AIC 2024
Bellagio Hotel
Las Vegas, NV
May 29 - June 1

Discerning complex associations between stressors and health outcomes with machine learning

Randall Reiserer[1], Zaid Ahmed[1,2], Beth Lambert[1], Josie Nelson[1], Chris D'Adamo[1,3]

1 Documenting Hope Project
2 University of North Carolina, Charlotte
3 University of Maryland School of Medicine

ABSTRACT

More than any other emerging technology, machine learning and artificial intelligence promise to advance our understanding of the root causes of chronic illness. Machine learning offers new methods for mining vast data sets and for exploring patterns that emerge through complex dependencies. Health maps onto biology across many stratified domains of structure, function, and performance, and these intricate relationships are subject to equally complex and hierarchical environmental governors. Building on previous IFM AIC reports (2020, 2021, 2023), we continue our investigation of childhood chronic diseases using the CHIRP™ (Child Health Inventory for Resilience and Prevention) Survey, a comprehensive instrument for inventorying potential stressors and for assessing whole-child health and resilience. In this report we make use of data mining techniques, such as hierarchical clustering and decision trees, to discern broad patterns of interaction across a large number of environmental stressors that map onto several childhood chronic conditions.

INTRODUCTION

Novel health metrics are essential to the study of integrative medicine, a discipline founded on complex systems biology. Complex systems are vast, multiplex, optimization functions with many degrees of freedom. Sadly, traditional statistical methods are unwieldy, if not ineffective, at discerning patterns distributed across large causally linked arrays.

In three previous reports (Nelson, et al., 2020; Reiserer et al., 2021; Reiserer et al., 2023), we used cutting-edge data aggregation methods and machine learning to explore and characterize survey data (CHIRP^{TM} Survey) comprised of information about family and child health, environmental exposures, medical interventions, and lifestyle choices. A detailed understanding of health will require that such qualitative domains be translated into quantitative factors for analysis. The Documenting Hope Project is dedicated to that effort.

Here we expand our work to include hierarchical clustering analysis as a means to deconstruct clinically-informed health and stressor indices into their component parts, so that we can discern the sources of signal in composite metrics.

Our Antibiotics index was of primary concern, since a previous report found that this index was a dominant factor in analyses of childhood chronic illnesses (Reiserer et al., 2023).

METHODS

We used Scikit-Learn Machine Learning Library and the Python programming environment to conduct hierarchical clustering analysis (HCA) on index variables that quantify features of health and health stressors (all derived from the CHIRP Survey™; *n* = 416 eligible participants). The clinically informed metrics for stressors represent a theoretical (experience-based) aggregation of many domains of potential influence on health, while the health metrics represent aggregated observational assessments of health, vitality, disease diagnoses, etc.

We used Euclidian distance and Ward's linkage criterion to merge data into clusters. Output graphs from HCA were used to examine the similarities among hundreds of variables to determine whether individual variables (both aggregated and disaggregated) represented net sources of noise or signal.

For this analysis, our Antibiotics Index was disaggregated into its component parts. Output graphs and distributions were then used to assess the value of different index components.

How Does HCA Work?

Hierarchical clustering analysis (HCA) is an unsupervised machine learning algorithm capable of processing large numbers of variables. Through extensive graphs and display options, HCA offers intuitive insight into complex patterns in large data models. It is particularly useful for models with many variables.

The primary graph is the dendrogram (tree), with branch lengths that represent the distances between nested clusters (Figure 1). The dendrogram is a bird's-eye view of data relationships, providing a sense for how tight the clusters are.

Bar graphs offer another means of visually scanning for patterns within clusters. These graphs display cluster value differences and error bars, allowing for rapid visual classification. Our analysis suggested, for example (see Figure 2), that children with more medical diagnoses (Fig. 2 A) tended to be older (Fig. 2 B), but that children with gastrointestinal symptoms (Fig. 2 C) were not significantly different between clusters (and, indeed, age groups). Furthermore, Children with a history of intensive antibiotics use (a topic we have addressed elsewhere [Reiserer, et al., 2023]) were strongly partitioned by the two clusters (Fig. 2 D), suggesting that they, as a cluster, tended to be older children who have several diagnoses of chronic illness.

Our analysis involved many other parameters that our brief demonstration cannot accommodate.

