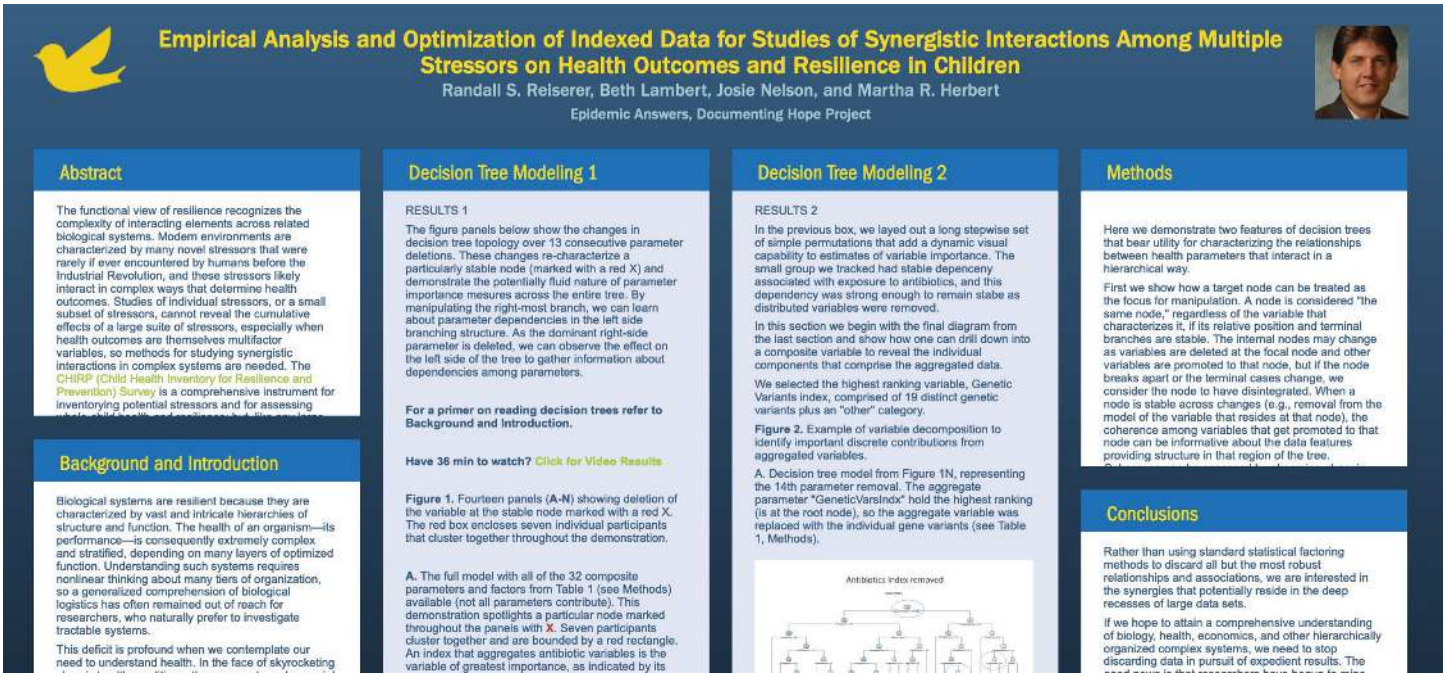
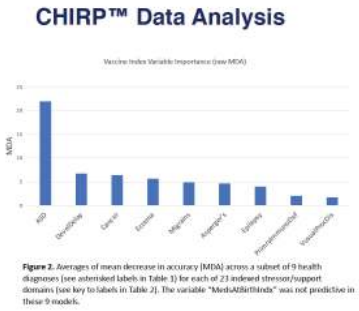
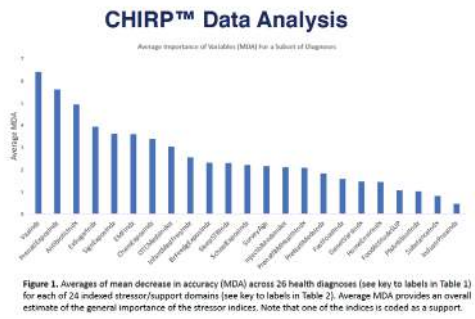
EMPIRICAL ANALYSIS AND OPTIMIZATION OF INDEXED DATA FOR STUDIES OF SYNERGISTIC INTERACTIONS AMONG MULTIPLE STRESSORS ON HEALTH OUTCOMES AND RESILIENCE IN CHILDREN

RANDALL S. REISERER, BETH LAMBERT, JOSIE NELSON, AND MARTHA R. HERBERT
EPIDEMIC ANSWERS, DOCUMENTING HOPE PROJECT







ABSTRACT

The functional view of resilience recognizes the complexity of interacting elements across related biological systems. Modern environments are characterized by many novel stressors that were rarely if ever encountered by humans before the Industrial Revolution, and these stressors likely interact in complex ways that determine health outcomes. Studies of individual stressors, or a small subset of stressors, cannot reveal the cumulative effects of a large suite of stressors, especially when health outcomes are themselves multifactor variables, so methods for studying synergistic interactions in complex systems are needed. The CHIRP (Child Health Inventory for Resilience and Prevention) Survey is a comprehensive instrument for inventorying potential stressors and for assessing whole-child health and resilience; but, like any large health and lifestyle survey, discovering patterns depends on the reliability of interpretive analytics. Data aggregation is commonly used in health research, especially in cumulative impact studies where many stressors contribute to multifaceted health outcomes, but aggregation methods rarely incorporate empirical integrity assessments.

Citation: *Reiserer, Randall S., Beth Lambert, Josie Nelson, and Martha Herbert (2020). 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).*

We previously reported on a novel system for aggregating data into hierarchically structured indices purpose-built for studying synergistic effects among and between large, stratified sets of potential environmental stressors and complex health outcomes (Nelson, et al., 2020). Our hierarchically structured indices effectively revealed cumulative correlations, supporting a "Total Load" model of chronic disease in children, but indices are hypotheses about relationships that need to be tested. To understand and optimize the effectiveness of data aggregation by our method, we systematically studied component data elements using predictive modeling approaches. Here we present and demonstrate our methods for systematically de-aggregating indexed data to fully interrogate its hierarchical structure and quantify the contributions of stratified data elements to various health outcomes. These modeling methods offer deep insight into synergistic couplings of many seemingly independent variables and allow indices to be empirically optimized across multiple outcome variables.

