IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 February 2024, accepted 1 May 2024, date of publication 8 May 2024, date of current version 16 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3398416

RESEARCH ARTICLE

Topological Risk-Landscape in Metric-Free Categorical Database

HSIEH FUSHING[1], **HONG-WEI KAO**[2], **AND ELIZABETH P. CHOU**[2]
[1]Department of Statistics, University of California at Davis, Davis, CA 95616, USA
[2]Department of Statistics, National Chengchi University, Taipei 116, Taiwan

Corresponding authors: Hsieh Fushing (fhsieh@ucdavis.edu) and Elizabeth P. Chou (eptchou@g.nccu.edu.tw)

This work was supported by the National Science and Technology Council (NSTC), Taiwan, under Grant NSTC 112-2118-M-004-001.

ABSTRACT The Entropy-based Categorical Exploratory Data Analysis (CEDA) paradigm is elaborately refined to algorithmically explore the intricate high-order directional associative relational patterns within the heterogeneous chronical disease dynamics captured by Behavioral Risk Factor Surveillance System (BRFSS) database. Operating on this imbalanced categorical dataset represented fully by its metric-free high-dimensional histogram, our algorithms conduct data-driven computations to investigate chronic disease mechanisms across four sub-populations along the age-axis, culminating in comprehensive systemic understandings. Upon this categorical data-world, CEDA first recognizes the category-oriented 1D histogram as the simplest form of a piece of explainable information. Then, utilizing Kolmogorov's randomness-proper-based reliability check, CEDA identifies and confirms collectives of 1D histograms as major feature-categories of varying orders within each sub-population. These confirmed major feature-categories' binary memberships are then arranged into a subject-vs-feature-category bipartite network heatmap, revealing serial horizontal and vertical blocks framed by clusters of similar subjects characterized by individual-risk-landscapes (IRL) against clusters of structurally dependent major feature-categories. Based on such block-series, sub-population-specific disease mechanisms emerge as collective high-order interacting effects, elucidating directional associative relationships from study subjects' topological neighborhoods to response-categories. Notably, the topological individual-risk-landscape offers profound insights into complex system dynamics and simultaneously exposes atypical subjects as explainable errors across all Machine Learning classifiers.

INDEX TERMS Behavioral risk factor surveillance system (BRFSS), bipartite network heatmap, categorical exploratory data analysis (CEDA), complex system, conditional entropy.

I. INTRODUCTION

The US agency Center for Disease Control and Prevention (CDC) conducts an annual phone survey with over 400K participants to construct a yearly Behavioral Risk Factor Surveillance System (BRFSS) database. Since 1984, the primary goal of this database has been to understand the dynamic and evolving linkages between multiple chronic diseases and their potential risk factors across the 50 states of US society over many years [1], [2], [3]. Each yearly BRFSS database, by design, encompasses all associative relations between multiple chronic diseases and many behavioral risk factors, referred to as feature-variables here, to sustain a complex system dynamics of chronic disease in American society for the year [4]. After 40 years, this yearly BRFSS complex system dynamics is, by and large, still unknown like a mystery, not to mention its evolution along the year-axis.

Could such a complex system dynamics of BRFSS be computationally extractable and explicitly displayable? To our limited knowledge, due to the seemingly boundless complexity and scope of such a system, comprehensive studies addressing this question have scarcely been conducted and rigorously reported in the literature. Nevertheless, positive and practical answers to this question are not only critical for the US but also for many countries that have developed

The associate editor coordinating the review of this manuscript and approving it for publication was Xianzhi Wang.

similar surveillance systems, as seen on the BRFSS website (https://www.cdc.gov/brfss/index.html).

In fact, this question holds a significant degree of universality across all sciences, and the potential impacts of its answers can extend far beyond the realm of the BRFSS. Why have there been hardly any comprehensive studies aimed at making complex systems dynamics readable and understandable? We are confident that the cause can be partly attributed to the fundamental barrier within data analysis. When handling a large system, data analysts encounter barriers stemming from two kinds of complexity embedded within data [5], [6]. The first kind of complexity is heterogeneity, which characterizes a large system by containing many heterogeneous local mechanisms. As described in [7], heterogeneity is expected to be observed through broken-symmetry patterns across different scales and localities within almost all large complex systems. This characteristic certainly is not limited to large physical or chemical systems. Indeed, the BRFSS has been shown to embrace heterogeneity characterized by (GenHL, Age) from the perspective of Heart Disease (HD) dynamics [8], where GenHL stands for the feature-variable called "general health."

The second kind of complexity pertains to the information content contained in large databases, which goes far beyond data visualization per se. Since the full information content in data (ICiD) includes all patterns of relational nature. Discovering such relational patterns, especially for high-order ones, requires a wide spectrum of genuine creativity and exploratory computing efforts. The information content channeled through high-order relational patterns is of particular scientific importance and practical interest because such directional associative relations "from a covariate feature-set toward another response feature-set" offer a unique window into essential mechanisms at a locality. However, such information has hardly ever been explored or even considered in data analysis due to its unknown functional form, making the ideas and practices of modeling unrealistic and incorrect. As a result, real-world associative relations of high orders, in general, are completely unknown even to domain scientists. This fact provides the brief background of this information complexity.

This information complexity is particularly evident in the BRFSS database because all its variables, from disease statuses to behavioral and demographic risk factors, are either categorical or categorized along certain axes. Thus, each yearly BRFSS database is entirely and completely represented by its high-dimensional histogram. Without losing any bit of information, this histogram forms a categorical data world of its own. Thus, the database's ICiD is perceived as consisting of all yet-to-be-discovered relational patterns of a wide spectrum of orders. In this paper, we delve deep into BRFSS's categorical data world, and set our primary goal as to graphically display BRFSS's computable and extractable information complexity. Conversely, this concrete and visible histogram would easily render man-made structures and assumptions as foreign objects. This is why almost all modeling-based results are likely unauthentic and obviously unscientific in this categorical data world.

The remainder of this Introduction section is dedicated to data descriptions, CEDA based computations and chief results for takeaways in three subsections, respectively. Subsection-A provides a detailed description of the Kaggle version of the BRFSS database, which serves two roles in this paper: as an illustrative example of our algorithmic CEDA computing and simultaneously as the database of scientific interest. Subsection-B explains how the computational CEDA paradigm generates the new concept of individual risk-landscape and its implications. We briefly outline the major results achieved in this paper in Subsection-C.

A. KAGGLE VERSION OF 2015 BRFSS DATABASE

Each yearly BRFSS database is complicated by the prevalence and various kinds of no-responses. Across more than one hundred questions in the survey, the no-response rates vary greatly, and the types of non-responses are diverse. However, it's important to note that "no-response" does not equate to "no information," as some subjects may choose not to answer certain questions due to sensitivity. Moreover, since many questions are highly related, such no-response data types persistently pose many difficulties and challenges when analyzing BRFSS databases.

Such difficulties and challenges pertaining to the specific 2015 BRFSS are avoided in its Kaggle version. It cleans out the majority of missing or non-response data points and significantly reduced the number of feature-variables, making it a popular database in Machine Learning literature. This Kaggle version of the 2015 BRFSS database consists of more than 250K subjects and 21 selected feature-variables. All 21 feature-variables, including several chronic diseases such as heart disease (HD), stroke (STK), and diabetes, among others, are categorical with symbolic codes. In this paper, as would be detailed below, some demographic and health variables, such as Age, Income, Mental Health..etc., of this Kaggle version are further regrouped to reduce their relative large numbers of categories. It is essential to note that these symbolic codes bear no sense of metric. For instance, code 1 may represent a diseased status, while code 0 indicates non-disease. Some feature-variables do bear ordinal senses among numerical codes. For example, five categories of both Age and GenHL are coded 1 to 5, where subjects with GenHL = 5 have the worst condition. However, the degree of "difference" between GenHL = 5 and GenHL = 4 is not necessarily equal to the difference between GenHL = 4 and GenHL = 3. As such, this Kaggle version categorical data is metric-free in nature.

In this paper, we continue to adopt the bivariate (GenHL, Age) as the defining axis of heterogeneity identified in [8]. Instead of focusing on one single chronic disease, here we designate the bivariate (Stroke (STK), Heart Disease (HD)) as the response (Re)-variable, with the remaining 17 one-dimensional feature-variables as covariate (Co)-variables.

The reason behind this choice of $\mathcal{Y}$ = (STK, HD) as the response variable is twofold. First, it better represents the real chronic disease dynamics of the complex system of interest than any single disease alone does. Secondly, it maintains a great degree of simplicity because all chronic diseases are structurally dependent. In this fashion, we utilize this Re-Co dynamics to represent the real chronic disease dynamics embraced by the 2015 BRFSS. Furthermore, all computational developments for $\mathcal{Y}$ = (STK, HD) can be easily expanded for any high-dimensional $\mathcal{Y}$.

To study this Re-Co dynamics, the entire 250K subjects are subdivided into 24 sub-populations with respect to 24 categories of (GenHL, Age). Notably, the category (GenHL, Age) = (5,2) is empty. Each subpopulation is postulated to embrace homogeneous disease mechanisms of (STK, HD). Therefore, the quest of analyzing the Kaggle version of 2015 BRFSS database is transformed into two steps: first, exploring each sub-population's ICiD thoroughly to enable a graphic display of its disease mechanisms; second, linking all 24 locally computed and fully represented disease mechanisms into a global disease dynamics. It is noted that, for the sake of length of this paper, we only demonstrate linkages among 4 sub-populations with GenHL = 5 and Age = 1, 3, 4, and 5. This synthesized disease dynamics of the poor health population along the Age-axis is of great scientific interest on its own right, while the full global disease dynamics is separately presented in a companion report.

The signature "imbalance phenomenon" of the BRFSS is retained in the Kaggle version as well. Here, the "imbalance phenomenon" refers to the highly uneven sample sizes among response-categories. Such a phenomenon is observed with significant unevenness of sample sizes across all 24 sub-populations. Specifically, the non-diseased category of (STK, HD) = (0,0) typically has a sample size many times that of the sample sizes of the three diseased categories (0,1), (1,0), (1,1) combined. This phenomenon is noteworthy because of its linkages to two technical fronts.

The first front pertains to recognizing why the marginal information of a feature-variable concerning a variable or a set of variables is imprecise and confusing. The second front is that this phenomenon has widely been attributed as the underlying cause of failures of many classifiers, such as various variants of Random Forest and Boosting, in Statistics and Machine Learning (ML) literatures. This phenomenon is even considered "intrinsic" [13]. Numerous remedial approaches have also been proposed without guaranteed successes [14], [15]. However, it is counterintuitive that an observed pattern of sample sizes of response-categories could become an intrinsic barrier hindering all classifiers. Additionally, it is equally counterintuitive regarding the merits of developing sampling schemes on observed data to improve the performance of classifiers per se without concerning the potential consequences of distorting ICiD. These two technical fronts are explicitly addressed in this paper, and their resolutions are outlined in the next two subsections, with further details provided in Section V. They indeed serve as two signatures of this paper.

At the end of this subsection, we describe our coding schemes for regrouping the following 5 variables of the Kaggle version of dataset.

1. [Age:] Age-1: 18 to 29; Age-2: 30 to 44; Age-3: 45 to 59; Age-4: 60-74; Age-5: 75 and above.
2. [BMI:] BMI-1: Body Mass Index less than 18.5; BMI-2: 18.5 to 24; BMI-3: 24 and above.
3. [Education:] EDU-1: highest grades less than grade 8; EDU-2: grade 9 to grade 12; EDU-3: 1-year college or more.
4. [Income:] Income-1: annual household income less than 25k; Income-2: 25K to 75K; Income-3: above 75K.
5. [Mental Health:] Mentlth-1: zero days during the past 30 days being not good in mental health; Mentlth-2: 1 to 9; Mentlth-3: 10 to 29; Mentlth-4: whole month.
6. [Physical Health:] Physhlth-1: zero days during the past 30 days being not good in physical health; Physhlth-2: 1 to 9; Physhlth-3: 10 to 29; Physhlth-4: all 30 days.

This regrouping scheme is necessary for computations conducted within sub-populations defined by (GenHL, Age). By so doing, the 24 sub-populations have sizes around several thousands. For the computational simplicity, all subjects with missing or no-responses within the Kaggle version are further excluded in this paper.

B. CEDA COMPUTING PARADIGM AND INDIVIDUAL RISK-LANDSCAPE

As mentioned earlier, the 21-dimensional histogram constructed from the Kaggle version of the BRFSS database as one whole is metric-free. Within this categorical data world, arithmetic operations and functional forms are meaningless. In this paper, we adapt the data-driven bottom-up computational paradigm, called Categorical Exploratory Data Analysis (CEDA), to effectively explore high-order associative relations that constitute and reveal the database's information complexity. These explorations are necessary due to our recognition of the 1D histogram as the simplest form of "a piece of explainable information" and Kolmogorov's randomness-proper on any contingency tables. Throughout this paper, a contingency table is typically constructed by arranging all categories of a covariate feature-set along the row-axis and the four categories of (STK, HD) along the column-axis. Any such a contingency table is simply a projection of the 21-dimensional histogram of the whole data set.

Information extraction from a contingency table is carried out by comparing one row-vector, representing a 1D histogram of the conditional variable of $\mathcal{Y}$ given a covariate feature-category, with the column-sum vector, representing a 1D histogram or marginal distribution of $\mathcal{Y}$. In a step-by-step manner, each individual comparison yields one piece of information regarding one aspect of the interacting effects of the covariate feature-set. Then, the collective comparisons reveal glimpses of potential interacting effects of the covariate feature-set on the response-variable $\mathcal{Y} =$ (STK, HD). No functional forms of interacting effects are needed in such comparisons.

In general, scientists are not capable of fully prescribing interacting effects because diverse asymmetric relational patterns are possible and potential among all involved categories. Therefore, it becomes not only necessary but also critical to be able to demonstrate and confirm all individual category-specific interacting effects. From this perspective of interacting effects, three standpoints of our elaborately refined CEDA here make evident differences from the original CEDA algorithms developed in a series of previous works [8], [10], [11], [12].

The first standpoint is that a feature-set's category-specific effect takes the central role, not its marginal effect, which is calculated via a weighted sum scheme. That is, a category-specific effect is demonstrated by comparing its corresponding conditional entropy of $\mathcal{Y}$ conditioning on the corresponding covariate category with the entropy of $\mathcal{Y}$ without involving with covariate information of any sort. This is a very unique standpoint taken in this paper.

The second standpoint is that this comparison must be conducted under equal "randomness" footings. This is where Kolmogorov's randomness-proper comes in to play its essential role through a contingency table platform [9]. Here, two versions of Kolmogorov's randomness-proper are respectively seen through the following two constructed ensembles: 1) an ensemble of mimicries of the observed contingency table, which share the same randomness embraced by the observed table; 2) another ensemble of simulated contingency tables only retain randomness embraced by the observed row-sum vector. Both ensembles are commonly subject to the column-sum vector, which represents the fixed sample sizes of the four categories of $\mathcal{Y}$. The conceptual differences of these two ensembles rest on the fact that the first ensemble genuinely reflects the data's intrinsic randomness, while the second ensemble embraces the hypothetical randomness as if the covariate feature of the row-axis is independent of $\mathcal{Y}$. The first ensemble gives rise to an alternative entropy distribution, while the second ensemble gives rise to a null entropy distribution. Thus, the aforementioned comparison is carried out by comparing alternative-vs-null entropy distributions resulting in the minimum sum of Type-I and Type-II errors or the two distributions' overlapping area.

