Sinha *et al.*
BMC Medical Informatics and Decision Making (2025) 25:27
https://doi.org/10.1186/s12911-025-02849-4

BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Causal analysis for multivariate integrated clinical and environmental exposures data

Meghamala Sinha[1*], Perry Haaland[2], Ashok Krishnamurthy[3,4], Bo Lan[5], Stephen A. Ramsey[1], Patrick L. Schmitt[3], Priya Sharma[3], Hao Xu[3] and Karamarie Fecho[3]

Abstract

Background Understanding the causal relationships between clinical outcomes and environmental exposures is critical for advancing public health interventions and personalized medicine. These causal relationships can be applied to augment medical decision-making or suggest hypotheses for healthcare research. In this study, we applied a causal inference algorithm to an EHR dataset on patients with asthma or related common respiratory conditions ($N = 14{,}937$).

Methods The EHR data were accessed via an open service named the Integrated Clinical and Environmental Service (ICEES). A multivariate feature table was extracted that included integrated data on features representing demographic factors, clinical measures, and environmental exposures; namely, sex, race, obesity, prednisone use, airborne particulate matter exposure, major roadway/highway exposure, residential density, and annual number of emergency department (ED) or inpatient hospital visits for respiratory issues, which we used as a proxy for asthma attacks. We estimated underlying causal relationships from the data by applying a Principal Component algorithm to identify significant causal relationships between the extracted features and asthma attacks. We also performed simulated interventions on the inferred causal network to detect the causal effects, in terms of shifts in the probability distribution for annual ED or inpatient hospital visits for respiratory issues.

Results We found that obesity and prednisone were causally related to annual ED or inpatient visits in our causal inference model, and sex and race were indirectly related to annual ED or inpatient visits via a causal relationship to obesity. We further found that interventions in which all patients are simulated as obese or using prednisone (but not female) caused a shift to the right in the probability distribution of annual ED or inpatient visits for respiratory issues, thus supporting the results of our causal analysis, which demonstrated direct effects of obesity and prednisone (but not sex) on asthma attacks.

Conclusions We successfully applied a causal model to the open ICEES service and identified direct causal relationships between prednisone and obesity on the frequency of asthma attacks, with indirect effects of sex and race by way of obesity. Our simulated interventions provided further support for our causal analysis by demonstrating a shift to the right in the probability of asthma attacks with interventions that assume all patients are using prednisone or obese.

Keywords Causal inference, Structured learning, Open clinical data, Asthma

*Correspondence:
Meghamala Sinha
meghamala.sinha@gmail.com
Full list of author information is available at the end of the article

Background

Causal inference [1–3] is an important tool in the domain of health sciences for informatics work such as finding causal effects of an adverse outcome or risk factors for a disease. Causality has traditionally been a core concept across all branches of medical science and considered when diagnosing patients based on their symptoms, effects of treatment, and years of historical evidence [4]. Indeed, the study of causal inference in health science research dates back to the 1970s and 1980s [5, 6]. Yet, non-causal models such as regression tend to be more commonly applied in health science research than causal models [7]. The goals and philosophy of causal inference differ from those of association-based predictions in several ways. For instance, with predictive models such as regression, one wants to measure the likelihood of occurrence of an event as a result of another event; for example, the occurrence of lung cancer based on exposure to smoke in the environment. However, such predictions may be subject to confounding; for instance, a researcher may find that a regional increase in the sale of matches is associated with lung cancer, but not necessarily causally associated if, for example, the association actually reflects a regional increase in match sales due to frequent blackouts, with a secondary unrelated association related to lung cancer and perhaps attributable to exposures such as workplace chemical exposures or lack of healthcare access. Most predictive models, unlike causal inference models, do not readily account for confounding variables and hence cannot differentiate causal versus spurious associations. Another aspect of causal inference that differentiates it from non-causal models is the ability to provide an explanation for the relationship between two events. For instance, causal inference can help to discern why a patient is sick and diagnose them or identify medications to treat them based on the underlying cause of their symptoms.