BACKGROUND AND INTRODUCTION

Biological systems are resilient because they are characterized by vast and intricate hierarchies of structure and function. The health of an organism—its performance—is consequently extremely complex and stratified, depending on many layers of optimized function. Understanding such systems requires nonlinear thinking about many tiers of organization, so a generalized comprehension of biological logistics has often remained out of reach for researchers, who naturally prefer to investigate tractable systems.

This deficit is profound when we contemplate our need to understand health. In the face of skyrocketing chronic health conditions, the urgency to make crucial connections cannot be overstated. Modern environments are so saturated with chemical, physical, and social stressors that understanding which ones contribute to declining health is a monumental feat.

The problem is manifold, but one area where progress can and has been made is in reducing many variables into a few. We have several methods for accomplishing this reduction, but we must usually grapple with a trade-off. How much information—and thus understanding—is lost in the process of condensing many variables into one? Whether we use statistical factor reductions or data aggregation methods, the answer is almost never known. If we are to optimize our environment, or at least improve it, we need to understand how multitudes of potential stressors interact so we might compensate for important losses of information.

A related issue is the question of what exactly is being represented in reduced variables. We hope that we are actually capturing a hidden variable, one that we cannot measure directly, but which materializes somewhere in the operations we perform when reducing variables to manageable proxies. Unfortunately, the notion that our aggregated data faithfully render the hidden variation can rarely be tested. We have an interest in remediating this deficit.

Machine learning and artificial intelligence (AI) offer a means to make real progress toward understanding complex systems and these useful algorithms are already widely used in clinical and research contexts. Large health data sets are often aggregated into indices, which are usually constructed based on logical relationships (those that make clinical or scientific sense). Often these relationships are so complex that we have no easy way of interrogating them to test their veracity (to validate them). It is important to remember that indices derived non-empirically are hypotheses that need to be tested before we rely on them to support scientific conclusions.

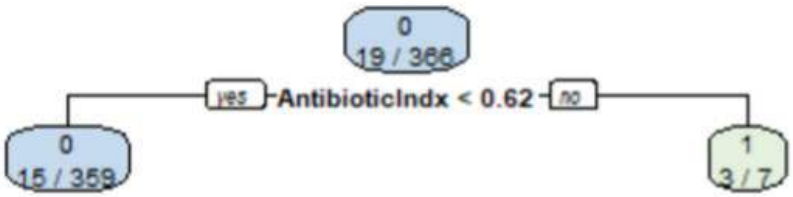
Here we offer a preview of methods we are developing to interrogate indexed data from a comprehensive survey of children's health, the CHIRP (Child Health Inventory for Resilience and Prevention) Survey. We previously reported on the construction and utility of our data indexing system (Nelson, et al., 2020). The clinically informed indices we generated are presently subjected to systematic deconstruction and characterization using decision trees, an accessible machine learning algorithm that offers visual comprehension.

This presentation is not a tutorial, but rather a methodological expose of one facet of the analytic procedures that we are developing to understand the landscape of health stressors and supports that impact the wellbeing of children and adults. This is an ongoing enterprise that we are happy to open-source. The urgency of understanding the crisis in public health is reason enough to share insights and encourage innovation from others.

Citation: *Reiserer, Randall S., Beth Lambert, Josie Nelson, and Martha Herbert (2020). 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).*

HOW TO INTERPRET OUR DECISION TREES

Decision trees are easy to read with little practice. The trees represent branching decision logic with starting and ending points. Unlike real trees and many other branching diagrams, decision trees are usually depicted upside down, with the root at the top spreading out toward leaves at the bottom. There are three types of nodes in decision trees, root, internal, and leaf (terminal branches). The following image shows a root node and two internal nodes.



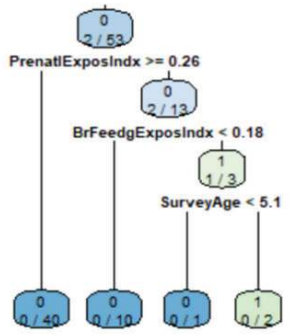
The numbers can be in one of several formats, but in this sample and in this presentation the numbers in all nodes are read as follows:

1. The top number is the reference for the binary condition for which the branching logic was generated. 0 = absent and 1 = present. The nodes are conveniently colored to depict the reference number that is dominant at the node, with blue corresponding to zero and green denoting one. In our examples 1 = presence of gastric reflux diagnosis.
2. The number on the left of the slash represents the number of cases that do not correspond to the reference number, so for the root node above there are 19 that are not zero (=19 ones).
3. The number right of the slash is the number of total cases of both ones and zeros at the node.

Below the root and internal nodes is a conditional statement, and this is the only tricky part of the diagram. To understand the decision point correctly, one must correctly interpret the logical statement. In the example above The logic says that the "AntibioticIndx" parameter is less than 0.62--easy enough--but then there is a left side labeled "yes" and a right side labeled "no". For the trees we present, the yes and no labels are always oriented in the same direction.

The tricky part is in understanding which side is high and which is low. For a greater than symbol (>), the yes represents the high side (is it greater than x? Yes it is greater than x). If, however, the logical statements contains a less than symbol (<) the high and low sides are reversed. As you can see, this is not all that tricky, but it can be confusing to the uninitiated.

The leaf nodes are much like the root and internal nodes, except they do not require a logical statement. The numbers are read in the same way as other nodes.



The four nodes at the bottom of the image above illustrate how things work. There are 53 cases represented. In fact, the top node tells you this. The next node down says there are thirteen cases below that node and three for the green internal node. The leaf nodes at the bottom tell you (on the left of the slash) how many cases are present after the final decision has been made. The zeros on the right of the slash tell you that there are no misclassifications (that section of the tree is fully resolved). If one or more of those right-side zeros were greater than one, that number would tell you how many misclassified cases were assigned to that leaf node.