This comparison plays a key role at the heart of this refined CEDA paradigm. Such a Kolmogorov's randomness-proper based comparison would be applied twice to select and confirm a major feature-category, instead of major feature-variable. Its first application is to a major feature-category candidate of given order, which is equal to the size of covariate feature-set. This application involves the entire samples belonging to the sub-population. Its second application is necessarily performed when the order of the potential major feature-category candidate is larger than one. Since we need to make sure this candidate is not redundant with respect to an already confirmed major feature-category of lower order. That is, a major feature-category of high order must provide extra-information (Extra-Info) on top of what a confirmed major feature-category of lower order can provide. Hence, this application involves only samples constituting this confirmed major feature-category of lower order. Subsequently, we build a graphic display based on a collection of selected high-order major feature-categories, which becomes the chief part of the ICiD within each sub-population.

The aforementioned feature-category based computational operations are newly developed here, offering contrasting differences with the original version based on marginal mutual information calculations in selecting major feature-variables. Such differences are especially evident and crucial when the database is subject to a high degree of "imbalance". Realistically speaking, the majority of real-world databases retain varying degrees of "imbalance".

The third standpoint relies on the capability of representing computed and confirmed major feature-categories through a graphic display in this new version of CEDA. As each selected major feature-category of any order has its own memberships due to its locality, each subject will be prescribed by a binary vector indicating its presence or absence with respect to all selected major feature-categories. This subject-specific binary vector sheds light on the positive and negative disease risks facing this subject. From this aspect, we term this binary vector of memberships of all selected major feature-categories the subject's individual risk-landscape (IRL).

Furthermore, based on the collective individual risk-landscapes, two significant sub-population specific characteristics can be derived. First, a topology is defined on the study-subject space with a natural choice of dissimilarity or similarity measure. The neighborhood system offered by this topological space explicitly reveal information about which subjects are close to which subjects, but far away from other subjects. A graphic display of the entire topological space pertaining to high-order major feature-categories in general is very informative regarding sub-population specific chronic disease dynamics and beyond, as would be clearly seen in Section V. For instance, this topological characteristic among subjects can serve as a critical basis for matching in causality study and optimal selection for the highest or lowest risk subject-groups.

The second significant sub-population specific characteristic is that each cluster of subjects' individual risk-landscapes will characterized by a horizontal blocks framed by a series of clusters of major feature-categories. When coupled with annotated response-categories, such a horizontal series of blocks provides "readable" and "visible" information defining this cluster of subjects. One piece of vital information is the explicit map of so-called "atypical subjects". Here, an "atypical subject" is referred to a study subject encoded with an annotated response-category is found belonging to an individual risk-landscape neighborhood sharing with several other study subjects, who are encoded with very different annotated response-categories from his/her. A large collective of "atypical subjects" allows us to fundamentally resolve the aforementioned "imbalance phenomenon" issue [13], [14], [15].

Putting together these two sub-population specific characteristics, we can further point to the fact that, as a byproduct, such a topological space of individual risk-landscapes is an informative platform for building variants of explainable inferential decision-makings, including prediction and classification.

C. MAJOR RESULTS OF THIS PAPER

With $\mathcal{Y} = (\text{STK}, \text{HD})$ as the response variable, we explore the Re-Co dynamics representing the chronic disease dynamics underlying the 2015 BRFSS database. Upon the Kaggle data set, our CEDA computations first illustrate why a feature-category specific 1D histogram is the simplest form of "a piece of information" within this categorical data world. We then explicitly demonstrate the unsuitability of the "marginal form of information". This simple fact establishes a wide spectrum of profound impacts that would be seen not only in Data Analysis as a scientific discipline [16], but also in all sciences. Since so far data analyzing methodologies from statistics and ML primarily rely on "operations of variables", such as all modeling based topics and methodologies, like variable selection among many others. That is, from the ICiD perspective, it is legitimate to seriously question the validity of methodologies developed in these two fields. Since a methodology employed in any data analysis for sciences needs to pass "the test of experience". This is the most crucial criterion underlying any scientific disciplines as advocated by John Tukey [16] in his 1962 paper with title: "The future of data analysis".

Secondly, the recognition of Kolmogorov's randomness-proper on any contingency table plays an instrumental role in our refined version of CEDA. In fact, this concept is essential and fundamental in its own right in Data Analysis beyond the categorical data world. Given that a histogram can very well approximate any quantitative variable's entire empirical distribution [17], so this concept is indeed applicable across all data types. Its importance is indeed self-evident for its capability of facilitating the pair of alternative-vs-null entropy distributions. It is extremely critical that Type-I and Type-II errors can be evaluated without assuming any man-made modeling structures and distributional assumptions in any data analysis.

Thirdly, we build algorithms to conduct CEDA computing step-by-step: from identifying and confirming major 1-feature-categories, major 2-feature-categories to major 3-feature-categories within each sub-population. Among major 2-feature-categories, we show diverse forms of asymmetry of order-2 interacting effects across a series of feature-pairs. Such diversity of asymmetric interacting effects is meant to reiterate a key point in data analysis: Invaluable knowledge of disease mechanisms is available to be discovered only if data analysts and domain scientists are willing to explore.

Fourthly, a sub-population's computed disease mechanisms are represented through the collection of all computed and confirmed order-3 interacting effects, so-called major 3-feature-categories with either positive or negative risks. The presence-absence memberships of this collection of major 3-feature-categories are compiled into a binary bipartite network matrix. Thus, this graphic display collectively reveals all involving subjects' individual risk-landscapes as a serial positive or negative disease risks exposures. After rearrangements via hierarchical clustering on the row-axis of subjects and column-axis of major 3-feature-categories, respectively, this block-pattern sustained heatmap reveals explicit relational patterns though the authentic topology of individual risk-landscape defined on the subject space and complex structured dependency on the collection of major 3-feature-categories. This heatmap allows us to figure out characteristics of the sub-population specific disease mechanisms via horizontal series of blocks discovered on the scale of category of $\mathcal{Y}$ and on a finer scale of clusters within category of $\mathcal{Y}$. That is, such a heatmap indeed sustains functionally critical and philosophically vital parts of information content in data (ICiD) pertaining to the sub-population under study.

Fifthly, along the heterogeneity-axis of $\text{GenHL} = 5$ and $\text{Age} = 1, 3, 4, \text{and } 5$, we then patch and link all relational patterns derived from the four sub-populations into a composite complex system. Such global functionality embraced by the linked four sub-populations provides one important aspect of understanding the whole complex system. The grand global view of chronic disease dynamics underlying 2015 BRFSS database would be separately presented in a companion study by embracing all 24 sub-populations.

Sixthly, the four sub-population specific heatmaps collectively and explicitly reveal the existence and prevalence of "atypical subjects" across diseased and non-diseased categories. These atypical subjects would definitely cause "errors" to whatever classifiers. From this standpoint, the "imbalance phenomenon" is indeed not the intrinsic cause of ill performances of all classifiers. On the other hand, if there are no "atypical subjects" present in a sub-population, this heatmap graphic display would allow perfect classifications even under the presence of a very severe "imbalance phenomenon". That is, the "imbalance phenomenon" is simply a fundamental misconception.

Finally, we conclude that the explicit demonstration of "atypical subjects" reflects the fact that building ICiD is indeed the ultimate goal of data analysis. Subsequently, any inferential operations must be performed strictly in accord with ICiD. This is the merit of pointing out this long-standing big mistake. On the other hand, we also emphasize here that a comprehensive study of complex systems must achieve ICiD.

From the technical perspective, the most far-reaching implication of our methodological developments in this paper is that this CEDA paradigm can, in fact, be at the heart of all structured data analysis of any data types. Since each quantitative data set can be categorized and simultaneously retain its chief part of ICiD. Then, from this ICiD perspective, the brand-new concept of individual risk-landscape truly provides authentic insights through its topological subject space. Through its block-sustained heatmap display, the comprehensive and vital pattern information pertaining to nature of complex chronic disease dynamics is explicitly summarized. As such it becomes rather unthinkable for any inferential decision-making without embracing insights of the subject's individual risk-landscape and its topological neighborhood structures.

We organize the rest of this paper as follows. In section II, we lay the foundations for the CEDA paradigm, including arguments for the simplest form of a piece of information and the directional associative relation from any feature-set toward the response variable $\mathcal{Y} = (STK, HD)$, as well as reliability checks based on Kolmogorov's randomness-proper. In section III, we develop the algorithm for CEDA computing major feature-categories of various orders and explain and visualize their rather convoluted interacting effects. Section IV is devoted to presenting diverse kinds of asymmetry found in order-2 interacting effects and their evolutions along the age-axis. In Section V, we show results centered around individual risk-landscapes and their heatmaps, along with consequent summarizing statistics. We also construct the global dynamics of $\mathcal{Y} = (STK, HD)$ under GenHL $= 5$ by combining results from the four sub-populations along the age-axis. In the conclusion section, we reflect on the potential impacts of our CEDA-enabled topological results and the induced issues within and beyond Data Analysis.

II. WHAT INFORMATION LOOKS LIKE?

Within the metric-free 21-dimensional categorical data world, the journey of data analysis naturally commences with addressing the simplest question: What does a piece of information look like? This inquiry is especially significant because such a piece of information remains invariant to all permutations along all dimensions of the histogram. Only after obtaining an answer to this question does it become feasible to address the subsequent critical and fundamental question: What is ICiD made of? In the following two subsections, we exemplify CEDA computations for selecting and confirming major feature-categories in a bottom-up data-driven fashion.

A. WHAT IS THE SIMPLEST FORM OF A PIECE OF INFORMATION?

Take anyone of the 21 categorical feature-variables. Does one of its 1D categorical data points mean anything? The answer is apparently negative. Since it is simply a label-code which can be arbitrarily encoded. A label-code only marks a "location" on the metric-free-axis of this feature-variable. Further, any aggregation of one single label-code is also meaningless in relation to this feature-variable's multiple locations. Furthermore, any missing aggregation of anyone label-code along this feature-variable's location-axis will distort the information formation. As such a label-code as a "location" apparently acts like an element in formatting a piece of information. And, the simplest form of a piece of information is delivered by an 1D histogram of an 1D categorical feature-variable. This is the answer to the first question.

When two 1D categorical feature-variables are observed or derived from the same system, this bivariate feature-variable owns an 2D contingency table or histogram. All aspects of relational information content in this 2D data set is fully described by "location-to-location" correspondences as being visibly laid out via the 2D contingency table. Likewise for all relational relations involving with more than 2 categorical feature-variable. As such the chief mechanism of formatting relational information within a categorical data world is still operated via the fundamental "location-to-location" correspondence within their histogram or so-called hyper-contingency table.

Nonetheless, when comparing two histograms, again it is conducted on the "location-to-location" basis. This comparison will not be altered by any permutations respectively applied on their common metric-free-axis. Thus, at least ideally, "location-to-location" basis still give rise to the full and meaningful information regarding comparing two histograms of any dimensionality. By the same argument, one effective way of comparing multiple histograms is simply done by adding an extra categorical ID-variable.

In summary, we term a "label-code" meaning a "location" of any 1D data point as "an element of information". It is understood that such an element of information bears with a feature-variable specific "location" message. With this concept of "element of information", we clearly see that a 1D histogram is indeed the most fundamental form of "a piece of information" in any categorical data worlds. Given that the "location-to-location" correspondence is the most fundamental mechanism of relational information formation, our next task is how to effectively extract all essential relational information content from data's very high dimensional histogram.

B. WHAT ICiD IS MADE OF?

Next we turn to the fundamental question: What ICiD is primarily made of? There are only two potential possible answers in sight: either marginal information of feature-variables, or 1D histogram of feature-category, in this categorical data world. As aforementioned, 1D histogram is the simplest form of a piece of information, while any feature-variable or a feature set's marginal information involves multiple 1D histograms arranged in a contingency table format. The first critical difference between these two possible answers rests on their different scales. An 1D histogram is of category-specific locality scale, while feature-variable's marginal information is of the global scale. The second critical differences is regarding their meanings. An 1D histogram is clear and explainable. In contrast, a feature-variable's marginal meaning can be confusing because of conflicting meanings derived from different localities. That is, individual category-specific meanings can be lost, even distorted, within its marginal version. This loss and distortion surely will miss out many important and essential feature-categories. We explicitly illustrate here why marginal information is not the fundamental format of information content representing categorical data. In fact, the feature-category is the right format for revealing aforementioned "broken symmetry" as another key characteristic of complex system [7]. Here this characteristic is seen through the drastically distinct pattern information emitted from a feature-variable or a feature-set's distinct categories.

In this subsection, our illustrating example is built upon the sub-population of GenHL $= 5$ and Age $= 1$, which consists of 1776 subjects. As aforementioned, we focus on the Re-Co dynamics with the categorical response variable is $\mathcal{Y} = (STK, HD)$ and the rest of 17 categorical features as covariate variables, including Diabetes, High Blood Pressure (HighBP), High Cholesterol (HighChol),.., etc. The response variable $\mathcal{Y}$ has four categories of bivariate disease-status: $\{(0, 0), (1, 0), (0, 1), (1, 1)\}$. Except BMI, Education (EDU) and Income are encoded with 3 categories, Mental health (MentHlth) and Physical Health (PhyHlth) encoded with four categories. The remaining covariate 12 feature-variables are all binary.

We begin by illustrating the associative relations between $\mathcal{Y} = $ (STK, HD) and HighBP through the 2×4 contingency Table 1. This table is denoted as $HCT[(STK, HD); HighBP]$. The two rows of this table are two 1D histograms with 4 bins pertaining to 1-feature-categories: $HighBP_0$ and $HighBP_1$, standing for two categories of subjects: not having and having high blood pressure, respectively. They are 1D histograms of $\mathcal{Y}$ conditioning on $HighBP = 0$ and $HighBP = 1$, respectively. In contrast, the column-sum vector pertaining to the four response categories is "constant" with respect to all covariate variables, which is the 1D histogram of $\mathcal{Y}$.

By listing the three histograms within Table 1, we intend to compare the individual 1D histograms of $\mathcal{Y}$ conditioning on $HighBP_0$ or $HighBP_1$ with the marginal 1D histogram of $\mathcal{Y}$. One meaningful comparison is performed by comparing row-wise conditional (Shannon) entropy with that of the column-sum vector [18]. The entropy of $\mathcal{Y} = (STK, HD)$ is calculated: $CE[\mathcal{Y}] = CE[(STK, HD)] = 0.7565$. Strikingly, given $HighBP_0$, the conditional entropy of $\mathcal{Y}$ is reduced to $CE[\mathcal{Y}|HighBP_0] = 0.5292$, while given $HighBP_1$, the conditional entropy of $\mathcal{Y}$ is increased to $CE[\mathcal{Y}|HighBP_1] = 0.9133$.

The entropy reduction of $CE[\mathcal{Y}|HighBP_0]$ is attributed to the observation of having relatively more subjects in the non-diseased $\mathcal{Y} = (0, 0)$ category and less subjects in diseased categories: $\mathcal{Y} \in \{(0, 1), (1, 0), (1, 1)$, in comparison with column-sum vector of proportion of $\mathcal{Y}$. While the entropy increase of $CE[\mathcal{Y}|HighBP_1]$ is attributed to the observation that non-diseased $\mathcal{Y} = (0, 0)$ category has a reduced proportion, but still keeps the majority of $HighBP_1$ subjects, while even though the proportions of diseased categories: $\mathcal{Y} \in \{(0, 1), (1, 0), (1, 1)$, have slight increases against the column-sum vector of proportion of $\mathcal{Y}$. In this fashion, the row vector of $HighBP_1$ becomes more evenly distributed among non-diseased and diseased categories than the column-sum vector of $\mathcal{Y}$. This is the phenomenon of "imbalance", which is underlying the somehow counterintuitive scenario of extra information of $HighBP_1$ indeed promoting more, not less, uncertainty of $\mathcal{Y}$.

