AIC 2023 Rosen Shingle Creek Orlando, FL **June 1-3**

Exploring the Impact of Antibiotic Use on General Health Status in Children by Aggregating Across Machine Learning Models

R. REISERER[1], J. NELSON[1], B. LAMBERT[1], H. TALLMAN RUHM[1], M. HERBERT[2,1]

1 Epidemic Answers, Windsor, CT, USA

2 Higher Synthesis Foundation, Cambridge, MA, USA

ABSTRACT

Antibiotics—perhaps the most consequential breakthrough in 20th century medicine—have saved countless lives and are credited with extending the average Western lifespan by at least two decades. The domain of utility for antibiotics is in the control of pathogenic microbes, but they can also damage the beneficial microbial ecosystems that humans depend on for healthy biochemistry. The overuse of antibiotics has become a concern because many pathogenic microbes have evolved antibiotic resistance. As our understanding of human-microbial-symbiosis has matured, many health professionals have come to suspect that harms from antibiotics extend much further, affecting general health in various ways. We used machine learning (decision trees) in a novel way to aggregate and analyze chronic conditions using hierarchical indices derived from a comprehensive survey of family lifestyle and childhood health. Our method makes use of classification models to reveal broad or dominant trends in data sets consisting of a multitude of metrics and data types. Such trends are difficult to glean from traditional statistical models. We generated decision tree models for many common disease states classified using 24 different stressor categories. We scored the dominance of each stressor category across all trees to get generalized measures of the association between stressors and chronic illnesses across our sample population. Antibiotics scored high and stood out as an important stressor across classification models that resolved chronic disease conditions by means of stressor categories. This work builds on IFM AIC reports from previous years (2020-2021).

INTRODUCTION

Antibiotics have been hailed as a triumph of Western medicine in an age-old war on germs and until recently few would argue against that position, but nowadays the accolades come with some caveats. Many problematic microbes can no longer be controlled effectively with antibiotics, owing to the evolution of resistant strains, and this antibiotic resistance has prompted calls for more judicious use of antibiotics. Perhaps even more problematic is the recognition that broad and indiscriminate use of antibiotics has altered the ecological balance of beneficial microbes on which we depend for good health. Only in the last few decades has science come to terms with the nuances of microbial symbiosis as one of the foundations of human and environmental health. Unfortunately, our once hailed panacea against harmful germs can also wreak havoc on the ecosystem of microbes that modulates human health.

The Documenting Hope Project (DHP) aims to better understand how modern technologies, such as antibiotics, may contribute to childhood chronic illnesses, which have reached epidemic proportions in recent times. Accordingly, DHP uses modern analytics and novel approaches to examine the complex systems involved in health, making use of our comprehensive health survey (CHIRP) to study the relationships between the environment, lifestyle, development, and health outcomes of children.

We previously reported on methods for aggregating hierarchically stratified data (Nelson et al., 2020) and for drilling down into aggregated data to investigate cumulative effects, synergies, and interactions among large sets of potential health stressors (Reiserer et al., 2021). Innovation in analytics, such as that afforded by machine learning, has been key to this effort, and we continue to explore novel ways to use these learning algorithms. One interesting option is to generate meta data from multiple machine learning models that are related thematically. To the extent that two or more models share an implicit meaning, summary information about those models may provide general insight that the individual models do not offer.

Decision trees are hierarchical models that display decision points as nodes on a branching (tree-like) diagram (see examples in Reiserer et al., 2021). The results of a series of decisions are displayed at terminal ("leaf") nodes, which typically display error statistics (correct and incorrect classifications). Internal nodes (decision points) display decision criteria and node statistics. The visual nature of decision trees makes them easy to interpret and compare to each other, and the hierarchical structure allows for assessment of a variable's relative importance at a glance. Decision trees are constructed algorithmically using calculated measures of parameter importance. Variables with high importance values appear at or close to the root node, while those toward the branch tips are characterized by more specific dependencies and lower overall importance values. This hierarchy of parameter importance can be scored in various ways to create derivative data that are not represented in standard decision tree metrics. For example, one can score a set of input variables according to their positions in a series of trees generated for any number of related target variables, thus deriving a data matrix that can be further analyzed or displayed in bar charts to examine certain relationships (see tables and figures below).

Here we present a derivative analysis that addresses a general question about antibiotic exposure across a diverse range of childhood chronic conditions. While performing an earlier analysis (Reiserer et al., 2021), we noted that our Antibiotics Index was a prominent predictor in many models of various chronic conditions, and that it frequently occupied high-importance positions in decision tree models. Several other input variables also stood out as potentially important predictors of poor health outcomes. The current presentation, however, focuses on antibiotics as one of the most prevalent predictors in our data set to date.

