Title:

Hypothesis Generation and Inference Using Emerging Properties of Causality

Authors: Abbas Shojaee [1,2] *, Isuru Ranasinghe [3], Alireza Ani [4]

Affiliations:
1. Pulmonary and Critical Care Section, Department of Internal Medicine, Yale University School of Medicine, New Haven, Connecticut
2. Center for Outcomes Research and Evaluation, Yale University, New Haven, Connecticut
3. Discipline of Medicine, The University of Adelaide, Adelaide, South Australia
4. Pegahsoft, Isfahan, Iran

*** Author for Correspondence:**

Dr. Abbas Shojaee,
Pulmonary and Critical Care Section, Department of Internal Medicine
Yale University School of Medicine, New Haven, Connecticut 06510
Email: abbas.shojaee@yale.edu;
Phone: 203 747-6914
Office: 203 785-3686
Fax: 203 785-3826

Disclosures
The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

Abstract:

Generating causal hypothesis is a key step in scientific research including in health sciences. The advent of massive omics and health observational data provides an opportunity for broad, expedient and inexpensive causal inference and to improve biomedical research efficiency and effectiveness. In this study, we introduce a top-down hypothesis generation concept and method named Causal Inference using Composition of Transactions" (CICT). CICT uses novel emergent properties of relevance network and Markov Chain data with machine learning(ML) methods to infer causality. CICT uses the idea that the stochastic distribution of transitions to and from a causal clinical condition is different than from a random event, and such difference is detectable in four unique distribution zones around each event. We applied CICT to a large-scale healthcare administrative data to infer causal relationships between clinical conditions and showed that without designing formal studies, or use of contextual clinical variables, CICT can efficiently produce highly accurate causal hypotheses at scale. Medical domain knowledge and an extensive set of validation methods used to ensure validity and veracity of the analysis. We explain the details of verification steps, and the results of empirical epidemiological studies conducted to assess the performance of CICT. Expedient producing of valid causal insights have the potential to prevent redundancy and accelerate research in health sciences in an era that the rate of data production exceeds our analysis capabilities.

One Sentence Summary: We introduce a novel concept and method for producing causal hypothesis in relevance networks and Markov Chain data and explain the details of the methods used to evaluate this causal inference method

Keywords:
causal inference; hypothesis generation; machine learning; relevance network; Markov Chain; network inference; artificial intelligence

Introduction

Generating causal hypothesis is a key step in scientific studies including biomedical research. However, causal hypothesis and relationships can be challenging to find and verify, in part because we have continued to rely on the inference of experts, experimental design, and statistical inference for causal reasoning. Insufficient understanding of potential causal relations is a source of cost, redundancy and slow progression of scientific research. For example, in biology and healthcare, we often stay at the level of correlation between phenomena for a long time as sufficient reliable data from convincing trial studies are sparse. The emergence of massive observational data in biology and healthcare provides an unprecedented opportunity to causal inference. Nevertheless, little success has been achieved in generating causal hypotheses in biomedical and health sciences.

Two common form of data in health sciences are undirected correlation or co-expression data known as relevance network (RN), and the one-step transition or Markov Chain network (MCN). Examples of MCN are patients transition between clinical conditions in population-level data or directed gene regulatory networks, and undirected gene expression data are RN network. The issue of producing causal hypothesis using RN or MCN is often faced in health sciences when trying to understand the interaction between different events and phenomena. Surprisingly few methods exist for direct causal inference in RN or MCN despite the popularity of these data forms (*1, 2*).

In the context of MCN an RN, conventional methods have restricted applicability (*3, 4*) for generating a causal hypothesis. For example, Potential Outcomes Framework (*5, 6*) starts with a basic understanding of the observed system and requires both a primary hypothesis and an environment of limited measurable confounders to create an experimental design or quasi-experimental design (*7-9*). By contrast, in RN and MCN a multitude of uncontrolled factors coexist and interact in all ways, including circular connections. A more recent and powerful concept of graphical structural models [3] and counterfactual analysis, spearheaded by Judea Pearl, also needs a comprehensive knowledge of the system's interactions before mapping the influence structure, evaluating a causal claim or measuring the effect size. Granger causality (*10, 11*), a significant advancement in causality inference, and techniques of state space reconstruction (*12-14*), conditional mutual information (*15, 16*), recurrence plots (*17, 18*) and information entropy transfer (*19*) are applicable only when time series data of sufficient length exist (*20*). These methods have been successfully used in studying relationships of recurring physical phenomenon of limited dimensions. However, in the context of systems medicine, the data is high-dimensional with no or limited repetition. Inspired by the mathematical notion of the graphs, methods such as Random Walk and its offspring such as PageRank (*21*), Walktrap (*22*), and MapEquation (*23*) were developed to understand the sub-segments or modules of a network and the direction of influence among them. However, these methods which work at the level of sub-segments rather than individual connections are not designed to infer causal relations.

In this work, we introduce a novel method for generating causal hypothesis and for identifying underlying network in MCN and RN. We use a large scale observational healthcare data (figure 1 A) to build an MCN of patients transitions between various clinical conditions Then we apply our causal inference method to this network to generate various causal hypotheses. Next, we evaluate

the generated hypotheses, using the established medical knowledge of an extensive set of causal relationships, as the gold standard. Also, we conduct various validation and evaluation tests to ensure the soundness of machine learning methods. We furthered the analysis by using epidemiological etiological reasoning means to evaluate CICT suggested hypothesis independently. Moreover, we provide references to some novel findings of our method that are already presented to various scientific communities. Throughout this manuscript, any reference to a causal relationship or causal hypothesis carries the meaning that the first phenomenon is a probabilistic component cause, precipitator or precursor of the second phenomenon.

Causal Inference using Composition of Transitions (CICT)

CICT is a novel concept and method for top-down(*24*), in-silico inference(*25*) and evaluation of multiple causal hypotheses in RN and MCN. CICT is developed based on our original observation that the composition of the events happening before and after a causal event is different from a random event in natural systems of stochastic processes when presented as RN or MCN. For example, observing a population, the set of events that happened to patients, before and after rheumatoid arthritis, which is cause for multiple comorbidities, is different from the set of events before or after influenza, which as a mostly random event may affect any individual. Such difference creates emergent patterns in four unique distribution zones that we define around each event. These emergent patterns can be revealed and learned through manual analysis or by supervised and unsupervised machine learning methods. Machine learning models, once trained on these patterns, can directly discriminate causal transitions from random transitions and predict the strength of unseen causal claims at the level of individual connections. Moreover, the features produced by CICT enables machine learning to predict the direction of causality in undirected relevance networks.

To demonstrate CICT, we use a large-scale administrative healthcare dataset to build Markov Chain network of patients transitions from one clinical condition to another (Figure 1 A, B). For each patient, we consider each two consequent patient's encounters as a transition of the patient between primary diagnosis of the first encounter to the primary diagnosis of the second encounter. By observing a population, and recording frequency of patients' transitions between various clinical conditions, we can build an MCN between all observed clinical conditions. Next, on each edge between two nodes i: source and j: target we define two probabilities confidence (Conf) and contribution (Contrib) as follow:

$$Conf_{ij} = \vec{ij}/i = probability\ of\ future\ transition\ to\ j\ conditioned\ on\ being\ in\ i$$

$$Contrib_{ij} = \vec{ij}/j = probability\ of\ a\ previous\ state\ of\ i\ conditioned\ on\ being\ in\ j$$

Hence, for each node i and the set of nodes j, we can define two distribution zones for transition edges $\vec{ij}$ and two distribution zones for transition edges $\vec{ji}$. Accordingly, for each node i we create four unique zones of transition distributions as shown in Figure 1 D. The first zone is the distribution of all confidences of node i to other nodes. Zone 2 is the contribution of node i to other nodes. Zone 3 is confidences of other nodes into node i. And Zone 4 is the contribution of other nodes into node i. Next, we estimated the probability density function (PDF) for each zone of each event by creating a histogram of values in each zone (for example Confidence in Zone 1).