Other graphs can be utilized to further classify clusters. For example, heat maps and histograms (Figure 3) are common output displays for HCA. Histograms display distributions for individual clusters. Figure 3 shows, for example, that age distributions are distinct between the two clusters, but the use of antibiotics is skewed and overlapping within these same clusters.

These visual displays are valuable for assessing relationships among clinically-informed index variables. In a previous study (Reiserer et al., 2023), we showed that heavy and frequent use of antibiotics was strongly associated with several common childhood chronic illnesses. Our Antibiotics Index captures 10 clinically relevant data points about 1) history of use, 2) durations of use, 3) antibiotic resistance, and 4) delivery modes. When we subjected all 10 variables to HCA, we learned that the delivery modes were not providing any significant signal to the models. Thus, we could eliminate them as sources of noise and improve the signal-to-noise-ratio of the trimmed index.

RESULTS

Numerous interesting associations emerged from the HCA, but we focus here on results associated with our Antibiotics Index. This antibiotics index consisted of 10 aggregated variables. After analysis using HCA, we detected that four variables related to modes of administration failed to contribute any significant signal to a model that resolved many other important health metrics. For example, the HCA model included number of diagnosed conditions, weight issues, general constitution, vitality, pain experienced, allergy severity, cardiovascular and pulmonary symptoms, gastrointestinal distress, chemical exposures and more, in addition to the antibiotics index.

A small sample of graphs from our HCA analysis is included here (Figures 1-4).

REFERENCES & DISCLOSURE

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).

Reiserer, Randall S, Nelson, Josie, , Beth Lambert, Heather Tallman Rhum, and Martha Herbert (2023). Exploring the impact of antibiotic use on general health status in children by aggregating across machine learning models (Poster presentation, IFM/AIC 2023).

The authors declare no conflicts of interest.

