Causal Inference in Microbiomes Using Intervention Calculus

Musfiqur Rahman Sazal[1], Vitalii Stebliankin[1], Kalai Mathee[2], Giri Narasimhan[1]

[1] Bioinformatics Research Group (BioRG), School of Computing & Information Sciences
[2] Herbert Wertheim College of Medicine
Florida International University, Miami, Florida, USA
{msaza001, vsteb002, matheek, giri}@fiu.edu

Abstract. Inferring causal effects is critically important in biomedical research as it allows us to move from the typical paradigm of associational studies to causal inference, and can impact treatments and therapeutics. Association patterns can be coincidental and may lead to wrong inferences in complex systems. Microbiomes are highly complex, diverse, and dynamic environments. Microbes are key players in health and diseases. Hence knowledge of genuine causal relationships among the entities in a microbiome, and the impact of internal and external factors on microbial abundance and interactions are essential for understanding disease mechanisms and making treatment recommendations.

In this paper, we investigate fundamental causal inference techniques to measure the causal effects of various entities in a microbiome. In particular, we show how to use these techniques on microbiome datasets to study the rise and impact of antibiotic-resistance in microbiomes. Our main contributions include the following. We introduce a novel pipeline for microbiome studies, new ideas for experimental design under weaker assumptions, and data augmentation by context embedding. Our pipeline is robust, different from traditional approaches, and able to predict interventional effects without any controlled experiments. Our work shows the advantages of causal inference in identifying potential pathogenic, beneficial, and antibiotic-resistant bacteria. We validate our results using results that were previously published.

Keywords: Causal Inference · Intervention · Microbiome · Antibiotic resistance · Causal effects.

1 Introduction

Inferring causality is the process of connecting a cause with an effect. Identifying even a single causal relationship from data is often worth more than observing dozens of correlations in a data set. The study of causality is not new in many areas of science, but in recent years with advances in causal calculus, data science, and machine learning, the focus is now on how to draw causal conclusions in a data-driven way. Given a sufficiently large and rich data set, the theoretical foundations of causality allows us to go beyond merely discovering statistical associations in data, to infer quantitative causal relationships and to explore "what-if" questions, thus profoundly impacting data-driven decision making in many domains. In the field of biomedicine, inferring causal relationships could impact treatment and therapy.

Causal inference can be achieved if a causal structure is readily available based on prior knowledge from experts (e.g., exercise reduces cholesterol). However, in most real-world situations, this is not known. Alternatively, causal inference is also possible if extensive experimentation is possible along with the ability to control all variables in play. In most applications, controlling all variables is impossible (e.g., setting the abundance of bacteria A or the concentration of metabolite M in a person's gut to specific values). If both options are unavailable, but extensive observational data is available, then we rely on the fact that we can test whether a causal model fits the data, even though no experimental manipulation has been carried out. Artificial intelligence already showed huge success in many domains, for example, computer vision [60, 22, 46] and speech recognition [20]. However, causal inference has been compared to human level intelligence [38] and recently has been successfully applied to data from education [38], economics [54], online advertising [5], medicine and epidemiology [25, 9], social sciences [37], natural language processing [59], policy evaluation [51], recommendation systems [4], and much more.

A *microbiome* is a community of microbes including bacteria, archaea, protists, fungi and viruses that share an environmental niche [41]. Microbiomes have been referred to as a *social network* because of the complex set of potential interactions between its members [15, 14], their products and their host. These interactions take the form of cellular communications, cooperation, competition, and much more. All animals and plants live in close association with communities of microbes [30]. These niche communities are highly dynamic. Human bodies harbor rich communities of microbes mostly in the gastrointestinal and reproductive tracts, and on cutaneous and mucosal surfaces such as the skin and the oral cavity [11, 40]. Bacteria (and microbes, in general) play an essential role in human health by helping in a variety of routine processes including digestion, immune responses, and synthesis of useful vitamins and other metabolites. Interactions between microbes in these communities can impact the genes they express and the metabolites they produce or utilize, and can therefore impact the health of the host or the environmental niche [13]. In a *symbiotic* microbiome, many microbial taxa play a useful role leading to a healthy ecosystem. An imbalance (dysbiosis) in the microbial community is strongly associated with a variety of human diseases [34], often by producing harmful metabolites or by preventing the production of sufficient quantities of necessary products [2]. Thus, inferring causal relationships among the entities of a microbiome and with their hosts are crucial for selection of treatments and recommendation of probiotics [7].

In this paper we investigate the causal relationships between microbes and other entities of the microbiome in subjects with *Inflammatory Bowel Disease* (IBD), with special emphasis on the causal effects of different antibiotics and on the resulting rise of antibiotic resistance in different taxa in the microbiome.

Dysbiosis of the gut microbiome is associated with IBD, colorectal cancer, obesity, and much more. However, the relationships between microbial taxa are complex and the experiments required to understand the causal mechanisms are expensive and time-consuming, and therefore remain poorly understood. Another major threat to public health is the rise of *antibiotic resistance* [56], resulting from the overuse, misuse and abuse of antibiotics. While there is no denying the value of antibiotic treatments to combat infectious diseases [3], the need to study antibiotic resistance as a microbial community characteristic is well recognized as a high priority [21].

Causality in microbiomes is a recent topic of research interest. Bourrat and Fishbach et al. discussed broad ideas about causal inference in microbiomes [6, 16]. Sanna et al. studied causality in microbiome using bidirectional Mendelian randomization [45]. Sazal et al. showed how to extract directional relationships among the taxa from oral microbiomes [48, 47]. Ramakrishnan et al. studied causal relationships in microbiomes related to upper airway diseases [42]. This paper approaches causality in microbiomes using a data-driven approach, drawing whenever possible from appropriate knowledgebases. The only other data-driven approach

we found on microbiomes was the work of Mainali et al. [27], where the authors focused on Granger causality, which infers causality from time series data.

