NETWORK SCIENCE

Optimal information networks: Application for data-driven integrated health in populations

Joseph L. Servadio[1] and Matteo Convertino[2,3,4]*

Development of composite indicators for integrated health in populations typically relies on a priori assumptions rather than model-free, data-driven evidence. Traditional variable selection processes tend not to consider relatedness and redundancy among variables, instead considering only individual correlations. In addition, a unified method for assessing integrated health statuses of populations is lacking, making systematic comparison among populations impossible. We propose the use of maximum entropy networks (MENets) that use transfer entropy to assess interrelatedness among selected variables considered for inclusion in a composite indicator. We also define optimal information networks (OINs) that are scale-invariant MENets, which use the information in constructed networks for optimal decision-making. Health outcome data from multiple cities in the United States are applied to this method to create a systemic health indicator, representing integrated health in a city.

INTRODUCTION

Multi-variable health index

Creating composite indicators from multiple variables holds the appeal of summarizing large quantities of information into a single numeric value. Rather than individually comparing multiple effects that may be similar in nature and possibly caused by similar or different factors, creating a single summary indicator is a concise way to draw conclusions that may consider the interrelatedness among the component variables. These summary indicators can be used to draw comparisons across different populations such as those living in different cities or countries or at any spatial and temporal scale of interest (*1*). The use of composite indicators can be applied to a variety of complex systems, such as the business cycle (*1*), ocean sustainability (*2*), and human health (*3*–*6*). However, in many cases, only some variables are selected to represent the complex system under analysis based on assumed hypotheses or preferences; it is rare to find composite indicators formed without any bias on the constituting variables. These indicators can be characterized by a probability distribution, and the forming variables may have different levels of importance from one place to another or from one period to another, leading to different health-based city design and population health strategies. The reliability of the indicator depends on the reliability of data, although the methods themselves do not.

Creating these indicators aids comparison of multiple entities (such as cities, population groups, or individuals) on various spatial and temporal scales because it creates a baseline range of values for the systemic indicator using the same method. Within the United States, comparing different cities is often of interest. Every year, different rankings of cities are formed to compare them according to their health but lack a unified definition of "health" and a shared method. Consequently, different criteria are used by different groups and very different rankings are formulated. This situation is often encountered in many other areas, including rankings of colleges, cars, and businesses. For cities, with wide diversity in demographics, economies, and organizational structures, drawing comparisons of outcomes is relevant for assessing efficacy of practices.

Within the scope of health outcomes, indicators are typically made using a variety of methods for variable selection and inclusion. Measures that focus on a more specific health field, such as cardiovascular disease, often include variables that are recommended by clinicians specializing in the topic of interest (*3*). Others are selected through statistical measures, including statistical significance in univariate models (*4*) or with significance in addition to a certain minimum magnitude of effect measure (*5*). Other methods use other statistical measures, such as Cronbach α coefficients (*6*), which are relatively popular in the public health community.

Although many works exist that create numeric indicators focused on particular facets of health, fewer scientifically developed indicators exist to measure the broad state of human health in a geographical area. Here, we aim to create a data-driven indicator that reflects the systemic health status of a population rather than a single-effect metric. This reflects the view of complexity science applied to health sciences, considering, for instance, the work on syndemics (*7*, *8*), exposomics (*9*), and complex genetically communicable diseases (*10*). Little work has been done to assess how an indicator can be developed that combines distinct outcomes to characterize the systemic health quality of a community. Some work has been done by Barabási *et al.* (*10*) to determine the human diseasome, the set of genetically related diseases, but nothing similar has been done at the population scale beyond using genetic information.

Characterizing overall health, rather than targeting more specific conditions, can be advantageous when a more comprehensive assessment of health, as well as an understanding of the role of the environment in health, is desired. Methods for creating such a metric are uncommon in scientific literature, especially in the public health literature where any systemic epidemiology purview is lacking (*11*). More broadly, this can provide an assessment of overall well-being, because human health is one important factor in measuring well-being.

Comparing cities or states in the United States is a motivator for creating an indicator for systemic health. Characterizing cities or states by their relative successes and failures regarding the quality of life of their inhabitants provides insight into possible improvements in these cities and across the nation. By identifying cities or states that have better overall outcomes, characteristics of those successful locations

[1]Division of Environmental Health Sciences, HumNat Lab, University of Minnesota School of Public Health, Minneapolis, MN 55455, USA. [2]Complexity Group, Information Communication Networks Lab, Division of Frontier Science and Media and Network Technologies, Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan. [3]Global Institution for Collaborative Research and Education (Gi-CoRE) Station for Big Data and Cybersecurity, Hokkaido University, Sapporo, Japan. [4]Department of Electronics and Information Engineering, Faculty of Engineering, Hokkaido University, Sapporo, Japan.
*Corresponding author. Email: matteo@ist.hokudai.ac.jp

can be assessed to see whether they can be applied to other locations to improve their own systemic health. Improvements in health are also related to improvements in the built environment; that is why it is also important to assess the robust network dependencies between observed disease and infrastructure that can be managed and designed (*12*). The built environment may affect multiple facets of human health, or it may affect areas of health that are not hypothesized a priori by known mechanisms. This is important to understand the differential effects on populations that both advance epidemiological knowledge and guide population-based personalized medicine.

Traditional and proposed methods for designing indicators

Current methods of selecting variables for inclusion in composite indicators can raise some concerns in relation to the deductive, reductionist, and hypothesis-driven design. Considering only a priori knowledge ignores data-driven techniques and information; moreover, this approach relies only on the expertise of those who are developing the indicator or on the existence of adequate relevant literature. Although this knowledge is important and should not be ignored in the process, including data-driven methods in an inductive or abductive design process is advantageous. It is certainly not recommended to only use data-driven methods without considering contextual relevance of considered variables (and without a posteriori validation of the inferred results). In this regard, care should be placed into the interpretation of an inferred correlation/causation because much of the prediction does not imply causality. It should also be noted that predictions are useful for guiding changes in population without the necessity to know all fine-scale causal processes. This is the purview in which this paper should be seen.

A data-driven technique to incorporate into the creation of an indicator should consider the possibility of correlations among variables. Relying on univariate models for selection neglects the possibility of strong correlations among included variables, which are actually fundamental in determining complex system patterns (*13*). As a result, the composite indicator may not serve its purpose as accurately as intended. Including variables that are highly correlated using classical statistical models can lead to an unintended overrepresentation of certain pieces of information. Pairing a data-driven method that can account for possible connections among variables with substantive knowledge will provide information from multiple sources to aid the decision of variable inclusion. These methods, such as the ones presented, include variable interdependence in a proper way for predicting robustly the systemic indicator variability.

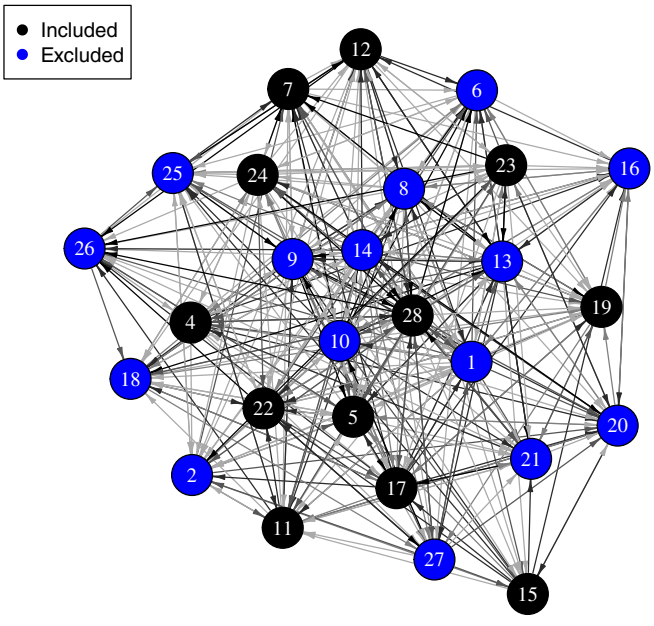

Fig. 1. MENet for all variables. The OIN is the MENet with the highest information content for predicting the integrated health (IH) indicator. Inclusion and exclusion indicated in the legend refers to inclusion in creating IH indicator. Variables identified by numeric labels are defined in Table 1. Darkness of edges represents amount of TE, with darker edges projecting greater amounts of TE. This network is directed due to the use of TE.