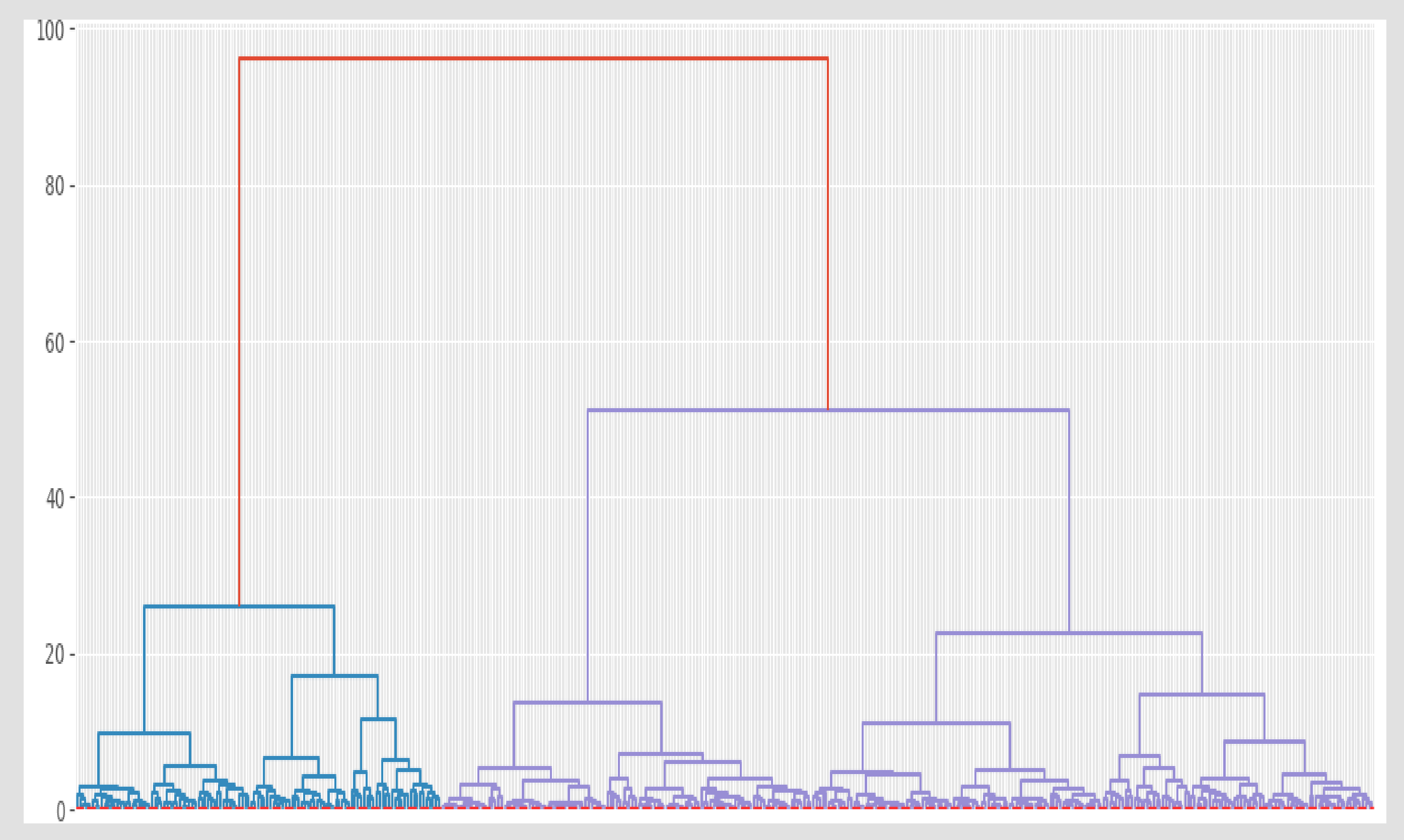

Figure 1. Dendrogram with two clusters (connected by red branches) from hierarchical clustering analysis (HCA). The branch tips represent 416 CHIRP Survey participants. Blue and purple branches denote the two main clusters (designated 0 and 1 respectively in other graphs).