The histogram is achieved by first mapping the entire range of values into a series of incremental bins and then counting frequency of values in each bin. We named these four PDFs as "composition of transitions.".

We hypothesized that the set of events before and after a causal event is different from a random event. Moreover, such difference should be identifiable in the composition of transitions for events. Figure 1-E demonstrates a real example of such differences in the composition of transitions. Rheumatoid arthritis, a well-known cause of multiple comorbidities such as arthropathies, is painted as a red density, and syncope, a known effect of multiple clinical conditions is shown in yellow. Also, influenza is used as a proxy for a random event that may affect people with a broad range of previous conditions. This example shows that the median of causal density (red) is about two order of magnitude different than for the random density (blue) in all for PDF zones. We hypothesized that composition of transitions for a causal event should exhibit identifiable generalized patterns. We called these generalized patterns a 'behavior' and suggested that ML methods can learn the differences in 'causal behavior' versus effect or random behavior to discern the causal or random nature of events and their transitions. Accordingly, to capture these behaviors and produce features for ML methods, we extracted moments of four PDFs of each event.

Next, we assumed that if a pair of events (e.g., i and j) have a causal relationship, the specific interplay of the two events should be reflected in the strength of their back ($\vec{ji}$) and forth ($\vec{ij}$) connections with regard to the behavior of the four corresponding PDF areas (Figure 1 F). For example, if i is the cause of j, the probability of a transition from i to j could be higher comparing to transitions to random events, after adjusting for prevalence. To train ML methods on the differences in behaviors of the cause, effect, and random events a set of features is required to represent that behaviour. Accordingly, we extracted moments of the distribution of confidences and contributions, in all the the four mentioned zones around each node. For example, we extracted mean, standard deviation, skewness, kurtosis, median absolute deviation(*26-28*) and L-moments(*29, 30*) for confidences and contributions. Median absolute deviation and L-moments are measures of distribution that are more stable and less sensitive to outliers(*29, 30*) compared to standard measures, such as mean or standard deviation. We suggested that this information about the behavior of source and target of a specific transition can guide ML methods to predict whether that transition is a cause, effect or random. Figure 1.F shows the eight influencing distribution zones that we considered for each specific edge. We added the measures of distributions of these 8 zones as features to each specific transition edge.

To improve training of ML models, we engineered features by creating equivalencies with physical systems such as electric circuits and closed hydraulic systems. For example, assuming edges are connections in an electrical circuit we calculated resistance ($^{abs(i-j)}/_{\vec{ij}}$), and taking the nodes and edges as a connected hydraulic system, we calculated output pressure ($^{(i-j)\,*\,\vec{ij}}/_{i}$). The feature engineering produced a total of 334 features for each transition edge on the network.

It is common knowledge that the right features are more important than the choice of model in the performance of machine learning (*31*). For example, the right features can help unsupervised methods to group similar data points into clusters that reflect real classes and improve the accuracy of prediction in supervised models. Here, our novel features enable both supervised and unsupervised models to discriminate causal relationships from random connections correctly and to predict possible causal relations in unlabeled connections.

Data Source and Preprocessing

To evaluate CICT through ML methods, we used a large-scale observational healthcare data to build an MCN of patients transitions. Deidentified observational all-payers claim data from Healthcare Utilization Project (HCUP) California State Inpatient Database (SID), and Emergency Department Database (ED)[25] from 2005 to 2011 was used for applying CICT and validating its results. HCUP is the result of a federal-state-industry partnership sponsored by the Agency for Healthcare Research and Quality, and each HCUP record represents a patient encounter [25]. The SID files contain all inpatient discharge records in community hospitals in the California state. The ED contains all encounters in the emergency department that do not result in a hospitalization. Both SID and ED include all patients covered by Medicare, Medicaid, private insurance, or uninsured and contains more than 97% (*32, 33*) of the patient population. The data include patients' demographic data, primary diagnosis, comorbidities, procedures, total costs of hospitalization, length of stay and more. Moreover, data has a pseudo patient-identifier that allows connecting all hospital visits of each patient across ED and SID dataset. This data capture longitudinal hospital visits information of all patients in California from 2005 to 2011.

We used International Classification of Diseases, Ninth Revision, Clinical Modification (ICD9CM) codes for the primary diagnosis and up to 24 comorbidities, to identify exposures or events. As our hypothesis generation method, uses transitions of patients between clinical conditions to build the MCN, we included patients with more than one observation. The final dataset contains 15,047,413 hospital admissions among 3,966,603 patients who had two or more hospitalizations during 2005 to 2011. We created MCN by taking each Diagnosis as a vertice on the network and mapping each patient transition to an edge between network vertices. For each pair of clinical events, we preserved frequency of observed clinical conditions as the node frequency and the frequency of transitions as the edge frequency. We used vertex and edge frequencies to calculate confidence and contribution parameters that we discussed in CICT method. The final MCN contained 10,192 clinical events as vertices and over 873,000 unique transitions edges between clinical events.

Next, we created a set of ground truth to label some of the data in our MCN so machine learning methods can use the data and the features that we built using CICT.

The Ground Truth

We require the ground truth to evaluate the results of unsupervised methods and to train supervised methods. Here the ground truth is the well established medical knowledge about a sufficient set of causal and random relations between pairs of clinical conditions. To prepare this set we used

Semantic MEDLINE Database (SemMedDB) from Semantic Knowledge Representation (SKR) project (*34*) that contains 82.2 million predicates between biomedical concepts extracted from all MEDLINE citations. We used SemMedDB to label a set of 267 causal relations from that we found a match for in our transition data. Two clinician internists as subject matter experts independently verified the correctness of identified causal relations (supplement table 1). In addition to causal relations, the predictive models require a set of random relations to learn the difference between causal and random relations. Accordingly, we chose a random sample of transitions from our transition graph; then, two subject matter experts manually tagged 267 relations as 'irrelevant-may coincide' denoting that a transition from the first clinical state to the second is most probably due to a random process and not a causal relation (Supplementary Table 2). Next, we labeled corresponding transitions on the graph as 'random' or 'causal' according to the ground truth.

We used multiple empirical analysis to evaluate CICT results against a set of established medical causal relations as the ground truth. Use of ground truth is the gold standard in system inference literature. The health domain provides a suitable testbed due to the availability of large-scale datasets and the benefit of well-established domain knowledge. Moreover, we avoided using the contextual knowledge that is specific to health domain into CICT design (e.g., in building models) to keep the findings as generalizable and straightforward as possible and to ensure applicability in other scientific fields. Also, we defined a minimum length model to show that in the absence of time series data and contextual information (e.g., Age, gender and laboratory data), simple one-step transitions frequencies, carry valuable information for causal inference.

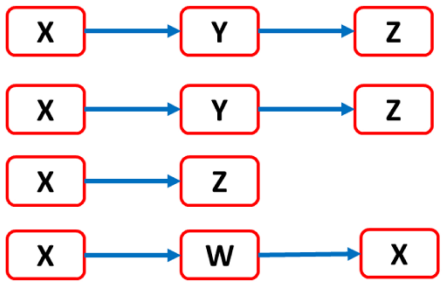
Results of Causal Inference using Composition of Transitions (CICT)

Often modeling and machine learning studies are focused on building new models and proving the validity of modeling presumptions. Here, we use standard, well-established ML models and show that the features that CICT suggests contain new information about causality that a standard model can use to predict the veracity of a causal claim. Thus we report the results of three experiments conducted using classification and clustering methods and evaluate the performance and results of these methods. Experiment **I** shows the high accuracy of CICT method in discriminating causal transitions from coexistence and random associations. Experiment **II** shows how much CICT can predict the causal direction in a bi-directional association. The third experiment ensures the elimination of possible subjective errors in the validation phase by applying the trained model on the set of previously unlabeled and randomly selected transitions and evaluating the results. In all experiments after optimizing and validating the predictive model, we estimated the discrimination power of the models using the area under the receivers operating characteristics curve (AUC of ROC) as a surrogate of the amount of causal knowledge that CICT learns. In line with this, we used the predicted probability of classification as the measure of causality referred as 'strength' in tables. Also, we describe the most important predictors of causality and provide interpretation for causal behavior.