Here, we propose an alternative method of variable selection and design of systemic indicators. The method infers a network of variables to use for the creation of a systemic health indicator by using a maximum entropy (MaxEnt) network (MENet). This method is used frequently in the fields of complexity and information sciences. The systemic indicator is created by using a decision theoretic model based on the MaxEnt-inferred network. Increased use of complexity science in health research is a newer concept that is being encouraged (*14*) for its ability to capture systems' interdependencies and for its abductive framework aimed also to simplify systems' complexity.

Although new to research in human health, the concept of MaxEnt is not new. It has previously been applied to fields such as image processing (*15*), hydrogeomorphology (*16*), distribution of animal species (*17*), and species interactions (*18*). Methodological research on this topic has also considered the dynamics of systems that are not in equilibrium (*19*), thus expanding its application to nonstationary systems. Because of its usefulness in these applications, there is interest in introducing MaxEnt to population health sciences. The methods proposed here are likely to be of interest to several applications in public health, because such comparisons across various regions can provide valuable information to public health professionals.

MENets with transfer entropy

MENets, based on the principle of MaxEnt (*17*, *20*, *21*), can be created using various forms of entropy as the systemic variable upon which to establish a threshold of significance for the considered problem. One type of entropy is transfer entropy (TE) (*20*). TE is a measure of directional dependence of one variable on another. It is derived using the estimated probability distributions of two variables, using both joint and conditional probabilities (*20*). It carries the advantages of determining dependencies of two variables without assuming a particular structure of the dependence (*22*). Its use in determining association network structure has been seen previously (*23*).

Putting the TEs of the various pairs of variables into a network structure facilitates communication of the connections between the variables. The MaxEnt model favors distributions with minimal assumptions that fit the available data by representing them as the distributions with the greatest entropy (leaving aside any subjectively driven observational methodologies) (*24*). That is, the MaxEnt model selects the simplest but most informative probability distribution function to represent a variable of interest such as the systemic health indicator, as it is the case in this paper. By using MENet, the network structure of variable relatedness with the highest total entropy can be considered the most plausible.

From this selected network, the characteristics of the variables within the network, such as the ability to represent the entire group of variables, can be examined to determine inclusion into an indicator of systemic health status. Thus, MENets can be considered optimal information

networks (OINs) that are functional networks derived from data defining the interdependence of different variables based on their directional (and maximized in a MaxEnt sense) mutual information. In particular, MENet is the simplest and most informative network to predict the designed systemic indicator. Concepts similar to OINs, which more precisely should be considered as the MENets in the scale-invariant regime, have been developed in other fields such as geomorphology, where optimal channel networks (*25*) are the MaxEnt and minimum spanning networks describing the observed river networks in the specified domain. In our context, OINs describe the observed variability of IH over the landscape accounting for all information variables contributing to it. Note that these concepts can apply to either functional networks derived from data (such as in our case) or structural networks that are visible over the landscape considered.

This paper is intended to demonstrate how MENets can be used in the variable selection process (also known as "metamodeling" in the field of model creation) (*26*) for creating composite indicators with basic science and practical purposes to understand and predict the systemic response of a complex system. The constructed directed network contains the variables of interest as nodes and the TEs as the weighted

Table 1. Variables included in the network shown in Figs. 1 and 5.

Node number	Variable name	In-degree	Out-degree	Weighted in-degree	Weighted out-degree
1	Adult binge drinking	14	24	3.333	1
2	Adult obesity	18	6	2.667	2
3	Adult physical activity levels	0	0	0	0
4	Adult seasonal flu vaccine	11	10	1.333	3.667
5	Adult smoking	17	18	1.333	6.333
6	AIDS diagnosis rate	16	10	4.333	0
7	All types of cancer mortality rate	13	14	0.667	6.333
8	Asthma emergency department visit rate	14	22	2.333	1.333
9	Child seasonal flu vaccine	11	26	4	0
10	Children's blood lead levels	11	26	2.333	0
11	Death rate (overall)	15	13	2.667	2.667
12	Diabetes mortality rate	13	13	1.667	4.667
13	*Escherichia coli* infections	13	22	2.333	1.333
14	Firearm-related emergency department visit rate	11	25	3.333	0.667
15	Firearm-related mortality rate	9	9	0	3.667
16	Heart disease mortality rate	16	7	3	1
17	HIV-related mortality rate	15	15	2.333	6
18	HIV diagnosis rate	14	10	1.333	0.333
19	Infant mortality rate	17	9	1	3.333
20	Life expectancy	18	8	5	0
21	Low–birth weight babies	17	8	3	1.333
22	Lung cancer mortality rate	15	17	1.667	4
23	Opioid-related overdose mortality rate	13	16	0.333	5
24	Persons living with HIV/AIDS rate	13	14	1.667	3
25	Pneumonia and influenza mortality rate	17	8	3.333	1
26	Pneumonia vaccine (age 65+)	16	4	3.667	0.333
27	Salmonella infections	14	9	2.667	1
28	Suicide rate	16	24	3	4.333
29	Teen binge drinking	0	0	0	0

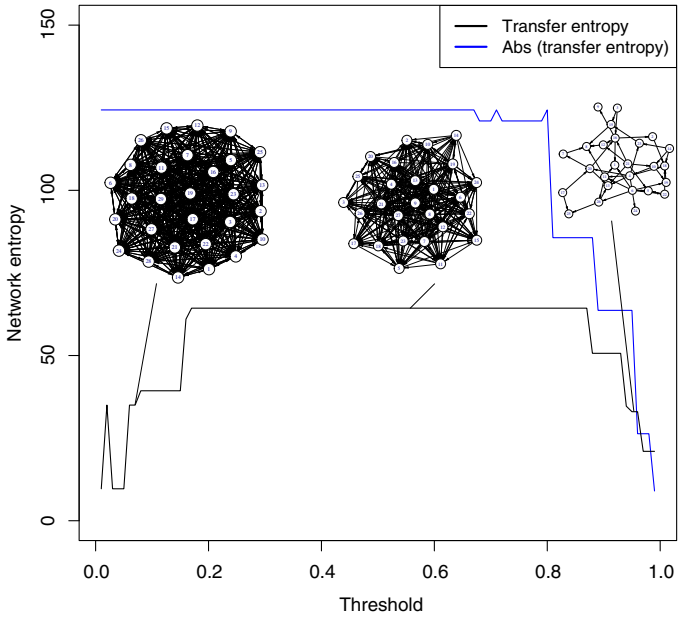

Fig. 2. Scale invariance analysis of MENets. The thresholds for edge inclusion are the value of the total TE in the network considering the sign or absolute values. Example networks at three thresholds are shown. For low and high thresholds, the network changes topology and the MCDA model provides a lower amount of information than the scale-invariant regime where the information entropy is the highest. This may lead to overfitting and underfitting when variables are too many or too few to describe the systemic indicator.

edges. That is, each edge weight is an information flux representing how much variability of one variable can explain the variability of another with an associated time delay, because TE is reliant on time series data and reflects characteristics of the time delay creating the history of a variable. The use of TE, a directed measurement, necessitates the construction of a directed network that allows vertices in both directions to exist between each pair of nodes.

Variables represented in this network will then be compared based on their ability to explain the group of variables while also considering the extent to which they are explained by the other variables in the network. The result is a composite indicator that is composed of variables selected through a data-driven process and integrated together via a multi-criteria decision analytical (MCDA) model, which refers broadly to a collection of methods that intake multiple pieces of information when making decisions (*27*). This data integration is done with less emphasis on a priori beliefs concerning variables' substantive significance or connections to a particular outcome, although beliefs can be included as weights of variables in the MCDA model. The selected variables will be the ones that are most explanatory of the aggregate collection. In a public health context, the proposed method reflects the shift in focus toward methods that account for connections between total exposure ("exposome") (*28*) and outcomes or the intrinsic relationships within each of them that has a more biological interest. The following innovative points introduced in the paper are worth mentioning:

(1) The combination of MCDA and TE-based models to create a systemic indicator where value functions are TEs used to assess variable importance; note that this puts a lot of emphasis on functional variable interdependencies rather than on independent variable importance as it should be for any complex system where interconnectedness is largely driving system's outcomes.