Electronic health records (EHRs) provide a potential source of structured clinical data such as diagnoses, medications, and laboratory results. Access to EHR data is thus critical for the advancement of health science research and clinical research. However, the many federal and institutional regulations that surround clinical data, while necessary to ensure patient privacy and protection of sensitive data, limit access to the data for research. In this study, we analyzed a patient-level dataset extracted from a regulatory-compliant open service called the Integrated Clinical and Environmental Exposures Service (ICEES) [8]. ICEES supports several use cases, including asthma, drug-induced liver injury, and coronavirus infection. The ICEES data are constructed by integrating clinical data elements derived from patient EHRs and environmental exposures data derived from a variety of public sources of environmental exposures data [9]. The data are then binned and de-identified by stripping all protected health information per the Safe Harbor method of the Health Insurance Portability and Accountability Act. The ICEES data are then exposed via an open application programming interface (OpenAPI). For our principal application use case, we focus on an existing ICEES cohort of patients with asthma or a related common pulmonary disorder (see [8] for details). We asked if there is a causal relationship between asthma attacks and the following features: sex, race, prednisone use, diagnosis of obesity, residential proximity to a major roadway or highway, residential density, and exposure to airborne pollutants. We focus on these features because published studies, including our prior work [8–10], have recognized one or more of them to be associated with asthma attacks. We consider the number of annual emergency department (ED) or inpatient hospital visits for respiratory issues as the primary outcome measure and indicator of asthma attacks, as we have done previously. We first generate a causal inference network. We then demonstrate simulated external interventions as an approach to validate the inferred causal network. We use subject matter expert knowledge and publication support as our ground truth to measure the correctness of our causal inference model. Finally, we discuss our findings, including the benefits and limitations of our causal inference model and approach.

Methods

Generation of multivariate ICEES table

We focused on an existing ICEES cohort of patients with asthma or another common pulmonary disorder and examined outcomes over a one-year study period (see [8] for details on the inclusion and exclusion criteria). In brief, the patients were included if they had at least one diagnosis of asthma and/or another common respiratory disorder, had a prescription or administration of a drug typically used to treat asthma and/or other common respiratory disorders, or had frequent ED visits during which an albuterol nebulizer was administered. The majority of patients included in the final dataset were 45 years of age or older, female, non-Hispanic white, and residing in a rural region.

We asked if there is a causal relationship between asthma attacks and the following features: sex, race, prescriptions for prednisone use, diagnosis of obesity, residential proximity to a major roadway or highway, residential density, and exposure to high levels of airborne pollutants. We selected these features because published studies, including our prior work [8–10], have recognized one or more of them to be associated with asthma attacks. The racial categories were self-reported,

as defined according to our hospital's EHR system, and included to capture potential racial disparities for further investigation into whether those are related to socioeconomic conditions, healthcare access, or other factors. We defined "asthma attacks" based on the annual number of ED or inpatient visits for respiratory issues. This is an acceptable clinical proxy for asthma exacerbations, one that we have applied successfully in prior work [8–10].

We queried the ICEES OpenAPI to generate a multivariate table. We focused on eight ICEES feature variables, namely, TotalEDInpatientVisits, Sex, Race, Prednisone, Obesity, PM2.5 Exposure, RoadwayExposure, and EstResidentialDensity, and TotalEDInpatientVisits, as defined in Table 1. The majority of patients did not have any ED or inpatient hospital visits over the one-year study period and were not active in the year of interest (data not shown), meaning that their EHR did not indicate any healthcare utilization. This finding was expected, but it introduced a skew in the distribution of TotalEDInpatientVisits, with the vast majority of patients grouped as TotalEDInpatientVisits = 0. To minimize the skew, we applied the filter "Active_In Year" before extracting the multivariate table, with Active_In_Year = 1 to select only patients who were active in 2010. In Fig. 1, we show the distribution of TotalEDInpatientVisits among the discrete categories of each feature variable after applying the Active_In_Year filter. With the Active_In_year filter in place, the distribution of TotalEDInpatientVisits indicated that most patients who were active in year 2010 visited the ED or had an inpatient hospital visit at least one time over the year of interest. There was an imbalance in TotalEDInpatientVisits across levels for some feature variables such as Prednisone, Obesity, Race, RoadwayExposure, and PM2.5Exposure. The final multivariate table in this work comprised data on 14,937 patients (i.e., rows represented individual patients in the asthma cohort, and columns represented feature variables). Figure 1 shows the number of TotalEDInpatientVisits across each level of the feature variables.