Citation: *Reiserer, Randall S., Beth Lambert, Josie Nelson, and Martha Herbert (2020). 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).*

DECISION TREE MODELING 1

RESULTS 1

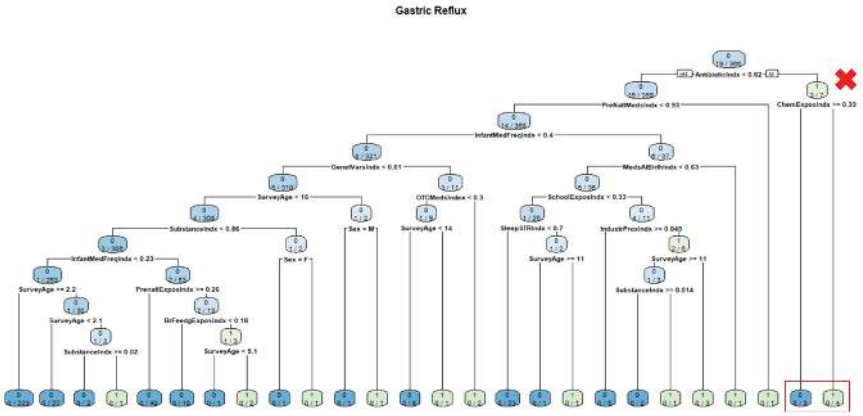
The figure panels below show the changes in decision tree topology over 13 consecutive parameter deletions. These changes re-characterize a particularly stable node (marked with a **X**) and demonstrate the potentially fluid nature of parameter importance mesures across the entire tree. By manipulating the right-most branch, we can learn about parameter dependencies in the left side branching structure. As the dominant right-side parameter is deleted, we can observe the effect on the left side of the tree to gather information about dependencies among parameters.

For a primer on reading decision trees refer to Background and Introduction. Have 36 min to watch? Click for Video Results

Figure 1. Fourteen panels (A-N) showing deletion of the variable at the stable node marked with **X**. The red box encloses seven individual participants that cluster together throughout the demonstration.

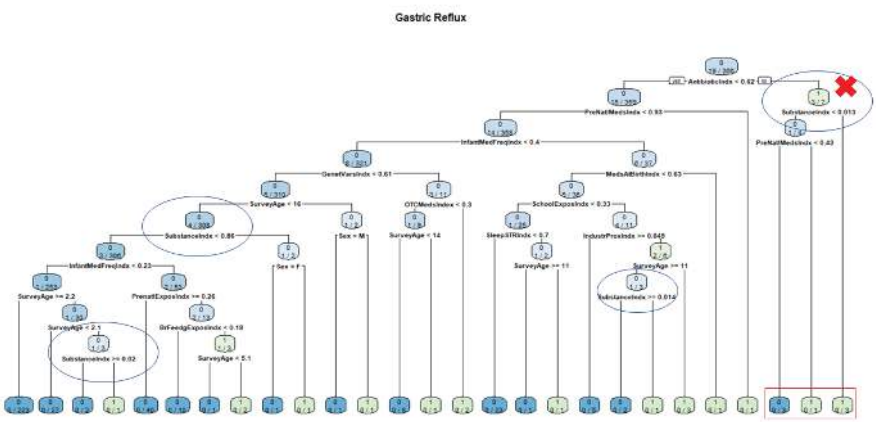
A. The full model with all of the 32 composite parameters and factors from Table 1 (see Methods) available (not all parameters contribute). This demonstration spotlights a particular node marked throughout the panels with **X**. Seven participants cluster together and are bounded by a red rectangle. An index that aggregates antibiotic variables is the variable of greatest importance, as indicated by its position at the root node (top of tree). Note the varables that occupy the left-side backbone of the tree and the higer level branchings. Changes in the prominence of these variables provide information about the importance of and interactions between variables that is not evident in a complete, static model.

Full model with all composite variables



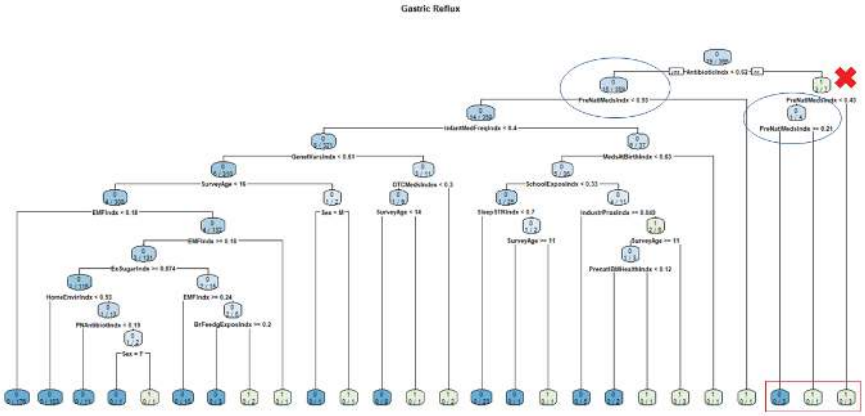
B. Model identical to A, but with Chemical Exposures Index removed. Note the circled nodes that denote the positions of the variable that has been newly promoted to the position marked **X**. The red box encloses the same participants as in A, but in an altered configuration.

Chemical Exposures Index removed



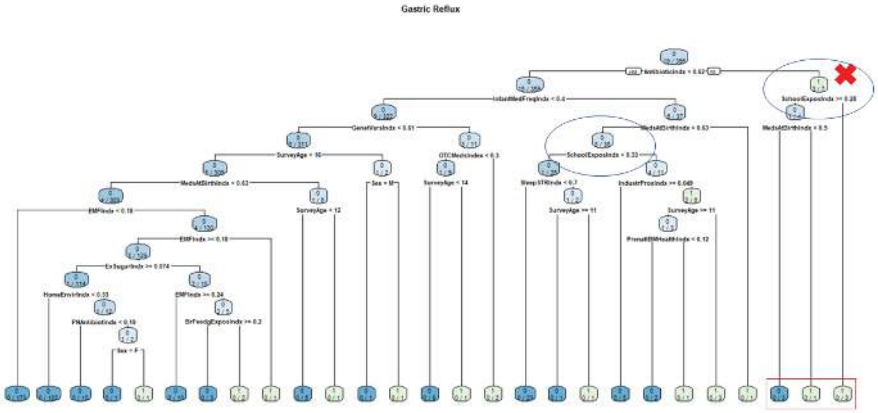
C. Model as in B minus mother's Substance Use Index. Circles highlight where the newly promoted variable (Prenatal Medications Index) is repeated in the tree. The Excess Sugar Index showed up for the first time in the tree, and will continue to play a role. The red box did not change in this iteration. Prenatal Medications Index dominates the two peultimate decision points.