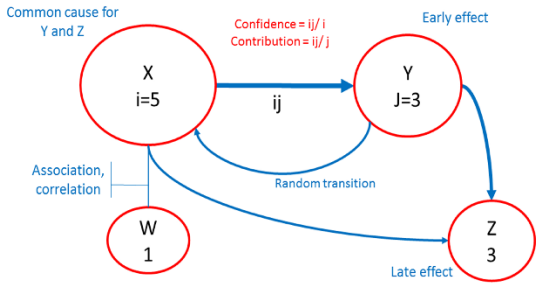
A

Admissions	**15,047,413**
Patients	3,966,603
Total transitions included	11,534,448
Unique transitions or Edges	873,761
Unique Clinical conditions	10,119

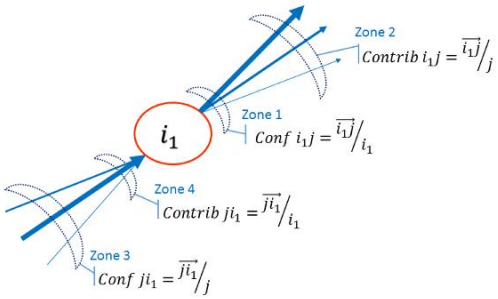
B



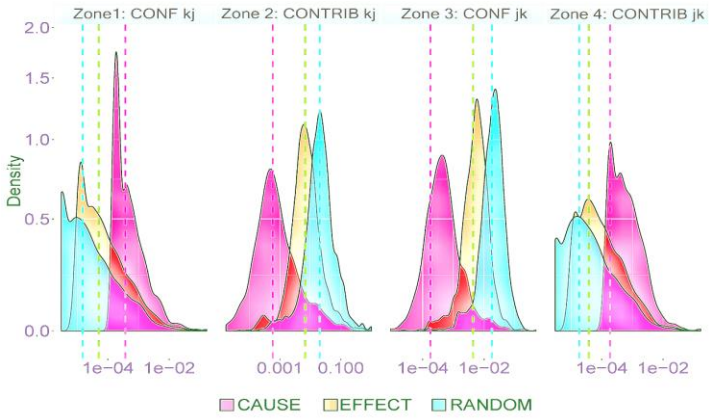
C



D



E



F

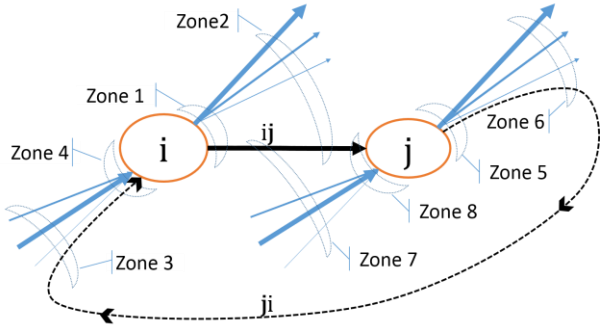


Figure 1: (A) Descriptive statistics on the data. (B) A set of transitions for four hypothetical patients. For example, the first patient is hospitalized with a principal diagnosis of condition X, and after a period is re-hospitalized with condition Y and so forth. If we start merging similar transitions, the result would be the transition graph shown in (C) Different types of transitions on a network. X is a common cause of Y and Z where Y is an early effect and Z is a late effect. X and W showed an association without an observable causal relation. Numbers represent hypothetical frequencies. (D) The 4 zones that carry different distributional information. It is important to note that the four areas are not overlapping and contain different information. Here **i** represents source and **j** represent destination. (E) The log-scale density graph shows the different distributions in four distribution zones for a cause: Rheumatoid Arthritis(red), an effect: Syncope (yellow) and a random event: Pneumonia (blue). Dotted lines show medians of correspondent density. (F) The eight distribution zones carry information relevant to the nature of the transition between source and destination (black edge). Zones 1,3,5,7 capture distribution of the parameters derived from Confidence calculation. Zones 2,4,6,8 capture distribution of the parameters derived from Contribution calculation.

Experiments and results

Evaluating CICT accuracy in discriminating random transitions from causal transitions in labeled data (Experiment I)

We used the ground truth to create a set of 267 random transitions and a set of 267 causal relations. Then we split this 534 transition-set into a 75% training subset and 25% validation set. Next, we trained a random forest(RF) model with 10-fold cross-validation on the training subset to separate random relations from non-random relations. RF is a well-studied machine learning method that works well in nonlinear and complex problem domains by aggregating the collective result of multiple decision trees as its output. To ensure the stability of results we repeated training and validation 50 times. RF predictive model shows an average discrimination power of AUC= 0.916 with Mean Square Error = 0.074 and R2 = 0.699 on outstanding validation sets. The model is well calibrated as evaluated by Hosmer-Lemeshow chi-square on ten deciles of risk (Chi-square = 6.846, P-value = 0.553). (Figure 2 A, B, C). RF converges in 3 decision trees of depth five which means a limited number of decisions over composition parameters can separate causal and random transitions. An area under curve greater than 0.9 for a model is an excellent discrimination power. The top 10 relations predicted by the model are shown in Table 1.A. All the top 10 relations are well-known casual associations.

The right features produced by CICT helps unsupervised methods to group similar data points into clusters that reflect real classes. Figure 2 D shows two clusters, as identified by unsupervised Partitioning Around Medoids (PAM)(*35*) clustering. The denser cluster (cyan polygon) mostly embodies cause-effect transitions (blue triangles) where the more significant scattered cluster (pink polygon) mainly contains random transitions (red dots). Graph axes are the first and second dimensions of Principal Component Analysis(PCA). We used these coordinates to show data points along their maximum variability extension to achieve a more precise visualization. Adjusted Rand Index (*36*) shows 0.468 agreement between clustering results and real classes.

Evaluating CICT accuracy in inferring direction of causality in bidirectional relations (Experiment II)

We hypothesized that if the composition of transitions contains information about causality, it should be able to predict the real direction of causation in the bidirectional association between pairs of clinical conditions. For example, if our observations show frequent both way transition between flu and pneumonia we expect a causal inference method to specify which of the two conditions is the cause or precipitating factor for the other. We used logistic regression and RF to predict the direction of causation in a bidirectional association. Using the ground truth established in Experiment **I**, we selected a set of 225 causal transitions (e.g. flu → pneumonia) and 225 reverse of causal relations (e.g. pneumonia → flu) (Supplementary Table 3). We then used a 75% random sample of this data for training and used the 25% remaining to test the model. We used 10-fold cross-validation on the training set and measured model performance on the 25% outstanding validation set. We conducted training and validation 50 times to ensure model stability. RF surpassed logistic regression. Best results achieved with RF using 30 trees with depth 5 and show an average discrimination power of AUC= 0.772 with Mean Square Error = 0.193 and R2 = 0.215 on the outstanding validation set. The model was well calibrated across ten deciles of risk (Hosmer-

Lemeshow Chi_square = 5.195, P-value = 0.736). (Figure 2 E, F, G). Top 10 cause-effect predicted relationships in bi-directional transitions, shown in Table 1.B, are well known causal relations in medicine. An AUC = 0.772 is an acceptable discrimination power as a model with an AUC of more than 0.7 is a fair model with practical applications(*37*).