As would be confirmed in the next subsection, both conditional entropies $CE[\mathcal{Y}|HighBP_0]$ and $CE[\mathcal{Y}|HighBP_1]$ pass the reliability checks of being significantly different from $CE[\mathcal{Y}]$. That is, both covariate categories are highly associated with $\mathcal{Y}$. However, if the predictive perspective is taken as the solo focus for associative relation, then these two 1-feature-categories give rise to two rather conflicting kinds of messages. The information of $HighBP_0$ is good for predictive relation with $\mathcal{Y}$, while the information $HighBP_1$ is not. That is, using the predictive capacity as a way of quantifying the strength of associative relationship between two variables is fundamentally improper, especially under the "imbalance phenomenon".

Further, from disease dynamics perspective, it is transparent that $HighBP_0$ strongly points to less risk of the bivariate disease, while $HighBP_1$ points to higher risk. In sharp contrast, the marginal conditional entropy: $CE[\mathcal{Y}|HighBP] = 0.73075$, which is calculated as the weighted sum of $CE[\mathcal{Y}|HighBP_0]$ and $CE[\mathcal{Y}|HighBP_1]$ with weights $\frac{843}{1776}$ and $\frac{933}{1776}$, respectively, does not convey either one of the two directional associative relations. That is, the effects on $\mathcal{Y}$ incurred by feature-variable: HighBP, can not be properly delivered by its marginal relationship with the response-variable $\mathcal{Y}$. In summary, the description of associative relation of HighBP-to-(STK, HD) is necessary of category-locality nature.

Based on this simple example, we are confident that 1D histogram is the answer to the question: What ICiD is made of? That is, the quest of data analysis is to extract all relevant pieces of relational information in a form of 1D histogram. We likewise conclude that all pattern information in ICiD is of locality nature. More evidences are seen through the interacting effects conveyed by categories of 1D covariate feature-pairs in Section IV. The implications of this somehow simplistic statement are far reaching. A major impact is that, under the shadow of "data's imbalance phenomenon", all Statistics and Machine Learning topics become by and large invalid because they solely rely on marginal information of global nature. Consequently, all modeling approaches in these two fields likely fail.

TABLE 1. Contingency table $HCT[(STK, HD); HighBP]$ through the perspective of heterogeneity (GenHL = 5, Age = 1).

HighBP/(STK, HD)	(0,0)	(1,0)	(0,1)	(1,1)	Row-sum
0	729	43	56	15	843
1	643	63	175	52	933
Col-sum	1372	106	231	67	1776

C. KOMOGOROV'S RANDOMNESS-PROPER AND RELIABILITY CHECK

Next, we discuss how to make sure that both pieces of information: the two 1D histograms of $\mathcal{Y}$ conditioning on $HighBP_0$ or $HighBP_1$ are significant by passing their reliability checks. For a piece of information, its reliability check is performed by precisely two observed versions: alternative-to-null, of Kolmogorov's randomness-proper pertaining to Table 1 [9]. The alternative version is column-wise randomness given only the corresponding column-sum. That is, each column vector is seen as a realization of Multinomial randomness given its column-sum and its observed column-specific vector of proportion. In contrast, the null version is the randomness of row-sum vector being equally imposed onto all columns. That is, each column vector is seen as a realization of Multinomial randomness given its column-sum and the common proportion vector of row-sums. These versions indeed cover all randomness observed within a contingency table, like Table 1, while the null version indeed bears no associative information regrading HighBP to $\mathcal{Y}$.

Here are the technicalities of the alternative and null versions of randomness underlying the 2×4 contingency table in Table 1. The 4-dim column-sum vector $(1372, 106, 231, 67)$ is fixed with a total 1776. They present four column-wise constraints in formatting the contingency Table 1. That is, both kinds of randomness are conditioning on this column-sum vector. Subsequently, the 2-dim row-sum vector $(r_0, r_1) = (843, 933)$ is an observed vector being specifically subject to randomness of covariate feature-variable HighBP as one whole under the constraint of total sum 1776. In other words, each row-sum's randomness is linked to its four components' randomness under the four column-wise constraints. With randomness-proper tied to all observed entries of Table 1 and (r_0, r_1), we can depict the all aspects of randomness-proper associated with Table 1 as follows.

Alternative Randomness: This randomness for the four observed columns of Table 1 is given as: $MN(n_y, P^a_y)$ with $y \in \{(0,0), (1,0), (0,1), (1,1)\}$ and $P^a_y = (n_y[0]/n_y, n_y[1]/n_y)$ with $(n_y[0], n_y[1])'$ being y-th column vector. Such multinomial randomness protocols constitute the randomness-proper underlying the alternative setting against the null setting described below.

Null Randomness: The column-wise null randomness is given by the multinomial distribution $MN(n_y, P^o)$ with $P^o = (r_0/1776, r_1/1776)$.

With the above alternative and null randomness specifications for Table 1, a generic form of simulated contingency table with respect to having alternative-effect and null-effect of HighBP, denoted by $HCT[\mathcal{Y}; HighBP]$ and $HCT[\mathcal{Y}; null-HighBP]$, respectively, is given in the Table 2.

TABLE 2. Generic form of simulated contingency table of $HCT[\mathcal{Y}; HighBP]$ and $HCT[\mathcal{Y}; null - HighBP]$ with respect to alternative randomness and null randomness based on the perspective of heterogeneity (GenHL = 5, Age = 1).

HighBP/(STK, HD)	(0,0)	(1,0)	(0,1)	(1,1)	Row-sum
0	$n'_{00}[0]$	$n'_{01}[0]$	$n'_{10}[0]$	$n'_{11}[0]$	r'_0
1	$n'_{00}[1]$	$n'_{01}[1]$	$n'_{10}[1]$	$n'_{11}[1]$	r'_1
Column-sum	1372	106	231	67	1776

We simulate two ensembles of 1000 contingency tables of $HCT[\mathcal{Y}; HighBP]$ and $HCT[\mathcal{Y}; null-HighBP]$, respectively. Upon the ensemble of $HCT[\mathcal{Y}; HighBP]$, we build the two alternative entropy distributions (orange colored) pertaining to $HighBP_0$ and $HighBP_1$ marked with observed red-colored vertical lines at $CE[\mathcal{Y}|HighBP_0](= 0.5292)$ and $CE[\mathcal{Y}|HighBP_1](= 0.9133)$, respectively, as seen in the two corresponding panels of Fig. 1. Likewise, we build the two null entropy distributions (blue colored) pertaining to $HighBP_0$ and $HighBP_1$.

From the two panels of Fig. 1, we clearly see that both pieces of information: the two 1D histograms of conditional entropy of $\mathcal{Y}$ given $HighBP_0$ and $HighBP_1$, respectively, have zero-sums of type-I and type-II errors. That is, both are confirmed being rather significant. In contrast, the marginal alternative entropy distribution is expected to be centered around $CE[\mathcal{Y}|HighBP] = 0.73075$ that would be heavily overlapping with the marginal null entropy distribution centered around $CE[\mathcal{Y}] = 0.7565562$. This example illustrated why the marginal evaluations of potential effect of any feature-variables have dangers of giving rise to misinformation.

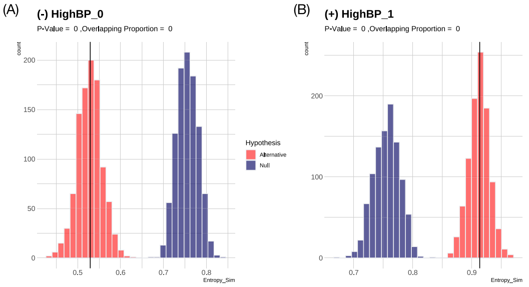

FIGURE 1. Two null (blue)-vs-alternative(orange) distributions of $HighBP_0$ and $HighBP_1$. See corresponding plots at https://github.com/CEDA2024/Metric-Free-Categorical-Database.

It is somehow critical for performing this reliability check based on the minimum sum of Type-I and Type-II errors, or the overlapping area of the alternative and null entropy distributions. Since the commonly used criterion based on P-value is simply too optimistic in a sense of too many false positive feature-categories being selected. This fact truly reflects the importance of Kolmogorov's randomness-proper when analyzing real world data.

III. DATA-DRIVEN BOTTOM-UP CEDA PARADIGM

After the description of reliability check in the previous subsection, it is essential to put the technical meaning of confirming both pieces of information of $HighBP_0$ and $HighBP_1$ pertaining to the dynamics of $\mathcal{Y}$ in perspective. Since this dynamics is limited to the sub-population defined by (GenHL, Age) = (5, 1). The structural dependency and de-associating operation discussed in details in [8] assure the fact that these two 1-feature-categories $HighBP_0$ and $HighBP_1$ indeed provide extra-information beyond what the bivariate-category (GenHL, Age) = (5, 1) can provide into the dynamics of $\mathcal{Y}$. This seemingly simple de-associating operation is critical for identifying true factors underlying $\mathcal{Y}$ from two fronts. First, its chief merit is to identify potential major feature-categories of various orders. Secondly, it provides a way of checking whether one feature-category indeed provides extra-info, instead of piggy-backing upon an already confirm major feature-category. Via these two fronts, we construct our CEDA bottom-up data-driven computational developments in this section.

A. MFCI ALGORITHM

As the first phase of developing CEDA paradigm, we build an algorithm for Major Feature-Category Identification (MFCI): from order-1 to higher orders. Before describing the MFCI algorithm, we first clarify the ‘‘identification’’ operation of CEDA paradigm that facilitate two types of tasks: MFC and Extra-info, in the MFCI algorithm given below.

1) FOR MFC

For identifying potential major k-feature-categories (MFC of order k) pertaining to a covariate feature-set A with cardinality k, we first build a hyper-contingency table $HCT[\mathcal{Y}; A]$ based on **the entire sub-population of data points**. Secondly, we perform reliability check upon each row of $HCT[\mathcal{Y}; A]$, respectively. Thirdly, a decision of identification is made with respect to a chosen threshold of minimum sum of Type-I and Type-II errors.

2) FOR EXTRA-INFO

For identifying whether 1-feature-category, says B_b, can provide extra-info upon an identified major k-feature-category, says A_a, we first build a hyper-contingency table $HCT[\mathcal{Y}; B|A_a]$ based on **the collection of data points belonging to A_a.** Secondly, we perform reliability check upon the $b-$th row of $HCT[\mathcal{Y}; B]$. Thirdly, again a decision of identification is made with respect to a chosen threshold of minimum sum of Type-I and Type-II errors.

Within the sub-population (GenHL, Age) = (5, 1), apparently, both tasks of identification for MFC and Extra-Info are so-called de-associating operations working on two different data-settings: one is the sub-population specified by (GenHL, Age) = (5, 1) and the other is specified by the targeted major feature-category, such as $HighBP_0$, see details in [8]. We now describe the MFCI algorithm below.

a: MAJOR FEATURE-CATEGORY IDENTIFICATION (MFCI) ALGORITHM:

MFCI-1. Identify and confirm effect of any major 1-feature-categories via minimum sum of type-I and type-II errors or overlapping area of alternative and null entropy distributions.

MFCI-2. Identify and confirming effects of any major 2-feature-categories in two steps. Step[2]-1 is to find out which 1-feature-categories can provide Extra-Info upon each confirmed major 1-feature-category; Step[2]-2 is to confirm the order-2 effect of any identified 1-feature-category in Step[2]-1 together with its corresponding confirmed major 1-feature-category via the criterion of minimum sum of type-I and type-II errors to identify a major 2-feature-category.

MFCI-3. Identify and confirming effects of any major 3-feature-categories in two steps. Step[3]-1 is to find out which 1-feature-categories can provide extra information upon each confirmed major 2-feature-category. Step[3]-2 is to confirm the order-3 effect of any identified 1-feature-category in Step[3]-1 together with its corresponding confirmed major 2-feature-category via the criterion of minimum sum of type-I and type-II errors to identify a major 3-feature-category.

Comp-4. Identify and confirming effects of any major higher-order-feature-categories in two steps exactly like the above MFCI-2 and MFCI-3. (Naturally, the finite sample size would force our computations to a stop at Step[k+1]-1 when no more 1-feature-categories can be found to provide Extra-Info upon all identified major k-feature-categories.)

The above description of MFCI algorithm is designed to cope with finite computing resource, in particular when facing a large number of 1D covariate feature-variables contained in data set. It might miss some major feature-categories of high orders with its 1D component-member-features being not involving in selected major 1-feature-categories. Such kinds feature-categories of high orders are relatively rare. On the other hand, if computing resource is large enough, then the orders of Step[2]-1 and Step[2]-2 can be switched. Then the concern of missing some order-2 interacting effect is resolved. Likewise for switching orders of Step[3]-1 and Step[3]-2. Nonetheless, if the MFCI-3 step is performed with its two steps in reversed order, then there would be $\binom{17}{3} = 680$ triplets of features and more than 5440 reliability checks. From the cost-benefit aspect, we do not switch the order of these two steps. Patterns reported in the next two sections seem to support this decision.

Here we report results from the first two steps of MFCI algorithm applied on the sub-population (GenHL, Age) = (5, 1). By applying MFCI-1 step of MFCI algorithm, we found 13 major 1-feature-categories with respect to a chosen 0.1 threshold value of minimum sum of Type-

I and Type-II error, *Error* $-$ *I&II* for short, see Fig. 2. From the column-perspective of this figure, it is worth noting that these 13 major 1-feature-categories consist one positive and one negative disease risk groups marked with "+" and "−" signs. Members of each group overlap with significant number of subjects. Such overlapping patterns clearly indicate the strong structural dependency among these 1D features. From the row-perspective, it is obvious that many subjects share the same or very similar memberships across 13 major 1-feature-categories, while they belong to very distinct response-categories. These patterns together promote the necessity of carrying out MFCI-2 step for more informative associative patterns.

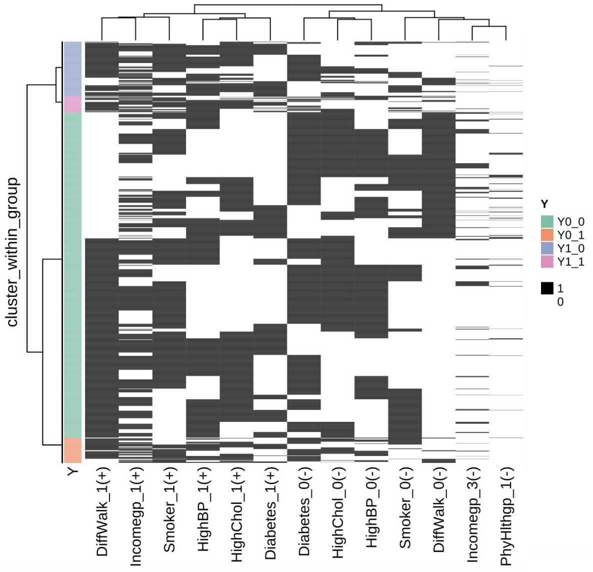

FIGURE 2. **Heatmap of confirmed major 1-feature-categories for dynamics of $\mathcal{Y} = (STK, HD)$ within the subpopulation GenHL=5 and Age=1.**

Upon applying Step[2]-1 of MFCI-2 step, we found 57 candidate 1-feature-categories that can provide extra-info upon the 13 major 1-feature-categories resulted from MFCI-1 step. Further, applying Step[2]-2 of MFCI-2 step, we identified and confirmed 31 major 2-feature-categories, see Fig. 3. Overall, the heatmap in Fig. 3 reveals much clear associative patterns from major 2-feature-categories of positive and negative disease risks than that in Fig. 2. Especially, it is striking to see that subject-members of the non-diseased category $\mathcal{Y} = (0, 0)$ are respectively separated into two obvious groups. One group consists of member-subjects having prevalent memberships among 14 major 2-feature-categories of positive disease-risk, and another group consists of member-subjects having prevalent memberships among 17 major 2-feature-categories of negative disease-risk.