Mother's Substance Use Index removed



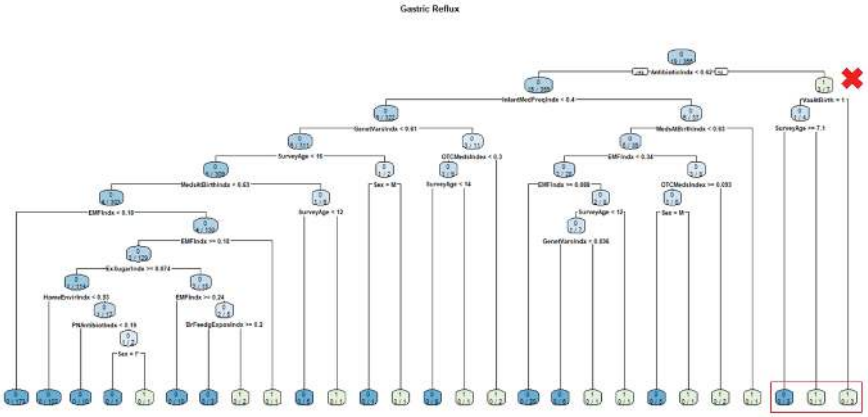
D. Model as in C minus Prenatal Medications Index. The newly promoted variable in the right-side branch is School Exposures Index. Note how persistent many of the variables are on the left. The deletion of a high-level variable caused global shifts in dependencies, but most of the same variables remained important. As variables are deleted, other variables shift around the tree, appearing and disappearing as variables gain and lose dominance over regions of the tree topology. Could measures of persistence during prescribed change sequences provide fresh perspectives on variable importance?

Prenatal Medications (used by mother) removed



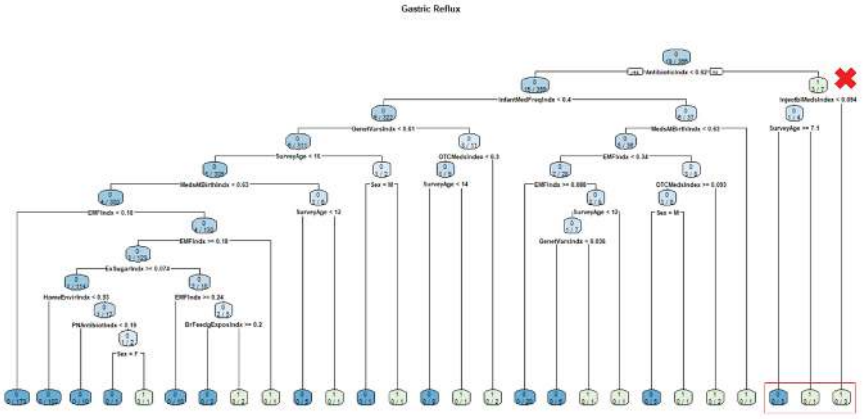
E. Model as in D except that the School Exposures Index has been removed. The newly promoted binary factor indicating presence or absence of a first-day-of-life vaccine appears only at the focal node, **X**. As we proceed, some variables make a brief appearance then disappear as relationships between variables change. The persistence of parameters provides greater confidence that those factors are robustly correlated with the condition under investigation.

Child School-related Exposures removed



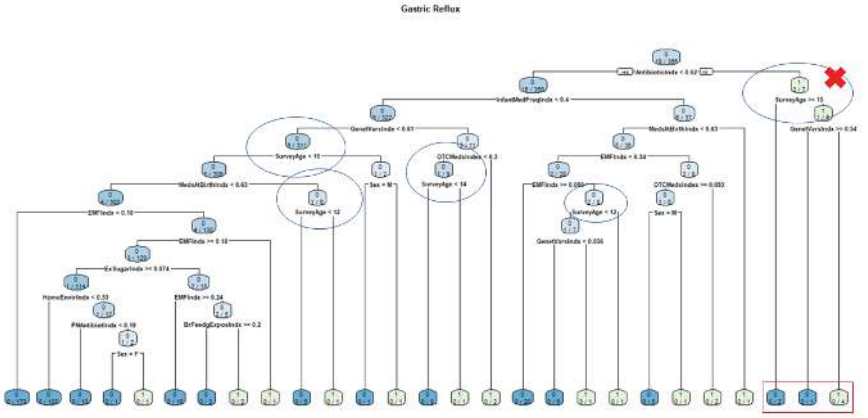
F. Model as in E minus the vaccine parameter. The newly promoted index concerns injectable medications and, despite its appearance near the root node, like its predecessor it appears to be of limited importance. These fleeting appearances underscore the value in a dynamic approach to assessing variable importance. Note again that the cases enclosed in the red box have not changed since the first parameter deletion, indicating correlated relationships with respect to gastric reflux diagnosis.

Child Vaccinated at Birth removed



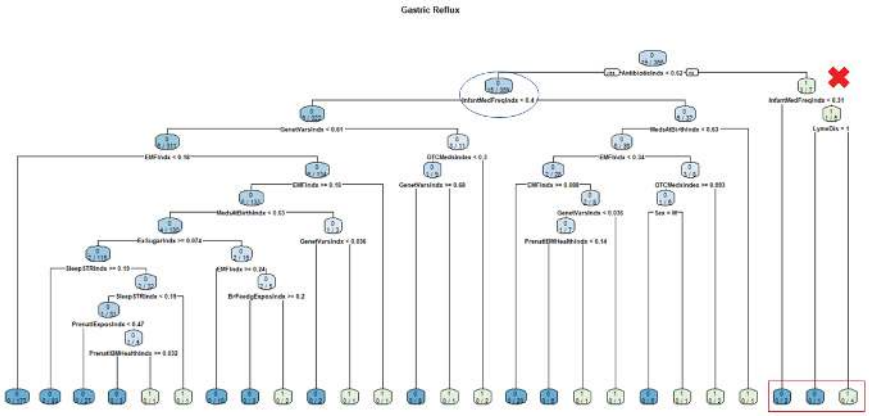
G. Model as in F minus Injectable Medications Index. The promoted numeric variable, Survey Age, has been present all along and continues to show a strong presence. This persistence might be important because, while age is sometimes an artifactual covariate of another variable, here it is persistent and consistent. It appears to be telling us that older children in this population are more susceptible to gastric reflux. Note now that the groupings in the red box have finally changed.

Child's Injected Medications removed



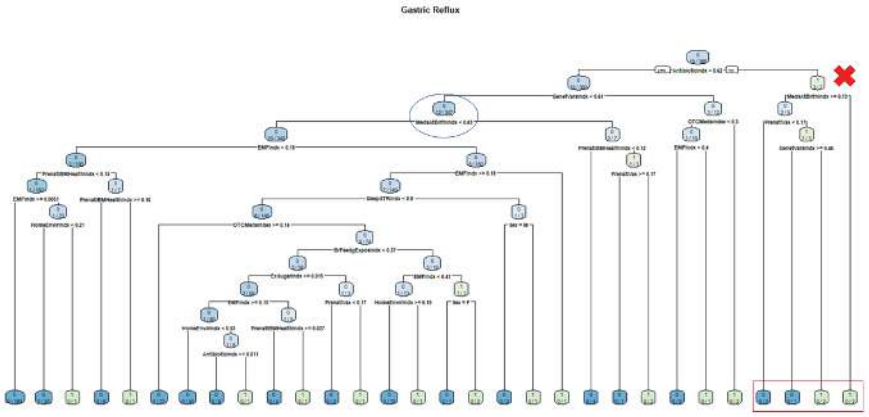
H. Model as in G but without Survey Age. The ascendant variable at the marked node is Infant Medications Frequency Index. This deletion had a global effect on the tree topology, but familiar variables are still informing the topology. The Lyme Disease factor is not informative about the reflux condition. The relationships circumscribed by the red box remain the same.