Figure 2 H shows two clusters, as identified by unsupervised Partitioning Around Medoids (PAM)(*35*) clustering method. The denser cluster (cyan polygon) mostly embodies cause-effect transitions (blue triangles) where the larger scattered cluster (pink polygon) mainly contains effect-cause relationships (red dots). Graph axes are the $1^{st,}$ and 2^{nd} dimensions of PCA and axis label shows the variability of data explained by each dimension. Adjusted Rand Index shows(*36*) 0.437 agreement between clustering results and the real classes reflecting the fact that even unsupervised algorithm can discriminate the direction of causality in association relationship considerably, using CICT features.

Evaluating CICT accuracy in finding causal hypothesis in unlabeled data (Experiment III)

To ensure that our results are not affected by design decisions in this experiment, we first empirically optimized training a predictive model using a set of 250 cause-effect relations as the positive class and a set of 90 effect-cause plus 840 random relations as the negative class. We then used all transitions with frequency > 20 among 873,761 total observed transitions, to create a random sample of size 1600. Next, we used a trained RF model to predict whether each of sampled transitions is a causal association or not. We used predicted value as model's measure of causality. For transitions that occurred in both directions in the results (like A→ B and B→ A), we retained transitions with higher predictive value. Next, we removed any transition with a predicted probability less than a threshold 0.535 and returned 75 relations. The choice of threshold made by applying Youden-Index(*38*) on prediction results of Experiment **II** to find the optimal cut-off point. This optimal threshold represents the best performance of discrimination when both effect-cause and cause-effect transitions exist. Next, we asked two clinician experts to evaluate the output independently. CICT did not report any random transition. Among the transitions identified, after removing 13 unexplainable relations due to coding or label ambiguity (e.g., CHF → CHF nonspecific), 62 causal transitions remained. Of these, 52 (p=0.764) were cause-effect and 10 (p=0.147) were effect → cause relations. The top 10 predicted relations are shown in Table 1C. We conducted this experiment five times on different random sets of transitions with little variation on prediction accuracy.

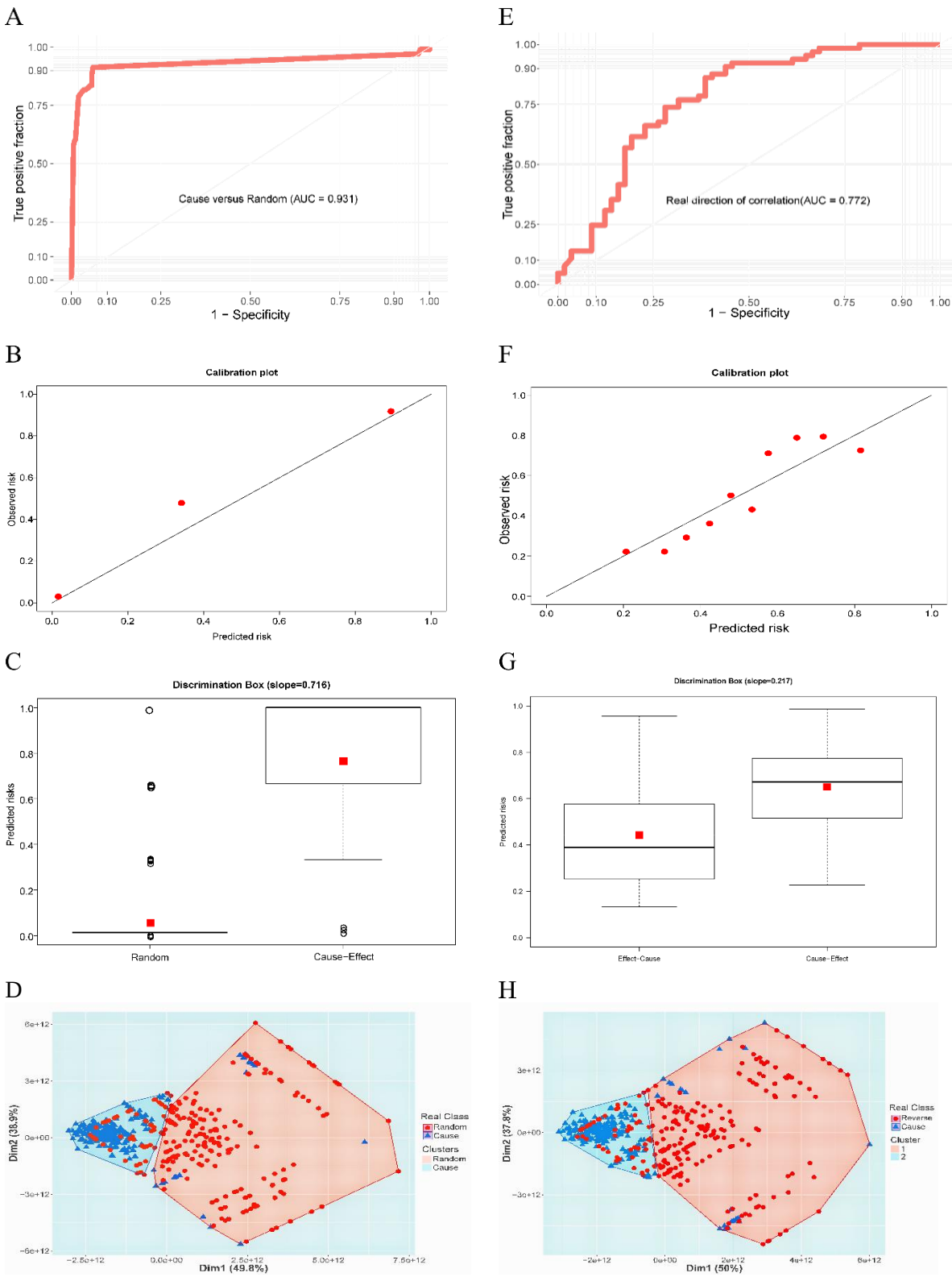


Figure 2. Left Column: CICT shows high accuracy in discriminating random transitions from associations. (A) ROC curve. (B) Calibration plot. (C) Discrimination box plot (D) Two clusters as identified by Partitioning Around Medoids along with the real class of data points. **Right Column:** CICT performs well in identifying direction of association: (E) ROC curve. (F) Calibration plot. (G) Discrimination box plot (H) Partitioning Around Medoids

A Top 10 predicted relationships in experiment **I**.

Strength	Source clinical condition	Target clinical condition
0.996	Hyperparathyroidism	Disorders of calcium metabolism
0.996	Peritonitis NOS	Abdominal pain
0.996	Pressure ulcer	Bacteremia
0.996	End stage renal disease	Anemia Not otherwise specified(NOS)
0.996	Polycystic ovaries	Overweight and obesity
0.996	Human immunodeficiency virus [HIV] disease	Bacteremia
0.996	Shock without mention of trauma	Transient alteration of awareness
0.996	Alcoholic cirrhosis of liver	Portal hypertension
0.996	Infectious mononucleosis	Splenomegaly
0.996	Calculus of ureter	Renal colic

B Top 10 Predicted relationships in experiment **II**

Strength	Source clinical condition	Target clinical condition
0.908	Systemic lupus erythematosus	Renal failure NOS
0.906	Hepatic encephalopathy	Transient alteration of awareness
0.887	Other pyelonephritis or pyonephrosis, not specified as acute or chronic	Renal failure NOS
0.858	Calculus of ureter	Leukocytosis NOS
0.845	Systemic lupus erythematosus	Thrombocytopenia NOS
0.815	Irritable bowel syndrome	Chronic pain
0.803	Meckel's diverticulum	Unspecified intestinal obstruction
0.798	Other specified disorders of circulatory system	Edema
0.777	Vesicoureteral reflux unspecified or without reflux nephropathy	Urinary tract infection, site not specified
0.769	Septicemia due to other gram-negative organisms	Shock without mention of trauma