This obvious improvement from major 1-feature-categories to major 2-feature-categories naturally motivates us to further carry out the MFCI-3 step. Before we report computational results from the MFCI-3 steps in Section V, we report four types of interacting effects of order 2 resulted from MFCI-2 step and illustrate their age-related evolving patterns across four sub-populations of (GenHL, Age) $=$ (5, k) with $k = 1, 3, 4, 5$ in the next Section IV. Discoveries of interacting effects and understanding of their evolutions across age-axis are especially important from the perspectives of societal chronical disease and individual risk dynamics. The scientific discoveries and understanding become self-evident when we face the diverse formats of asymmetry among all involving 2-feature-categories. In contrast, we would see heatmap-based displays of individual risk-landscape topologies and their age-related evolution along the same age-axis in Section V.

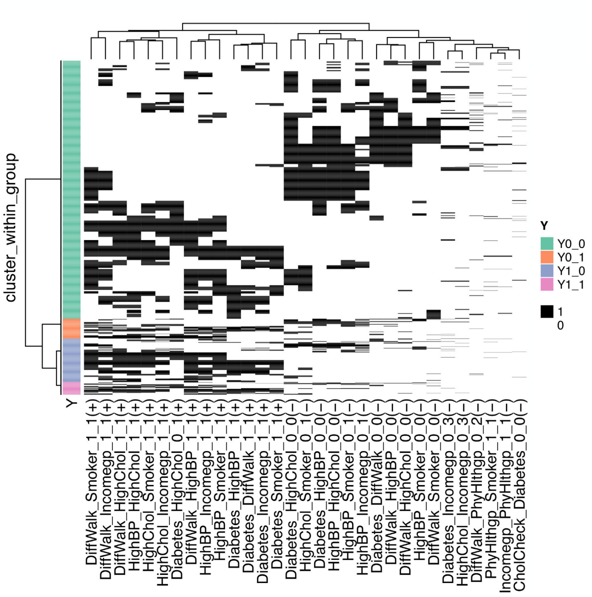

FIGURE 3. **Heatmap of confirmed major 2-feature-categories for dynamics of $\mathcal{Y} = (STK, HD)$ within the subpopulation GenHL=5 and Age=1.**

IV. DIVERSE TYPES OF ORDER-2 INTERACTING EFFECTS AND THEIR EVOLUTIONS

As an 1-feature-category is found to provide Extra-Info to any major 1-feature-category, this fact indeed signals the potentials of discovering essential and important interacting effects of order-2. In particular, when such discoveries are arranged with respective to age-axis, we figure out their evolutions. Such evolutions are rather interesting and critical. Here, four types of asymmetry of order-2 interacting effects would be illustrated centering around HighBP coupled with four 1D binary features. Their four types of interacting effects are characterized by their diverse relations with HighBP given as follows: 1) "DiffWalk being independently equal"; 2) "Diabetes being nearly complete dominated"; 3) "HighChol being highly dependently equal", and 4) "Smoker being seemingly irrelevant, but strikingly modified just at one locality". Each type gives to one format of asymmetry as displayed in a figure format with double-scale panels. At the age-scale, such a figure consists of four age-panels with increasing age-category. At the bivariate-category scale for patterns of interacting effects, each age-panel consist of 4 panels: (0,0), (0,1), (1,0) and (1,1). Such discovered asymmetric patterns of interacting effects in general give rise to clear senses of information complexity. Computationally, we would clearly see the merits of (Step[2]-1, Step[2]-2) in MFCI-2 in the MFCI algorithm in this section.

All interacting effects of order-2 are uploaded into the GitHub with address listed in the caption. It is also noted that, in fact, such explorations can and should be likewise done for interacting effects of any orders. Here, it is necessary to reiterate that authentic interacting effects of order-2 and higher-order ones will contribute to our true understanding on bivariate-disease dynamics of $\mathcal{Y}$.

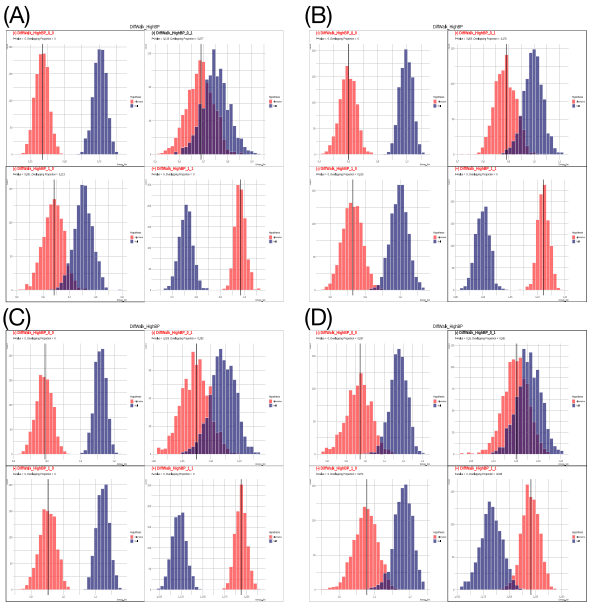

FIGURE 4. **Four increasing-age-panels: (A) to (D), with 2×2 bivariate-category-panels of interacting effects of bivariate-feature (Diffwalk, HighBP) and their reliability check via simulated alternative (in orange color) and null (in blue color) entropy distributions. See corresponding plots at https://github.com/CEDA2024/Metric-Free-Categorical-Database.**

A. TYPE-0

From the Fig. 4, we discuss our discoveries in a fashion with respect to each of 2×2 panels of bivariate-feature (Diffwalk, HighBP) across the four age-categories.

1. On the (0,0)-panel, among the four age-categories, three observed pairs of CE-distributions of 2-feature-category {DiffWalk-HighBP = [0,0] } achieve zero $Error - I\&II$, except the one of Age = 5, which achieves a less than the threshold 0.1 $Error - I\&II$ value. Such significant results can be achieved from either of the two possible directions. Direction-1: 1-feature-category {DiffWalk = 0} provides extra-info upon {HighBP = 0} to achieve the significant CE value. Direction-2:1-feature-category {HighBP = 0} provides extra-info upon {DiffWalk = 0} to result this significant CE value. In reality, we confirm that {DiffWalk = 0} and {HighBP = 0} indeed provide each other Extra-info in age-categories: Age = 1 and Age = 3 but neither directions being confirmed in Age = 4 and Age = 5. It is also evidently that the four mode-locations of their alternative CE-distributions reveal sizeable CE-reductions from the minimums of mode-locations of alternative CE-distributions pertaining to 1-featue-categories {DiffWalk = 0} and {HighBP = 0}. These are computed interacting effects bivariate-feature (Diffwalk, HighBP) = (0,0) from the age-axis perspective.
2. On the (0,1)-panel, among the four age-categories, the four observed pairs of CE-distributions of 2-feature-category {DiffWalk-HighBP = [0,1] } achieve rather large $Error - I\&II$ values comparing with the threshold. One common pattern among these four alternative-vs-null distributions is that the alternative one is on the left hand side of the null one. That is, the alternative one is stochastic smaller than the null one. Such a pattern of stochastic comparison is more evident in Age = 3 and Age = 4 than in Age = 1 and Age = 5. This pattern indicates that the category {DiffWalk = 0} seems somehow dominant over the category {HighBP = 1}, which has an opposite relational pattern of its alternative-vs-null distributions. This is one notable pattern of computed interacting effects of bivariate-feature (Diffwalk, HighBP) = (0,1).
3. On the (1, 0)-panel, among the four age-categories, the four observed pairs of CE-distributions of 2-feature-category {DiffWalk-HighBP = [0,1] } reveal even more evident common pattern among these four alternative-vs-null distributions as seen in (0,1)-panels: the alternative one is on the left hand side of the null one. The pairs Age = 3, Age = 4 and Age = 5 achieve rather small $Error - I\&II$ values comparing with the threshold. That is, the alternative one is evidently stochastic smaller than the null one. This pattern strongly indicates that the category {DiffWalk = 1} is dominated by the category {HighBP = 0}. This is an essential pattern of computed interacting effects of bivariate-feature (Diffwalk, HighBP)= (1,0). The meaning of this interacting effects is important from both societal and individual perspectives.
4. On the (1, 1)-panel, among the four age-categories, like in the panel-(0,0), the four observed pairs of CE-distributions of 2-feature-category {DiffWalk-HighBP = [1,1] } achieve zero $Error - I\&II$, except the one of Age = 5, which achieves a less than the threshold 0.1 $Error - I\&II$ value. Again, we confirm that {DiffWalk = 1} and {HighBP = 1} indeed provide each other Extra-info in age-categories: Age = 1, Age = 3 and

Age = 4, but neither directions being confirmed in Age=5. Again, it is also evidently that the four mode-locations of their alternative CE-distributions reveal sizeable CE-increments from the maximums of mode-locations of alternative CE-distributions pertaining to 1-feature-categories {DiffWalk = 1} and {HighBP = 1}. These are computed interacting effects bivariate-feature (Diffwalk, HighBP) = (1,1) from the age-axis perspective.

The evolution of non-linear interacting effects of bivariate-feature (Diffwalk, HighBP) across the four Age-panels is exhibited through the above computational patterns and results. Via its graphic display in the Fig. 4, this evolution strongly indicates that {HighBP} and {DiffWalk} play somehow equal roles along the age-axis. And both {DiffWalk = 0} and {HighBP = 0} retain dominant effects over {HighBP = 1} and {DiffWalk = 1}, respectively. This dominance manifestation of interacting effects of bivariate-feature (Diffwalk, HighBP) sends multiple strong medical and scientific messages about the dynamics underlying $\mathcal{Y} = (STK, HD)$ with respect to Age. Here, such messages precisely refer to the potential benefits of changing status: from {DiffWalk = 1} to {DiffWalk = 0} and {HighBP = 1} to {HighBP = 0} in individual and societal levels.

Nonetheless, if changes can't be done on both{DiffWalk = 1} and {HighBP = 1}, one change would also create significant impacts. This is one of the key merits of figuring out the interacting effects of bivariate-feature (Diffwalk, HighBP). On the other hand, the Fig. 4 is indeed a graphic-display for demonstrating the necessity of employing a data-driven bottom-up computational paradigm like CEDA for authentic information contained in data.

B. TYPE-I

From the Fig. 5, we continue discussing patterns of interacting effects in the same fashion with respect to each of 2 × 2 panels of bivariate-feature (Diabetes, HighBP) across the four age-categories. The computed patterns here embrace some intrinsic differences from that found in the above Type-0 of bivariate-feature (Diffwalk, HighBP), in particular, in (0,1)- and (1,0)-panels.

1. On the (0,0)-panel, among the four age-categories, three observed pairs of CE-distributions of 2-feature-category {Diabetes-HighBP = [0,0] } achieve zero $Error - I\&II$, except the one of Age = 5, which achieves a less than the threshold 0.1 $Error - I\&II$ value. The four mode-locations of their alternative CE-distributions reveal sizeable CE-reductions from the minimums of mode-locations of alternative CE-distributions pertaining to 1-featue-categories {Diabetes = 0} and {HighBP = 0}. However, we confirm that {Diabetes = 0} only provide extra-info upon {HighBP = 0 for Age = 4 categories, while {HighBP = 0} indeed provides Extra-Info upon {Diabetes = 0} with $Error - I\&II \leq 0.1$ in all Age categories.
2. On the (0,1)-panel,{Diabetes-HighBP = [0,1] } fails to be a major 2-feature-category across all four age-categories. Nonetheless, we confirmed the directional effect: {HighBP = 1} provides Extra-Info upon {Diabetes = 0}, but not the other way, across all four age-categories. Such Extra-info results are reflected on the fact that the alternative entropy distribution appeared on the right-hand side of null entropy distribution as being coherent with the pattern of {HighBP = 1}only in the Age = 1, while the three pairs of alternative-vs-null CE distributions are nearly completely overlapping in Age = 3, Age = 4 and Age = 5. This evolution of patterns of interacting effects of bivariate-feature (Diffwalk, HighBP) seemingly indicates that the the category {Diabetes = 0} in fact has varying capacity of reducing the disease risk from that of {HighBP = 1} with respect to age-categories.
3. On the (1,0)-panel, again {Diabetes-HighBP = [1, 0] } fails to be a major 2-feature-category across all four age-categories. Though, we confirmed the one-directional effect: {HighBP = 0} provides extra-info upon {Diabetes = 1}, but not the other way, the alternative entropy distribution appeared on the left-hand side of null entropy distribution same as the pattern of {HighBP = 0}, but the opposite of the pattern of {Diabetes = 1}, also across all four age-categories. Thus, patterns of interacting effects

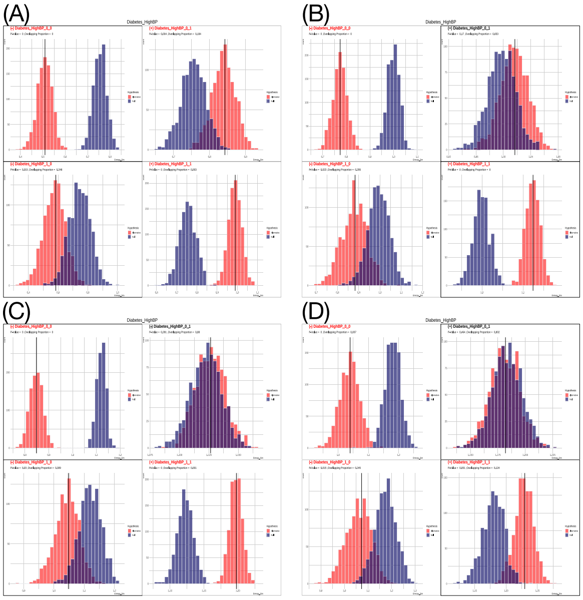

FIGURE 5. Four increasing-age-panels: (A) to (D), with 2 × 2 panels of interacting effects of bivariate-feature (Diabetes, HighBP) and their reliability check via simulated alternative (in orange color) and null (in blue color) entropy distributions. See corresponding plots at https://github.com/CEDA2024/Metric-Free-Categorical-Database.

of bivariate-feature (Diffwalk, HighBP) found in this panel and the above panel together indicate the dominance of feature {HighBP} over {Diabetes} in their interacting effects.

4. On the (1,1)-panel, {Diabetes-HighBP = [1, 1] } is confirmed as a major 2-feature-category in Age = 1, Age = 3, and Age = 4, but not in Age = 5. Also we confirmed the one-directional effect: {HighBP = 1} provides extra-info upon {Diabetes = 1} in Age = 1 and Age = 3, but not the other way, in the first three age-categories. Such one-directional effects reflects on varying amounts of CE increments of {Diabetes-HighBP = [1, 1] } over the CEs of {HighBP = 1} and {Diabetes = 1}.