(2) The use of an information theoretic global sensitivity and uncertainty analysis (GSUA), where TE (measured via the Karskov estimator that eliminates biases related to binning choices), rather than mutual information, is used to better characterize the second-order importance indices of variable for the predicted systemic indicator.

(3) The inference of OINs based on a thresholding criterion that considers the maximum information (within an information-invariant range) and the least complex functional networks of variables constructed by using the TE; typical network selection methods make use of topological criteria, such as the node degree, to select networks and without considering how the overall information content is changing.

RESULTS AND DISCUSSION

Here, we introduce the concept of IH as a systemic health indicator of cumulated health effects over space and time, hence the use of the word "integrated." This concept is based on the definition of MENets that define how the integrated variables are interdependent with each other. Within these MENets, we define OINs that are minimum spanning functional networks useful for predicting the whole variability of the integrated variable. OINs are derived after a topological selection method that neglects redundant nodes and links and preserves the MaxEnt criteria. OINs maximize information flow (function) by preserving the topological structure of the network and selecting one integrated variable to predict (that is, a system service) (*29*, *30*).

Making a financial analogy, IH can be thought of as a portfolio indicator (say a "payoff") resulting from the potential investment on the assemblage of available stocks. MaxEnt solutions are the ones that maximize the payoff conditional to a cost function. In this situation, the payoff is the total TE of the network, because the intent was to create a MENet. The cost in this situation is based on feasibility constraints; values of the input data that are not possible constrain the values of the edge weights in the network. The cost function can alter the cost of the investment; one wants to have the minimum cost for the highest investment. This is analogous to the topological threshold for MENets that guarantees the highest accuracy in information predictability. Equivalently, OINs maximize the average information accuracy while minimizing its variance as a classical Pareto optimal solution.

The constructed MENet, shown in Fig. 1, displays the directions and the strengths of the TE projections. The directions of the edges in the graph are based on the directions of the TE projections. Establishing a threshold of TE for edge inclusion in the network relies on ranking the TEs and comparing the total network entropy with the cutoff for inclusion, represented as a percentile. The highest network entropy was a constant value for TE thresholds ranging from the 17th percentile through the 87th percentile (Fig. 2). Near the middle of this range, at the 53rd percentile, lies the cutoff to only include edges in the network that represent TEs strictly greater than zero. Before this cutoff, the included edges from selecting cutoffs between the 17th and 53rd percentiles included edges representing very small values of TE. By selecting this threshold at the 53rd percentile, the network contains a total of 387 edges among the 29 nodes. Two of the nodes, denoted with the numbers 3 and 29, are not present in the figure because they did not connect to any other nodes at this threshold. These nodes may not be highly influential for a population.

The temporal stability of the network relates to the stability of the variables contributing to it. Because the variables used here include many that do not change quickly, such as mortality rates and disease incidence rates, the structure of the network remains stable for long

periods of time. This contrasts with short-term variables with seasonal patterns such as influenza incidence. In situations such as these, there may be interest in observing how the MENet changes over time, either cyclically or acyclically (*31*). For cyclic networks, the predictability of integrated function (such as IH) is generally smaller than for acyclic-directed networks (such as scale-free networks) whose stability increases predictability. In contexts such as influenza outbreaks, MENet may fluctuate seasonally within a year, so a composite indicator used in this context may vary depending on the time of year, possibly following a yearlong cycle.

On the basis of the connectedness of the MENet, it most closely resembles a small-world network rather than a random network, scale-free network, or regular network (*32*). The graph is highly connected, lacking regularity in its shape but certainly being far from other network topologies. Strong clusters or subgraphs in the network with all included nodes do not appear to exist.

Nearly a third of the pairs of variables had TE values of zero. This high number of pairs is likely a result of the short time series. Having four observations in a time series is much less than typically expected from time series data, which may lead to difficulty in establishing TE. Having access to a longer time series with fewer missing data values could lead to more pairs with nonzero TE. Despite this challenge, our results show how this method can be applied while still maintaining the ability to highlight its advantages.

The selection of the threshold for edge inclusion is illustrated in Fig. 2. This threshold applies to Eq. 4 when selecting E_{MENet}. The threshold is a percentile above which edges are included. The figure visually shows which threshold of inclusion produces a network with the greatest total TE. The graph shows various thresholds for edge inclusion and the total network entropy associated with those thresholds. The peak of the curve shows a constant total entropy level for a wide range of thresholds. This is not only a result of the high number of pairs of variables with zero TE but also a sign of the topological invariance of the network across a wide range of information entropy (or network complexity, equivalently). Changing the threshold for those values of variables did not change the entropy cutoff for inclusion. The invariance of the network allows one to draw reliable conclusions about the relationships of different variables. The scale invariance (where scale is here defined by the information content played by the entropy threshold) is a manifestation of the nontrivial relationships between population variables, yet of the nonregular but small-world character of the resulting OIN.

The use of total absolute TE using the absolute value of TE rather than its actual value was also considered. It is not surprising that, as Fig. 2 displays, including all possible edges leads to the greatest absolute value of the total TE, and raising the threshold simply reduces this total. This is not desirable for our purposes, because the purpose of this network is to identify a subset of variables that can be used to predict IH, which highlights the need to consider directionality in the interdependence of data. Thus, this is to show the power of our model that assesses directional dependencies. The direction of dependencies is not relevant for structural networks whose flow (such as water or blood) is known and cumulating from upstream nodes to downstream nodes, although, also in this case, coarse-graining effects for low and high thresholds appear. For low thresholds, "noise" induced by not relevant variables obscures the backbone network, whereas for high thresholds, the backbone structure is lost because relevant links and nodes are removed.

Examining the various thresholds and total network entropies demonstrates the care that should be taken to assure that a threshold is selected that will lead to the highest total entropy in the network. Selecting