C Top 10 predicted relationships in experiment **III**

Strength	Source clinical condition	Target clinical condition
1.00	Unspecified hypertensive heart disease without heart failure	CHF NOS
1.00	Other and unspecified rheumatic heart diseases	CHF NOS
1.00	Other primary cardiomyopathies	CHF NOS
1.00	Overweight and obesity	Localized adiposity
1.00	Mitral valve disorder	CHF NOS
1.00	Secondary malignant neoplasm of retroperitoneum and peritoneum	Malignant neoplasm of ovary
1.00	Infection and inflammatory reaction due to internal prosthetic device, implant, and graft	Acquired deformities of hip
1.00	Malignant essential hypertension	Hypertension NOS
1.00	Diabetes mellitus complicating pregnancy, childbirth, or the puerperium	Abnormal glucose tolerance of mother, complicating pregnancy, childbirth, or the puerperium
1.00	Unspecified intracranial hemorrhage	Intracerebral hemorrhage

Table 1: In all tables, strength stands for predicted strength of causal hypothesis (A) Top 10 predicted association relationships in a mixture of random and causal transitions. For 9 out of 10, the causal direction is predicted correctly. (B) Top 10 predicted the direction of causality in bi-directional association transitions (C) Top 10 predicted cause-effect relationship in a random unlabeled set of transitions. NOS stands for "Not Otherwise Specified."

Evaluating CICT results with time to event analysis

In health sciences, the two important assessment for evaluating risk and causal hypothesis using observational data is measuring the effect size and establishing the time arrow between potential cause and its possible effect. Here we use three carefully designed conventional analyses to evaluate one of the CICT suggested novel causal hypothesis. We investigated controversial relationship between allergic asthma (ICD9CM: 493.0) and emphysema(ICD9CM: 492) and malignancies using CICT. CICT predicts that emphysema could cause lung cancer, a fact that has been established with multiple studies (*39*) (*40*). Additionally, CICT predicts that asthma could cause 'benign neoplasm of the pituitary gland and craniopharyngeal duct' (ICD9CM: 227.3, N=6232, male =39.2%, age~=54.5). We evaluated this new hypothesis using epidemiological methods of causal inference in observational data. Accordingly, we used our data to define a population study of 9,398,589 patients 18 years and older and considered patients exposed to allergic asthma as the case group. To minimize the effects of confounding and obtaining an unbiased estimate of odds ratios, we matched the control group with the exposed patients using coarsened exact matching (CEM)(*41, 42*). The imbalance of case and control group dropped significantly judging by multivariate L1 metric (*43*) from 0.31 to 0.09. The association between asthma and cancers was estimated using the Cochran- Mantel-Haenszel (CMH) (*44, 45*) common odds ratio (CMH OR: 2.43, CI: 1.9-3.11) over the whole cohort. Odds ratio analysis suggest a strong association between the presence of the two conditions. To establish the direction of time and to measure the add risk, we created a cohort of all patients who developed the specific neoplasm and compared the group differences between cases exposed to allergic asthma and those who were not exposed. The direction of time arrow was evaluated using Non-parametric maximum likelihood estimator (NPLME) and parametric multivariate Cox proportional hazards (CPH) models. NPMLE chart shown in Figure 3 shows that the exposed group develop benign neoplasm of pituitary glands some seven years earlier than the unexposed group judging by 50% disease free on the curve of the survival function. The multivariate CPH model, adjusted for smoking, gender, and age also confirms that exposure to allergic asthma increases the risk of 'benign neoplasm of the pituitary gland and craniopharyngeal duct' with a hazard ratio equal to 1.71 ($p < 0.01$).

In summary, the evaluation of the effect size and direction of influence confirms CICT results. In addition to extensive evaluation of the machine learning methods using a set of ground truth through three set of independent evaluations we used standard epidemiological reasoning methods to evaluate results of CICT (*46-48*).

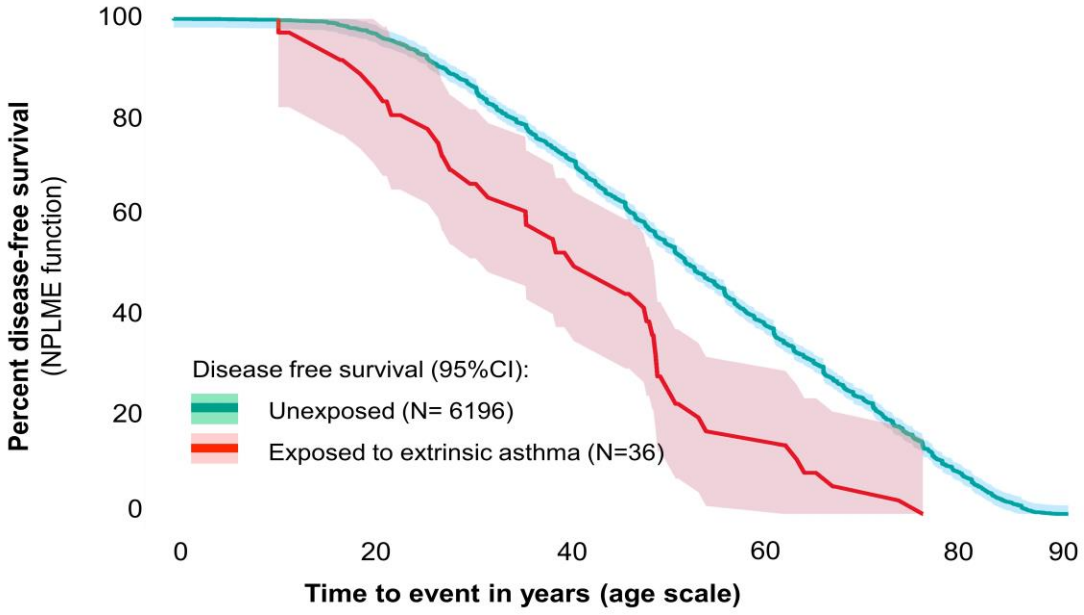


Figure 3. Nonparametric maximum likelihood estimation (NPMLE) function confirms CICT hypothesis and shows that patients with allergic asthma develop benign neoplasms of pituitary gland and craniopharyngeal duct years before those without exposure to allergic asthma.

Synthetic predictors of causal relations

Here, we evaluate model variables to understand which of the over 300 computed features are important predictors able to discriminate a causal relation from a random one. We used the model in Experiment **I** and calculated relative importance (RI)(*49*) of variables on random forests to rank predictors. Then we kept the top 8 predictors of the model with relative importance between 1.0 and 0.043 (figure 3 A). Figure 3 B represents histogram and density graphs of the top predictors in log scale and shows that the distributions of predictors for causal and random transitions are the result of two different generative processes. We evaluated the significance of differences by non-parametric two-sample Kolmogorov-Smirnov test (p-value for top 6 predictors tends to zero, for intvl_median $p < 0.001$).

The most important predictor was MADN_CNFSO (Fig 4), which measures the median absolute deviation of normalized confidences of outputs from the source. The median of the distribution of MADN_CNFSO for the source of causal edges (dashed red line) is three orders of magnitude larger than that of random edges (dashed blue line). In simple terms, this suggests that after adjusting for target probabilities, the probability of a target conditioned on the source is higher for causal relations, which is intuitive. The interesting observation is that without adjusting confidence = P(target | source) for the frequency of a target, a pure conditional probability is insignificant

among the predictors with relative importance (RI) <0.001 and cannot differentiate signal from noise. This observation characterizes the emitting behavior of the source in a causal relation.

The second predictor is M_CNTSO, which measures the median of contributions of the target to other nodes. We defined contribution as the probability of being previously in a specific primary state once we are in a second state. For example, knowing that a patient has pneumonia, what is the probability that he had influenza beforehand? Judging the distribution of M_CNTSO by their medians (vertical dashed red and blue lines), it is one order of magnitude lower for targets of causal transitions comparing to random transitions. This difference means that effects (targets of a causal transition) usually have a lower rate of contribution to a broader range of others comparing to targets of random transitions. A plausible interpretation is that effects do not contribute significantly to specific others, so transitions from them to other events tend to be random. This observation characterizes the influence of the effects of causal transitions, on their following transitions.