From the above results of the four panel displayed in Fig. 5, we can see that {HighBP} is apparently dominant over {Diabetes}. Also, we found that both {HighBP = 0} and {HighBP = 1}indeed bring extra information upon both {Diabetes = 0} and {Diabetes = 1}, but not the other way around. Both results conclude non-linear interacting effects for bivariate-feature ({HighBP}, {Diabetes}) coupled with somehow sophisticated dominance within the dynamics underlying $\mathcal{Y} = (STK, HD)$.

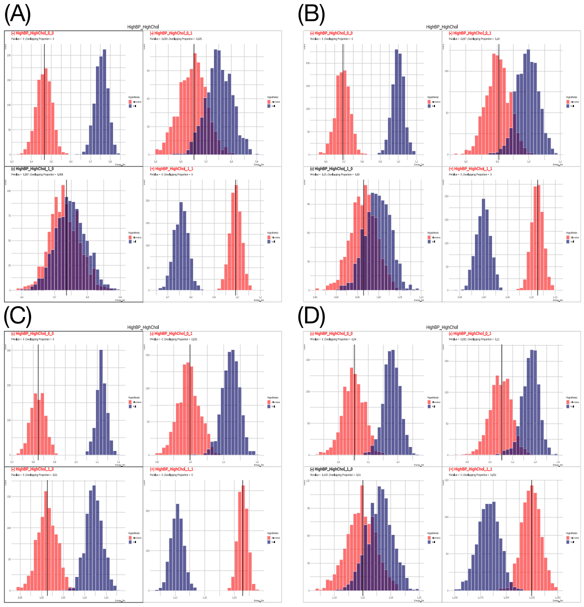

FIGURE 6. Four increasing-age-panels: (A) to (D), with 2 × 2 panels of interacting effects of bivariate-feature (HighBP, HighChol) and their reliability check via simulated alternative (in orange color) and null (in blue color) entropy distributions. See corresponding plots at https://github.com/CEDA2024/Metric-Free-Categorical-Database.

C. TYPE-II

The three 1D features: {DiffWalk}, {HighBP} and {HighChol}, are key risk factors of disease dynamics underlying $\mathcal{Y} = (STK, HD)$. At Age = 1, 3 and 4, these three binary factors mutually provide Extra-Info among their categories, while such mutual relations disappear in Age = 5. Nonetheless, the evolutions of patterns of order-2 interacting effects of bivariate-feature ({HighBP}, {DiffWalk}) and ({HighBP}, {HighChol}) are somehow distinct. As would be seen below through Fig. 6 and panel-based summary, the bivariate-feature ({HighBP}, {HighChol}) reveal some extents of "asymmetry", which is not exactly identical the asymmetric patterns found in bivariate-feature ({HighBP}, {DiffWalk}).

1. On the (0,0)- and (1,1) panels, {HighBP-HighChol = [0, 0] } and {HighBP-HighChol = [1, 1] } are all confirmed as a major 2-feature-category at Age = 1, 3 and 4, but not Age = 5.
2. On the (0,1)-panel, {HighBP-HighChol = [0, 1] } is confirmed as a major 2-feature-category at Age = 3 and 4, but not Age = 1 and 5. The alternative entropy distribution is located on the left-hand side of null entropy distribution with sizable overlapping at Age = 1 and 5, but having near-zero overlapping at Age = 3 and 4. The interpretation of this evolving pattern over age is that the status {HighBP = 0} is more important than status {HighChol = 1} in terms of subject's disease risk.
3. On the (1, 0)-panel, {HighBP-HighChol = [1, 0] } is also confirmed as a major 2-feature-category only at Age = 4, but not at Age = 1, 3 and 5. The evolving pattern of relative position of the alternative entropy distribution toward the null entropy distributions has gone from almost entirely overlapping to entirely separated and back to heavily overlapping from Age = 1 to Age = 5. It means that {HighChol = 0} is more important than {HighBP = 1} only at Age = 4.

Though {HighBP} and {HighChol} seem to play equal roles, their interacting effects revealed in (0,1)- and (1,0)-panels are highly asymmetric across all age-categories. Such evolving asymmetry is hardly known in any priori fashion. That is, the evolution of asymmetric order-2 interacting effects of {HighBP} and {HighChol} can only be described precisely on the category-locality, not on global or marginal scale. The computational approach for patterns of such nature needs to be data-driven and bottom-up like CEDA paradigm. And these computational and observed facts further enhance that ICiD is consisting of 1D histograms of feature-categories.

D. TYPE-III

Next, we consider the evolving order-2 interacting effects of bivariate-feature (HighBP, Smoker) across the four age-categories. Upon the Fig. 7, we would see two significant evolving patterns. The first pattern is that, through the (0,0)-, (0,1)- and (1,1)-panels, we see the two categories of {Smoker} have nearly zero interacting effects with categories of {HighBP} at Age = 1, 3 and 4, while this pattern of interacting effect disappears at Age = 5 in a fashion that the alternative and null entropy distributions become heavily overlapping. The second pattern is seen through (1,0)-panels

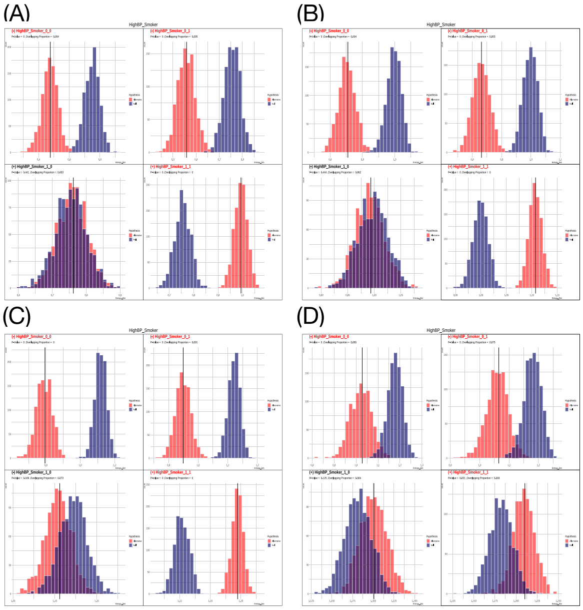

FIGURE 7. Four increasing-age-panels: (A) to (D), with 2 × 2 panels of interacting effects of bivariate-feature (HighBP, Smoker) and their reliability check via simulated alternative (in orange color) and null (in blue color) entropy distributions. See corresponding plots at https://github.com/CEDA2024/Metric-Free-Categorical-Database.

across age-categories. The pair of alternative and null entropy distributions is nearly overlapping each other at Age = 1. Then, the alternative one shifts to the left of the null one at Age = 3 and 4. At the end, the alternative one shifts to the right of the null one at Age = 5. This evolving pattern means that interacting effects of {Smoker = 0} and {HighBP = 1} are visible, but deceasing to a great extent at Age = 5 subpopulation.

In summary, these two evolving patterns indeed bear significant scientific impacts on understanding chronical diseases. Hence, it is worth reiterating that not only {HighBP} plays a dominant role over {Smoker} with highly asymmetric effects, but also their relational patterns do change along the age-axis. In fact, as would be seen in the next section, {Smoker} does play important role through its interacting effects with {DiffWalk}, {HighBP} and {HighChol}. This is one authentic and scientific, but very different way of describing effects of smoking in our society. This is indeed rather striking. In contrast, {Veggies} don't have similar effects at all.

Though the evolution of the above four types of order-2 interacting effects are not unthinkable if we take a retrospective viewpoint, the existence of such seemingly all natural types interacting effects emphasizes one simple fact that the diversity of functional forms of interacting effects can be too complex to be modeled realistically. As such we again emphasize the fact that these natural and explainable patterns are possible and visible only when we adopt bottom-up data-driven computational paradigm, like CEDA. This simple fact is tied to the categorical-locality nature.

V. TOPOLOGICAL INDIVIDUAL RISK-LANDSCAPES AND THEIR EVOLUTIONS VIA MFCI

Four heatmaps of confirmed major 1-feature-categories for the four age-categories are resulted from applying MFCI-1 step of algorithm MFCI respectively and shown in Fig. 8. The memberships of the four sets of major 1-feature-categories (or 1-risk-factor-categories) are highly overlapping. The common risk factors along the age-axis are:{DiffWalk, Diabetes, HighBP, HighChol}. Interestingly, risk-factors {Smoker, Income} are present from Age = 1 to Age = 3, but drops out at Age = 4 and are replaced by {NoDocbcCost}. This evolution coherently confirms the expected fact that the effects of these two groups of risk-factors are highly age-dependent. It is evidently noted that there is an evolutionary break-down seen in the Age = 5 panel, which consists only 3 major 1-feature-categories. The discussion of this phenomenon is given in the subsection just before the Conclusions.

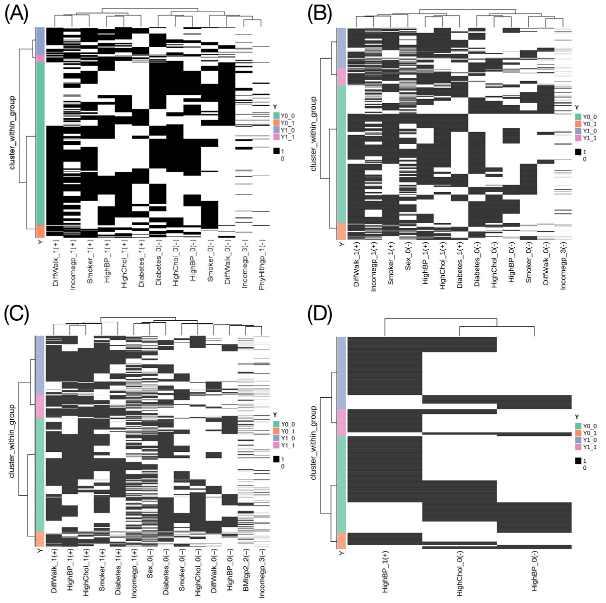

FIGURE 8. Four heatmaps of binary bipartite network matrices of major 1-feature-categories selected with respect to the threshold $Error - I\&II \leq 0.1$ across the four age-categories.

Three heatmaps of confirmed major 2-feature-categories for the age-categories: Age = 1, 3 and 4, are resulted from applying MFCI-2 step of algorithm MFCI respectively and shown in Fig. 9. At Age = 5, we don't find any 1-feature-category being able to provide Extra-Info for all confirmed major 1-feature-categories found through MFCI-1 step. However, we present those 2-feature-categories that merely satisfy the threshold $Error - I\&II \leq 0.1$.

Across the four heatmaps in Fig. 9 along the age-axis, almost all major 2-feature-categories are primary interacting pairs of major 1-feature-categories. The chief implication of such an evident pattern is that higher order interacting effects are highly potential at least in age-categories: Age = 1, 3 and 4. On one hand, since 2-feature-categories narrated

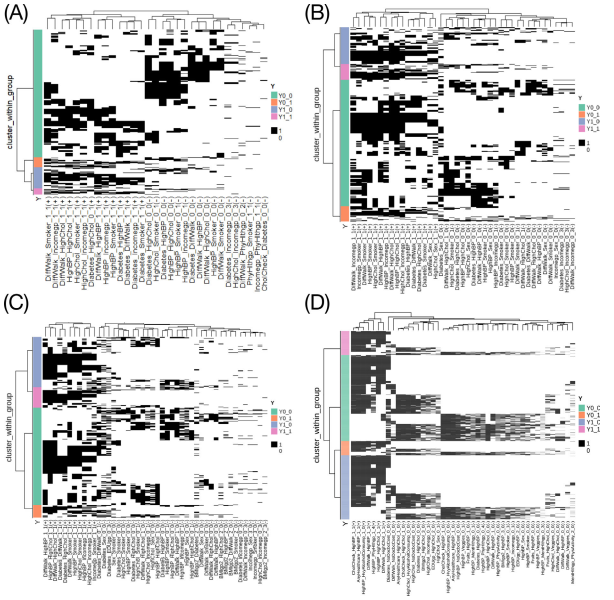

FIGURE 9. Four heatmaps of binary bipartite network matrices of major 2-feature-categories selected with respect to the threshold *Error* − *I&II* ≤ 0.1 across the four age-categories. The panel of Age = 5 consists of unconfirmed 2-feature-categories with respect to the Step[2]-1 of MFCI-2 step.

in (0,1)- and (1,0)-panels in the previous section are most not major 2-feature-categories found in Fig. 9, diseased-vs-non-diseased subjects belonging to these two 2-feature-categories need to be further separated by at least one more 1-feature-categories. On the other hand, subjects in those confirmed major 2-feature-categories narrated in (0,0)- and (1,1)-panels in the previous section could be further separated to achieve better diseased-vs-non-diseased separation beyond 2-feature-categories, as would clearly be seen in next two subsections.

A. RESULTS OF MFCI-3 STEP AT AGE = 1

In this subsection, we report computed patterns through various heatmaps of major 3-feature-categories within the subpopulation (GenHL, Age) = (5, 1), while similar resultant patterns of subpopulations at Age = 3 and 4 are reported in the next subsection. In this section, one key idea of individual risk-landscape would be introduced. And all subjects' individual risk-landscapes are collectively displayed through three versions of heatmaps. We then construct summarizing contingency tables to confirm that such individual risk-landscapes based heatmaps contain significant amounts of pattern information content in data (ICiD). At the end, we demonstrate the apparently existing so-called "atypical subjects" in both diseased and non-diseased response categories.

Upon the Step[2]-2 of MFCI-2 step, we identified and confirmed 31 major 2-feature-categories as seen in Fig. 3. We then further perform the MFCI-3 step of the MFCI algorithm. Upon applying Step[3]-1 of MFCI-3 step, we found 65 major and non-major 1-feature-categories that can provide extra-info upon the 31 major 2-feature-categories resulted from MFCI-2 step. Further, applying Step[3]-2 of MFCI-3 step, we identified and confirmed 31 major 3-feature-categories, see Fig. 10. There will be 41 confirmed if the confirmation criterion is switched to P-value being less than 0.05, see Fig. 11. This heatmap is presented here to indicate the potential fact that selection criterions based only on P-values, not involving with alternative distributions, are likely over-optimistic. All subsequent analyses are to be based on results contained in Fig. 10.

Here, here by having the binary bipartite network's matrix lattice as a platform, we only collect all the major 3-feature-categories and arrange them onto the column-axis in Fig. 10. Memberships of each major 3-feature-category is represented by the corresponding binary column-vector. The column-axis is framed by a hierarchical clustering (HC) tree derived by using Euclidean distance, while the row-axis is also rearranged in a response-category specific fashion. That is, subjects belonging to the same response-category are arranged by its own HC-tree. So subjects of different response-categories do not fix together. Such a heatmap is created purely for easy visualization purpose. We wish to convey that response-category specific patterns in such heatmaps could help shed light on how major 3-feature-categories collectively work out their roles for the dynamics of $\mathcal{Y}$.

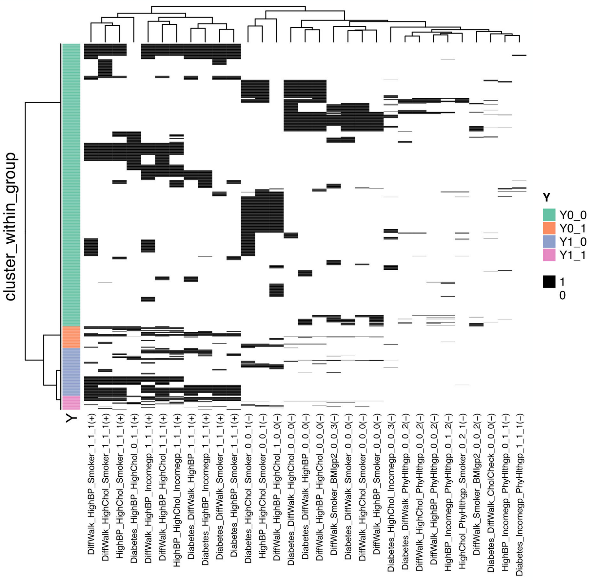

FIGURE 10. Heatmap of binary bipartite network matrix 1776 × 31 with 31 major 3-feature-categories selected with respect to the threshold *Error* − *I&II* ≤ 0.1.