2 Causal Inference

The first step in inferring causality is to learn the *causal relationships* (also called causal discovery or causal search), which entails discovering the structure of the relationships. The second step is to use the structure to infer the *causal effects*, i.e., the magnitude of the strength of causal relationships.

2.1 Causal Discovery

The goal of causal discovery is to establish causal relationships between the entities from observed data or using domain knowledge. A particular type of Bayesian network (BN) is often used to encode such relationships. A BN, sometimes called a belief network or causal network, is a *Probabilistic Graphical Models* (PGMs) that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG). A causal network is a BN where the edges correspond to direct causal relationships. In a causal network or causal BN, the parents of each vertex are its presumed direct causes. The direct (and indirect) causes of X_i are the variables that, when varied, will change the distribution of X_i [35].

Formally, we define *causal structures* (CS) (or *causal Bayesian networks*) as a class of *PGMs* [36, 24] where each node represents one of n random variable from a set, $\mathbf{X} = \{X_i, i = 1, \ldots, n\}$, and each edge represents a direct causal relationship. These structures are represented as a graph $G = (V, E)$, where each vertex in V represents a random variable from $\mathbf{X}$, and E is the set of edges. Although undirected edges are used in cases where the direction cannot be reliably determined or when both directions appear to be valid, the graph G is often "manipulated" to be a Directed Acyclic Graph (DAG). Each random variable X_i has an associated probability distribution. A directed edge in E between two vertices represents direct stochastic dependencies. Therefore, if there is no edge connecting two vertices, the corresponding variables are either marginally independent or conditionally independent (conditional on the rest of the variables, or some subset thereof). The "local" probability distribution of a variable X_i depends only on itself and its parents (i.e., the vertices with directed edges into the node X_i); the "global" probability distribution, $P(\mathbf{X})$ is the product of all local probabilities, i.e., a joint distribution [49], given by

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i | Parents(X_i)). \tag{1}$$

Note that the equation is simpler when the causal structure is sparser. Thus, an important step in our pipeline is to identify all independent pairs of random variables. More importantly, we also identify as many conditionally independent pairs as possible since these represent indirect or non-causal relationships.

All local structures in a causal structure can be classified into three sub-categories: *chains*, *forks*, and *colliders*. In a chain, two variables X and Y are conditionally independent given Z, if there is only one unidirectional path between X and Y, and Z is the set of variables that intercepts that path. In a fork, variable Z is a "common cause" for variables X and Y; this happens when there is no directed path between X and Y, and they are independent conditional on Z. Finally, variable Z is a "collider" node between X and Y, if it is the "common-effect". In a collider, as in the fork, there is no directed path between X and Y. However, the difference is that X and Y are unconditionally independent, but become dependent when conditioned on Z and any descendants of Z.

In general, causal models can be very complex. A pair of variables can be connected through multiple chains, forks, and colliders, making it non-trivial to determine the dependency between two arbitrary variables. *Directional separation* (or, just *d-separation*) is a useful concept in this context [18] because covariance terms corresponding to d-separated variables are equal to 0. In a directed graph, G, two vertices x and y are d-connected if and only if G has a collider-free path connecting x and y. More generally, if X, Y and Z are disjoint sets of vertices, then X and Y are d-connected by Z if and only if G has a path P between some vertex in X and some vertex in Y such that for every collider C on P, either C or a descendant C is in Z, and no non-collider on P is in Z. X and Y are d-separated by Z in G if and only if they are not d-connected by Z in G. The concept of d-separation allows for more edges to be eliminated in a causal structure.

2.2 Intervention

Intervention measures the impact of an action and can be thought of as the effect of "doing/intervening." It helps to answer interventional questions of the type: "if a person consumes a specific antibiotic, how will the abundance of taxon A in her gut change?" or "what is the expected abundance of *B. longum* if the relative abundance of *C. difficile* is fixed at 0.1?" Note that a controlled experiment can potentially answer such interventional questions, but may be either prohibitively expensive, impossible, or unethical to perform. Causal calculus allows us to answer such interventional questions in an *in silico* manner. We clarify that data collected from research studies (e.g., a microbiome study) are observational data, and not the result of controlled interventions, which require that variables be artificially held at specific values. Conditional expectation is given by $E[Y|X = x]$, while intervention is given by $E[Y|\text{do}(X = x)]$, which is the expectation of Y if every sample in the population had variable X fixed at value x. Observational distribution $P(y|x)$ is different from interventional distribution $P(y|\text{do}(x))$. Observational distribution describes that the distribution of Y given that variable X takes value x is observed. On the other hand, interventional distribution of Y is what we would observe if we intervened in the data generating process by artificially forcing the variable X to take value x, but data of other variables remain same. Pearl showed how to compute interventions in a causal model [39]. This is done by "mutilating" the model – to achieve $\text{do}(X = x)$, delete all incoming edges to node X, fix its value at x, and then perform computations on the resulting network.

