	5
took_ibuprofen_3m	took ibuprofen at 3 months	129	6	159	6
took_antibacterial_3m	took antbacterial at 3 months	130	7	160	7
num_antibacterial_3m	number of antibacterial at 3 months	131	8	161	8
took_antifungal_3m	took antifungal at 3 months	132	9	162	9
took_anthelminthic_3m	took antihelminthic at 3 months	133	10	163	10
num_anthelminthic_3m	number of antihelminthic at 3 months	134	11	164	11
took_antiallergic_3m	took antiallergic at 3 months	135	12	165	12
num_antiallergic_3m	number of antiallergic at 3 months	136	13	166	13
took_acetaminophen_6m	took acetaminophen at 6 months	137	14	167	14
num_acetaminophen_6m	number of acetaminophen at 6 months	138	15	168	15
took_ibuprofen_6m	took ibuprofen at 6 months	139	16	169	16
num_ibuprofen_6m	number of ibuprofen at 6 months	140	17	170	17
took_antibacterial_6m	took antibacterial at 6 months	141	18	171	18
num_antibacterial_6m	number of antibacterial at 6 months	142	19	172	19
took_antifungal_6m	took antifungal at 6 months	143	20	173	20
num_antifungal_6m	number of antifungals at 6 months	144	21	174	21
took_antiallergic_6m	took antiallergic at 6 months	145	22	175	22
num_antiallergic_6m	number of antiallergic at 6 months	146	23	176	23
took_acetaminophen_1y	took acetaminophen at 1 year	147	24	177	24
num_acetaminophen_1y	number of acetaminophen at 1 year	148	25	178	25

Table B.8. Top 25 Medication Variables from Gradient Boosting Machine for the outcome the Enterobacteriaceae-to-Bacteroidaceae ratio at 1 Year. Results shown for the 90% and 85% variable response models. Within those, the overall rank when considering all variable domains in shown, along with the ranking within the medication domain itself. Table Organized by medication rank for 90%.

Variable	Description	Overall Rank	Medication Rank	Overall Rank	Medication Rank
X1_year_benzocaine	benzocaine at 1 year	83	1	156	2
num_acetaminophen_6m	number of acetaminophen at 6 months	98	2	162	3
took_antibacterial_3m	took antbacterial at 3 months	107	3	168	4
num_ibuprofen_6m	number of ibuprofen at 6 months	170	4	202	5
X1_year_zinc.oxide	zinc oxide at 1 year	171	5	203	6
took_acetaminophen_6m	took acetaminophen at 6 months	176	6	207	7
num_antifungal_6m	number of antifungals at 6 months	179	7	209	8
took_acetaminophen_1y	took acetaminophen at 1 year	182	8	116	1
num_antibacterial_1y	number of antibacterial at 1 year	183	9	210	9
took_antibacterial_1y	took antibacterial at 1 year	184	10	211	10
X1_year_acetaminophen	acetaminophen at 1 year	185	11	212	11
vitamins_supplements_1_Year	vitamins and supplements at 1 year	188	12	214	12
X1_year_sulfamethoxazole	sulfamethoxazole at 1 year	189	13	215	13
X6_months_hydrocortisone	hydrocortisone at 6 months	190	14	216	14
X1_year_Herbal.Remedy.Supplement	herbal remedy supplement at 1 year	191	15	217	15
X1_year_sodium.chloride	sodium chloride at 1 year	192	16	218	16
X6_months_Herbal.Remedy.Supplement	herbal remedy supplement at 6 months	193	17	219	17
medications_rash_1_Year	rash medications at 1 year	196	18	222	18
num_antifungal_1y	number of antifungal at 1 year	198	19	224	19
medications_second_cold_1_Year	medications for second cold at 1 year	200	20	226	20
vitamins_supplements_other_1_Year	other vitamins and supplements at 1 year	201	21	227	21
X6_months_albuterol	albuterol at 6 months	203	22	229	22
X1_year_clarithromycin	clarithromycin at 1 year	204	23	230	23
X6_months_drug.cream	drug cream at 6 months	205	24	231	24
X6_months_amoxicillin	amoxicillin at 6 months	206	25	232	25

Table B.9. Top 25 Medication Variables from Gradient Boosting Machine for the outcome Firmicutes-to-Bacteroidetes ratio at 3 months. Results shown for the 90% and 85% variable response models. Within those, the overall rank when considering all variable domains in shown, along with the ranking within the medication domain itself. Table Organized by medication rank for 90%.