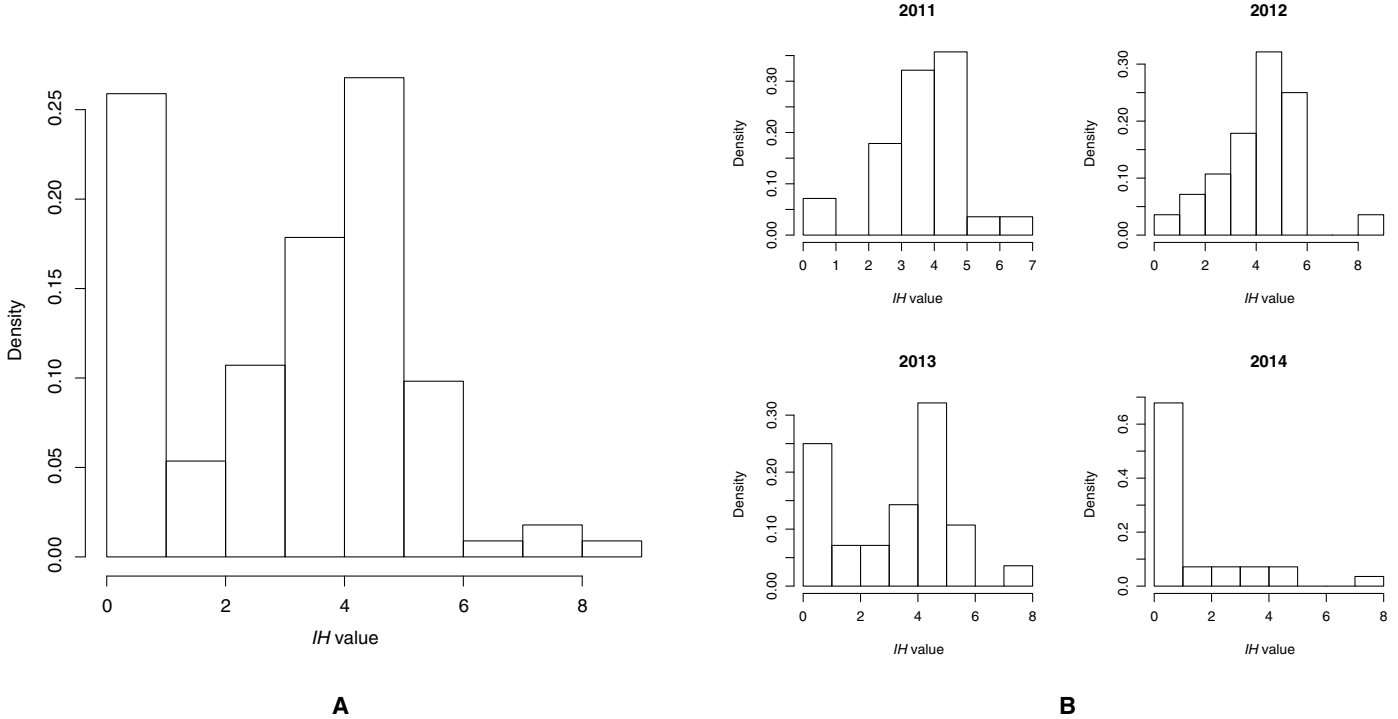

Fig. 3. Probability density functions of IH. (**A**) Among all cities across all years. (**B**) Within individual years of observation. The probability density function (pdf) is bimodal with some years when it is more like a normal distribution (2011 and 2012), bimodal in 2013, and with a long tail in 2014. With more accurate data, it would be interesting to see how the shape of IH changes as a function of some driving factors.

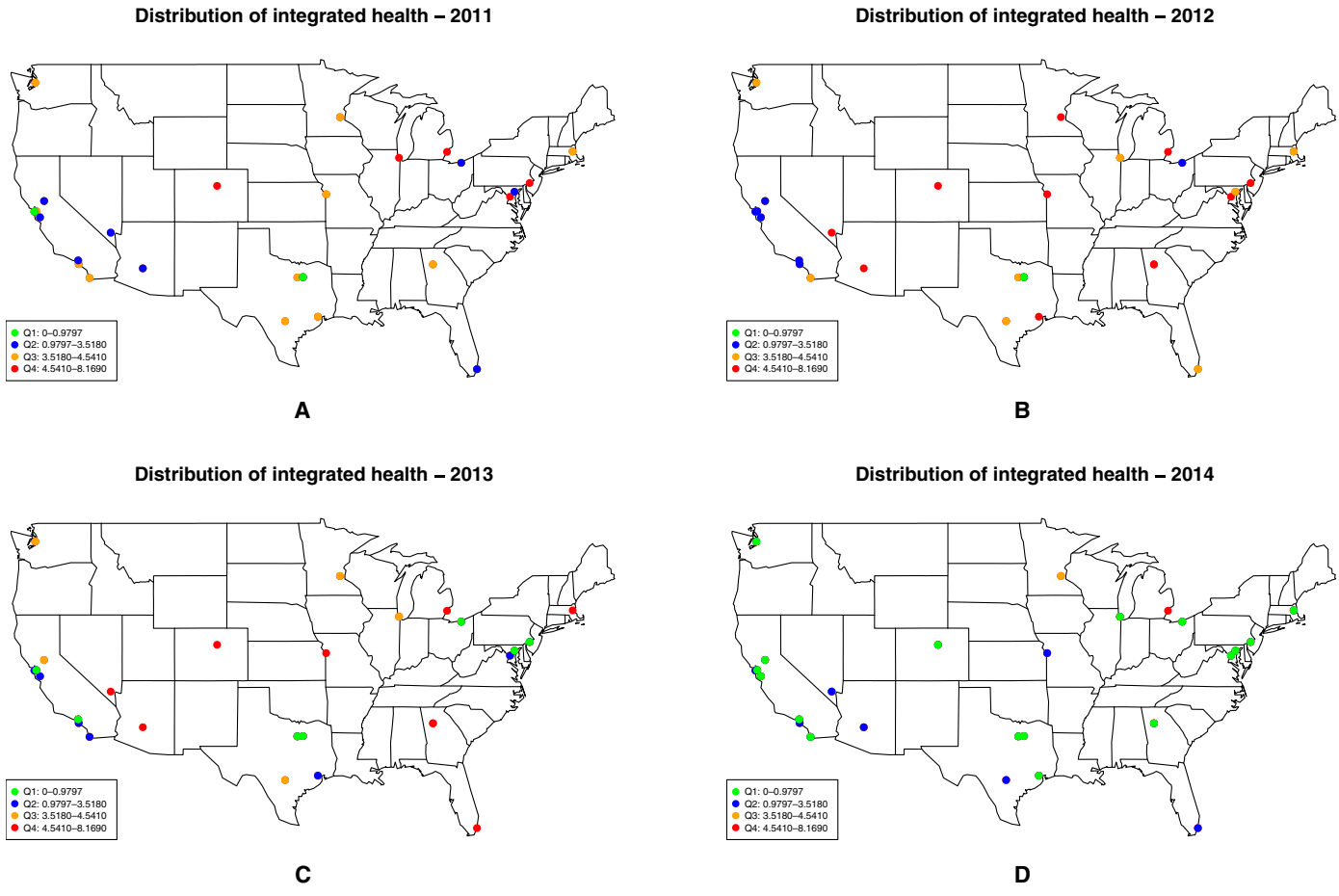

Fig. 4. Patterns of integrated health. (**A**) Distribution for 2011. (**B**) Distribution for 2012. (**C**) Distribution for 2013. (**D**) Distribution for 2014. The dots are for the cities included, and their color is proportional to the quartile they belong to for IH. From green to red, IH increases in value. With the data we have, we observe that IH is initially improving until 2013 and is better in 2014; however, this may be a data quality issue. Spatially, we observe higher IH values for the midwestern cities and the lowest for southwestern cities.

a threshold that leads to the inclusion of too many edges can lead to overfitting after variable selection. These additional variables will not contribute adequate information to the model and will be extraneous. Rather than contributing useful information, they will instead lead to an indicator that overrepresents certain pieces of information but not the fundamental signal. This could lead to a high likelihood of creating a biased measure that is not adequately representative. The signals seen in multiple variables included in the indicator will be inappropriately overrepresented. As a result, there will be bias in the resulting indicator.

Alternately, selecting a threshold that leads to the inclusion of too few variables will leave much of the variability in the system unexplained. The small set of variables will not be sufficient for characterizing the overall health status. Comparing potential thresholds with their associated total network entropies will assure that entropy is maximized in the system, which will result in an indicator that explains adequate variability of the system without overfitting and that is likely representative of the epidemiology of the system considered. Larger ranges of scale invariance lead to more reliable linkages between observed patterns and processes shown by the data. With these thresholds and associated total entropies, Fig. 2 shows examples of three networks. The first is an example of a threshold that is too low. There are a large number of edges in the graph, which may not accurately reflect the causal patterns among the group of variables. The second network contains an edge configuration that leads to the maximum total TE in the network. Variable selection is based on the connectivity pattern that is the most plausible by the measure of TE. The third network has a threshold that is too high, leading to the exclusion of edges that would benefit the overall explanatory ability of the network. Variable selection from this network would omit information.

From these results, the IH indicator is constructed by including the variables that best explained the entire set of variables. As described previously, these variables were selected as those with a weighted in-degree higher than their weighted out-degree. Weighting of the degrees was selected by the magnitude of the TE associated with that edge. As a result, variables that project the greatest amounts of TE onto the other variables without having large amounts of TE projected onto them were included.