The third predictor MADN_CNFS, which measures the median absolute deviation of the normalized contribution of nodes into the target, is significantly higher for causal transitions. This higher contribution means that after adjusting for the source prevalence, on average the influence of inputs into an effect (the target of causal transitions) is higher than in random transitions. This observation characterizes the receiving behavior of the effect in a causal relation.

The fourth predictor is LM_CNFOS, is the L-mean of confidences of transitions from target to other nodes. The median of the distribution of LM_CNFOS is lower for random relations than for causal relations. This lower median, suggests that a 'random target' on average transits to more conditions at lower rates compared to an 'effect target', which transmits to a lower number of conditions each with higher confidence. The width of distribution demonstrates that 'effect nodes' show a wide range of transition behaviors. One interpretation is that some of the effects act as sinks or modulators that put patients on common subsequent care pathways. This observation characterizes the emitting behavior of the effects of causal transitions.

Interestingly, we can see that causal transitions have a more extensive range of time intervals compared to random transitions judging based on intvl_median which is the distribution of medians of time intervals for each transition edge. The distribution shows that causal transition may happen as soon as one day or as late as 1700 days with a median of 67.5 days. However, random transitions are occurring mostly after ten days with a median of approximately 150 days. The graph of intvl_median shows that the median of intervals for causal transitions is higher than the median of intervals of random transitions.

Another interesting finding is that 6 of the eight most significant predictors are related to node receiving-emitting compositions and just two low-rank predictors 'SZ_CNFCNT',(RI = 0.18) and intvl_median(RI= 0.08) are features of the specific edge. A plausible intuitive interpretation could be that some phenomena are by their very nature causal events and some are effects regardless of any other circumstances. Also, it is noteworthy that none of the features that we created to measure the asymmetry of bidirectional transitions showed up as significant predictors of causality.

Accordingly, the most crucial factor in determining whether a specific transition is causal or random is the nature of source and target of the transitions.

Evidently, a supervised method can gain a considerable amount of knowledge about the causal nature of each phenomenon from the composition of its previous (input or receiving) and subsequent (output or emitting) events. Also, it is the nature of source and target that mostly specifies the type of transition between them. We can conclude that to understand whether a specific transition is causal or random depends on a higher order or meta-structure of inputs and outputs to source and target. These results suggest three noteworthy findings: (1) standard Markov chains contain implicit hidden structure that is richer in information than was previously known; (2) the fact that some of the essential predictors are rates of emissions of other nodes into source of an edge (e.g., MADN_CNFSO), or influences of target into other nodes(e.g., M_CNTSO); suggests that a higher level or meta-structure exist in Markov chain data. This finding, which is also reported in another recent study on Markov Chain data, deserves further evaluation as it continues to exist despite the fundamental rule of formation of Markov Chains as a memoryless stochastic process; and (3) analysis of the composition of input and outputs can produce true causal hypotheses at scale, using real life, uncontrolled and noisy data.

CICT Time and space complexity

CICT has a space-time complexity follows the space-time complexity of the specific unsupervised or supervised method used for machine learning. For example for random forests, it is $O(e^* \log_e)$ for training on a subset of edges ‘e’ which is usually a small subset of the whole edge sets and O(E) for prediction over the set of all edges E.

(A) Important discriminating predictors of Causal versus Random relations

	Variable	scaled importance	Description
1	MADN_CNFSO	1.000	Median absolute deviation of normalized confidences emitted from target
2	M_CNTSO	0.977	Median of contributions of source into other nodes
3	MADN_CNFS	0.402	Median absolute deviation of normalized confidence of other nodes into source
4	LM_CNFOS	0.367	L-Mean of confidences emitted from source to other nodes
5	SDP_ST	0.272	Standard deviation of power of all inputs into the target of a transition
6	SZ_CNFCNT	0.184	Sum of Z-score of confidence and contribution of edge
8	K_CNTOS	0.090	Kurtosis of others contribution into the source of a transition
9	intvl_median	0.086	Median of time intervals of observed transitions between source and target

B)

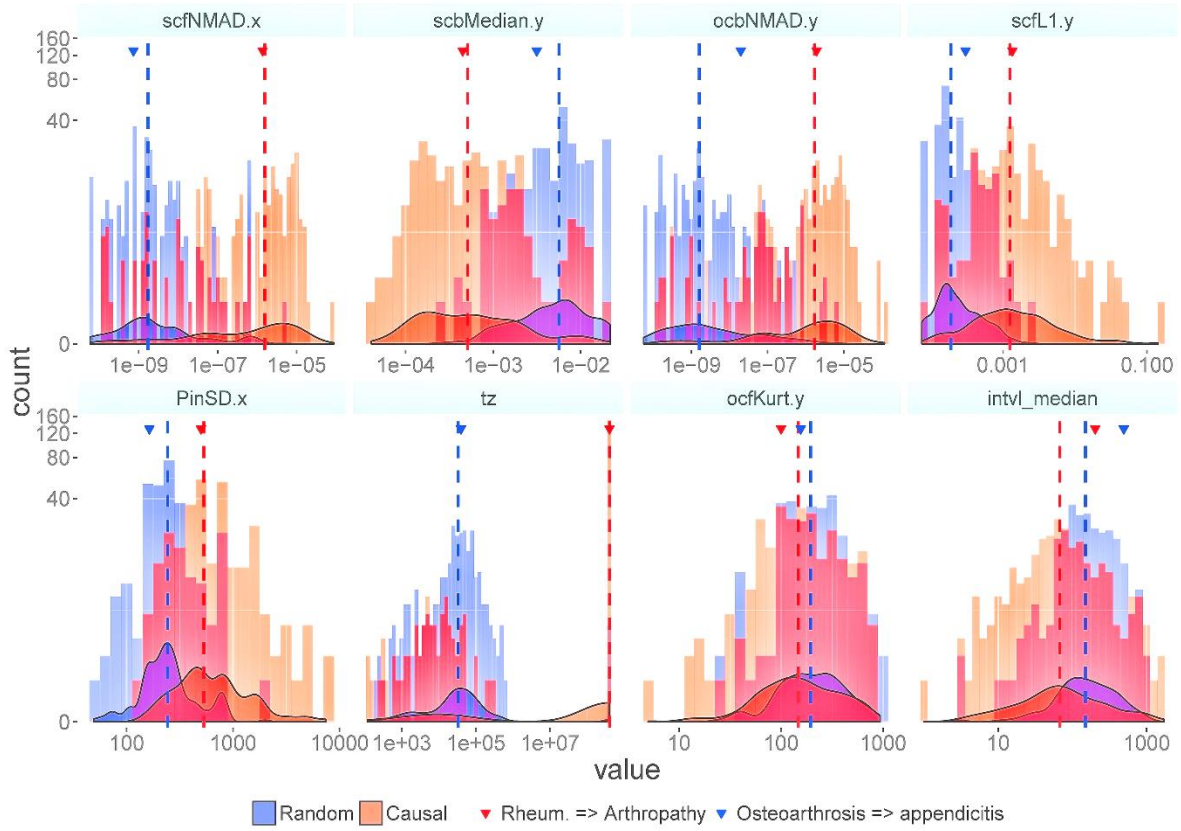


Figure 4: (A) Important discriminating predictors of Causal versus Random relations (**B)** Distribution of most important discriminating predictors of causal transitions (red) from random ones (blue). A logarithmic scale used for x and y. Histogram and density area graphs are superimposed to reflect differences in distribution better. Vertical dash lines show the median of the distribution. Red point shows the location of a causal relation 'Rheumatoid arthritis → Arthropathy, unspecified, pelvic region and thigh" and blue triangle shows the location of random transition 'Osteoarthrosis → Acute appendicitis". The significant differences of distribution suggest that they are the result of different generative processes, confirmed by a two-sample Kolmogorov-Smirnov test.