2.3 Intervention Calculus

Consider the n random variables $X_1, \ldots, X_n$ and let pa_j denote the parents of X_j. Any distribution that is generated from a causal structure can be factorized as

$$f\left(x_1, \ldots, x_n\right) = \prod_{j=1}^{n} f\left(x_j|\text{pa}_j\right) \tag{2}$$

A distribution generated from a DAG with independent error terms results in a Markovian model for which an intervention $do(X_i = x)$ on the set of variables $X_i, \ldots, X_n$ is given by the following formula

$$f(x_1, \ldots, x_n|do(X_i = x)) = \begin{cases} \prod_{j=1, j\neq i}^{n} f(x_j|pa_j)|_{x_i=x}, & \text{if } x_i = x \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $f(x_j|pa_j)$ are pre-intervention conditional distribution. The above formula uses the causal structure to write interventional distribution on the left-hand side in terms of pre-intervention conditional distributions on the right hand side. It is possible to summarize the distribution generated by an intervention by its mean

$$E[X_n|do(X_i = x)] = \begin{cases} E[X_n], & \text{if } X_n \in pa_i \\ \displaystyle\int E[X_n|x, pa_i] f(pa_i) \mathrm{d}pa_i, & \text{if } X_n \notin pa_i \end{cases} \tag{4}$$

Assuming that the joint distribution of n random variables is Gaussian, the causal effect of X_i on X_n is given as

$$C(X_i, X_n) = \frac{\partial}{\partial x} E[X_n|\text{do}(X_i = x)], \tag{5}$$

and γ becomes constant because of linearity assumption. Since normality implies that $E(X_n|pa_i, X_i = x)$ is linear in x and pa_i we can express the expectation value using the following equation

$$E\left(X_n|pa_i, X_i = x\right) = \gamma_0 + \gamma_i x + \gamma_{pa_i}^T pa_i, \tag{6}$$

for some values $\gamma_0, \gamma_i \in \mathbf{R}$. The causal effect of X_i on X_n with $X_n \notin pa_i$ is denoted by $C(X_i, X_n)$ and equals the regression coefficient of x_i above. Thus,

$$C(X_i, X_n) = \gamma_i, \tag{7}$$

where γ_i is as dictated by Eq. (6).

3 Methods and Experiments

In our novel approach, we applied intervention calculus or do-calculus to the causal network constructed from microbiome data sets to (a) determine the causal effects of each microbial taxa on other microbial taxa, and (b) to determine the effects of antibiotics on different microbial taxa. Our proposed method is as follows: (1) learn a causal graph, (2) compute causal effects, (3) analyze the role of causally significant microbial taxa.

3.1 Problem Formulation

The problem formulation for causal effects among taxa is as follows: Let $T = \{B_1, B_2, \ldots, B_n\}$ be the set of microbial taxa present in the cohort of healthy or disease samples with abundance values $\{b_1, b_2, \ldots, b_n\}$. For two taxa $\{B_i, B_j\} \in T$, the causal effect of B_i on B_j is given by

$$C(B_i, B_j) = \frac{\partial}{\partial b} E[b_j | \text{do}(b_i = b)].$$

We computed causal effects for all pairs in T and ranked all taxa according to the sum of absolute values of causal effect on all the other taxa, with the hope of identifying the most influential taxa in the microbiome. Using the above ranking, we considered the top 30% of taxa for further analysis.

Similarly, to study the causal effects of different antibiotics on the taxa, we let $T = \{A_1, A_2, \ldots, A_n\}$ be the set of antibiotics applied, and let $O = \{b_1, b_2, \ldots, b_n\}$ be the set of abundances of the microbial taxa $\{B_1, B_2, \ldots, B_n\}$ for those samples. (T is for treatment or interventional variable, O is for outcome variable in this context.) Causal effect of an antibiotic A_i on the abundance of a microbial taxon b_i is $C(A_i, b_i) = \frac{\partial}{\partial t} E[b_i | \text{do}(A_i = t)]$. Since causal effects can be positive or negative, we computed causal effects for all pairs in $T \times O$ and separately ranked the pairs with positive and negative causal effects for further analysis.

3.2 Data

We analyzed five data sets related to IBD: three from Integrative Human Microbiome Project (iHMP) [1] and two from MicrobiomeHD database [12]. The iHMP IBD data set includes multiomics data from subjects with Crohn's Disease (CD), ulcerative colitis (UC), and non-IBD (i.e., healthy), all of which were used in this study. MicrobiomeHD database includes 28 published case-control gut microbiome studies spanning ten diseases, from which we chose data sets associated with *C. difficile* infections and enteric diarrhea. These choices were made because the role of many taxa for those diseases are reasonably well established. Table 1 gives a summary of IBD related data sets. For each data set, we computed the total causal effect of each microbial taxon on all other taxa.

Table 1. Description of IBD-related data set

Database	**Data Set**	**# of Samples**
iHMP	Ulcerative colitis (UC)	459
	Crohn's disease (CD)	749
	non-IBD (healthy)	429
Data (MicrobiomeHD)	**Cases**	**Controls**
EDD (Enteric diarrhea)	201	82
CDI (*C. difficile* infection)	93	154

To explore the effect of antibiotics on antibiotic-resistant taxa, we analyzed a dataset from Gibson et al. [19]. It consists of 401 stool metagenomic samples from 84 premature infants that were sampled in multiple time points. All but two infants received antibiotic therapy within the first 24 hours. Sixty-one percent of the infants received additional antibiotic treatments ("Antibiotic" cohort) between 1–10 weeks of life. The remaining 39 percent formed the "Control" group. Each treatment consisted of one or more antibiotics. We considered each measurement as a separate sample from a distribution, and did not take into account the temporal aspect of the measurement as showed in [17].

3.3 Experiments

For each data set, we generated a causal structure by applying the PC-stable algorithm [10] and we computed the causal effect of each microbial taxon on all other taxa. We also computed the change in causal effects in healthy (non-IBD) versus diseased states.

To understand the causal role of microbial taxa in disease mechanisms we augmented the data set by merging healthy and disease abundance matrices and by adding an extra node named 'disease'; we call this process 'context embedding'. Context embedding is important for causal inference because in different contexts, the same event can be interpreted differently. For the healthy state, the value of disease node is 0, and for the disease state its value is 1. Thus the disease node becomes a binary random variable. We computed the causal effect of all taxa on disease, and vice versa.