Selecting variables by the criterion of having a weighted out-degree greater than weighted in-degree assures that the amount of information contributed by the variables is greater than the amount of information that the other variables contribute for it. The weighted out-degree is a measurement of the amount of information a node (variable) contributes toward explaining the group. The information in this context is represented in the form of TE. The in-degree can be viewed as a measurement of the amount of information that the other variables contribute regarding that variable. By using only the variables that have

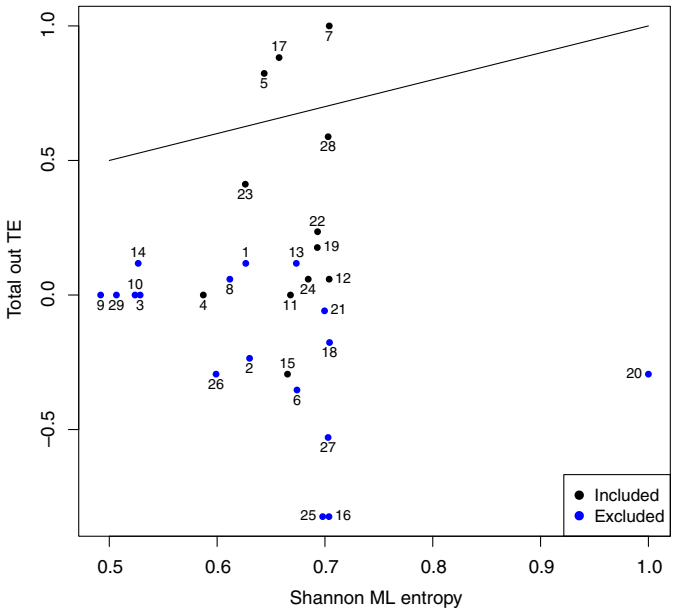

Fig. 5. Information theoretic global sensitivity analysis for variables included in MENets. The total information outflow is how much each node affects the others, whereas the Shannon entropy is the intrinsic variable information content (for example, in predicting IH) (*y* and *x* axes, respectively). Note that the most important variables have high total information outflow and Shannon entropy. The continuous line identifies nodes that are predominantly interacting with all others (*5*, *7*, *17*). Variables represented by numeric labels are defined in Table 1.

higher out-degrees, the included variables in the indicator are the ones that contribute the most information regarding the entire collection of variables present in the network.

Of the 33 variables initially considered, 12 were included in the final IH indicator. The distribution of the indicator, shown in Fig. 3, is bimodal, with a lower mode with low variability and a higher mode with high variability. The points separating the two modes, that is, when the first derivative of the pdf changes sign, are transition points between states with different health averages (related to each mode). Some cities, such as Detroit, have high values for all 4 years, indicating worse outcomes. Others, such as San Francisco, have low values for all 4 years, indicating more favorable health outcomes. Other cities showed variability in the time period. Philadelphia, for example, had values of 5.3 and 5.7 in 2011 and 2012, respectively, followed by values less than 1 for 2013 and 2014.

To draw broad comparisons of how this IH measure differs between cities and regions during the 4 years of observation, the values of IH were binned into quartiles and mapped yearly. These maps can be seen in Fig. 4. The maps within each year can be used to compare cities. Because the number of nonmissing variables is sensitive to the year of observation, these maps are best examined as cross-sectional comparisons.

Cross-sectional comparisons within years show tendencies for cities in California to perform better than other parts of the country. Within years, midwestern cities including Denver, Minneapolis, and Chicago appear to perform poorly compared to other regions of the nation contrarily to what national city rankings show (*33*, *34*). This is also seen in the northeastern part of the country.

Previous work comparing relative health statuses of cities is predominantly generated by media outlets rather than by the scientific community. These rankings of U.S. cities for overall health status tend to rank the west coast, the midwestern cities of Denver and Minneapolis, and the northeast as the healthiest cities. The methodologies are often unclear but indicate reliance on measurements of physical activity opportunities and prevalence of healthier food. These are not the only factors to consider, and not the only ones to assess IH and well-being in a population. Factors such as the incidence of infectious diseases, drug and alcohol use, or mortality rates are not commonly mentioned (*33*, *34*).

Figure 5 shows how the Shannon entropies of the variables compare to the TEs. The figure shows that the values of Shannon entropy and TE appear not to relate to each other. More specifically, IH seems to be driven by the interactions among variables and only one variable ("20," that is, life expectancy) is highly responsible for explaining the whole variance of IH. This variable is "life expectancy" that is frequently taken as the indicator of city health and well-being for good reasons. Note that this variable is excluded by the OIN because of the topological redundancy criteria. This means that, as expected, many other variables explain its variability and yet the total variability of IH. Figure 1 shows how node 20 has a very high out-degree. Thus, to predict IH, one can consider life expectancy by itself or considering all other dependent variables. This shows the double importance of MENets to both highlight systems' drivers and formulate simple models for making accurate prediction of patterns of interest.

By using the exposed method for variable selection, the process becomes more data-driven rather than being solely motivated by a priori knowledge. This method reduces the subjectivity involved, although that should not be ignored entirely and can be incorporated in the MCDA model in the form of subjective information or weights. A data-driven approach carries the advantage of being broadly applicable to numerous fields. Although the present network and composite indicator are used for public health purposes, the method used can be applied to create a composite indicator for any application. Once data are acquired to answer a question of interest, and the variables collected are initially screened for broad applicability, this method can be applied in the same way to determine the inclusion of variables into the indicator. Extensive substantive knowledge is not required, but the application of findings requires collaboration with topic experts and managers.

Expansions of this method can incorporate locations and individual patterns of different cities. Rather than creating one network to represent all cities, each city can be represented in its own network and intersect the population outcome networks (or "diseasome" at the population scale) with infrastructure networks. This would require data that have fewer missing values than those seen in the Big Cities Health Inventory (BCHI). These city-level networks can then be connected by geography to incorporate potential spatial effects. Different ways of formulating this network can lead to different implications regarding the information that the created indicator can convey.

This work can be expanded to other applications within public health. As stated previously, there is a notable advantage to using this method, and expanding the contexts of its use is beneficial. In the context of environmental justice, multiple factors can be considered that contribute to this notion. By establishing MENet to connect multiple factors relating to environmental justice, certain populations can be identified that are more or less likely to be exposed to environmental hazards and be susceptible to their associated health outcomes. Factors examined can include air pollution exposure, water quality, food security, or other local environmental factors of interest. On the basis of available data, these conclusions can be drawn at the community or neighborhood level.

CONCLUSIONS

This study provides a general modeling framework for defining integrated indicators reflecting systemic health in defined geographical areas such as cities. Our model is a robust framework for defining integrated functions; however, the physical interpretation of results must always take into account the quality of data as well as stakeholders' preferences for any criteria used in forming the systemic indicator. Here, the following items are worth mentioning.

(1) The definition of MENet and the inference of the OIN as the MENets that explain the spatiotemporal variability of a multi-criteria IH function. These networks define the robust (scale-invariant across an information gradient) interdependencies of variables, such as the ones between population health outcomes in cities. The formulation of IH allows stakeholders to define a baseline of health in cities and to compare different cities considering the same indicators versus current ranking systems.

(2) The analysis of the scale invariance of IH based on the information content related to the data available. This allows one to determine the range of the largest information available as well as the reliability of data across different cities, potentially. The entropy threshold is the uncertainty (or better complexity of information) related to the network of population outcomes, which can be tested for different cities; the variability of the network may manifest differences in data quality or differences in population outcome importance.

(3) The detection of subnetworks of highly interdependent variables for explaining the majority of variability of IH via GSUA. In addition, the topological out-degree network simplification allows one to reduce the complexity of the network, to reduce the redundancy of metrics forming an indicator (which can lead to overfitting), and to detect highly important and interdependent nodes such as life expectancy as shown in our case study. This allows stakeholders to direct public health prevention, environmental management, and city design toward directions that improve population health systemically.

This study considers only population outcome data, which can be thought of as population-scale diseasome information. However, using the proposed methodology, other networks can be built at the population scale such as exposome networks [see, for instance, the study of Patel and Manrai (*35*)] and functional networks such as mobility networks related to city infrastructure. It can be applied to multiplex connections between infrastructure networks and population outcomes, which can be highly important for understanding the epidemiology of complex environmentally communicable diseases and the design of cities to maximize health and well-being. More broadly, the study advances network inference techniques (*36*) and frames those into a decision analytical perspective for making results usable by stakeholders.