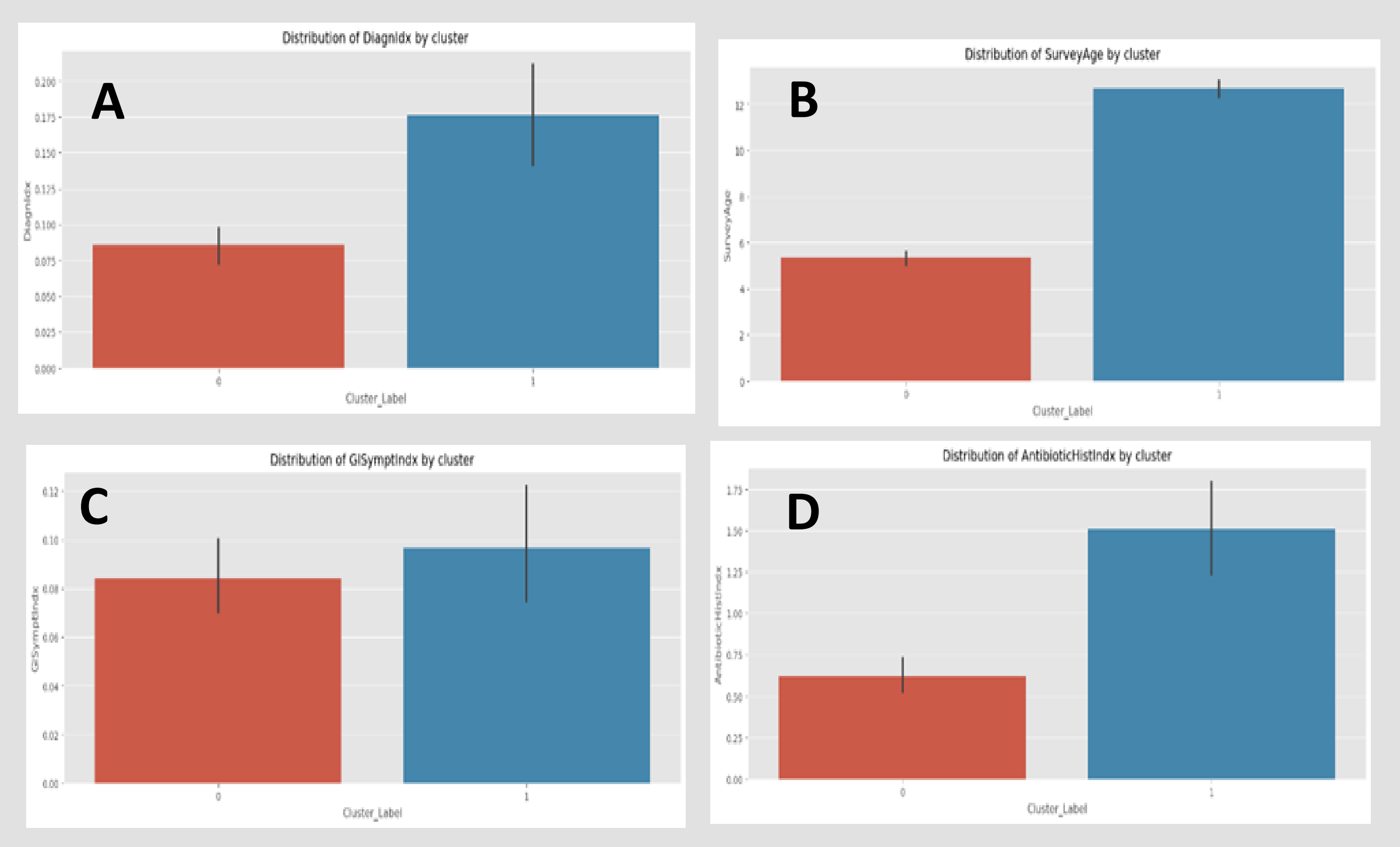

Figure 2. Bar graphs generated by HCA. Graphs include number of diagnoses (A), survey age (B), gastrointestinal symptoms (C), and Antibiotics History Index (D). A, B, and D share similarities in their distributions with respect to clusters, whereas C does not.