RESULTS

The simple count of occurrences (Pvar; Figure 1) showed that all but one input variable appeared in multiple tree models. The Antibiotics Index (AntibioticIndx, Table 2, Figure 1) occurred in 60% of the trees (and, therefore, was predictive for 60% of the chronic conditions). Four other composite variables occurred more often across the 30 trees than Antibiotics Index and four additional variables occurred as often (Figure 1). The Antibiotics Index was likewise of intermediate prevalence when all occurrences were counted (Figure 2). Only ten of the 24 index variables appeared at the root nodes (the primary decision point) across the 30 trees (Figure 3), with the Antibiotics Index appearing most often (six out of 30 models). The Antibiotics Index was also most prevalent at the second-level decision point (Figure 4), occurring eight times in that position. Totaled together, the AntibioticIndx was the most frequent index variable at the first and second decision points (Figure 5). See derivative variables in Table 2.

Table 2. Derivative data variable definitions. from decision tree analysis (see Figures 1-5). See linked table for definitions of abbreviations (Reiserer et al., 2021) and Methods for derivative

Index	Pvar (≥1 occurance per tree)	Percent (of 30 trees)	P1 (root node)	P2 (2nd position)	P1 + P2	NT (total occurrences in all trees)
SubstanceIndx	25	83.3	1	6	7	55
PrenatIBMHealthIndx	24	80.0	4	2	6	52
ChemExposIndx	20	66.7	0	3	3	42
SchoolExposIndx	19	63.3	2	1	3	37
AntibioticIndx	18	60.0	6	8	14	36
PreNatlMedsIndx	18	60.0	0	3	3	35
PrenatlExposIndx	18	60.0	0	1	1	34
EMFIndx	17	56.7	1	1	2	33
BrFeedgExposIndx	18	60.0	0	2	2	32
VaxIndx	16	53.3	5	4	9	27
SleepSTRIndx	18	60.0	0	5	5	26
SignExposIndx	15	50.0	1	2	3	26
FoodAttitudeSUP	16	53.3	0	2	2	21
InfantMedFreqIndx	14	46.7	4	0	4	20
ScreenTimeIndx	13	43.3	0	2	2	18
PNAntibiotIndx	14	46.7	0	1	1	17
InjectblMedsIndex	14	46.7	4	3	7	16
HomeEnvirIndx	12	40.0	0	2	2	16
OTCMedsIndex	10	33.3	1	0	1	15
FastFoodIndx	10	33.3	0	1	1	14
MedsAtBirthIndx	10	33.3	0	1	1	12
ExSugarIndx	10	33.3	0	0	0	12
IndustrProxIndx	8	26.7	0	1	1	9
PrenatlVax	0	0.0	0	0	0	0

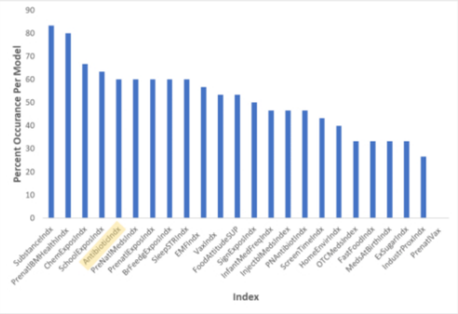

Figure 1. Percentage of decision trees in which a variable occurs (Pvar). Index variables need only occur once to be included.

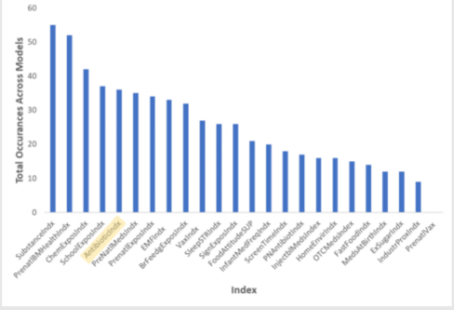

Figure 2. Total number of occurrences (NT) registered across all decision tree models.

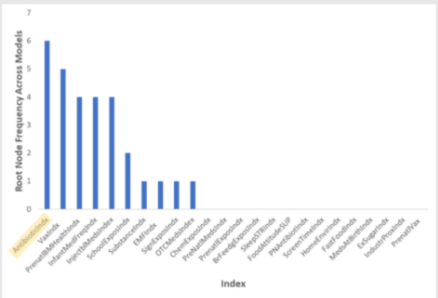

Figure 3. Occurrence of index variables at the root node (first decision point, P1) across 30 decision tree models. See Table 1 in Methods for list of the chronic conditions that comprise target variables

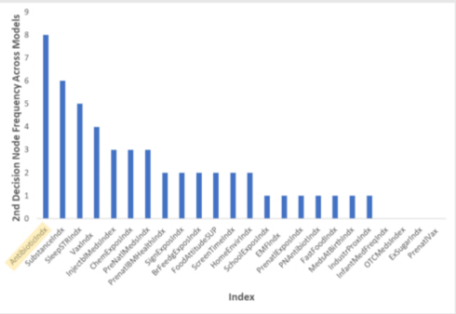

Figure 4. Occurrence of index variables at secondary decision nodes (P2). See Table 1 in Methods for list of the chronic conditions that comprise target variables

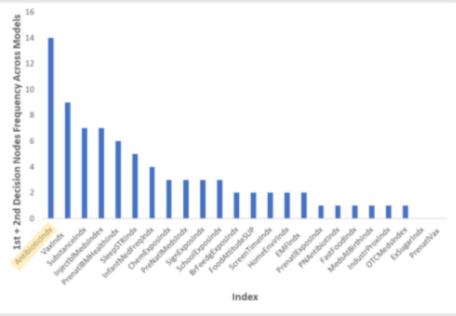

Figure 5. Occurrence of index variables at first (root) and second decision nodes (P1 + P2). See Table 1 in Methods for list of the chronic conditions that comprise target variables.