Child's Survey Age removed



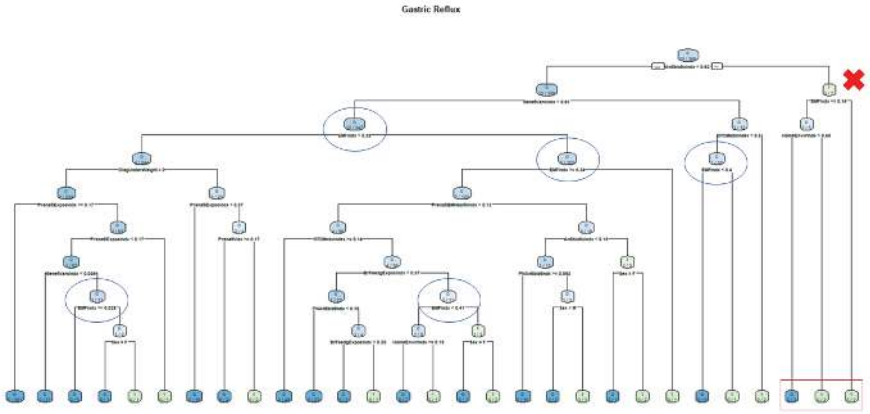
I. In this model, Infant Medication Frequency Index is replaced at the focal node by a similar index that aggregates medication variables. The high-level deletion made global changes on both sides of the tree. The red-boxed relationships have assumed a new configuration, but represent the original cases.

Medication Frequency in Infancy removed



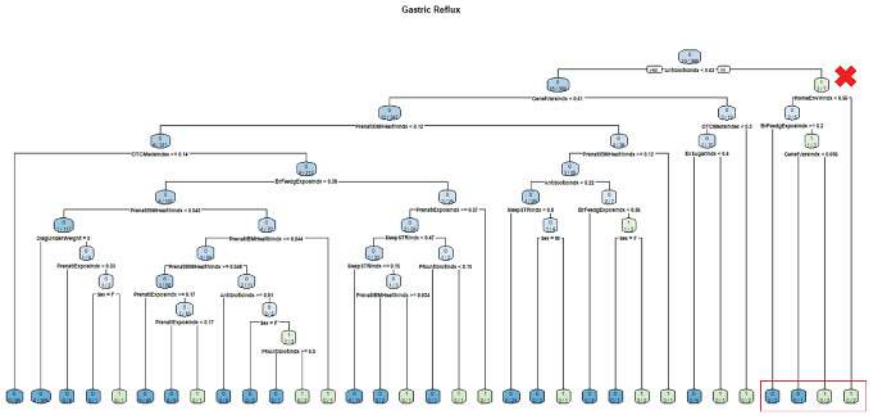
J. This model replaces Medications at birth and Electromagnetic Frequency Index rises to replace it at the focal node. This index is well represented in previous topologies, raising some interesting questions about the role of EMF in gastric health. When a variable starts out at a low level and continues to be promoted up the tree, appearing in new places as dependencies change, it indicates that the variable is important, but we would not know this from a static but complete model. The relationships in the red box are again in a new configuration.

Medications at Birth removed



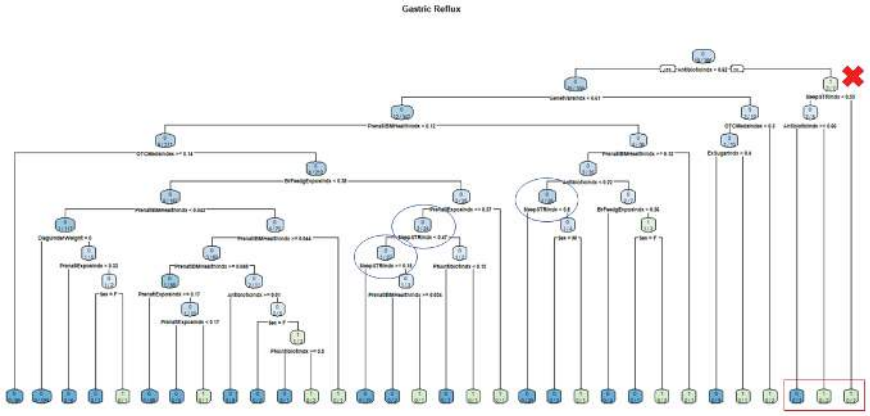
K. Removal of EMF Index promotes another new-comer to the focal node. The Home Environmental Index relates to dust mites, asbestos, radon, and other home hazards. It does not occur in the complete model, but shows up on the left after recent deletions, indicating that it might be more important than we would conclude without performing these operations. The terminal and internodal relationships in the red box repeat the topology from I above.

Electromagnetic Radiation Index removed



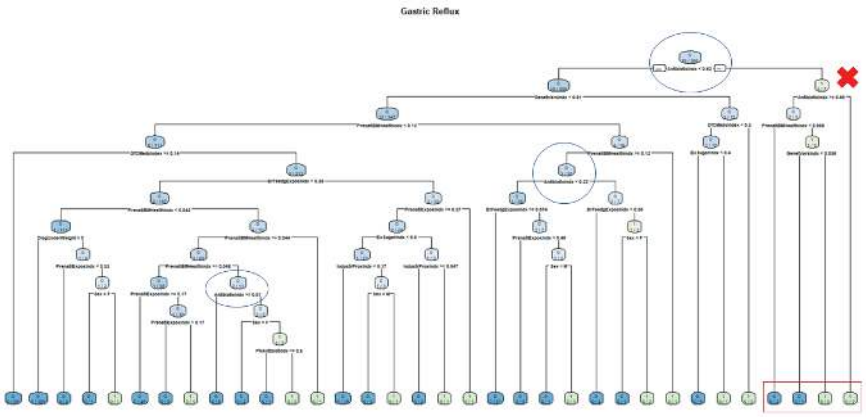
L. Removing Home Environmental Index promotes a familiar variable to the focal node. The Sleep Stressor Index has been a fairly steady player from the start. The red boxed relationships are also familiar.