MATERIALS AND METHODS

Data source

To show how the proposed method can be used for a city-level health indicator, longitudinal data were needed at the city level for a diverse set of health outcomes. The data used were from the BCHI (*37*), which was composed of data for numerous health outcomes in several major cities in the United States. The BCHI effort is currently an ongoing process, adding and improving data continually. Outcomes measured contained up to 5 years of data between 2010 and 2014. A total of 28 U.S. cities were represented in the data, each having up to 37 health outcomes plus 16 demographic variables. Data were particularly sparse in 2010. Some of the variables were broken down by specific populations, such as adults, children, or elderly populations, and others were similar in nature, such as HIV and AIDS diagnosis rates (*37*). All variables were adjusted to account for population size either as rates per given population size or as percentages of the population.

Although useful for the breadth of variables and organized compilation, this data set has limitations that would make the MENet more challenging to implement for the purpose of creating a city-level health indicator. Many cities have missing data in patterns that were not uniform across years. There exist variables where a particular city may have data for 2012, but not 2011 or 2010. This can be problematic for creating an indicator with multiple variables due to inconsistencies in data availability. Another limitation in the data is that they were aggregated yearly and only available for up to 5 years. TE is typically applied to time series data with more than five time points. Despite these limitations, the diversity of variables and availability of data in multiple cities made this data set well suited to demonstrate the use of MENets with TE for variable selection. Data sets with extended time spans and more complete data are desired for a better suited use of MENet, because this would allow a better causality assessment of the relationships between different population health outcomes. However, the purpose of this paper was to present the MENet methods rather than investigating the epidemiology of "city health."

Transfer entropy

TE is an information theoretic variable that assesses the information flow between pairs of variables, considering both flow magnitude and time delay. TE has advantages over other measurements of information because it is both directed and nonparametric (*20*, *22*, *38*). It was used for establishing potential causal connections between two random variables (*20*, *22*), and it has been shown to perform much better than other traditional methods such as Granger causality methods and convergent cross mapping (*39*, *40*). The computation of TE was based on the distributions of the two variables of interest conditioned on their histories and the history of the other variable (*20*, *38*, *41*). Comparing the conditional probability of the variable on its own history with the conditional probability of the variable on both its own history and the history of a predictor variable provides asymmetry in determining predictive abilities of one variable onto another. Directed TE of two time series variables, denoted as X_i and X_j, was calculated as

$$T_{X_i \to X_j} = \sum p(X_{j,t}, X_{j,\tau}, X_{i,\tau}) \cdot \log_2 \left(\frac{p(X_{j,t} | X_{j,\tau}, X_{i,\tau})}{p(X_{j,t} | X_{j,\tau})} \right) \quad (1)$$

where $X_{i,\tau}$ and $X_{j,\tau}$ denote the respective histories of X_i and X_j at time t (*38*, *41*). These methods have been adapted to calculate the TE for continuous variables (*42*). The present paper uses continuous data and therefore uses these adaptations as well as an alternate definition for the calculation of TE using mutual information (*43*). TE in the present paper was calculated as

$$\mathrm{TE}_{X_i \to X_j} = \mathrm{MI}(X_i, X_{i,\tau} \otimes X_{j,\tau}) - \mathrm{MI}(X_i, X_{i,\tau}) \quad (2)$$

where $X_{i,\tau} \otimes X_{j,\tau}$ refers to the joint distribution of the histories of the two variables (*41*–*44*). Mutual information was computed using the k-nearest neighbor approach proposed by Kraskov *et al.* (*43*) and Torbati *et al.* (*45*).

This can be computed by using the TransferEntropy package in R (*45*). Note that this definition of TE assumes that the processes analyzed obey a Markov model; this implies that future states depend only on the current state, not on the events that occurred before it. Most of the time, this is not true, especially for slowly varying processes (such as chronic diseases); however, this constraint can be relaxed by choosing temporal lags that were small to focus on short-term interdependencies not related to long memories of the underlying processes (*46*).

Maximum entropy networks

The MaxEnt theory favors probability distribution functions with MaxEnt as the most general distributions that fit the observed data (*24*, *47*). This theory can be applied to a network structure where edge weights were based on entropy. The network structure with the greatest total entropy can be similarly favored as the most general network structure that fits the observed data. The network structure considers all possible pairs of variables in both directions for predicting a system pattern of interest. The edges that comprise the network with the greatest total TE were then included. Selecting the edges that contribute to the greatest amounts of TE, according to the MaxEnt theory, produces the network that most accurately describes the "causal" patterns among the included variables.

The creation of this network provides the necessary information for constructing a data-driven multi-criteria indicator. Examining the most likely network of potential causal relationships informs variable selection or weighting. Variables that do well to predict others can be used to substitute the variables that they predict or be weighted more heavily for their ability to represent multiple variables. Very similar approaches have been used in finance (*1*), genetics (*48*), and ecology predominantly. In the ecosystems considered by Lezon *et al.* (*48*), a very large network of trophic interaction exists, but the whole network was not necessary to predict many ecosystem patterns such as ecosystem diversity; a subset suffices. This was in line with the philosophical foundation of information theory that seeks to produce the simplest and most accurate models to predict population patterns. The MaxEnt theory has been also used to determine relevant ecosystem species interactions at different scales for single metapopulations to characterize local habitat and dispersal corridors (*49*). Recent works emphasize that the MaxEnt principle provides a bridge between statistical mechanics models for collective behavior in ecological/biological networks and experiments on networks of real ecological and biological systems. Most of this work has focused on capturing the measured correlations among pairs of species and biomarkers. Many of the most interesting phenomena of life are collective, emerging from interactions among many elements, and physicists have long hoped that these collective biological phenomena could be described within the framework of statistical mechanics (*50*). In our context, we extend this vision to public health.

Using MENet in a multi-criteria decision analysis context

Most data in BCHI represent negative health outcomes such as mortality rates or disease diagnosis rates. Others represent positive health outcomes such as percentage of the population compliant with habits such as engaging in adequate physical activity or receiving vaccinations. In this context, the terms "positive" and "negative" refer to the expected influences on human health, not to their classifications as positive or negative real numbers. To make the data more consistent for the purpose of a summary indicator, the positive health indicators were reversed. Rather than representing the percentage of the population that was compliant with positive habits, they were changed to represent the percentage that was not compliant with positive habits. Another variable that needed to be changed was life expectancy. Life expectancy was changed to show how much less each city's life expectancy was compared to the national average in that year. As a result of these changes, larger values represent a worse state of health for all variables.

For initial simplicity, the composite IH measure was defined by summing the various health variables, weighted equally, in an MCDA approach (*51*, *52*). Because the different variables were measured on various scales that were not comparable, all variables were standardized to allow a sensible combination. To standardize the data, the observations of each variable were divided by that variable's maximum observed value, resulting in all data being valued between zero and one, with the exception of life expectancy, which can have values less than zero. This allows equal weighting for all variables initially. The scaled variables represent the level of each variable relative to the city-year with the worst outcome of that variable.

The process of MCDA is broad and can be adapted to a variety of applications (*51*, *52*). It is approached by optimizing a utility function that considers multiple factors. A utility function can be thought here as a value function multiplied by stakeholder preferences, where value functions are the population outcome considered. This optimization is subject to feasibility constraints. The process can be described as maximizing the following expression

$$U[f_{1,1}(X), f_{1,2}(X), \ldots, f_{2,1}(X), \ldots, f_{n,n}(X)], X \in \Omega \quad (3)$$

where Ω represents feasibility restrictions, the $f_{i,j}(X)$ functions represent factors that were considered in the analysis, and $U[\cdot]$ is the utility function that considers these factors (*53*). The function U is defined by the contextual application, as are the f functions that are in the arguments of U. In the context of the present goal of creating an indicator, $f_{i,j}$ were defined as

$$f_{i,j}(X) = \begin{cases} T_{X_i \to X_j}, \{X_i, X_j\} \in E_{\text{MENet}} \\ 0, \{X_i, X_j\} \notin E_{\text{MENet}} \end{cases} \quad (4)$$

where $T_{X_i \to X_j}$ is defined in Eq. 1, $\{X_i, X_j\}$ represents the directed edge connecting X_i to X_j, and E_{MENet} represents the set of directed edges in the network with the maximum total TE. The selection of edges to be included in the network was determined by finding the network with the greatest total TE. In the present context, the function U is defined as the total TE of the network, and it was maximized by selection of the $f_{i,j}$ functions. To the best of our knowledge, this is the first time that TE was framed in a decision analytical model and a network entropy threshold model was used to determine the MENet.