Evaluation of feature importance

We evaluated the importance of each feature using a tree-based machine learning model: random forest. The random forest analysis was conducted to provide a comparison with the causal network analysis. We leveraged the caret R package [11] to evaluate the feature importance. We controlled the parameters for training by using the *repeatedcv* method to divide our dataset into ten-folds cross-validation and repeated three times.

Causal network analysis

Most of the naturally occurring trends that we come across are simply passive observations of events occurring in the world that are either coincidental or unexplained associations. For example, statements like "drinking beer everyday increases the chance of prostate cancer" are common in the news and scientific reporting and in our day-to-day personal beliefs. These associations can be easily mistaken as causation, making us susceptible to logical fallacies without knowing the real underlying cause. Causal inference is the science of distinguishing cause from effect [1–3]. It is an important field of research because it helps us eradicate spurious correlation [12–14]. The primary aim of inferring causal relations from data is to discover interactions between different entities in the form of $V_i \rightarrow V_j$, where V_i and V_j are observable features in a domain and the arrow indicates that the state of V_i influences the state of V_j. Causal inference can be either discovered through observational measurements (seeing) or from measurements after performing some external manipulation/intervention (doing). A causal network [1–3, 15] can be represented

Table 1 Feature variables used to generate the multivariate table

Feature variable	Variable definition and enumeration
Sex	Male (0), Female (1)
Race	White, Black African American, Asian, Native Hawaiian/Pacific Islander, American/Alaskan Native, Other, Unknown
Prednisone	Common medication for asthma-like conditions (1 = Yes, 0 = No)
Obesity	Diagnostic code for obesity anytime over 'study' period (1 = Yes, 0 = No)
Airborne particulate exposure	Abbreviated herein as "PM2.5Exposure". US Environmental Protection Agency estimated maximum daily exposure to particulate matter ≤ 2.5-microns in diameter over 'study' period, binned using pandas.cut
Roadway exposure	Abbreviated herein as "RoadwayExposure". US Department of Transportation estimated distance in meters from residential household to nearest major roadway or highway (1 = 0–49, 2 = 50–99, 3 = 100–149, 4 = 150–199, 5 = 200–249, 6 = ≥ 250 m)
Residential density	Abbreviated herein as "EstResidentialDensity". US Census Bureau American Community Survey 2007–2011 estimated total population [block group], binned according to US Census Bureau definitions
Emergency Department or inpatient visits	Abbreviated herein as "TotalEDInpatientVisits". Total number of emergency department or inpatient visits for respiratory issue(s) over the 'study' period (0, 1, 2, 3, …)