Medication Variable	Description	Overall Rank	Medication Rank	Overall Rank	Medication Rank
X1_year_hydrocortisone.valerate	hydrocortisone valerate at 1 year	61	1	98	2
num_antifungal_3m	number of antifungal at 3 months	69	2	67	1
X1_year_nystatin	nystatin at 1 year	75	3	106	3
took_acetaminophen_3m	took acetaminophen at 3 months	127	4	157	4
num_acetaminophen_3m	number of acetaminophen at 3 months	128	5	158	5
took_ibuprofen_3m	took ibuprofen at 3 months	129	6	159	6
took_antibacterial_3m	took antbacterial at 3 months	130	7	160	7
num_antibacterial_3m	number of antibacterial at 3 months	131	8	161	8
took_antifungal_3m	took antifungal at 3 months	132	9	162	9
took_anthelminthic_3m	took antihelminthic at 3 months	133	10	163	10
num_anthelminthic_3m	number of antihelminthic at 3 months	134	11	164	11
took_antiallergic_3m	took antiallergic at 3 months	135	12	165	12
num_antiallergic_3m	number of antiallergic at 3 months	136	13	166	13
took_acetaminophen_6m	took acetaminophen at 6 months	137	14	167	14
num_acetaminophen_6m	number of acetaminophen at 6 months	138	15	168	15
took_ibuprofen_6m	took ibuprofen at 6 months	139	16	169	16
num_ibuprofen_6m	number of ibuprofen at 6 months	140	17	170	17
took_antibacterial_6m	took antibacterial at 6 months	141	18	171	18
num_antibacterial_6m	number of antibacterial at 6 months	142	19	172	19
took_antifungal_6m	took antifungal at 6 months	143	20	173	20
num_antifungal_6m	number of antifungals at 6 months	144	21	174	21
took_antiallergic_6m	took antiallergic at 6 months	145	22	175	22
num_antiallergic_6m	number of antiallergic at 6 months	146	23	176	23
took_acetaminophen_1y	took acetaminophen at 1 year	147	24	177	24
num_acetaminophen_1y	number of acetaminophen at 1 year	148	25	178	25

Table B.10. Top 25 Medication Variables from Gradient Boosting Machine for the outcome Firmicutes-to-Bacteroidetes ratio at 1 Year. Results shown for the 90% and 85% variable response models. Within those, the overall rank when considering all variable domains in shown, along with the ranking within the medication domain itself. Table Organized by medication rank for 90%.

Medication Variable		Overall Rank	Medication Rank	Medication Variable	Overall Rank
X6_months_drug.solution	drug solution at 6 months	43	1	43	2
vitamins_supplements_1_Year	vitamins and supplements at 1 year	122	2	140	3
num_antifungal_1y	number of antifungal at 1 year	132	3	149	4
num_acetaminophen_3m	number of acetaminophen at 3 months	137	4	154	5
took_antifungal_1y	took antifungal at 1 year	139	5	156	6
X1_year_vitamin.d	vitamin D at 1 year	140	6	157	7
num_anti_inflammatory	number of anti inflammatory	150	7	167	8
took_acetaminophen_3m	took acetaminophen at 3 months	153	8	170	9
took_antibacterial_1y	took antibacterial at 1 year	156	9	172	10
num_antibacterial_1y	number of antibacterial at 1 year	158	10	174	11
num_analgesic	number of analgesic	217	11	20	1
took_ibuprofen_3m	took ibuprofen at 3 months	190	12	205	12
took_antibacterial_3m	took antbacterial at 3 months	191	13	206	13
num_antibacterial_3m	number of antibacterial at 3 months	192	14	207	14
took_antifungal_3m	took antifungal at 3 months	193	15	208	15
num_antifungal_3m	number of antifungal at 3 months	194	16	209	16
took_anthelminthic_3m	took antihelminthic at 3 months	195	17	210	17
num_anthelminthic_3m	number of antihelminthic at 3 months	196	18	211	18
took_antiallergic_3m	took antiallergic at 3 months	197	19	212	19
num_antiallergic_3m	number of antiallergic at 3 months	198	20	213	20
took_acetaminophen_6m	took acetaminophen at 6 months	199	21	214	21
num_acetaminophen_6m	number of acetaminophen at 6 months	200	22	215	22
took_ibuprofen_6m	took ibuprofen at 6 months	201	23	216	23
num_ibuprofen_6m	number of ibuprofen at 6 months	202	24	217	24
took_antibacterial_6m	took antibacterial at 6 months	203	25	218	25

Table B.11. Table Showing Variable names used for ANCOM-BC analysis for Antimicrobials. Short Name in CHILD refers to the name found in the data files provided by the CHILD cohort Study. The Name or group for ANCOM-BC is the name or grouping assigned or the analysis.