TE was computed in R version 3.3.2 using the TransferEntropy package (*45*, *54*). Each variable has a 4-year time series for each city in the data set. The R function to compute the TE from Eq. 2 assumes that both variables contain a single time series. To transform each variable into a single time series, rather than a time series for each of the cities present, the average value for each variable in each year was calculated using the available data for each variable in each year. This averaging was motivated by differing patterns of missing data among the various cities as well as a desire to create a network representing national trends. As a result, this MENet was designed to reflect the patterns of all cities with available data on average. The results from this network were then applied to the individual cities to compare values of the IH

metric. Depending on the availability of data or other objectives, TE can be calculated differently, for example, by calculating that for different cities separately and averaging those TEs to get national statistics.

The TEs of the yearly averages for each variable were used to construct the MENet by first assessing a threshold for edge inclusion. The threshold represents a percentile of TE, and only the edges representing TE above this threshold were included in the network. In selecting this threshold, two selection methods were considered. The first included the edges with the greatest TEs. The second included the edges with the greatest absolute value of the TEs. The second method considers the possibility of negative TE and sought to include the edges with the TEs of greatest magnitude rather than greatest value. Negative values of TE can be computed by using the method described in Eq. 2, because it computes the difference between two mutual information quantities. The entropy threshold should be considered as the amount of information made available from data; thus, it defines the level of interdependency and the amount of variables necessary to predict steadily the same network topology or integrated metric resulting from that network.

After the function U from Eq. 3 was maximized, the IH indicator was constructed by selecting variables based on the selected $f_{i,j}$ functions. To reduce redundancy in creating a MENet, variables that were causally predicted strongly by other variables were excluded. This was done by evaluating the weighted in-degree and out-degree of each node in the network. Nodes with a greater weighted out-degree than in-degree were included in MENet. These nodes are strongly predicting the variability of other nodes, thus the overall network dynamics. These steps are shown in Eqs. 5 and 6. Variable selection was defined by a function $g(X_i)$, defined as follows

$$g(X_i) = \begin{cases} 1, \sum_j f_{i,j}(X) > \sum_j f_{j,i}(X) \\ 0, \sum_j f_{i,j}(X) \le \sum_j f_{j,i}(X) \end{cases} \quad (5)$$

so that variable inclusion depends on the comparison of the TE projected by the variable X_i onto the other variables and the TE projected by the other variables onto X_i. In this way, the MENet inference was based on information theoretic and topological criteria to screen (i) the necessary and sufficient information and (ii) the non-redundant information. The defined function g was then used to create the IH indicator

$$\mathrm{IH}_t = \sum_i x_{i,t} \cdot g(x_{i,t}) \quad (6)$$

which represents the sum of all of the variables that were included by the structure of MENet in a multi-criteria value function. This IH indicator represents a simple use of the results from this variable selection method rather than an optimal characterization of city health.

Information theoretic GSUAs

To perform GSUAs for the defined IH, we used an information theoretic approach (*55*). TEs of variables used in the inference of MENet and Shannon entropies were used to calculate second- and first-order sensitivity indices. The Shannon entropy was calculated using the entropy package in R (*56*) by using the maximum likelihood (ML) method for estimating pdfs of variables. When finding the total entropy of a composite indicator (such as IH), the information balance equation that defines the total entropy was given by the sum of the Shannon ML entropies of all input variables as considered alone in the variability of IH and the sum of their TEs that is assessing variable interdependence for the variability of IH. Thus, the total entropy of IH the can be written as

$$H(\mathrm{IH}) \approx \sum_i H(x_i) + \sum_i \sum_{j \neq i} \mathrm{TE}_i(x_i, x_j) + \sigma(\mathrm{IH}) \quad (7)$$

where x_i denote the variables that contribute to the IH indicator. In this equation, $H(\cdot)$ denotes Shannon entropy, and $\mathrm{TE}(\cdot, \cdot)$ denotes TE from the first variable to the second variable. The sum of TEs was a proxy of the mutual information of a variable, thus considering the whole set of variable interdependencies. Comparing the total TE to Shannon entropy compares information contained within each variable to the sum of information projected to other variables and information other variables project to that variable. That is, Eq. 7 describes how information of IH was contained in single variables by themselves in isolation and related to other variables present in MENet that were synergistically interacting with each other (*57*). The term $\sigma(\mathrm{IH})$ represents unexplained noise that is representative of unexplained variables, discretization artifacts, and numerical methods used in the calculation. The ratios $\mu = \sum_i \frac{H(x_i)}{H(\mathrm{IH})}$ and $\sigma = \sum_i \sum_{j \neq i} \frac{\mathrm{TE}_i(x_i, x_j)}{H(\mathrm{IH})}$ are the first- and second-order sensitivity indices (*55*).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/4/2/e1701088/DC1

Excel file

REFERENCES AND NOTES

1. N. Xi, R. Muneepeerakul, S. Azaele, Y. Wang, Maximum entropy model for business cycle synchronization. *Phys. A* **413**, 189–194 (2014).
2. B. S. Halpern, C. Longo, D. Hardy, K. L. McLeod, J. F. Samhouri, S. K. Katona, K. Kleisner, S. E. Lester, J. O'Leary, M. Ranelletti, A. A. Rosenberg, C. Scarborough, E. R. Selig, B. D. Best, D. R. Brumbaugh, F. S. Chapin, L. B. Crowder, K. L. Daly, S. C. Doney, C. Elfes, M. J. Fogarty, S. D. Gaines, K. I. Jacobsen, L. B. Karrer, H. M. Leslie, E. Neeley, D. Pauly, S. Polasky, B. Ris, K. S. St. Martin, G. S. Stone, U. R. Sumaila, D. Zeller, An index to assess the health and benefits of the global ocean. *Nature* **488**, 615–620 (2012).
3. J. Y. Kim, Y. J. Ko, C. W. Rhee, B. J. Park, D. H. Kim, J. M. Bae, M. H. Shin, M. S. Lee, Z. M. Li, Y. O. Ahn, Cardiovascular health metrics and all-cause and cardiovascular disease mortality among middle-aged men in Korea: The Seoul male cohort study. *J. Prev. Med. Public Health* **46**, 319–328 (2013).
4. E. M. Magnan, D. M. Bolt, R. T. Greenlee, J. Fink, M. A. Smith, Stratifying patients with diabetes into clinically relevant groups by combination of chronic conditions to identify gaps in quality of care. *Health Serv. Res.* (2016).
5. J. P. Hirdes, D. H. Frijters, G. F. Teare, The MDS-CHESS scale: A new measure to predict mortality in institutionalized older people. *J. Am. Geriatr. Soc.* **51**, 96–100 (2003).
6. D. M. Sletten, G. A. Suarez, P. A. Low, J. Mandrekar, W. Singer, COMPASS 31: A refined and abbreviated composite autonomic symptom score. *Mayo Clin. Proc.* **87**, 1196–1201 (2012).
7. N. Freudenberg, M. Fahs, S. Galea, A. Greenberg, The impact of New York City's 1975 fiscal crisis on the tuberculosis, HIV, and homicide syndemic. *Am. J. Public Health* **96**, 424–434 (2006).
8. M. Singer, S. Clair, Syndemics and public health: Reconceptualizing disease in bio-social context. *Med. Anthropol. Q.* **17**, 423–441 (2003).
9. C. J. Patel, J. Bhattacharya, A. J. Butte, An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLOS ONE* **5**, e10746 (2010).
10. A.-L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
11. D. A. Luke, K. A. Stamatakis, Systems science methods in public health: Dynamics, networks, and agents. *Annu. Rev. Public Health* **33**, 357–376 (2012).
12. A. Abbott, City living marks the brain. *Nature* **474**, 429 (2011).
13. D. Helbing, *Social Self-Organization* (Springer, 2012).
14. S. Jayasinghe, Conceptualising population health: From mechanistic thinking to complexity science. *Emerg. Themes Epidemiol.* **8**, 2 (2011).
15. S. F. Gull, J. Skilling, Maximum entropy method in image processing. *IEE Proc. F* **131**, 646–659 (1984).