For the antibiotic data set, as part of preprocessing, we profiled metagenomic reads against 14506 complete bacterial, archaeal, and viral genome sequences from RefSeq v.92 using FLINT [53] framework. Reference genomes were obtained from a repository hosted by the Kraken [58] tool. After obtaining abundance matrix, we created a causal network using recently used antibiotics and relative abundance of bacterial taxa. We computed causal effects of each antibiotic on each taxon and vice versa, to be used for further analysis.

4 Results and Analysis

IBD and non-IBD Data Fig. 1 shows a causal structure inferred from ulcerative colitis (UC) samples. Fig. 2 represents a causal network combining data from non-IBD and UC samples, but with an additional "disease" node (colored blue). Fig. 3 shows a causal network with nodes representing antibiotics and microbial taxa. Some more networks are shown and explained in the Appendix. In each network, nodes represent random variables for relative abundance of taxa, disease status, or antibiotic dosages. Edges represent their conditional relationships. The size of each node is proportional to sum of relative abundance and the color of the edges represents the sign of the correlation between the node variables. In a causal structure, directed edges suggest potentially direct causal effect between the connected variables. The absence of an edge suggests that there is no direct causal effect, although indirect causal effects may exist. An inferred causal structure may contain undirected edges if the data are not enough to support an edge orientation. Those undirected edges remain causally "uninterpretable".

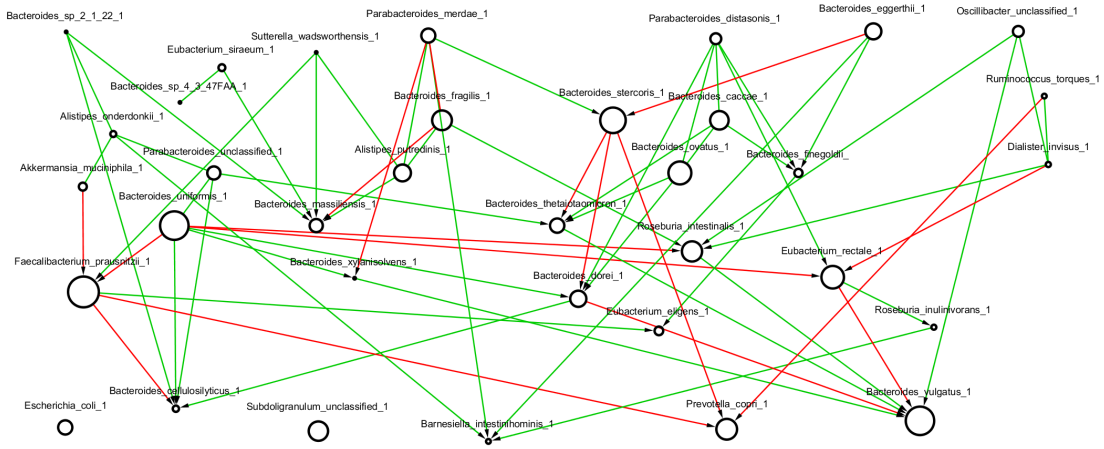

Fig. 1. Causal network from Ulcerative Colitis (UC) data. All nodes represent taxa abundance.

We looked at the distribution of causal effects for each data set as shown in Fig. 4. Most of the causal effects are relatively small (see peak centered at 0). Approximately 30% of the causal effects are relatively large. The top 15% (shown in green rectangle) and bottom 15% (shown in red rectangle) are zoomed in for

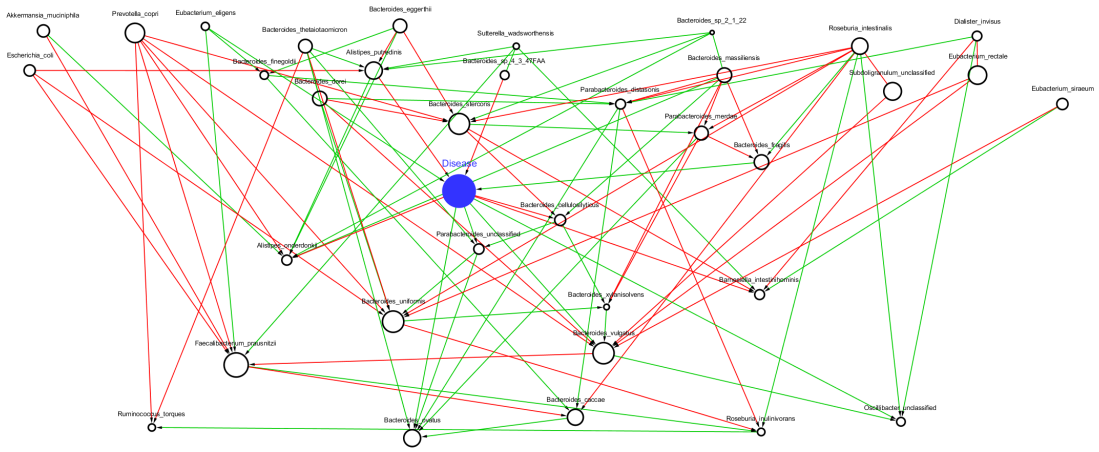

Fig. 2. Causal network after introducing a "disease" node and using data from UC and non-IBD samples. Disease node is shown as a filled blue node.

clarity. Based on sum of absolute values of causal effects we ranked the taxa as shown in Fig. 5. The hope is that this list shows the taxa that play a significantly key role in health and/or disease.