Home Environmental Factors Index removed



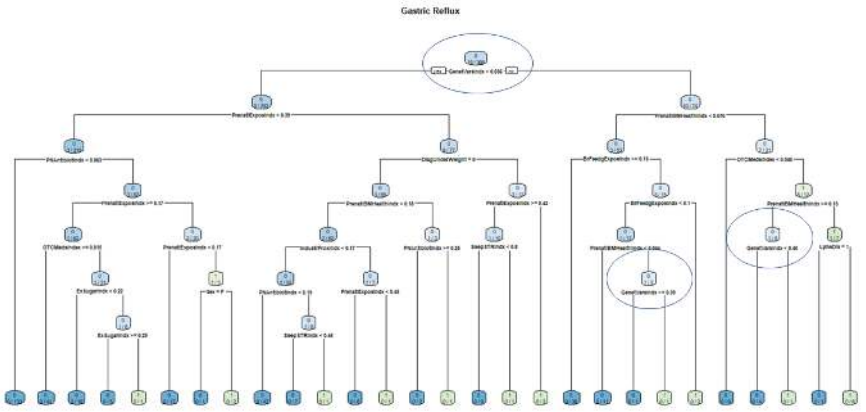
M. Removal of the Sleep Stressor Index places the Antibiotic index at both the ultimate and focal nodes. The red box still encompasses the same individuals, which have remained subject to the focal, **X**, node for 13 parameter removals, some of which dramatically reorganized other parts of the tree. The stability is due to the dominance of the Antibiotics Index.

Child's Sleep Stressors Index removed



N. The 14th parameter removal disintegrated the focal node and reorganized the tree in other ways. Despite deleting 14 out of 36 variables, all terminal nodes remained resolved. We could have chosen to focus on the ultimate node, or any other node of interest. By selecting the right-side penultimate node, we were able to explore the dependencies that affected both the right and left sides of the tree. Other operations would provide different information, and by combining different exploratory paths, we can gain insight about the broad dependencies that influence a health outcome. The alternative is to conclude whatever the complete model indicates, but we would miss details that might be crucial to health outcomes.

Antibiotics Index removed



DECISION TREE MODELING 2

RESULTS 2

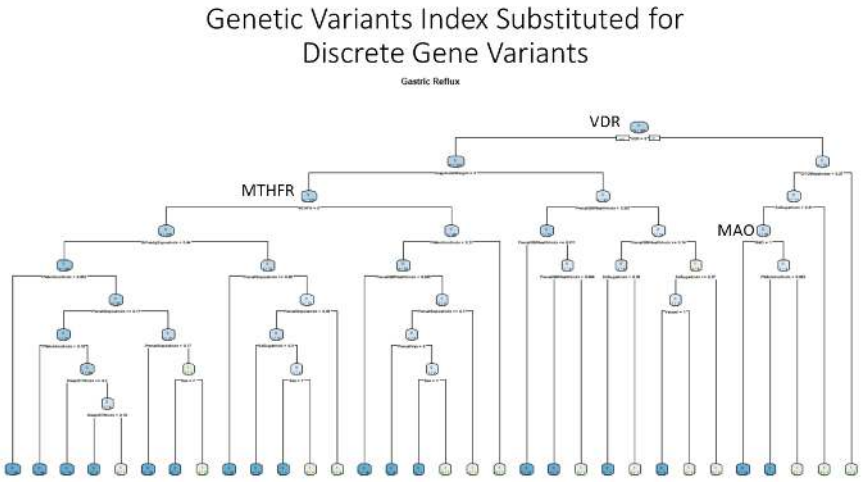
In the previous box, we layed out a long stepwise set of simple permutations that add a dynamic visual capability to estimates of variable importance. The small group we tracked had stable depenceny associated with exposure to antibiotics, and this dependency was strong enough to remain stabe as distributed variables were removed.

In this section we begin with the final diagram from the last section and show how one can drill down into a composite variable to reveal the individual components that comprise the aggregated data.

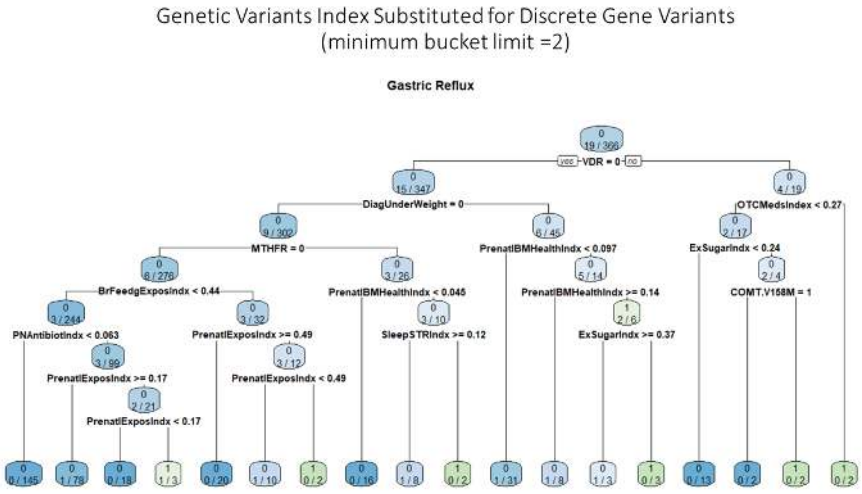
We selected the highest ranking variable, Genetic Variants index, comprised of 19 distinct genetic variants plus an "other" category.

Figure 2. Example of variable decomposition to identify important discrete contributions from aggregated variables.

A. Decision tree model from Figure 1N, representing the 14th parameter removal. The aggregate parameter "GeneticVarsIndx" hold the highest ranking (is at the root node), so the aggregate variable was replaced with the individual gene variants (see Table 1, Methods).



B. Modeled tree after replacement of aggregate variable with discrete gene variants (binary data).



METHODS

Here we demonstrate two features of decision trees that bear utility for characterizing the relationships between health parameters that interact in a hierarchical way.

First we show how a target node can be treated as the focus for manipulation. A node is considered "the same node," regardless of the variable that characterizes it, if its relative position and terminal branches are stable. The internal nodes may change as variables are deleted at the focal node and other variables are promoted to that node, but if the node breaks apart or the terminal cases change, we consider the node to have disintegrated. When a node is stable across changes (e.g., removal from the model of the variable that resides at that node), the coherence among variables that get promoted to that node can be informative about the data features providing structure in that region of the tree. Coherence can be assessed by observing changing patterns of dependency in other parts of the tree. This procedure can be performed at any internal node or the root node. It is most easily interpreted when the terminal (leaf) nodes are resolved (i.e., fully classified), or when cases can be identified as they shift in the tree topology.