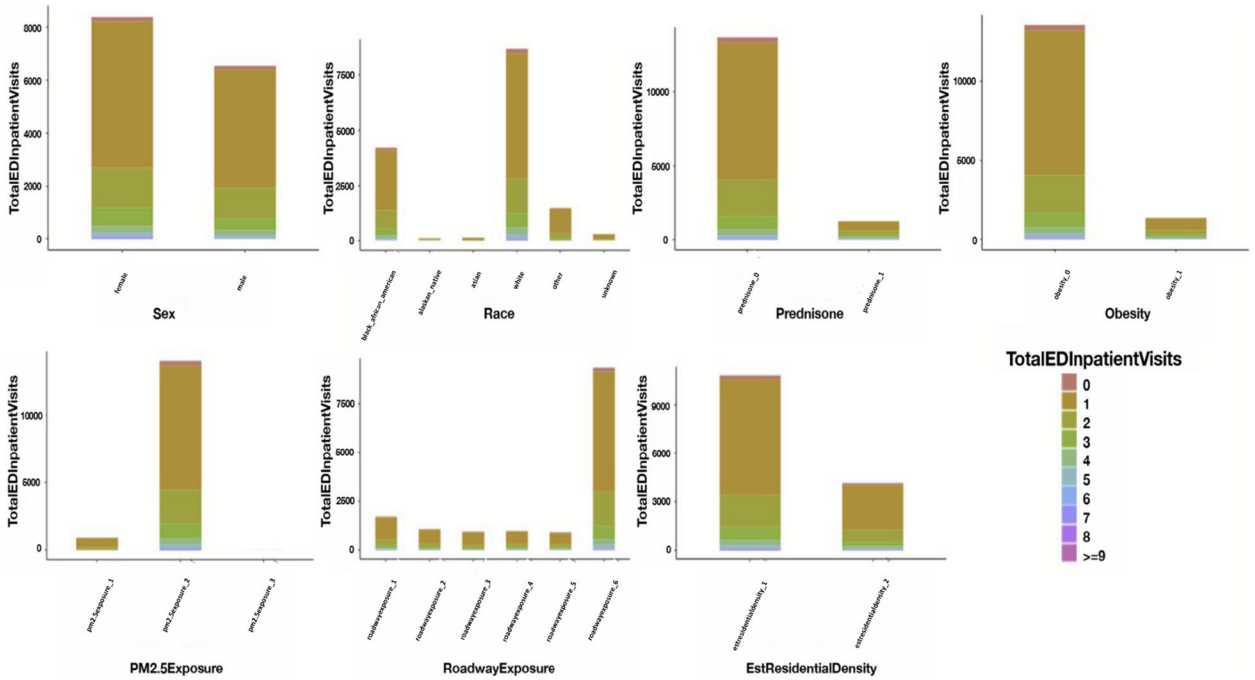

Fig. 1 Stacked bar chart representing the number of TotalEDInpatientVisits across each level of the feature variables. See Table 1 for feature variable definitions

with a directed acyclic graph (DAG) $G = (V, E)$, where $V = V_1, \ldots, V_n$ denotes the set of features and $E \in (V \times V)$ denotes the set of edges that are causal in nature. For a causal edge (V_i, V_j), we say that V_i is a cause (parent) of V_j, and V_j is the resulting effect (child) of V_i. Let $pa(V_i)$ denote the set of parents of V_i. The conditional probability distribution P_i defines the probability of Vi given the state of its parents $pa(V_i)$. A causal network represents a joint distribution P over variables V as long as it satisfies two main assumptions:

(a) Causal Markov assumption: Any given variable V_i is independent of its non-descendants, conditioned on all of its direct causes (parents). This implies that the joint distribution $P(V)$ can be factored as:
$p(V) = \prod^{n}_{i=1} p_i\,(V_i \mid \mathrm{Pa}(V_i))$.
(b) Faithfulness assumption: The joint distribution $p(V_1,\ldots, V_n)$ is faithful to G if every conditional independence relation in the probability distribution P is entailed by the Markov assumption applied to G [16].

To reconstruct a causal graph from data, we generally start by finding an approximation of the graph, given V, and then optimize based on conditions on data. The two main approaches used for causal network inference are:

1. Score-based: This is based on a Bayesian scoring function $S(G \mid D)$, which estimates the goodness-of-fit of graph G to the data D [17], as objective functions to maximize, while favoring simpler structures. The score function is usually combined with a search heuristic that explores the space of all possible graphs. Score-based methods are robust and can be extended to include interventional studies (if available), but they are not scalable as network or data size increases.
2. Constraint-based: This method is based on estimating some of the conditional (in)dependencies in the distribution P from the data D by performing hypothesis tests of conditional independence. Constraint based methods usually start with a fully connected, undirected graph and progressively remove edges whenever a new conditional independence relation is discovered, while satisfying the corresponding d separation statements.