1) INDIVIDUAL RISK-LANDSCAPE INTERPRETATION

The chief merit of employing a heatmap here is to explicitly reveal the concept of individual risk-landscape and foster authentic understanding from collective patterns of such risk-landscapes. We first recall that the heatmap shown in Fig. 10 has a format of 1776 × 31 matrix

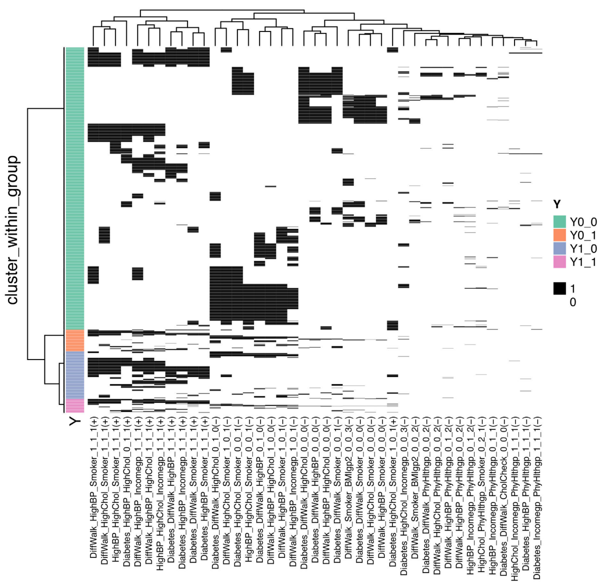

FIGURE 11. Heatmap of binary bipartite network matrix 1776 × 41 with 41 major 3-feature-categories selected with respect to the threshold P-value $\leq$ 0.05.

with 31 identified and confirmed major 3-feature-categories arranged along the column-axis. The binary memberships among the 1776 subjects each of 3-feature-category is listed as one column. As such each subject is represented by a 31-dim binary row-vector across the 31 major 3-feature-categories. Such a 31-dim binary vector indeed explicitly indicates what kinds of positive or negative disease risks this subject is facing simultaneously. Specifically speaking, along the column-axis with 31 columns, each subject within this sub-population (GenHL $= 5$ and Age $= 1$) embraces potential positive disease risk via memberships of 11 major 3-feature-categories marked with "+" signs and potential negative disease risk via memberships of 20 major 3-feature-categories marked with negative"−" signs.

As such a 31-dim binary vector endorses a subject's individual-risk-landscape that becomes the most critical information pertaining to this individual's health. In comparison with the 31 major 2-feature-categories presented in Fig. 3, all major 3-feature-categories in Fig. 10 by-and-large have higher disease-to-non-disease odds. Such higher odds would render clearer and more informative disease-related mechanistic patterns as would be derived below.

All 1776 subjects' individual 31-dim individual risk landscapes indeed collectively constitute visible patterns of various scales. The heatmap shown in Fig. 10 is framed by a hierarchical clustering (HC) tree on column-axis and 4 color-coded bivariate diseases categories. At its top internal node, HC-tree splits into left and right branches, coded as L1 vs. R1, respectively. The L1 branch consists all 11 positive disease risk major 3-feature-categories, while the R1 branch consists all 20 negative disease risk 3-feature-categories. Branch L1 further splits into L1L2 and L1R2 subbranches which are color-coded gray on 4 and green on 7 major 3-feature-categories of positive disease risk, respectively. Likewise R1 splits into R1L2 and R1R2 subbranches color-coded red on 3 and blue on 17 major 3-feature-categories negative disease risk, respectively.

These four subbranches indeed embrace their characteristics due to their distinct compositions of major 3-feature-categories. It becomes natural to take these four characteristics into considerations when thinking about the similarity or dissimilarity among study subjects. That is, we make the membership-sums of these four subbranches into four extra feature-variables. The 35-dim Euclidean distance is used to remake an extended version of heatmap as shown in Fig. 12. This heatmap embraces more evident blocks than the one in Fig. 10.

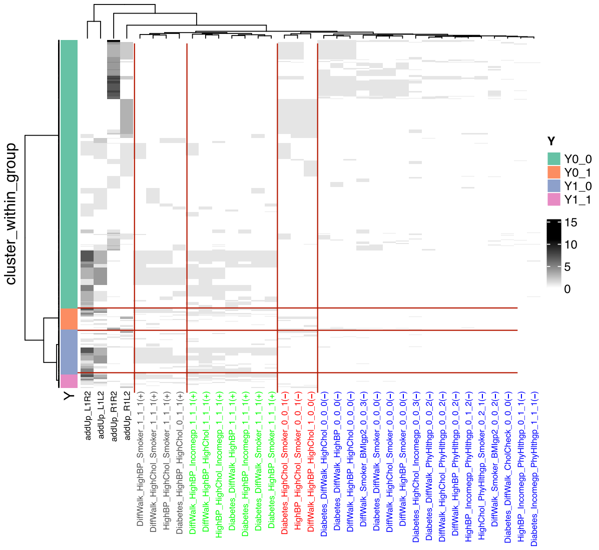

FIGURE 12. Extended version heatmap with 35-dim Euclidean distance from the orignal version based on the binary bipartite network matrix 1776 × 31 with 31 major 3-feature-categories selected with respect to the threshold *Error* − *I&II* $\leq$ 0.1.

This revised heatmap clearly reveal block-patterns as demonstrated in Fig. 12. Apparently, each block in is jointly framed by membership-cluster of a subbranch of major 3-feature-categories, which are more or less constant, and a cluster of study subjects, who are rather similar for their individual risk-landscapes. These blocks collectively convey explicit pattern-dynamics underlying $\mathcal{Y}$. In particular, a series of horizontally displayed blocks will characterize a cluster of study subjects with visible and explainable pattern information. In the next subsection, we elaborate merits and importance of such characterization in details.

As each horizontal series of blocks induced a well-defined neighborhood for all study subject participating in this block, this heatmap can be taken as an informative display of "topology" defined on the collection of 1776 study subjects. In mathematical term, this topological space here is equipped with the 35-dim Euclidean distance that defines neighborhoods for all study subjects. In covariate information term, a subject's neighborhood is meant to be a set of subjects having very similar individual risk-landscapes. This heatmap as a topological display is one of our chief summarizing statistics. The reasons underlying this statement are given as follows. The relational pattern information regarding dynamics of $\mathcal{Y}$ is in full display: [response-category]-vs -[individual risk-landscape], in the heatmap. Further, the topological insights will have profound impacts on Data Analysis as a scientific discipline. In contrast, from this topological perspective, many statistical and machine learning topics: ranging from classification to clustering, are not rigorously formulated when facing real-world complex systems, also see [8].

2) SUMMARIZING STATISTICS BASED ON SUBJECTS' TOPOLOGY

To explicitly see merits of the topology-bearing heatmap from another aspect of summarizing statistics, we specifically look at one positive and one negative disease risk subbranches: L1R2 and R1R2, respectively, among the four aforementioned subbranches on column-axis. And for expositional simplicity, we interpret the disease risk pertaining to $\mathcal{Y} \in \{(1, 0), (0, 1), (1, 1)\}$ against $\mathcal{Y} = (0, 0)$ via the odds. The baseline or overall disease odds in this (GenHL $= 5$ and Age $= 1$) subpopulation as $\frac{404}{1372} = 0.2944$.

Upon the L1R2 subbranch, which consists of 7 major 3-feature-categories of positive disease risk, each subject's total memberships of this subbranch ranges from 0 to 7. The odds for subjects, who accumulate 4 up to 7 memberships, is $\frac{109}{119} = 0.9160$. This odds indicates that a subject having 4 or more memberships within this subbranch has probability of belonging to the diseased with probability nearly 0.5. The odds-ratio is calculated as $\frac{0.9160}{0.2944} = 3.1114$. This ratio indicates that these subjects are at least 3 times more likely to be diseased: either Stroke or Heart disease, than subjects in the entire subpopulation in general. In sharp contrast, the odds for subjects, who accumulate 1 membership up to 3 memberships is $\frac{168}{397} = 0.4232$. And the odds-ratio is calculated as $\frac{0.4232}{0.2944} = 1.4375$. This ratio indicates that these subjects nearly 1.5 times likely to be diseased as the subjects in this subpopulation in general. Further, subjects have zero memberships on this subbranch have an odds $\frac{127}{856} = 0.1484$. That is, such subjects have an odds-ratio $\frac{0.1484}{0.2944} = 0.5041$, that is, subjects in this sub-population in general are twice as likely to be diseased as such subjects with zero memberships in this subbranch. These three spreading widely odds-ratios indicates the informativeness of L1R2 subbranch on the positive disease risk.

However, the heatmap shown in Fig. 12 reveal much more important visual patterns beyond the above three widely spreading odds-ratios and their interpretations based on results associated with L1R2 subbranch. Here are the essences of three implications derived from the visual patterns:

1. The superficial disease-imbalance phenomenon indeed is embedded with somehow surprising hidden structural causes: "atypical subjects", as would be described below. Such causes render any predictive approaches unsustainable because of not only having very high error-rates, but also being neither informative nor scientifically correct.
2. The precise and drastically distinct multi-scale block-patterns of topological individual risk-landscapes of all involving subjects are critical for understanding the dynamics of $\mathcal{Y} = (STK, HD)$.
3. The high vs zero intensities of memberships within each block across positive and negative disease risks of major 3-feature-categories pave ways for distinguishing high risk subjects against lower risk ones.

These are three chief findings in our CEDA based data analysis and chief characteristics of resultant ICiD of locality nature.

The above three chief finding of CEDA data analysis are even more evident via interpretations of results from the branch R1R2. This branch consists of 17 major 3-feature-categories of negative disease risk. There are total 51 subjects having 8 or more memberships. Strikingly, this group of subject have zero odds. This result is indeed striking. There are 280 subjects who have at least 3, but no more than 7 memberships. This group of subjects' odds is $\frac{13}{267} = 0.0487$, and odds-ratio $\frac{0.0487}{0.2944} = 0.1654$. The probability of being diseased for subjects in this group is as low as 0.05, and its relative risk of this group to the whole sub-population is less than one fifth. There are 253 subjects have one or two memberships. This group's odds is $\frac{30}{223} = 0.1345$, so its subjects' probability of being disease is less than $\frac{1}{8}$. Its odds-ratio is less than $\frac{1}{2}$. Finally, there are 1192 subjects who do not own any memberships out of these 20 major 3-feature-categories. There are 360 having diseases among these 1192 subjects, that is, the odds is $\frac{361}{1192-361} = \frac{361}{831} = 0.4344$, and the probability is 0.3029. The odds-ratio is 1.4755. That is, subjects with zero memberships on R1R2 branch will have 1.5 times of the relative risk of subjects within sub-population in general.

The above topological risk-landscape based findings of positive risk based on the branch L1R2 and of negative risk on the branch R1R2 together clearly spell out the essential merits of identified and confirmed high orders effects of feature-categories. It is somehow revealing to see such significant results and informative patterns via such simplistic computations. More revealing is that the amount of branch-memberships becomes a synthesized variable. That is, L1R2 and R1R2 can be transformed into two very informative variables that more precisely prescribe positive and negative disease risks, respectively, than any feature-sets. Such consequential synthesizing mechanism of risk factors is amazingly achieved without any man-made structures.

We respectively transform the membership in branch L1R2 and in branch R1R2 into two new variables: $Syn[L1R2]$ and $Syn[R1R2]$, in the following fashions. Denote a subject's total memberships on branch L1R2 and R1R2 as two variables: $\#[L1R2]$ and $\#[R1R2]$, respectively.

1. [On branch L1R2:] $Syn[L1R2] = 4+$ if $\#[L1R2] \geq 4$; $Syn[L1R2] = 1+$ if $1 \leq \#[L1R2] \leq 3$; $Syn[L1R2] = 0+$ if $\#[L1R2] = 0$.
2. [On branch R1R2:] $Syn[R1R2] = 8-$ if $\#[R1R2] \geq 8$; $Syn[R1R2] = 3-$ if $3 \leq \#[R1R2] \leq 7$; $Syn[R1R2] = 1-$ if $1 \leq \#[R1R2] \leq 2$; $Syn[R1R2] = 0-$ if $\#[R1R2] = 0$.

We then build the following odds and odds-ratio table. Let $n_{4+,8-} = d_{4+,8-} + nond_{4+,8-}$ be the number of subjects belonging to the $(Syn[L1R2], Syn[R1R2]) = (4+, 8-)$, which is the sum of $d_{4+,8-}$ as the number of diseased subjects and $nd_{4+,8-}$ as the number of non-diseased subjects. This table reveals that the bivariate $(Syn[L1R2], Syn[R1R2])$ is capable of a wide spectrum of odds, so it is rather informative to dynamics underlying $\mathcal{Y}$. That is to say that the topological individual risk-landscape via heatmap, shown in Fig. 12, indeed captures the very essential associative patterns regrading this Re-Co dynamics.

TABLE 3. Odds table of $Syn[L1R2] - vs - Syn[R1R2]$ division of the sub-population of heterogeneity (GenHL = 5, Age = 1).

$Syn[L1R2]/Syn[R1R2]$	8-	3-	1-	0-	row-wise
4+	$\frac{d_{4+,8-}}{nd_{4+,8-}} = \frac{0}{0}$	$\frac{d_{4+,3-}}{nd_{4+,3-}} = \frac{0}{0}$	$\frac{d_{4+,1-}}{nd_{4+,1-}} = \frac{1}{6}$	$\frac{d_{4+,0-}}{nd_{4+,0-}} = \frac{108}{113}$	$\frac{109}{119}$
1+	$\frac{d_{1+,8-}}{nd_{1+,8-}} = \frac{0}{0}$	$\frac{d_{1+,3-}}{nd_{1+,3-}} = \frac{0}{7}$	$\frac{d_{1+,1-}}{nd_{1+,1-}} = \frac{6}{33}$	$\frac{d_{1+,0-}}{nd_{1+,0-}} = \frac{162}{357}$	$\frac{168}{397}$
0+	$\frac{d_{0+,8-}}{nd_{0+,8-}} = \frac{0}{51}$	$\frac{d_{0+,3-}}{nd_{0+,3-}} = \frac{13}{260}$	$\frac{d_{0+,1-}}{nd_{0+,1-}} = \frac{23}{184}$	$\frac{d_{0+,0-}}{nd_{0+,0-}} = \frac{91}{361}$	$\frac{127}{856}$
col-wise	$\frac{0}{51}$	$\frac{13}{267}$	$\frac{30}{223}$	$\frac{361}{831}$	$\frac{404}{1372}$

Likewise we make two synthesized variables: $Syn[L1]$ and $Syn[R1]$, based on the two major branches L1 and R1. The odds table of $Syn[L1] - vs - Syn[R1]$ is given in Table 4. The information content is relatively similar with that in Table 3.

TABLE 4. Odds table of $Syn[L1] - vs - Syn[R1]$ division of the sub-population of heterogeneity (GenHL = 5, Age = 1).