Short Name in CHILD	Name or Group for ANCOM-BC
took_antibacterial_3m	took_antibacterial_3m
num_antibacterial_3m	num_antibacterial_3m
took_antibacterial_6m	took_antibacterial_6m
num_antibacterial_6m	num_antibacterial_6m
took_antibacterial_1y	took_antibacterial_1y
num_antibacterial_1y	num_antibacterial_1y
num_antibacterial	num_antibacterial
x1_year_antibacterial.drug	x1_year_antibacterial.drug
x6_months_antibacterial.drug	x6_months_antibacterial.drug
antibacterial_by_mother_3m	antibacterial_by_mother_3m
prenatal_antibacterial	prenatal_antibacterial
sum_analgesic_birth	sum_analgesic_birth
sum_analgesic_3_months	sum_analgesic_3_months
sum_analgesic_1_year	sum_analgesic_1_year
ampicillin_birth	penicillins_birth
ampicillin_3_months	penicillins_3m
vancomycin_1_year	glycopeptide_1y
ceftriaxone_1_year	cephalosporin_1y
cefotaxime_1_year	cephalosporin_1y
cefprozil_1_year	cephalosporin_1y
amoxicillin_1_year	penicillins_1y
cefixime_1_year	cephalosporin_1y
clarithromycin_1_year	macrolide_1y
cefprozil_3_months	cephalosporin_3m
azithromycin_1_year	macrolide_1y
antibacterial_drug_1_year	antibacterial_1y
clarithromycin_3_months	macrolide_3m
amoxicillin_3_months	penicillins_3m
bacitracin_1_year	polypeptide_1y
tobramycin_1_year	aminoglycoside_1y
bacitracin_3_months	polypeptide_3m
mupirocin_3_months	carboxylicacidclass_3m
silver_sulfadiazine_1_year	sulfa_1y
clavulanic_acid_1_year	beta_lacatamase_1y
sulfisoxazole_1_year	sulfonamide_1y
trimethoprim_1_year	trimethoprim_1y
sulfamethoxazole_1_year	sulfonamide_1y

Short Name in CHILD	Name or Group for ANCOM-BC
trimethoprim_3_months	trimethoprim_3m
cefotaxime_birth	cephalosporin_birth
meropenem_birth	carbapenem_birth
antibacterial_drug_3_months	antibacterial_3m
cefaclor_1_year	cephalosporin_1y
mupirocin_1_year	carboxylicacidclass_1y
cephalexin_1_year	cephalosporin_1y
tobramycin_3_months	aminoglycoside_3m
cephalexin_3_months	cephalosporin_3m
crystal_violet_3_months	crystalviolet_3m
crystal_violet_1_year	crystalviolet_1y
cefotaxime_3_months	cephalosporin_3m
cefazolin_3_months	cephalosporin_3m
sulfamethoxazole_3_months	sulfonamide_3m
ceftriaxone_3_months	cephalosporin_3m

Table B.12 Table Showing Variable names used for ANCOM-BC analysis for Antimicrobials. Short Name in CHILD refers to the name found in the data files provided by the CHILD cohort Study. The Description is a laymen's description of the variable name.

Short Name In CHILD	Variable Description
1Year_vitaminD	Sum of times vitamin D was taken (6m-1y)
6months_vitamin D	Sum of times vitamin D was taken (birth-6m)
pmffqq146	vitamin d (calciferol) (mcg)
pmffqq147	vitamin d2 (ergocalciferol) (mcg)
Pmffqq148	vitamin d3 (cholecalciferol) (mcg)
msuppl18wq1_7	How often did you take vitamin D
vitamin_d_taken	Vitamin D usage at 3 months based on 3-month nutrition and 3-month medication questionnaires.
NUTR1YQ7_1a	Are you giving your child any vitamins or supplements? If Yes, which are used? Vitamin D:

Appendix C.

Supplemental Data Files

Supplementary Data 1 Description: Top 25 variable importance scores for each domain from Random Forest and Gradient Boosting Models with using datasets with 85% and 95% data completeness.

Filename:

supplementary_data_1.xlsx

Supplementary Data 2 Description: All variable importance scores for each domain from Random Forest models with 90% and 85% variable completeness. The first column contains the outcome variable, the second column contains the variable, the third column contains the domain of information the variable belonged to, and the final two columns contain the variable importance scores.

Filename:

supplementary_data_2.csv

Supplementary Data 3 Description: All variable importance scores for each domain from Gradient Boosting models with 90% and 85% variable completeness. The first column contains the outcome variable, the second column contains the variable, the third column contains the domain of information the variable belonged to, and the final two columns contain the variable importance scores.

Filename:

supplementary_data_3.csv