More interesting patterns are visible if we look at the changes in total causal effects between non-IBD (healthy) and UC taxa. Fig. 5 shows the ten taxa with the highest change in total causal effects. Green bars indicate higher total causal effect values in healthy, while red bars indicate higher values in UC, suggesting that the taxa on the left of the chart are potentially playing a beneficial role in health individuals while the taxa on the right of the chart are playing a harmful role in UC. Thus, in non-IBD subjects, the bacterial taxa *B. xylanisolvens, E. eligens, B. finegoldii, B. ovatus* and some species of *Oscillobacter* have more causal impact on the remaining taxa than others. These claims are supported by published literature, which show those taxa are potentially beneficial [28, 52, 32, 57]. On the other hand, in the diseased state (UC), other taxa including *R. torques, B. massiliensis, P. distasonis*, and *D. invisus* are more impactful. Again, the published literature supports the above claims [29, 26, 55, 33]. Thus our methods allow us to identify potentially beneficial and pathogenic bacteria in microbiomes.

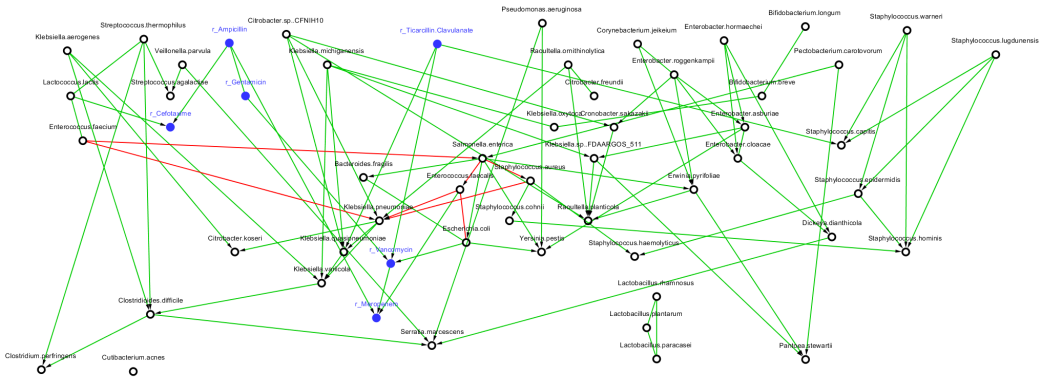

Fig. 3. Causal network with antibiotics and taxa from samples obtained during or immediately after the antibiotic treatment. Blue nodes represent antibiotic doasges and black nodes represent taxa abundance.

Disease networks We measured causal effects of each taxa on the specially identified "disease node". We sorted the absolute value of the effects as shown in Table 2. When we queried the published literature on this topic, we discovered that barring two, all the taxa listed in Table 2 are known to be either potentially pathogenic or beneficial, again supporting the claim that our approach helps to identify pathogenic and beneficial bacteria in healthy and diseased patients. We also computed total causal effects of disease on individual taxa. Most of the causal effects values are nearly zero, with values in the range of $[-0.02, 0.02]$. We then listed taxa with values greater than 0.01 or less than -0.01. Our findings suggest that the disease affects some commensal bacteria to become more abundant, while shrinking the abundance of potentially beneficial bacteria. This claim is supported by published literature as summarized in Table 3.

Table 2. Taxa with highest causal effect on the *ulcerative colitis* "disease" node.

Cause	Effects on disease node	Potential behavior
S. wadsworthensis	6.348413325	Beneficial
B. xylanisolvens	5.612681215	Beneficial
B. intestinihominis	3.645354302	Beneficial
B. ovatus	3.002307827	Beneficial
P. unclassified	2.821679346	?
D. invisus	2.807382178	Pathogenic
B. cellulosilyticus	2.577919214	Beneficial
A. putredinis	2.131188842	Beneficial
R. torques	2.002078229	Pathogenic
P. distasonis	1.967066432	?

Table 3. Taxa most impacted by the *ulcerative colitis* "disease" node

Taxa name	Behavior	References
B. uniformis	Commensal	[43]
B. ovatus	Commensal	[44]
E. coli	Commensal	[50]
A. muciniphila	Beneficial	[61]
P. copri	Beneficial	[8]
B. stercoris	Beneficial	[31]
A. putredinis	Beneficial	[23]

Antibiotics Data From the antibiotic data set, we learned a causal network and computed causal effects of each antibiotic to each taxa. The top seven positive causal effects and top seven negative causal effects are shown in Table 4 since all other values were close to the background noise. A positive causal effect of an antibiotic on a taxon suggests that either the antibiotic is inappropriate for that particular taxon or the taxon is resistant to the antibiotic. Similarly, a negative causal effect suggests that the antibiotic was effective against the taxon.

The significance of the causal graph shown in Figure 3 is that we are able to separate the effect of each antibiotic on a taxon. In general, doctors often administer combinations of antibiotics. In our study, 17 unique combinations of antibiotics were used to treat infants (one, two, or three types at a time). Therefore, it is difficult to identify which particular drug is associated with the rise of antibiotic-resistant taxa, or which antibiotic is most effective against pathogenic bacteria. Our proposed methods appear to be able to deconvolve these effects.