$Syn[L1]/Syn[R1]$	8-	3-	1-	0-	row-wise
4+	$\frac{0}{0}$	$\frac{0}{0}$	$\frac{1}{7}$	$\frac{191}{255}$	$\frac{192}{262}$
1+	$\frac{0}{0}$	$\frac{0}{16}$	$\frac{27}{147}$	$\frac{87}{202}$	$\frac{114}{365}$
0+	$\frac{0}{55}$	$\frac{44}{437}$	$\frac{37}{188}$	$\frac{17}{65}$	$\frac{98}{745}$
col-wise	$\frac{0}{55}$	$\frac{44}{453}$	$\frac{65}{342}$	$\frac{295}{522}$	$\frac{404}{1372}$

3) ATYPICAL SUBJECTS

As the [response-category]-vs-[individual risk-landscape] relational pattern information of dynamics of $\mathcal{Y}$ being in full display via block-patterns embedded within the heatmap shown in Fig. 12, we clearly see that there are 3 types of non-diseased subjects in the category $\mathcal{Y} = (0, 0)$. The three types are: 1) zero positive risk with prevalent memberships in R1 branch; 2) some positive and some negative risk; 3) zero negative risk with prevalent memberships in L1 branch across the 31 major 3-feature-categories. These three types are seemingly seen within the diseased subjects in the categories $\mathcal{Y} = \{(0, 1), (1, 0), (1, 1)\}$ as well. Immediately, we realize contradicting mechanisms through the simultaneous presence of these three types of [response-category]-vs -[individual risk-landscape] relational patterns in both diseased and non-diseased categories.

Intuitively and ideally speaking, the category $\mathcal{Y} = (0, 0)$ should be full of the type-1 subjects, or at most include some type-2, and the categories $\mathcal{Y} = \{(0, 1), (1, 0), (1, 1)\}$ should be full of type-3 subjects. But, apparently, this is not case in the heatmap in Fig. 12. Counter-intuitive and non-ideal situations are observed: a large group of type-3 subjects in the category $\mathcal{Y} = (0, 0)$ and a group of type-1 subjects in the categories $\mathcal{Y} = \{(0, 1), (1, 0), (1, 1)\}$.

In particular, a large group of type-3 subjects in the category $\mathcal{Y} = (0, 0)$ are those who resist the trend to go much higher odds-ratios than 3 in the bivariate cell $(Syn[L1R2], Syn[R1R2]) = (4, 0)$ in Table 3 or $(Syn[L1], Syn[R1]) = (4, 0)$ in Table 4. For this reason, we term such subjects: ''atypical subjects'' in $\mathcal{Y} = (0, 0)$. Likewise, we have ''atypical subjects'' in the categories $\mathcal{Y} = \{(0, 1), (1, 0), (1, 1)\}$. To better visualize the presence of such atypical subjects, we build another heatmap by lifting off the response-category constraint in Fig. 13. It is clearly seen that subjects with similar individual risk-landscapes are grouped together, while their response-category-marks are mixed.

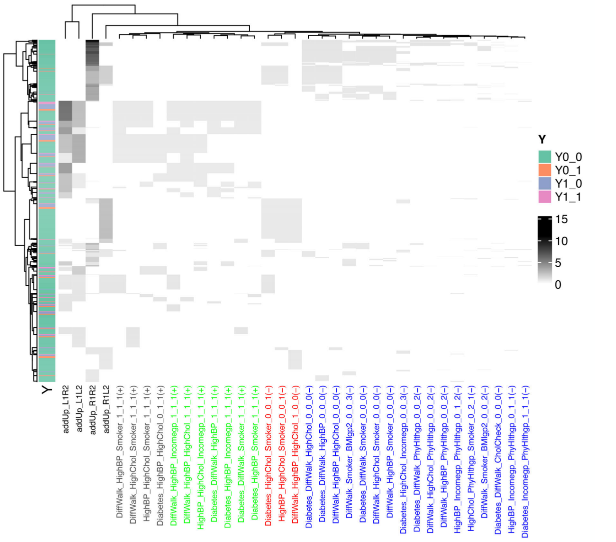

FIGURE 13. Extended version heatmap with 35-dim euclidean distance and without response-categories constraints on the row-axis.

At the end of this subsection, we mention the following obvious implications of ''atypical subjects'' within *calY*-vs-IRL topological relations. Though the nature of ''atypical subjects'' is to be discovered, the graphic displays of mapping out and displaying all such ''atypical subjects'' within the heatmap indeed reveal full [response-category]-vs-[individual risk-landscape] relations. Thus, Fig. 12 and Fig. 13 are essential computational results in data analysis.

The existence of "atypical subjects" also points out the superficial nature of so-called "imbalance" phenomenon when analyzing this Kaggle data set in Statistics and Machine Learning literatures. Further, any validity of "re-balancing" operations as seen in [13], [14], and [15], which are created to remedy or lessen the impact of such "imbalance" between non-diseased and diseased, are by-and-large questionable.

B. RESULTS OF MFCI-3 STEP AT AGE = 3 AND 4

In this subsection, we present computed relational patterns from Age = 3 and 4 and compare them in a side-by-side fashion. Such representations are designed for the purposes of convenient comparisons to figure out evolving changes between these two sub-populations.

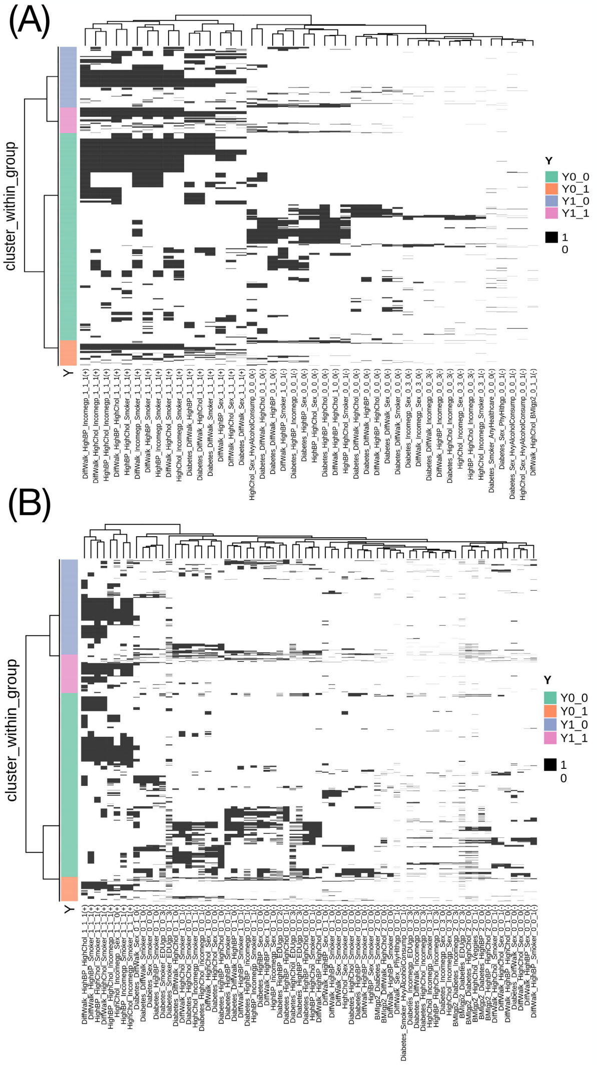

FIGURE 14. Two heatmaps of binary bipartite network matrices of major 3-feature-categories selected with respect to the threshold *Error* − *I&II* ≤ 0.1: (a) 3244 × 44 for Age = 3; (b) 4819 × 70 for Age = 4.

By applying Step[3]-2 of MFCI-3 step of the MFCI algorithm upon the two sub-populations of Age = 3 and 4, we respectively identified and confirmed 44 (= 28(−)+16(+)) and 70 (= 62(−)+8(+)) major 3-feature-categories, see two panels of Fig. 14: (a) for Age = 3 and (b) Age = 4.

Upon the 3244 × 44 heatmap of Age = 3, almost all 16 out of 44 are positive-diseased risk oriented major 3-feature-categories. The three feature-members of these 16 major 3-feature-categories primarily consisting two features from {DiffWalk, Diabetes, HighBP, HighChol} coupled with one feature from {Income (group-1), Smoker, Sex (male)}. One worth noting major 3-feature-category is {(HighChol, Income, Smoker) = (1,1,1)} in this young-adult sub-population. As for the 28 negative-diseased risk oriented major 3-feature-categories, their members of three selected features maintain the same structure as the positive risk ones, while the category of {Sex} is female and category of {Income} is the group 3.

In contrast, upon the 4819 × 70 heatmap of Age = 4, only 8 out of 70 major 3-feature-categories are positive-diseased risk oriented. Interestingly, the three feature-members of these 13 major 3-feature-categories primarily consist 3 features from {DiffWalk, HighBP, HighChol, Income, Smoker}, excluding {Diabetes}. In particular, the female gender is coupling with the lowest income-group: {Income = 1} at this late-adult sub-population. As for the 62 negative-diseased risk oriented major 3-feature-categories, their members of three selected features again maintain the same structure as the positive risk ones, while the category of {Sex} remains female and category of {Income} is the group 3. Another evident change in Age = 4 is that categories: {Edu = 3} and {BMI = 2}, are involved in more than 10 selected major 3-feature-categories on the side of negative disease risk.

Further evident and significant patterns are found in Age = 3 and Age = 4. First, the major 1-feature-categories of negative disease risk never involve in major 3-feature-categories of positive disease risk. This pattern is also seen in major 2-feature-categories presented in the previous section. The second pattern is that major 1-feature-categories of positive disease risk are involved in many major 3-feature-categories of negative disease risk. This observation strongly indicates the complexity embraced by the disease dynamics of response variable $\mathcal{Y} = (STK, HD)$.

Furthermore, since the positive disease risk effects are possible to be compensated by multiple negative disease risks. Such a possibility indeed offers for individuals to alter their disease risks by making some behavioral changes. We take this possibility is one of the chief merit of displaying computed and confirmed pieces of information in a format of heatmap of individual risk-landscape. This pattern is seen much more prevalent in Age = 4 than Age = 3. One interpretation of this age-related difference is tentatively attributed to more complex disease dynamics of response variable $\mathcal{Y} = (STK, HD)$ in Age = 4 than that in Age = 3.

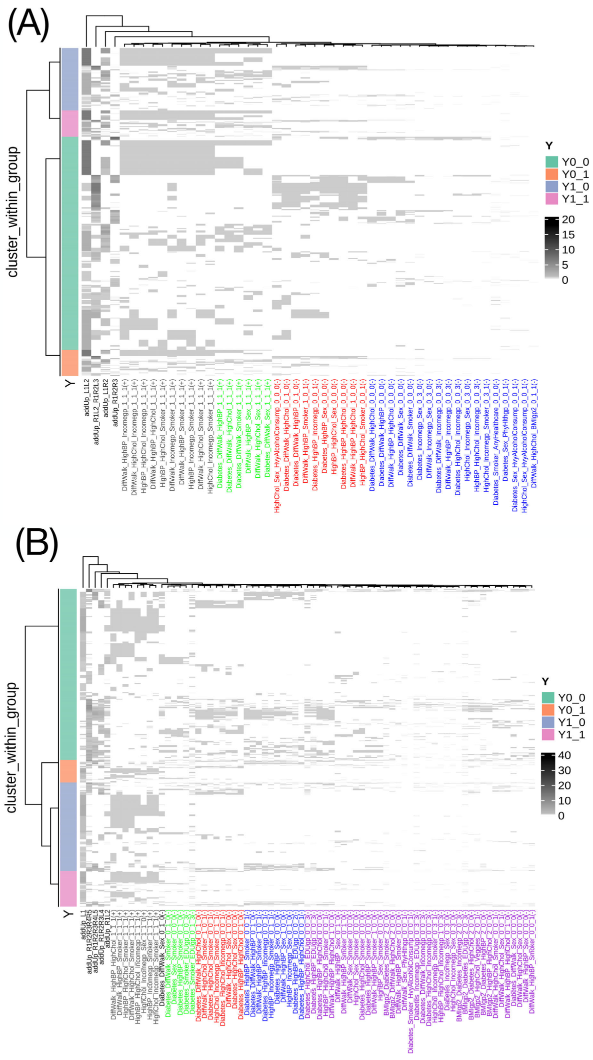

FIGURE 15. Two heatmaps of binary bipartite network matrices of major 3-feature-categories with extra dimensions and modified Euclidean distances: (a) 3244 × 44 for Age = 3 with 4 extra dimensions; (b) 4819 × 70 for Age = 4 with 5 extra dimensions.

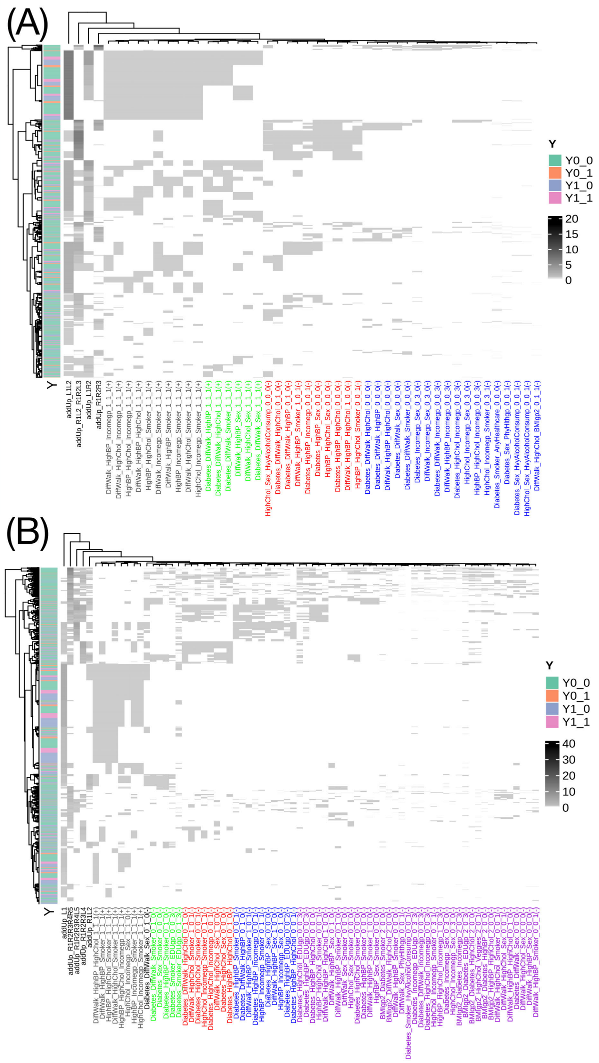

FIGURE 16. Two heatmaps of binary bipartite network matrices of major 3-feature-categories with extra dimensions and modified Euclidean distances, but without constraints on response categories on row-axis: (a) 3244 × 44 for Age = 3 with 4 extra dimensions; (b) 4819 × 70 for Age =4 with 5 extra dimensions.

From the perspective of "atypical subject", the Fig. 15 and Fig. 16 together reveal its existential evidence in diseased and non-diseased categories within both Age = 3 and Age = 4. Such an existence of "atypical subjects" strongly indicates the importance of recognizing the true goal of data analysis as constructing graphic displays that can explicitly exhibit detailed individual risk-landscape. Upon these heatmaps, we can visualize the topological relations pertaining to each subject's neighborhoods under the same platform, with which subject-specific similarity and dissimilarity become natural and obvious. Such topologies immediately link to many essential scientific issues, such as how to design randomized trials, how to design experiments for finding extreme high (or low) disease risk subjects and how to properly understand causal effects under observational study and many others.