Second, we demonstrate how decomposition of a composite variable can provide insight about that variable's influence on overall tree topology. In the case illustrated, we elected to "decompose" genetic variants. By decomposition, we mean that an aggregated variable is substituted in the model for its de-aggregated components. The example we present is uncomplicated, but similar procedures (as well as other relational operations) can be used at any node characterized by a composite variable.

We used the statistical software R (R Core Team, 2013) and the package rpart (Therneau and Atkinson, 2019) to construct decision trees. Data were either binary, ranked, or numeric, but all were scaled to the range 0-1. Index structure and construction, as well as the instrument, is described elsewhere (Nelson, et al. 2020). Briefly, the structure of indices reflect natural hierarchies in the structure of health data. Many of the survey questions capture a binary response (yes/no, present/absent), but they belong to logical categories such that many binary questions can be aggregated into a numeric index. Other questions were Likert-style ranked responses, or time-referencing responses, and still others were numeric before aggregation. The scaling step was used to render all variables equivalent at a given level of aggregation. While there was clinically derived logic in the hierarchical structuring of components (questions, scores, subindices, and indices), this type of aggregation is clearly a hypothesis. The methods herein were formulated to examine explicitly hierarchical hypotheses.

The analysis began with stressor-variable subindices, high-level components comprised of two to four levels of integration, as well as other variables deemed likely to covary with stressors (see Table 1). The dependent variable was presence vs. absence of gastric reflux. Gene variants were initially aggregated into the variable named "GenetVarsIndx," and only incorporated into the model for last analysis.

Decision trees were generated using the rpart package in R, with model parameters set to maximize resolution at leaf nodes (unless noted). The complete model included all composite variables and all factors except for individual genes. Variables were then excluded from the model one at a time, starting with the penultimate variable in the selected branch (right-most branch in trees). This branch characterized seven participants, four of whom had been diagnosed with gastric reflux. Upon deletion of the target variable, the algorithm was rerun with no additional changes and the output graph was captured for further analysis (see Decision Tree Modeling section). This selective variable deletion was used to assess the associations between variables and detect relative dependencies among variables by examining changes in dependencies and variable prevalence in the rest of the tree.

Variables often appeared at more than one node in the tree. The criterion for deletion, however, was the presence of a particular variable at the selected node (upper right of the tree). As long as any variable characterized the selected node, the deletion routine continued, but when the node disappeared (i.e., the tree structure became fundamentally rearranged), the deletion routine was halted. Another halt criterion specified that no misclassifications could occur in the tree. That is, all terminal (leaf) nodes (those at the bottom of the tree) had to be fully resolved. This halt criterion was not met, as all leaf nodes remained resolved throughout the deletion routine (i.e., the first criterion was met before the second occurred).

Decision trees used the GINI index to rank variable importance. These values can be recovered from the rpart model. The GINI values are specific to the modeled nodes, so they must be systematically gathered for each tree generated. We did not include them here, but they can be used to characterize relative changes in parameter importance within and between similar trees. GINI values can, therefore, be used to automate this analysis.

For this presentation, the final tree was generated by substituting the composite variable GenetVarsIndx (genetic variants index) for all of the individual genetic variants to see which genes were influencing the tree topology.

Note: We are careful in our interpretations, so because our sample was relatively small (366) we do not make any claims about the reliability of the results displayed. The data were used for demonstration purposes. The CHIRP Study is ongoing and we expect to generate robust results as additional data are gathered. Nevertheless, we consider decision tree models to be more reliable toward the root node (top node), especially when large samples characterize a given node. Generally, the top half of a tree is more reliable than the bottom half, because sample sizes decrease toward the leaf nodes.

Table 1. Variables used in decision tree analysis. Variables were included or excluded depending on analysis objectives.

Variable Name	Variable type	Brief Description
Sex	binary factor	Biological sex
SurveyAge	number	Age at survey completion
GastReflux	binary factor	Gastric reflux (dependent variable)
DiagUnderWeight	binary factor	Diagnosed underweight
DiagOverWeight	binary factor	Diagnosed over weight
DiagObese	binary factor	Diagnosed obese
DiagMorbObese	binary factor	Diagnosed morbidly obese
SubstanceIndx	number	Substance index (smoking, alcohol, drugs)
PreNatlMedsIndx	number	Prenatal medications index
MedsAtBirthIndx	number	Medications at birth index
PNAntibiotIndx	number	Prenatal antibiotics index
PrenatlVax	multi-level factor	mother's vaccinations during pregnancy score
InjMedIndx	multi-level factor	Mother's injectable medications score
PrenatlBMHealthIndx	number	Mother's pre-birth health scored as stressors
PrenatlExposIndx	number	Mother's pre-birth exposures to a variety of agents
BrFeedgExposIndx	number	Mother's exposures to a variety of agents while nursing
ChemExposIndx	number	Child's direct exposures to chemical stressors
EMFIndx	number	Child's exposure to electromagnetic radiation
IndustrProxIndx	number	Family home's proximity to a variety of industries
HomeEnvirIndx	number	Includes potential environmental hazards in the home
SchoolExposIndx	number	Includes potential environmental hazards in the child's school
SignExposIndx	number	Significant exposures not categorized above
ScreenTimeIndx	number	Child's electronic device screen time
FastFoodIndx	number	Child's fast food habits/access
ExSugarIndx	number	Child's excess sugar consumption
SleepSTRIndx	number	Sleep stressors index
OTCMedsIndex	number	Child's use of over-the-counter medications
InfantMedFreqIndx	number	Child's frequency and variety of medications
Vaxxed	binary factor	Whether vaccinated or not
VaxAtBirth	binary factor	Whether vaccinated at birth
InjectblMedsIndex	number	Child's injectable medications other than vaccines and antibiotics
AntibioticIndx	number	Child's composite antibiotics index
GenetVarsIndx	number	Genetic variants, a scaled summation of all genes below
ACAT	binary factor	Acetyl-CoA acetyltransferase
AHCY	binary factor	Adenosylhomocysteinase
BHMT	binary factor	Betaine-Homocysteine S-Methyltransferase
CBS	binary factor	Cystathionine-beta-synthase
CBS.A360A	binary factor	Cystathionine-beta-synthase A360A
CBS.C699T	binary factor	Cystathionine-beta-synthase C699T
COMT	binary factor	Catechol-O-methyltransferase
COMT.V158M	binary factor	catechol-O-methyltransferase V158M
CYP	binary factor	Cytochrome P450 variants
GST	binary factor	Glutathione S-transferases
MAO	binary factor	Monoamine oxidase
MTHFR	binary factor	Methylenetetrahydrofolate reductase
MTHFR.A1298C	binary factor	Methylenetetrahydrofolate reductase A1298C
MTHFR.C677T	binary factor	Methylenetetrahydrofolate reductase C677T
MTR_MTRR	binary factor	5-methyltetrahydrofolate-homocysteine methyltransferase/reductase
NOS	binary factor	Nitric oxide synthase
SHMT	binary factor	Serine Hydroxymethyltransferase variants
SUOX	binary factor	sulfite oxidase
VDR	binary factor	Vitamin D receptor
Other	binary factor	diagnosed with a gene variant not listed