As mentioned above, the top 7 positive and negative causeal effect values of antibiotics on taxa are shown in Table 4. Except for three (out of 14) cases, our causal effect values can explain the average change in relative abundance that was computed by the previous study of Gibson et al. [19]. In particular, if the causal-effect of an antibiotic on taxa is positive, the average abundance of this taxa is increased and vice

Table 4. Causal effects of antibiotics on taxa

Cause	Effects	Magnitude of Causal Effects	Average Δ abundance
Ticarcillin Clavulanate	*K.pneumoniae*	0.041646824	+
Cefotaxime	*E.coli*	0.037287043	-
Meropenem	*E.faecalis*	0.021974783	+
Vancomycin	*E.coli*	0.016519979	0
Gentamicin	*E.coli*	0.01563053	+
Meropenem	*S.aureus*	0.013329356	NA
Ampicillin	*E.coli*	0.010371501	+
Meropenem	*K.pneumoniae*	-0.023983874	-
Cefotaxime	*E.faecalis*	-0.017125561	+
Cefotaxime	*K.pneumoniae*	-0.01689829	-
Cefotaxime	*E.faecium*	-0.015538356	-
Ticarcillin Clavulanate	*E.faecalis*	-0.009494291	-
Vancomycin	*E.faecalis*	-0.009451166	+
Ticarcillin Clavulanate	*S.aureus*	-0.009099242	NA

versa. The contradictory results may lead to new insights about antibiotic effectiveness. For example, even though after administering Cefotaxime and Vancomycin the relative abundance of *E. faecalis* on average tended to increase, our causal effect graph suggests that these antibiotics were effective against these taxa and that some other factors may be causing their increased abundance.

5 Conclusion

Causal inference shows promising results in analyzing microbiome data, especially in the identification of potentially pathogenic, beneficial, and antibiotic-resistant bacteria. Thus, in future, this process can allow us to evaluate the efficacy of probiotics and prebiotics. Moreover, causal inference from purely observational data is important to prioritize in picking wet-lab experiments for further anaysis. *Intervention* techniques can be used to quantify the average causal impact of one entity on another. The next challenge is to study the causal effect of one entity on another within a single sample.

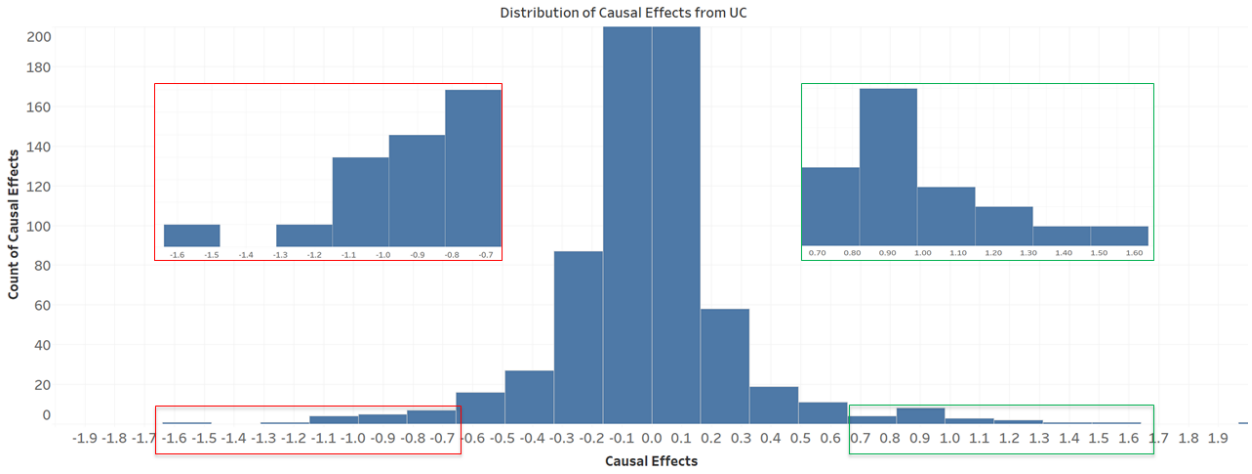

Fig. 4. Histogram of causal effect values in UC. The top (green) and bottom (red) 15% are zoomed in for details.

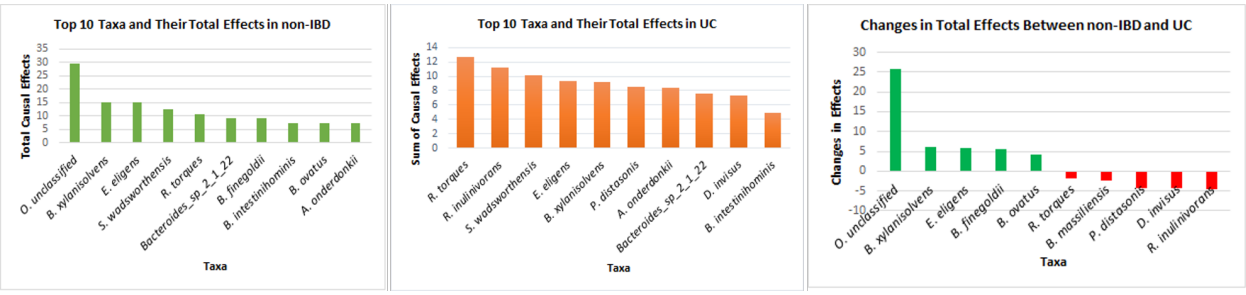

Fig. 5. Top ten causally significant taxa from non-IBD (left), UC (middle), and Top 10 changes in causal effects from UC to non-IBD (right)