At the end of this subsection, we again present the synthesized bivariate $(Syn[L1], Syn[R1])$ as an identified informative summarizing 2D statistics for complex disease dynamics of $\mathcal{Y} = (STK, HD)$ within the sub-populations Age = 3 and 4, respectively. Two corresponding contingency tables: Table 5 for Age = 3 and Table 6 for Age = 4, are seen to capture a wide spectrum of disease risk potentials and characteristics: from very low disease risk to rather high disease risk in terms of cell-specific odds comparing with the odds pertaining to the sub-population. It is evident that the so-called imbalance phenomenon is no longer present in these two sub-populations. However, the presence of "atypical subjects" remains. That is, any predictive approaches are to suffer very high error rates and to be seen as being impractical. Once again, this presence of "atypical subjects" is a clear and strong reminder regarding the fact that some important risk-factors might still be missing in this Kaggle version of BRFSS data set.

TABLE 5. Odds Table of *Syn*[*L*1] – *vs* – *Syn*[*R*1] division of the sub-population of heterogeneity (GenHL = 5, Age = 3).

$Syn[L1]/Syn[R1]$	8-	3-	1-	0-	row-wise
4+	$\frac{0}{0}$	$\frac{42}{91}$	$\frac{118}{211}$	$\frac{605}{698}$	$\frac{765}{1000}$
1+	$\frac{28}{81}$	$\frac{60}{179}$	$\frac{92}{222}$	$\frac{78}{151}$	$\frac{258}{633}$
0+	$\frac{34}{190}$	$\frac{39}{189}$	$\frac{22}{66}$	$\frac{17}{31}$	$\frac{112}{476}$
col-wise	$\frac{62}{271}$	$\frac{141}{459}$	$\frac{232}{499}$	$\frac{700}{880}$	$\frac{1135}{2109}$

TABLE 6. Odds Table of *Syn*[*L*1] – *vs* – *Syn*[*R*1] division of the sub-population of heterogeneity (GenHL = 5, Age = 4).

$Syn[L1]/Syn[R1]$	8-	3-	1-	0-	row-wise
4+	$\frac{0}{0}$	$\frac{25}{22}$	$\frac{270}{212}$	$\frac{674}{381}$	$\frac{965}{615}$
1+	$\frac{165}{283}$	$\frac{257}{379}$	$\frac{145}{136}$	$\frac{305}{223}$	$\frac{872}{1021}$
0+	$\frac{196}{646}$	$\frac{110}{209}$	$\frac{57}{87}$	$\frac{21}{16}$	$\frac{384}{958}$
col-wise	$\frac{361}{929}$	$\frac{392}{610}$	$\frac{472}{435}$	$\frac{1000}{620}$	$\frac{2225}{2594}$

C. FROM REFLECTIONS IN AGE = 5 TO EVOLUTION OF $\mathcal{Y} = (STK, HD)$

The sub-population Age = 5 consists of 2242 subjects, which is larger the size of Age = 1. However, we encounter very striking phenomenons via our CEDA computations within this sub-populations. The first phenomenon is that the numbers of positive and negative disease risk major 1-feature-categories have suddenly shrunk to 3, as seen in the Age = 5 panel of Fig. 8. The three major 1-feature-categories are: $\{$HighBP $= 1(+)$, HighBP $= 0(-)$, HighChol $= 1(-)\}$. Even more dramatic is the second phenomenon that no major 2-feature-categories or 3-feature-categories could be found and confirmed. That is, all behavioral risk factors can offer Extra-Info to these three major 1-feature-categories, even {HighBP} and {HighChol} can not mutually offer Extra-Info to each other, like what they do in Age = 1, 3 and 4. In other words, beside these two risk factors, the remaining 15 behavioral risk factors do not matter for the complex disease dynamics of $\mathcal{Y} = (STK, HD)$.

This is a very strange and striking phenomenon observed in Age = 5. We collectively term this phenomena in Age = 5 an "Information Break-down". This phenomenon is worth further looking into from information evolution perspective of complex disease dynamics along the age-axis.

Finally, we make several concluding remarks on the evolution of complex disease dynamics of $\mathcal{Y} = (STK, HD)$ along the age-axis under the constant GenHL = 5 category. As demonstrated through the series of heatmaps of individual risk-landscape, the spectrum of major 3-feature-categories of positive disease risk is expanded from Age = 1 to Age = 3, then shrinks toward Age = 4 and completely disappears in Age = 5, while the spectrum of major 3-feature-categories of negative disease risk is greatly expanded from Age = 1 to Age = 4, then suddenly disappears in Age = 5. However, the contents of individual risk-landscape based patterns found within the three heatmaps, like the presence of "atypical subjects", are more or less constant from Age = 1 to Age = 4.

Further, the disease risk evaluations evolve rather distinctively from sub-population (GenHL, Age) = (5,1) to (GenHL, Age) = (5,4). Since the series of odds tables pertaining to ($Syn[L1]$, $Syn[R1]$) evolves rather drastically. In fact, the information content within the series of sub-populations from (GenHL, Age) = (5,1) to (GenHL, Age) = (5,4) changes dramatically. For instance, in comparison of the three tables of $Syn[L1] - vs - Syn[R1]$ across three age-categories: Table 4 for Age = 1, Table 5 for Age = 3 and Table 6 for Age = 4, we first look at the cell with the highest risk: $(Syn[L1], Syn[R1]) = (4, 0)$. For Age = 1, the cell's sub-population specific odds-ratio is calculated as $\frac{191}{255}/\frac{404}{1372} = 2.5437$. For Age = 3, the cell's sub-population specific odds-ratio is calculated as $\frac{605}{698}/\frac{1135}{2109} = 1.6105$. Correspondingly, for Age = 4, the the cell's sub-population specific odds-ratio is calculated as $\frac{674}{381}/\frac{2225}{2594} = 2.0624$. Thus, from the highest disease risk aspect, we see very non-linear evolution of complex disease dynamics of $\mathcal{Y} = (STK, HD)$. Non-linearity is seen along other evolving patterns for the lowest disease risk as well, among many others.

VI. CONCLUSION

We have refined our CEDA algorithmic computing paradigm, beginning with the recognition of the "element of information" inherent in categorical data points and confirming the 1D histogram of any feature-category as its simplest form of "a piece of explainable information". Subsequently, we employ the conditional entropy of the response variable given a covariate feature-category to unravel directional associative relationships of locality nature. The CEDA algorithm then utilizes newly developed algorithms based on de-associating operations and the concept of Kolmogorov's randomness-proper to identify and confirm major feature-factor-categories from order-1 to higher orders.

We adopt a heatmap platform as a graphic display of a binary bipartite network matrix recording all presence-absence memberships of major feature-categories of high order. Subjects are arranged along the row-axis, while major feature-categories are arranged along the column-axis, respectively. Each binary subject-specific row vector not only provides a dissimilarity measure for its neighborhood but also facilitates a highly interpretable individual risk-landscape. The significance of subjects' individual risk-landscape extends beyond the scope of our discussion, influencing real-world issues such as optimal selection and causality studies.

Moreover, this heatmap undergoes rearrangement operations by superimposing its row- and column-axes with two hierarchical clustering trees based on simple and natural Euclidean distances, respectively. The resulting block-pattern sustained heatmap serves as a graphic display, illustrating a topological space annotated with memberships of response-categories (or labels). Each block represents a relational construct, showcasing a group of similar subjects against a cluster of major feature-categories of order-3. Consequently, any cluster of subjects can be precisely described by its corresponding horizontal series of blocks. This block-series collectively reveals intricate relational information of the response-variable $\mathcal{Y}$ =(STK, HD). Conversely, any cluster of high-order feature-categories unveils complex structural dependency among all involving feature-variables. Such complexity in structural dependency remains unexplored in the literature.

Additionally, this heatmap-based individual risk-landscape topology highlights typical subjects versus ‘‘atypical subjects’’ within each response-category. The contrasting presence of typical versus atypical subjects across distinct response-categories underscores the importance of computing and displaying pattern information as the primary objective of data analysis. This elucidates why errors may occur in predictive approaches within the fields of ML and Statistics.

Throughout this paper, we employ CEDA to address the extremely important and fundamental data analysis issue: What is the information content in data (ICiD)? We propose a primary approach to answer this question: A heatmap of individual risk-landscape of high orders, which serves as the chief component of sub-population specific ICiD in this paper. We also endeavor to synthesize all findings concerning the four sub-populations of GenHl=5 and Age=1, 3, 4, and 5 to gain a true understanding of the joint disease dynamics of multiple chronic diseases. With such results in hand, we are confident that our computational approach is a critical method for studying the BRFSS as a complex system and its evolution over many years.

Although this paper focuses on discussions within a categorical data world, the entire computational framework is applicable to all structured databases. Any database represented in a matrix format inherently contains a categorical data world. Specifically, any quantitative variable can be categorized through its histogram as an approximately sufficient statistic. Their joint high-dimensional histogram would retain almost all essential information content in data (ICiD). Thus, by accepting a slight amount of information loss when relinquishing ‘‘smoothness’’, the gains from applying CEDA are tremendous from many perspectives. The foremost perspective is the explicit interpretability of ICiD, which is completely free from all man-made structures and assumptions. Therefore, all CEDA results are authentic. This fact leads to another essential perspective in scientific data analysis: Unlike symmetry-based correlation, which may provide a distorted marginal version of associative information, our directional associative patterns are of a local nature. These patterns are visible, explicit, realistic, intuitive, and most importantly explainable.

In conclusion, all heatmaps presented in this paper explicitly underscore the central role of classification in the study of complex systems. The scientific value of a complex system lies in comprehensive explanations of its dynamic nature, which is expressed through all study subjects. Therefore, any classification task must be realistically approached by revealing and showcasing intrinsic information related to each individual study subject. In this manner, we convincingly demonstrate in this paper that our CEDA paradigm can effectively explore complex systems and uncover their dynamics captured in ICiD. Furthermore, we illustrate that CEDA is capable of accommodating highly complex response variables and relatively small sample sizes.

ACKNOWLEDGMENT

The authors would like to thank Kaggle for making this cleaned version of 2015 BRFSS dataset available for their research.

REFERENCES

[1] C. Pierannunzi, S. S. Hu, and L. Balluz, ‘‘A systematic review of publications assessing reliability and validity of the behavioral risk factor surveillance system (BRFSS), 2004–2011,’’ *BMC Med. Res. Methodol.*, vol. 13, no. 1, p. 49, Dec. 2013.

[2] D. E. Nelson, E. Powell-Griner, M. Town, and M. G. Kovar, ‘‘A comparison of national estimates from the national health interview survey and the behavioral risk factor surveillance system,’’ *Amer. J. Public Health*, vol. 93, no. 8, pp. 1335–1341, Aug. 2003.

[3] A. H. Mokdad, D. F. Stroup, and W. H. Giles, ‘‘Public health surveillance for behavioral risk factors in a changing environment: Recommendations from the Behavioral Risk Factor Surveillance team,’’ *MMWR Recomm Rep.*, vol. 52, no. 9, pp. 1–12, 2003.

[4] K. Turner and D. Wolpert, *Collectives and the Design of Complex Systems*. Cham, Switzerland: Springer, 2004.

[5] M. Gell-Mann, ‘‘What is complexity?’’ *Complexity*, vol. 1, pp. 16–19, Dec. 1995.

[6] C. Adami, ‘‘What is complexity?’’ *BioEssays*, vol. 24, no. 12, pp. 1085–1094, 2002.

[7] P. W. Anderson, ‘‘More is different: Broken symmetry and the nature of the hierarchical structure of science,’’ *Science*, vol. 177, pp. 393–396, Aug. 1972.

[8] H. Fushing, E. P. Chou, and T.-L. Chen, ‘‘Multiscale major factor selections for complex system data with structural dependency and heterogeneity,’’ *Phys. A, Stat. Mech. Appl.*, vol. 630, Nov. 2023, Art. no. 129227.

[9] A. N. Kolmogorov, ‘‘On logical foundations of probability theory,’’ in *Probability Theory and Mathematical Statistics* (Lecture Notes in Mathematics), vol. 1021, J. V. Prokhorov and K. Itô, Eds. Berlin, Germany: Springer, 1983, doi: 10.1007/BFb0072897.

[10] T.-L. Chen, E. P. Chou, and H. Fushing, ‘‘Categorical nature of major factor selection via information theoretic measurements,’’ *Entropy*, vol. 23, no. 12, p. 1684, Dec. 2021.

[11] E. P. Chou, T.-L. Chen, and H. Fushing, ‘‘Unraveling hidden major factors by breaking heterogeneity into homogeneous parts within many-system problems,’’ *Entropy*, vol. 24, no. 2, p. 170, Jan. 2022.

[12] Chen, T-L., Fushing Hsieh and Chou, E, p. 2022, ‘‘Learned practical guidelines for evaluating Conditional Entropy and Mutual Information in discovering major factors of response-vs-covariate dynamics,’’ *Entropy*, vol. 24, no. 10, 1382.

[13] F. Provost, ‘‘Machine learning from imbalanced data sets 101,’’ in *Proc. AAAI Workshop Imbalanced Data Sets*, vol. 68, no. 2000. AAAI Press, 2000.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, ‘‘SMOTE: Synthetic minority over-sampling technique,’’ *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[15] H. He, Y. Bai, E. A. Garcia, and S. Li, ‘‘ADASYN: Adaptive synthetic sampling approach for imbalanced learning,’’ in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 1322–1328.

[16] J. W. Tukey, ‘‘The future of data analysis,’’ *Ann. Math. Statist.*, vol. 33, pp. 1–67, Mar. 1962.

[17] H. Fushing and T. Roy, ‘‘Complexity of possibly gapped histogram and analysis of histogram,’’ *Roy. Soc. Open Sci.*, vol. 5, no. 2, Feb. 2018, Art. no. 171026.

[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.

[19] C. Chen and H. Fushing, "Multiscale community geometry in a network and its application," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 86, no. 4, Oct. 2012, Art. no. 041120.

[20] H. Fushing and C. Chen, "Data mechanics and coupling geometry on binary bipartite networks," *PLoS ONE*, vol. 9, no. 8, Aug. 2014, Art. no. e106154.

[21] F. Hsieh, E. P. Chou, and T.-L. Chen, "Mimicking complexity of structured data Matrix's information content: Categorical exploratory data analysis," *Entropy*, vol. 23, no. 5, p. 594, May 2021.

HSIEH FUSHING received the Ph.D. degree in statistics from Cornell University, Ithaca, NY, USA, in 1990. He is currently a Professor of statistics with the University of California at Davis, Davis, CA, USA. He develops conditional entropy and mutual information based on categorical exploratory data analysis (CEDA) and major factor selection as foundations for analyzing data from complex systems with or without rhythm.

HONG-WEI KAO received the B.S. degree in statistics from National Chengchi University, Taipei, Taiwan, in 2022, where he is currently pursuing the master's degree in statistics with the Research Laboratory under the supervision of Dr. Elizabeth P. Chou.

During the second semester of his first year of graduate studies, he was a Research Assistant for the industry-academia collaboration project. In the first semester of his second year, he assumed the role of a Teaching Assistant for R programming courses. His research interest includes metric-free categorical data analysis.

ELIZABETH P. CHOU received the B.A. degree from National Chengchi University, Taipei, Taiwan, in 2007, the M.A. degree from Columbia University, USA, in 2008, and the Ph.D. degree from UC Davis, USA, in 2014. Currently, she is an Associate Professor with National Chengchi University. Her main research interests include statistics and machine learning.