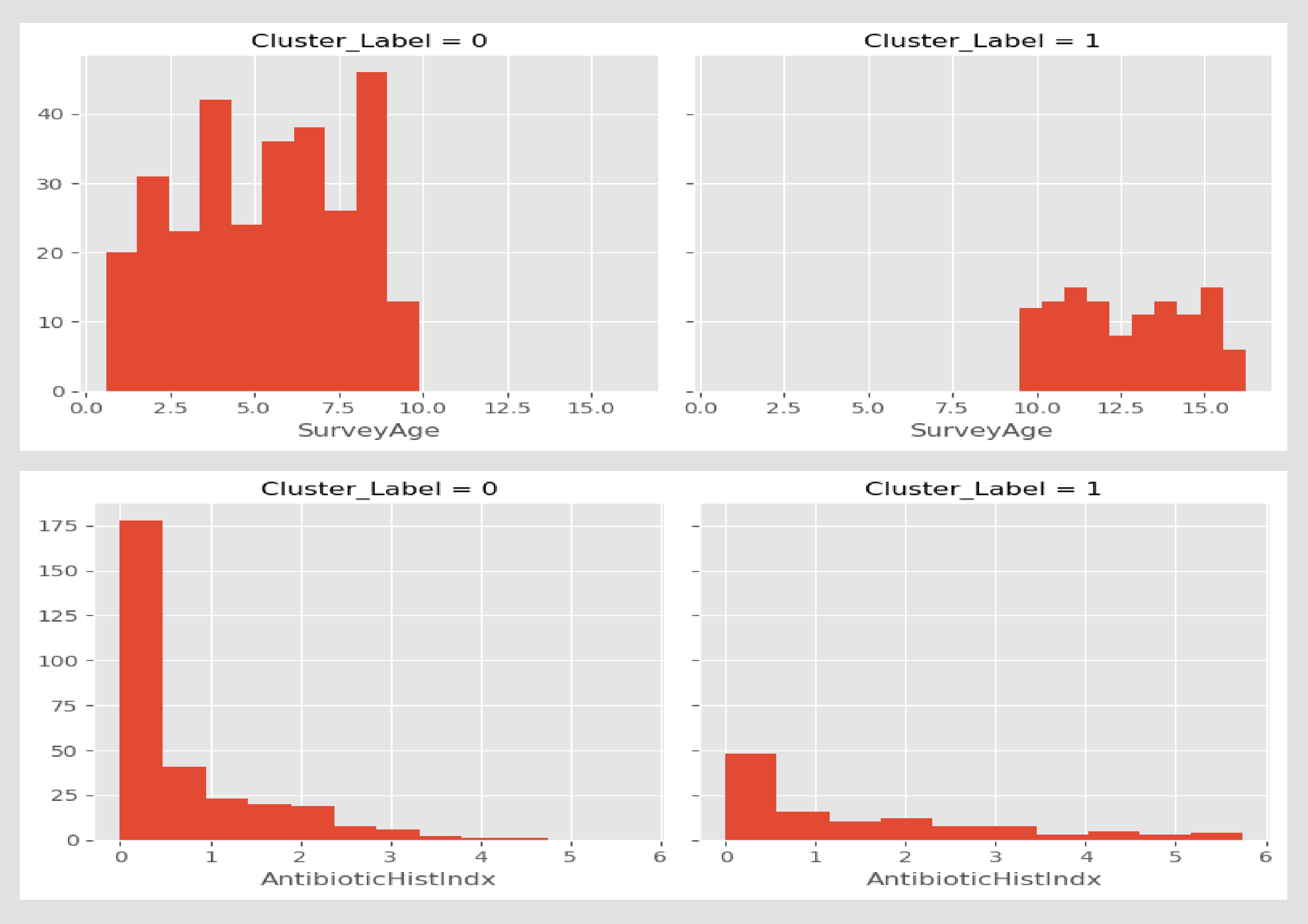

Figure 3. Histograms generated by HCA, showing the distributions of two metrics—survey age and antibiotics history—in each of two clusters. Survey age fell into distinct clusters, while the antibiotics history metric shows a broad and skewed distribution in both clusters.

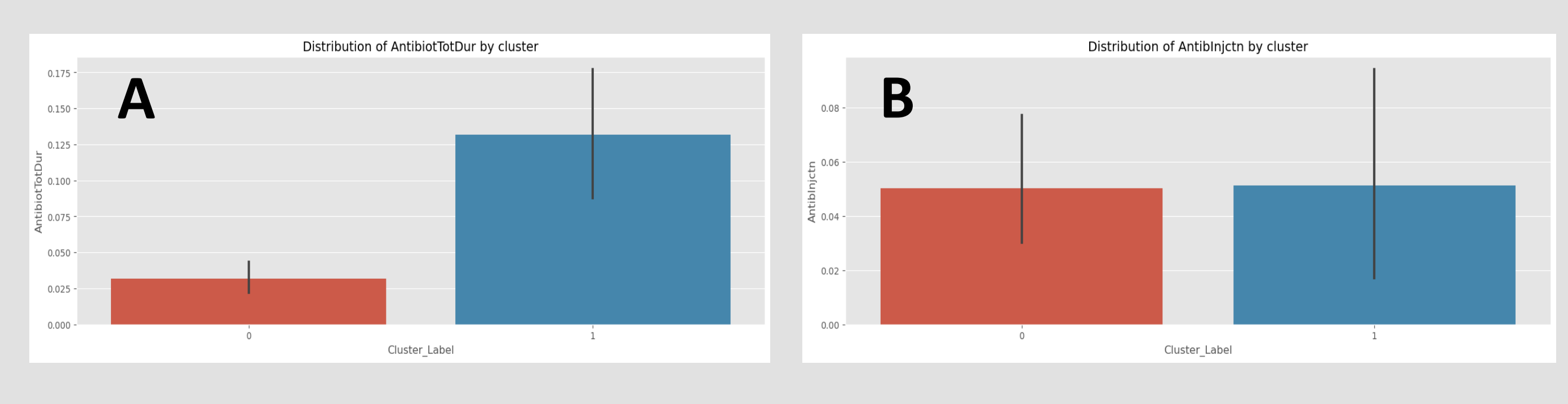

Figure 4. Bar graphs generated by HCA. Total duration of antibiotics (A) is shown for comparison with antibiotics delivered by injection (B). Other delivery modes had distributions similar to B.

CONCLUSIONS

Hierarchical clustering analysis (HCA) is a valuable tool for discerning patterns in large arrays of data. HCA allowed us to refine our Antibiotics Index by reducing noisiness contributed by four out of ten factors.

Our index variables are explicitly stated hypotheses, and thus can be refined using evidence. We were able to eliminate factors in our Antibiotics Index, an aggregate variable constructed using clinically relevant information. Using HCA, we could discern that routes of administration contributed nothing to the utility if the index for discerning patterns in health-related outcomes. Our new Antibiotics Index is thus informed by both clinical and empirical evidence.