CONCLUSIONS

Rather than using standard statistical factoring methods to discard all but the most robust relationships and associations, we are interested in the synergies that potentially reside in the deep recesses of large data sets.

If we hope to attain a comprehensive understanding of biology, health, economics, and other hierarchically organized complex systems, we need to stop discarding data in pursuit of expedient results. The good news is that researchers have begun to mine rather than scrape their data sets, and machine learning has made these advances possible.

In our first demonstration we, delved deep into a single node that robustly characterized a total of seven individuals. This node was stabilized by high exposure to antibiotics, and the Antibiotic Index dominated the entire tree.

Other relationships that appeared important include chemical exposures both pre- and post-natally, pre- and post-natal medications, substance use by mothers, electromagnetic radiation, Home hazards, sleep stressors, antibiotics and genetic variants. The operations we performed using rules for the focal node told us more about the rest of the tree than about the small sample under the focal node. We were able to observe how removal of parameters that we knew were of penultimate importance on the right shuffled parameters in the whole tree. In addition to assessing the dynamics of variables originally present on the left side of the decision tree, we detected variables that were not in the complete model. For example, school-related exposures, home environmental exposures, and excess sugar appeared enough to warrant curiosity and further study.

As much as we might learn from the explicit associations, we seem also to have lessons to learn from aggregate variables that apparently do not influence gastric reflux in children. One of the biggest surprises was that our Fast Food Index was silent on a GI disorder, and body weight diagnoses were nearly as quiet. Only one weight-related variable showed up in the analysis (Diagnosed Underweight), and it only appeared when individual genetic variants were considered.

Concerning genetic associations, the vitamin D receptor gene might predispose children to GI problems, and two other genes, MTHFR (which appears to be associated with underweight diagnoses) and MAO deserve further scrutiny.

A next logical move might be to reassemble the stripped model in a stepwise fashion, leaving the genetic variants decomposed. We could also add the variables back to a model in reverse of the deletion order, or just add each back independently to see how they get distributed in the tree without the dependencies from other deleted variables. The R programming language makes such operations quite accessible and we plan to add automated steps in the future.

The methods we presented lend themselves well to automations. In the future, we might be able to strip and reassemble aggregated indices in minutes to reveal a web of causal and associative interconnectivity.

If you have read this far, you see some of the potential we see and you want to know more about where this type of analysis might lead. Imagine a system that automatically completes the operations we demonstrated, but records in a database the ranks, dependencies, sample sizes, and number of occurrences of variables across a series of specific manipulations. The program might systematically expand and delete nodes in a decision model while writing metadata about each parameter that meets performance criteria. Then, it might use those measurements to perform a network analysis on the strengths of associations for all parameters that made it into the manipulated model. It could display the parameter importance using variable-sized circles, for example, with thick connecting lines between highly associated parameters. It might be programmed to identify super-associations based on threshold criteria and to organize them into domains. With such a system, we could readily assess truly complex sets of dependencies, and perhaps begin to get a handle on complex health problems.

One thing strikes us as certain: if we continue to throw data away because it failed to make it into the kitchen sink model, we will only ever identify the strongest signals and we will lose sight of the other factors that accumulate to threshold levels of synergy. After all, what is resilience if not remaining north of a threshold defined by many important synergies.

DISCLOSURES

The authors declare no conflicts of interest.

ACKNOWLEDGEMENTS

We wish to recognize the contributions of the following donors, sponsors, and collaborators

LivingMatrix

Pure Encapsulations

The John Brown Cook Foundation

The Lattner Foundation

The Westreich Foundation

Patricia S. Lemer

Stephanie Marango

Citation: *Reiserer, Randall S., Beth Lambert, Josie Nelson, and Martha Herbert (2020). 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).*

ABSTRACT

The functional view of resilience recognizes the complexity of interacting elements across related biological systems. Modern environments are characterized by many novel stressors that were rarely if ever encountered by humans before the Industrial Revolution, and these stressors likely interact in complex ways that determine health outcomes. Studies of individual stressors, or a small subset of stressors, cannot reveal the cumulative effects of a large suite of stressors, especially when health outcomes are themselves multifactor variables, so methods for studying synergistic interactions in complex systems are needed. The CHIRP (Child Health Inventory for Resilience and Prevention) Survey is a comprehensive instrument for inventorying potential stressors and for assessing whole-child health and resilience, but, like any large health and lifestyle survey, discovering patterns depends on the reliability of interpretive analytics. Data aggregation is commonly used in health research, especially in cumulative impact studies where many stressors contribute to multifaceted health outcomes, but aggregation methods rarely incorporate empirical integrity assessments. We previously reported on a novel system for aggregating data into hierarchically structured indices purpose-built for studying synergistic effects among and between large, stratified sets of potential environmental stressors and complex health outcomes (Nelson, et al., 2020). Our hierarchically structured indices effectively revealed cumulative correlations, supporting a "Total Load" model of chronic disease in children, but indices are hypotheses about relationships that need to be tested. To understand and optimize the effectiveness of data aggregation by our method, we systematically studied component data elements using predictive modeling approaches. Here we present and demonstrate our methods for systematically de-aggregating indexed data to fully interrogate its hierarchical structure and quantify the contributions of stratified data elements to various health outcomes. These modeling methods offer deep insight into synergistic couplings of many seemingly independent variables and allow indices to be empirically optimized across multiple outcome variables.

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).

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Therneau, Terry and Beth Atkinson (2019). rpart: Recursive Partitioning and Regression Trees.

URL https://CRAN.R-project.org/package=rpart/.