References

1. NIH (Integrative Human Microbiome Project), www.hmpdacc.org/ihmp/, (Date last accessed 06-Aug-2019)
2. Althani, A.A., Marei, H.E., Hamdi, W.S., Nasrallah, G.K., El Zowalaty, M.E., Al Khodor, S., Al-Asmakh, M., Abdel-Aziz, H., Cenciarelli, C.: Human microbiome and its association with health and diseases. Journal of cellular physiology **231**(8), 1688–1694 (2016)
3. Aslam, B., Wang, W., Arshad, M.I., Khurshid, M., Muzammil, S., Rasool, M.H., Nisar, M.A., Alvi, R.F., Aslam, M.A., Qamar, M.U., et al.: Antibiotic resistance: a rundown of a global crisis. Infection and drug resistance **11**, 1645 (2018)
4. Bonner, S., Vasile, F.: Causal embeddings for recommendation. In: Proceedings of the 12th ACM Conference on Recommender Systems. pp. 104–112. ACM (2018)
5. Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D.X., Chickering, D.M., Portugaly, E., Ray, D., Simard, P., Snelson, E.: Counterfactual reasoning and learning systems: The example of computational advertising. The Journal of Machine Learning Research **14**(1), 3207–3260 (2013)
6. Bourrat, P.: Have causal claims about the gut microbiome been over-hyped? BioEssays **40**(12), 1800178 (2018)
7. Brown, J.M., Hazen, S.L.: Targeting of microbe-derived metabolites to improve human health: The next frontier for drug discovery. Journal of biological chemistry **292**(21), 8560–8568 (2017)
8. Cani, P.D.: Human gut microbiome: hopes, threats and promises. Gut **67**(9), 1716–1725 (2018)
9. Cinelli, C., Pearl, J.: On the utility of causal diagrams in modeling attrition: a practical example. Epidemiology **29**, e50–e51 (2018)
10. Colombo, D., Maathuis, M.H.: Order-independent constraint-based causal structure learning. The Journal of Machine Learning Research **15**(1), 3741–3782 (2014)
11. Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I., Knight, R.: Bacterial community variation in human body habitats across space and time. Science **326**(5960), 1694–1697 (2009)
12. Duvallet, C., Gibbons, S.M., Gurry, T., Irizarry, R.A., Alm, E.J.: Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nature communications **8**(1), 1784 (2017)
13. Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., Huttenhower, C.: Microbial co-occurrence relationships in the human microbiome. PLoS Comp Bio **8**(7), e1002606 (2012)
14. Fernandez, M., Aguiar-Pulido, V., Riveros, J., Huang, W., Segal, J., Zeng, E., Campos, M., Mathee, K., Narasimhan, G.: Microbiome analysis: State of the art and future trends. Computational Methods for Next Generation Sequencing Data Analysis pp. 401–424 (2016)
15. Fernandez, M., Riveros, J.D., Campos, M., Mathee, K., Narasimhan, G.: Microbial "social networks". BMC Genomics **16**(11), 1 (2015)
16. Fischbach, M.A.: Microbiome: focus on causation and mechanism. Cell **174**(4), 785–790 (2018)
17. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian networks to analyze expression data. Journal of computational biology **7**(3-4), 601–620 (2000)
18. Geiger, D., Verma, T., Pearl, J.: d-separation: From theorems to algorithms. In: Machine Intelligence and Pattern Recognition, vol. 10, pp. 139–148. Elsevier (1990)
19. Gibson, M.K., Wang, B., Ahmadi, S., Burnham, C.A.D., Tarr, P.I., Warner, B.B., Dantas, G.: Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. Nature Microbiology **1**(4), 16024 (Mar 2016). https://doi.org/10.1038/nmicrobiol.2016.24, http://www.nature.com/articles/nmicrobiol201624
20. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing. pp. 6645–6649. IEEE (2013)