DISCUSSION

When considering occurrences across all decision tree models, the Antibiotics Index was of moderate importance, being equal to or less prevalent than eight other index variables. The Antibiotics Index was, however, most prevalent at both the root node and the second decision point nodes, suggesting that the use of antibiotics is a significant factor in the modulation of a diverse set of health conditions. We consider the first and second decision points to be more informative than overall prevalence because these top-level decision points reflect objective parameter importance measures determined by the decision tree model, whereas overall prevalence reflects varying dependencies on model complexity. Decision tree models vary in number of nodes, with some decision paths involving many nodes while others involve few. All of our models had at least two decision points and were thus comparable.

The methods we have demonstrated here are useful for understanding summary characteristics of machine learning models (especially decision trees) that are related thematically. Analytic techniques—indeed, numbers in general—only have utility because they integrate across domains of meaning. For example, the number two represents all sets of things that can be operationally (meaningfully) organized into pairs. This is not a tautology, but rather a definition coded in numerical symbology. Analytical symbology is merely a more complex application of definitions and consistent logic. The analysis that we present herein derives meaning from the potential relatedness of diverse chronic health conditions and aggregates that meaning via occurrences and positions of occurrence (parameter importance measures) across empirical models.

It is tempting to speculate about the relationship between the prevalence of antibiotic use and serious adverse health outcomes, but we consider these results to be preliminary, suggesting further research rather than speculation or conclusion.

.

REFERENCES

Nelson, Josie, Randall S. Reiserer, Beth Lambert, Stephanie Marango, and Martha R. Herbert (2020). Modeling cumulative environmental stressors using a consilience approach supports a Total Load model of chronic childhood health conditions (Poster presentation, IFM/AIC 2020).

Reiserer, Randall S., Beth Lambert, Josie Nelson, and Martha R. Herbert (2021). Empirical Analysis and Optimization of Indexed Data for Studies of Synergistic Interactions Among Multiple Stressors on Health Outcomes and Resilience in Children (Poster presentation, IFM/AIC 2021)

METHODS

Data came from the DHP CHIRP Survey. We used decision trees (a common machine learning analysis) to classify 366 subjects using 24 scaled index variables that capture information about the magnitudes of various environmental stressors (see Nelson et al., 2020 and Table 1 in Reiserer et al., 2021,). As with the other stressor indices, the Antibiotics Index is a composite quantification of several CHIRP Survey questions that capture information about the history, timing, duration, circumstances, adverse reactions, and magnitude of exposure to antibiotics, including prenatal exposure. Each element in an index is first quantified then scaled such that it contributes an equal part to a cumulative score (see Nelson et al., 2020). The index is thus a hypothesis that can be interrogated using alternate scoring schemes and tests of statistical relationships.

Target variables were binary, representing presence or absence of 30 chronic condition diagnoses (see Table 1 below). Thirty individual decision trees were generated (one for each diagnosis) using Rattle software and the R statistical environment. Individual trees were used to score derivative variables that reflect the prevalence and relative importance of stressor categories across the set of chronic conditions examined. The following derivative variables were defined: 1) Pvar: variable occurs in tree model at least once, 2) P1: presence of variable at the root node (first decision point), 3) P2: presence of variable at the second node (second decision point[s]), and 4) NT: number of total occurrences across all models. Binary and count variables were totaled across all tree models to derive summary variables (see Results, Table 2).

Table 1. Abbreviations of chronic conditions. These variables are coded as binary (present/absent).

Abbreviation	Chronic Condition
GIDis	Gastrointestinal disorder
MetabolCond	Metabolic condition
ADD-ADHD	Attention deficit [hyperactivity] disorder
Allergy	Allergy
AutoImmDis	Autoimmune disorder
Anemia	Anemia
Asperger	Asperger's Syndrome
Asthma	Asthma
AuditProcDis	Auditory processing disorder
ASD	Autism Spectrum Disorder
Cancer	Cancer
DevelDelay	Developmental delay
Eczema	Eczema
Epilepsy	Epilepsy
GastReflux	Gastric reflux
Hypospadia	Hypospadia
Jaundice	Jaundice
LearnDis	Learning disability
Migraines	Migraines
MitochondDis	Mitochondrial disorder
MoodDis	Mood disorder
MultChemSens	Multiple Chemical Sensitivity
PANDAS	Pediatric Autoimmune Neuropsychiatric Disorders Associated with Streptococcal Infections
PrimryImmunodef	Primary immunodeficiency
SensryProcDis	Sensory Processing Disorder
SleepApn	Sleep Apnea
ThyrdCond	Thyroid condition
VisualProcDis	Visual Processing Disorder
DiagUnderWeight	Diagnosed underweight
DiagOverWeight	Diagnosed overweight