16. M. Convertino, A. Troccoli, and F. Catani, Detecting fingerprints of landslide drivers: A MaxEnt model. *J. Geophys. Res. Earth Surf.* **118**, 1367–1386 (2013).
17. S. J. Phillips, R. P. Anderson, R. E. Schapired, Maximum entropy modeling of species geographic distributions. *Ecol. Model.* **190**, 231–259 (2006).
18. I. Volkov, J. R. Banavar, S. P. Hubbell, A. Maritan, Inferring species interactions in tropical forests. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 13854–13859 (2009).
19. R. Dewar, Information theory explanation of the fluctuation theorem, maximum entropy production and self-organized criticality in non-equilibrium stationary states. *J. Phys. A Math. Gen.* **36**, 631–641 (2003).
20. T. Schreiber, Measuring information transfer. *Phys. Rev. Lett.* **85**, 461–464 (2000).
21. C. E. Shannon, A mathematical theory of communication. *Bell Labs Tech. J.* **27**, 379–423 (1948).
22. R. Vicente, M. Wibral, M. Lindner, G. Pipa, Transfer entropy—A model-free measure of effective connectivity for the neurosciences. *J. Comput. Neurosci.* **30**, 45–67 (2011).
23. Y. Hu, H. Zhao, X. Ai, Inferring weighted directed association network from multivariate time series with a synthetic method of partial symbolic transfer entropy spectrum and granger causality. *PLOS ONE* **11**, e0166084 (2016).
24. E. T. Jaynes, Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630 (1957).
25. I. Rodriguez-Iturbe, A. Rinaldo, *Fractal River Basins: Chance and Self-Organization* (Cambridge Univ. Press, 2001).
26. M. Ratto, A. Pagano,. State dependent regressions: From sensitivity analysis to meta-modeling, in *System Identification, Environmental Modelling, and Control System Design*, L. Wang, H. Garnier, Eds. (Springer, 2012).
27. V. Belton, T. Steward, *Multiple Criteria Decision Analysis: An Integrated Approach* (Kluwer, 2002).
28. S. M. Rappaport, Implications of the exposome for exposure science. *J. Expo. Sci. Environ. Epidemiol.* **21**, 5–9 (2011).
29. B. Barzel, Y.-Y. Liu, A. L. Barabsi, Constructing minimal models for complex system dynamics. *Nat. Commun.* **6**, 7186 (2015).
30. F. Vafaee Using multi-objective optimization to identify dynamical network biomarkers as early-warning signals of complex diseases. *Sci. Rep.* **6**, 22023 (2016).
31. H.-J. Kim, J. M. Kim, Cyclic topology in complex network. *Phys. Rev. E* **72**, 036109 (2005).
32. D. Easley, J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World* (Cambridge Univ. Press, 2010).
33. K. McSpadden, "These are the healthiest (and unhealthiest) cities in America," *Time*, 19 May 2015.
34. M. Haiken, "America's top 20 healthiest cities," *Forbes*, 13 September 2011.
35. C. J. Patel, A. K. Manrai, Development of exposome correlation globes to map out environment-wide associations. *Pac. Symp. Biocomput.* 231–242 (2015).
36. M. Nitzan, J. Casadiego, M. Timme, Revealing physical interaction networks from statistics of collective dynamics. *Sci. Adv.* **3**, e1600396 (2017).
37. Big City Health, www.bigcitieshealth.org/city-data/ [accessed September 2017].
38. H. Kantz, T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge Univ. Press, ed. 2, 2004).
39. P. Wollstadt, M. Martínez-Zarzuela, R. Vicente, F. J. Díaz-Pernas, M. Wibral, Efficient transfer entropy analysis of non-stationary neural time series. *PLOS ONE* **9**, e102833 (2014).
40. X. S. Liang, Information flow and causality as rigorous notions ab initio. *Phys. Rev. E* **94**, 052201 (2016).
41. D. Gencaga, K. H. Knuth, W. B. Rossow, A recipe for the estimation of information flow in a dynamical system. *Entropy* **17**, 438–470 (2015).
42. A. Kaiser, T. Schreiber, Information transfer in continuous processes. *Phys. D* **166**, 43–62 (2002).
43. A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information. *Phys. Rev. E* **69**, 066138 (2004).
44. A. F. Villaverde, J. Ross, F. Morán, J. R. Banga, MIDER: Network inference with mutual information distance and entropy reduction. *PLOS ONE* **9**, e96732 (2014).
45. G. H. Torbati, G. Lawyer, D. Mount, S. Arya, TransferEntropy: The Transfer Entropy Package (2016); https://CRAN.R-project.org/package=TransferEntropy
46. S. Ito, Backward transfer entropy: Informational measure for detecting hidden Markov models and its interpretations in thermodynamics, gambling and causality. *Sci. Rep.* **6**, 36831 (2016).
47. K. Friedman, A. Shimony, Jaynes's maximum entropy prescription and probability theory. *J. Stat. Phys.* **3**, 381–384 (1971).
48. T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, N. V. Fedoroff, Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19033–19038 (2006).
49. M. E. Aiello-Lammens, M. L. Chu-Agor, M. Convertino, R. A. Fischer, I. Linkov, H. R. Akçakaya, The impact of sea-level rise on Snowy Plovers in Florida: Integrating geomorphological, habitat, and metapopulation models. *Glob. Chang. Biol.* **17**, 3644–3654 (2011).
50. G. Tkačik, O. Marre, T. Mora, D. Amodei, M. J. Berry II, W. Bialek, The simplest maximum entropy model for collective behavior in a neural network. *J. Stat. Mech.* P03011 (2013).
51. I. B. Huang, J. Keisler, I. Linkov, Multi-criteria decision analysis in environmental sciences: Ten years of applications and trends. *Sci. Total Environ.* **409**, 3578–3594 (2011).
52. M. Convertino, K. M. Baker, J. T. Vogel, C. Lu, B. Suedel, I. Linkov, Multi-criteria decision analysis to select metrics for design and monitoring of sustainable ecosystem restorations. *Ecol. Indic.* **26**, 76–86 (2013).
53. A. M. Geoffrion, J. S. Dyer, A. Feinberg, An interactive approach for multi-criterion optimization, with an application to the operation of an academic department. *Manag. Sci.* **19**, 357–368 (1972).
54. R Core Team, R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, 2016); www.R-project.org/.
55. N. Lüdtke, S. Panzeri, M. Brown, D. S. Broomhead, J. Knowles, M. A. Montemurro, D. B. Kell, Information-theoretic sensitivity analysis: A general method for credit assignment in complex networks. *J. R. Soc. Interface* **5**, 223–235 (2008).
56. J. Hausser, K. Strimmer, entropy: Estimation of entropy, mutual information and related quantities (2014); https://CRAN.R-project.org/package=entropy.
57. R. Quax, O. Har-Shemesh, P. M. A. Sloot, Quantifying synergistic information using intermediate stochastic variables. *Entropy* **19**, 85–112 (2017).

Acknowledgments

Funding: J.L.S. and M.C. acknowledge funding from the NSF SRN project (no. 1444745, "SRN: Integrated Urban Infrastructure Solutions for Environmentally Sustainable, Healthy, and Livable Cities"; www.sustainablehealthycities.org). M.C. acknowledges funding from the MNDrive program at the University of Minnesota and funding from the Global Institution for Collaborative Research and Education Initiative on Big-Data and Cybersecurity at Hokkaido University, Sapporo, Japan.
Author contributions: J.L.S. acquired data, applied methods to data, and prepared the manuscript. M.C. conceptualized methodological framework, supervised J.L.S., and prepared the manuscript.
Competing interests: All authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors or found at https://github.com/matteoconvertino.

Submitted 12 April 2017
Accepted 5 January 2018
Published 2 February 2018
10.1126/sciadv.1701088

Citation: J. L. Servadio, M. Convertino, Optimal information networks: Application for data-driven integrated health in populations. *Sci. Adv.* **4**, e1701088 (2018).

Science Advances

Optimal information networks: Application for data-driven integrated health in populations

Joseph L. Servadio and Matteo Convertino

Sci Adv **4** (2), e1701088.
DOI: 10.1126/sciadv.1701088