21. Gupta, S., Arango-Argoty, G., Zhang, L., Pruden, A., Vikesland, P.: Identification of discriminatory antibiotic resistance genes among environmental resistomes using extremely randomized tree algorithm. Microbiome **7**(1), 123 (2019)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
23. Ke, X., Walker, A., Haange, S.B., Lagkouvardos, I., Liu, Y., Schmitt-Kopplin, P., Von Bergen, M., Jehmlich, N., He, X., Clavel, T., et al.: Synbiotic-driven improvement of metabolic disturbances is associated with changes in the gut microbiome in diet-induced obese mice. Molecular metabolism **22**, 96–109 (2019)
24. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press (2009)
25. Lam, K.: Effect of a two-stage intervention package on the cesarean section rate in guangzhou, china: a before-and-after study. PLoS Medicine (2019)
26. Lucke, K., Miehlke, S., Jacobs, E., Schuppler, M.: Prevalence of bacteroides and prevotella spp. in ulcerative colitis. Journal of medical microbiology **55**(5), 617–624 (2006)
27. Mainali, K., Bewick, S., Vecchio-Pagan, B., Karig, D., Fagan, W.F.: Detecting interaction networks in the human microbiome with conditional granger causality. PLoS computational biology **15**(5), e1007037 (2019)
28. Man, S.M., Kaakoush, N.O., Mitchell, H.M.: The role of bacteria and pattern-recognition receptors in crohn's disease. Nature reviews Gastroenterology & hepatology **8**(3), 152 (2011)
29. Matsuoka, K., Kanai, T.: The gut microbiota and inflammatory bowel disease. In: Seminars in immunopathology. vol. 37, pp. 47–55. Springer (2015)
30. Mendes, R., Raaijmakers, J.M.: Cross-kingdom similarities in microbiome functions. The ISME journal **9**(9), 1905 (2015)
31. Microbewiki: Bacteroides, https://microbewiki.kenyon.edu/index.php/Bacteroides
32. Moore, W., Moore, L.H.: Intestinal floras of populations that have a high risk of colon cancer. Appl. Environ. Microbiol. **61**(9), 3202–3207 (1995)
33. Morio, F., Jean-Pierre, H., Dubreuil, L., Jumas-Bilak, E., Calvet, L., Mercier, G., Devine, R., Marchandin, H.: Antimicrobial susceptibilities and clinical sources of dialister species. Antimicrobial agents and chemotherapy **51**(12), 4498–4501 (2007)
34. Nadal, I., Donant, E., Ribes-Koninckx, C., Calabuig, M., Sanz, Y.: Imbalance in the composition of the duodenal microbiota of children with coeliac disease. Journal of medical microbiology **56**(12), 1669–1674 (2007)
35. Pearl, J.: Causality: models, reasoning and inference, vol. 29. Springer (2000)
36. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Elsevier (2014)
37. Pearl, J.: Detecting latent heterogeneity. Sociological Methods & Research **46**(3), 370–389 (2017)
38. Pearl, J.: Theoretical impediments to machine learning with seven sparks from the causal revolution. arXiv preprint arXiv:1801.04016 (2018)
39. Pearl, J., Glymour, M., Jewell, N.P.: Causal inference in statistics: A primer. John Wiley & Sons (2016)
40. Peleg, A.Y., Hogan, D.A., Mylonakis, E.: Medically important bacterial–fungal interactions. Nature Reviews Microbiology **8**(5), 340 (2010)
41. Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J.A., Bonazzi, V., McEwen, J.E., Wetterstrand, K.A., Deal, C., et al.: The NIH Human Microbiome Project. Genome research **19**(12), 2317–2323 (2009)
42. Ramakrishnan, V.R., Frank, D.N.: Microbiome in patients with upper airway disease: moving from taxonomic findings to mechanisms and causality. Journal of Allergy and Clinical Immunology **142**(1), 73–75 (2018)
43. Rapozo, D.C., Bernardazzi, C., de Souza, H.S.P.: Diet and microbiota in inflammatory bowel disease: The gut in disharmony. World journal of gastroenterology **23**(12), 2124 (2017)
44. Saitoh, S., Noda, S., Aiba, Y., Takagi, A., Sakamoto, M., Benno, Y., Koga, Y.: Bacteroides ovatus as the predominant commensal intestinal microbe causing a systemic antibody response in inflammatory bowel disease. Clin. Diagn. Lab. Immunol. **9**(1), 54–59 (2002)
45. Sanna, S., van Zuydam, N.R., Mahajan, A., Kurilshikov, A., Vila, A.V., Võsa, U., Mujagic, Z., Masclee, A.A., Jonkers, D.M., Oosting, M., et al.: Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. Nature genetics p. 1 (2019)
46. Sazal, M.M.R., Biswas, S.K., Amin, M.F., Murase, K.: Bangla handwritten character recognition using deep belief network. In: 2013 International Conference on Electrical Information and Communication Technology (EICT). pp. 1–5. IEEE (2014)
47. Sazal, M.R., Mathee, K., Ruiz-Perez, D., Cickovski, T., Narasimhan, G.: Inferring directional relationships in microbial communities using signed bayesian networks. bioRxiv (2020)
48. Sazal, M.R., Ruiz-Perez, D., Cickovski, T., Narasimhan, G.: Inferring relationships in microbiomes from signed bayesian networks. In: 2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS). pp. 1–1. IEEE (2018)
49. Scutari, M.: Bayesian network constraint-based structure learning algorithms: Parallel and optimised implementations in the bnlearn r package. arXiv preprint arXiv:1406.7648 (2014)

50. Singh, V., San Yeoh, B., Xiao, X., Kumar, M., Bachman, M., Borregaard, N., Joe, B., Vijay-Kumar, M.: Interplay between enterobactin, myeloperoxidase and lipocalin 2 regulates e. coli survival in the inflamed gut. Nature communications **6**, 7113 (2015)
51. Swaminathan, A., Joachims, T.: Counterfactual risk minimization: Learning from logged bandit feedback. In: International Conference on Machine Learning. pp. 814–823 (2015)
52. Ulsemer, P., Toutounian, K., Schmidt, J., Karsten, U., Goletz, S.: Preliminary safety evaluation of a new bacteroides xylanisolvens isolate. Appl. Environ. Microbiol. **78**(2), 528–535 (2012)
53. Valdes, C., Stebliankin, V., Narasimhan, G.: Large scale microbiome profiling in the cloud. Bioinformatics **35**(14), i13–i22 (07 2019). https://doi.org/10.1093/bioinformatics/btz356, https://doi.org/10.1093/bioinformatics/btz356
54. Varian, H.R.: Causal inference in economics and marketing. Proceedings of the National Academy of Sciences **113**(27), 7310–7315 (2016)
55. Wang, K., Liao, M., Zhou, N., Bao, L., Ma, K., Zheng, Z., Wang, Y., Liu, C., Wang, W., Wang, J., et al.: Parabacteroides distasonis alleviates obesity and metabolic dysfunctions via production of succinate and secondary bile acids. Cell reports **26**(1), 222–235 (2019)
56. Wang, S., George, D., Purych, D., Patrick, D.: Antibiotic resistance: a global threat to public health. BCMJ, 2014, 6: 295-296—BC Centre for Disease Control (2014)
57. Woloszynek, S., Pastor, S., Mell, J., Nandi, N., Sokhansanj, B., Rosen, G.: Engineering human microbiota: influencing cellular and community dynamics for therapeutic applications. In: International review of cell and molecular biology, vol. 324, pp. 67–124. Elsevier (2016)
58. Wood, D.E., Salzberg, S.L.: Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology **15**(3), R46 (Mar 2014)
59. Wood-Doughty, Z., Shpitser, I., Dredze, M.: Challenges of using text classifiers for causal inference. arXiv preprint arXiv:1810.00956 (2018)
60. Wu, Z., Shen, C., Van Den Hengel, A.: Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recognition **90**, 119–133 (2019)
61. Zhai, R., Xue, X., Zhang, L., Yang, X., Zhao, L., Zhang, C.: Strain-specific anti-inflammatory properties of two akkermansia muciniphila strains on chronic colitis in mice. Frontiers in Cellular and Infection Microbiology **9**, 239 (2019)