Discussion

In the literature for causal inference, it is traditionally thought that short-term data cannot provide enough information to infer a causal relation(*50*) and that almost all data-driven causality inference methods need time-series data of sufficient length (usually more than 25 points).

Here, We introduced causal inference using composition of transitions (CICT) as a novel and general analytic method for efficiently producing accurate causal hypotheses and for top-down network inference in Markov chain networks. We show that causal and random events, are results of different generative processes and we use the discriminating characteristics of them for phenotyping causal and random associations. MC network is frequent in many real-world scenarios in different disciplines where only short-term one-step transition data exist. These complex scenarios frequently happen, such as in econometrics or high throughput biological data(*51*), as well as in physics, web page ranking, molecular and higher order phenotypes, and epidemiology(*23*). Network inference in MC data is considered as a hard and computationally expensive problem due to the exponential increase of candidate networks given the number of nodes. In such scenarios, CICT can reveal the underlying system or dependency structure efficiently and make further analysis tractable.

Most of the causal inference methods are built around certain simplifying assumptions that are problematic for real-world data, especially in complex and interconnected domains. For example, the common modeling assumption that the data would follow a Gaussian or Poison distribution. Similarly, constraints on the internal structure of data such as the existence of time series data with sufficient length or consistent sequence of cause and effect. The third group, are suppositions about the underlying causal structure such as being acyclic or non-recursive as in Bayesian networks or separable cause and effect in Granger method. Being free from these three groups of constraints CICT can be applied in a range of contexts, for various objectives and in combination with existing causal inference methods. Methods that make minimal assumptions about data structure and causal structure, such as CICT, can help to identify useful causal relations using real-world observational data in an expedient, non-resource intensive manner.

The idea of using compositions of inputs/outputs in CICT introduces a rich set of features and reveals some previously unknown facets of causality. The distributional facets of causality continue to persist even when preconditioning, such as cohort definition, filters the observations or the labeling of events changes, as long as the process is consistent and reflective of real phenomena. Another significant finding is that the asymmetry of the back, and forth transitions between two events are insignificant comparing to distributional features. The lower importance of transition asymmetry has implications for developing directed network inference models. Also, relying on distributional features makes CICT applicable to relevance network structures where a concept of time does not exist. In RN the identifiable zones around each event shrinks to two zones. Moreover, we suggest that CICT is resilient against adding or dropping parts of information as it extracts features from stable measures of distributions like median absolute deviation(*26, 27*) and L-moments(*29, 52*). Such quality keeps CICT robust in the presence of unmeasured factors and noise. CICT remains stable with a randomly selected sample of data to further reduce computation in massive datasets. The robustness, simplicity, generalizability and low time-space complexity of

CICT let it be used solely or in combination with other methods for the analysis of large and dense graph data across disciplines.

From a healthcare research point of view, to the best of our knowledge, this study is the first to describe methods to drive broad causal hypothesis generation using administrative data which has previously been considered unfit for causal inference due to low clinical content and coding errors. Another significance of CICT is the departure from the conventional experimental or observational study design paradigm for identifying and measuring correlations and causal relations in healthcare. In their seminal paper "Causation and causal inference in epidemiology" K. J. Rothman et al. state, "Philosophers agree that causal propositions cannot be proved, and find flaws or practical limitations in all philosophies of causal inference. Hence, the role of logic, belief, and observation in evaluating causal propositions is not settled. Causal inference in epidemiology is better viewed as an exercise in the measurement of an effect rather than as a criterion-guided process for deciding whether an effect is present or not"(*53*). Despite their limitations, observational studies are often the only way to address many important causal questions(*54*). Thus, observational studies are a necessary part of our causal toolbox. Here we show how observational data provides the simple transition rates between clinical conditions and carry valuable information to reveal causal relations even without using contextual information such as age, gender, race or clinical factors.

The possibility of identifying causal networks from their compositional behavior reveals new facets of causality and serves as an additional tool for system identification in readily available and low-cost Markov Chain data. The use of novel emergent properties of network data for causal inference expands our understanding of causality.

References

1. A. Alyass, M. Turcotte, D. Meyre, From big data analysis to personalized medicine for all: challenges and opportunities. *BMC medical genomics* **8**, 33 (2015).
2. S. Green, H. Vogt, Personalizing medicine in silico and in socio. (2016).
3. A. Ward, Causal criteria and the problem of complex causation. *Medicine, Health Care and Philosophy* **12**, 333-343 (2009).
4. J. P. Vandenbroucke *et al.*, Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. *PLoS Medicine* **4**, e297 (2007).
5. J. Pearl, Causal inference in statistics: An overview. *Statistics surveys* **3**, 96-146 (2009).
6. J. P. Vandenbroucke, A. Broadbent, N. Pearce, Causality and causal inference in epidemiology: the need for a pluralistic approach. *International Journal of Epidemiology* **45**, 1776-1786 (2016).
7. W. R. Shadish, T. D. Cook, D. T. Campbell, *Experimental and quasi-experimental designs for generalized causal inference*. (Wadsworth Cengage learning, 2002).
8. D. T. Campbell, H. Riecken, Quasi-experimental design. *International encyclopedia of the social sciences* **5**, 259-263 (1968).
9. T. D. Cook, Quasi-experimental design. *Wiley Encyclopedia of Management*, (2015).
10. Y. Chen, Rangarajan, G., Feng, J. & Ding, M. Phys. , Analyzing multiple nonlinear time series with extended Granger causality. 26–35 (2004).
11. A. Nicola *et al.*, Radial basis function approach to nonlinear Granger causality of time series. *Physical Review E* **70**, 056221 (2004).
12. G. S. Munch *et al.*, Detecting Causality in Complex Ecosystems. (2012).
13. S. J. Schiff *et al.*, Detecting dynamical interdependence and generalized synchrony through mutual prediction in a neural ensemble. *Physical Review E* **54**, 6708 (1996).
14. K. Hlavackovaschindler, M. Palus, M. Vejmelka, J. Bhattacharya, Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports* **441**, 1-46 (2007).
15. M. Paluš *et al.*, Synchronization as adjustment of information rates: Detection from bivariate time series. *Physical Review E* **63**, 046211 (2001).
16. M. Vejmelka, A. o. S. o. t. C. R. Institute of Computer Science, Praha, Czech Republic and Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University, Praha, Czech Republic, M. Paluš, A. o. S. o. t. C. R. Institute of Computer Science, Praha, Czech Republic, Inferring the directionality of coupling with conditional mutual information. *Physical Review E* **77**, 026214 (2008).
17. J. H. Feldhoff, R. V. Donner, J. F. Donges, N. Marwan, J. Kurths, Geometric detection of coupling directions by means of inter-system recurrence networks. *Physics Letters A* **376**, 3504-3513 (2012).
18. Y. Hirata, T. U. o. T. Institute of Industrial Science, Tokyo 153-8505, Japan, K. Aihara, T. U. o. T. Institute of Industrial Science, Tokyo 153-8505, Japan, Identifying hidden common causes from bivariate time series: A method using recurrence plots. *Physical Review E* **81**, 016203 (2010).
19. T. Schreiber, N. S. Max Planck Institute for the Physics of Complex Systems, 01187 Dresden, Germany, Measuring Information Transfer. *Physical Review Letters* **85**, 461 (2000).

20. H. Ma, K. Aihara, L. Chen, Detecting causality from nonlinear dynamics with short-term time series. *Scientific Reports, Nature* **4**, 7464 (2014).
21. L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web., (1999).
22. P. Pons, M. Latapy, Computing communities in large networks using random walks (long version). (2005).
23. M. Rosvall, C. T. Bergstrom, Maps of random walks on complex networks reveal community structure. (2008).
24. F. J. Bruggeman, H. V. Westerhoff, The nature of systems biology. *TRENDS in Microbiology* **15**, 45-50 (2007).
25. M. L. Petersen, S. E. Sinisi, M. J. van der Laan, Estimation of direct causal effects. *Epidemiology* **17**, 276-284 (2006).
26. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. **49**, 764–766 (2013).
27. T. Pham-Gia, T. L. Hung, The mean and median absolute deviations. *Mathematical and Computer Modelling* **34**, 921-936 (2001).
28. A. Azzalini, T. D. Cappello, S. Kotz, Log-Skew-Normal and Log-Skew-t Distributions as Models for Family Income Data. *11*, (2007).
29. J. R. M. Hosking, L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *Journal of the Royal Statistical Society. Series B (Methodological)* **52**, 105-124 (1990).
30. Estimation of quantile mixtures via L-moments and trimmed L-moments. **51**, 947–959 (2006).
31. P. Domingos, A few useful things to know about machine learning. *Communications of the ACM* **55**, 78-87 (2012).
32. .
33. SEDD Database Documentation. Healthcare Cost and Utilization Project (HCUP). December 2017. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/db/state/sedddbdocumentation.jsp; Last modified 9/13/17.
34. H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, T. C. Rindflesch, SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics (Oxford, England)* **28**, 3158-3160 (2012).
35. L. Kaufman, P. J. Rousseeuw, in *Finding Groups in Data*. (John Wiley & Sons, Inc., 2008), pp. 68-125.
36. J. M. Santos, M. Embrechts, in *Artificial Neural Networks – ICANN 2009: 19th International Conference, Limassol, Cyprus, September 14-17, 2009, Proceedings, Part II*, C. Alippi, M. Polycarpou, C. Panayiotou, G. Ellinas, Eds. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2009), pp. 175-184.
37. Verification, validation, and confirmation of numerical mode. *Science* **263**, (1994).
38. R. Fluss, D. Faraggi, B. Reiser, Estimation of the Youden Index and its Associated Cutoff Point. *Biometrical Journal* **47**, 458-472 (2005).
39. H. Yao, I. Rahman, Current concepts on the role of inflammation in COPD and lung cancer. *Current opinion in pharmacology* **9**, 375-383 (2009).
40. J. P. de-Torres *et al.*, Lung cancer in patients with chronic obstructive pulmonary disease. Development and validation of the COPD Lung Cancer Screening Score. *American journal of respiratory and critical care medicine* **191**, 285-291 (2015).

41. S. M. Iacus, G. King, G. Porro, CEM: software for coarsened exact matching. *Journal of statistical Software* **30**, 1-27 (2009).
42. S. M. Iacus, G. King, G. Porro, J. N. Katz, Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 1-24 (2012).
43. S. M. Iacus, G. King, G. Porro, CEM: software for coarsened exact matching. *Journal of Statistical Software* **30**, 1-27 (2009).
44. N. Mantel, W. Haenszel, Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute* **22**, 719-748 (1959).
45. W. G. Cochran, Some methods for strengthening the common χ 2 tests. *Biometrics* **10**, 417-451 (1954).
46. A. SHOJAEE, N. Kaminski, M. A. Pisani, Y. Liu, J. M. Siner, in *C46. CRITICAL CARE: HUMAN TOUCH - PATIENT AND FAMILY ENGAGEMENT AND PALLIATIVE CARE*. pp. A5081-A5081.
47. A. SHOJAEE *et al.*, in *A53. ASTHMA EPIDEMIOLOGY*. pp. A1913-A1913.
48. A. Zinchuk, K. Yaggi, Z. Wenlan, J. M. Siner, A. Shojaee, in *A77. EFFECT OF SLEEP DISORDERED BREATHING ON CARDIOMETABOLIC AND NEUROCOGNITIVE OUTCOMES*. pp. A7694-A7694.
49. S. Ciss, Variable Importance in Random Uniform Forests. (2015).
50. H. Ma, K. Aihara, L. Chen, Detecting Causality from Nonlinear Dynamics with Short-term Time Series. *Scientific Reports* **4**, 7464 (2014).
51. C. Sima, J. Hua, S. Jung, Inference of gene regulatory networks using time-series data: a survey. *Current genomics* **10**, 416-429 (2009).
52. J. Karvanen, Estimation of quantile mixtures via L-moments and trimmed L-moments. *Computational Statistics & Data Analysis* **51**, 947-959 (2006).
53. K. J. Rothman, S. Greenland, Causation and Causal Inference in Epidemiology. *http://dx.doi.org/10.2105/AJPH.2004.059204*, (2011).
54. J. Reiter, Using statistics to determine causal relationships. *The American Mathematical Monthly* **107**, 24-32 (2000).

Code Availability: The required code for training and testing machine learning methods used in this research and for reproducing the results will be available on Github and on a specific website that is under design for this project.

Data Availability: The patient transition network datasets generated and analyzed in this study will be available on '*figshare'* repository.

Acknowledgments: We thank Naftali Kaminsky, Harlan Krumholz, Ronald Coifman, Shu-Xia Li, Paul Horak, Andreas Coppi, and Sudhakar Nuti for their comments and Helen Arjmandi for visualizations and assistance with this work. The section for Pulmonary, Critical Care & Sleep Medicine of internal medicine department and the Center for Outcomes Research and Evaluation, Yale University supported this research.

Figure Legend

Figure 1: General concepts and data used for CICT method

(**A**) Descriptive statistics on the data. (**B**) A set of transitions for four hypothetical patients. For example, the first patient is hospitalized with a principal diagnosis of condition X, and after a period is rehospitalized with condition Y and so forth. If we start merging similar transitions, the result would be the transition graph shown in (**C**) Different types of transitions on a network. X is a common cause of Y and Z where Y is an early effect and Z is a late effect. X and W showed an association without an observable causal relation. Numbers represent hypothetical frequencies. (**D**) 4 zones that carry different distributional information. It is important to note that the four areas are not overlapping and contain different information. Here **i** represents source and **j** represent destination. (**E**) The log-scale density graph shows the different distributions in four distribution zones for a cause: Rheumatoid Arthritis(red), an effect: Syncope (yellow) and a random event: Pneumonia (blue). (**F**) The eight distribution zones that are identified above carry information relevant to the nature of the transition between source and destination. Zones 1,3,5,7 capture distribution of the parameters that are derived from Confidence calculation. Zones 2,4,6,8 capture distribution of the parameters that are derived from Contribution calculation.

Figure 2: Evaluation results of Experiments using CICT

Left Column: CICT shows high accuracy in discriminating random transitions from associations. (**A**) ROC curve. (**B**) Calibration plot. (**C**) Discrimination box plot (**D**) Two clusters as identified by Partitioning Around Medoids along with the real class of data points. **Right Column:** CICT performs well in identifying direction of association: (**E**) ROC curve. (**F**) Calibration plot. (**G**) Discrimination box plot (**H**) Partitioning Around Medoids

Figure 3: Top 8 important predictors of causal transitions

(A) Important discriminating predictors of Causal versus Random relations (**B)** Distribution of most important discriminating predictors of causal transitions (red) from random ones (blue). A logarithmic scale used for x and y. Histogram and density area graphs are superimposed to better reflect differences in distribution. Vertical dash lines show the median of the distribution. Red point shows the location of a causal relation ‘Rheumatoid arthritis → Arthropathy, unspecified, pelvic region and thigh” and blue triangle shows the location of irrelevant transition ‘Osteoarthrosis → Acute appendicitis”. The significant differences of distribution suggest that they are the result of different generative processes, confirmed by a two-sample Kolmogorov-Smirnov test.