In this work, we used a constraint-based approach called the Principal Component (PC) algorithm, given that the dataset was observational. To infer the causal graph from data, we learned the equivalence class of a directed acyclic graph (DAG) from data with the

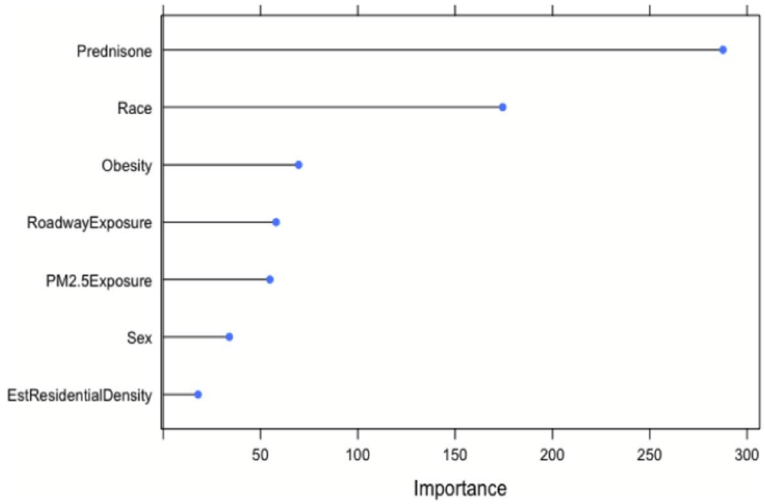

Fig. 2 Relative feature importance for all features with respect to TotalEDInpatientVisits. See Table 1 for feature variable definitions

traditional constraint-based PC algorithm proposed by [15]. Given a dataset D having n features V_i,....., V_n, we conducted the following steps. We started with a complete undirected graph given n features. We then eliminated edges between variables that are unconditionally independent. For each pair of variables (V_i, V_j) with an edge between them, and for each variable V_k with an edge connected to either of them, we eliminated the edge between V_i and V_j if $V_i \perp\!\!\!\perp V_j \mid V_k$. For each pair of variables V_i, V_j having an edge between them, and for each pair of variables V_k, V_l with edges both connected to V_i or both connected to V_j, we eliminated the edge between V_i and V_j if $V_i \perp\!\!\!\perp V_j \mid V_k, V_l$. We continued to check independencies conditional on subsets of variables of increasing size n until there were no more adjacent pairs (V_i, V_j) such that there was a subset of variables of size n in which all of the variables in the subset were adjacent to V_i or adjacent to V_j. For each triple of variables (V_i, V_j, V_k) such that V_i and V_j were adjacent, V_j and V_k were adjacent, and V_i and V_k were not adjacent, we oriented the edges $V_i - - V_j - - V_k$ as $V_i \rightarrow V_j \leftarrow V_k$, if V_j was not in the set conditioning on which V_i and V_k became independent and the edge between them was accordingly eliminated. We called such a triple of variables a v-structure. For each triple of variables such that $V_i \rightarrow V_j - - V_k$, and V_j and V_k were not adjacent, we oriented the edge $V_j - - V_k$ as $V_j \rightarrow V_k$ (i.e., orientation propagation).

We applied a causal model based on the eight feature variables included in our random forest analysis (section "Evaluation of feature importance"). We compared our model output with a model of expected edges based on subject matter expertise (e.g., a distinguished professor, practicing physician, and expert on pulmonary disorders) and the published literature [18–26]. Thus, both sources were used to generate a model of expected edges.

2.4 Simulated interventions.

We used the eight-feature causal model generated as described in section "Causal network analysis" to answer relevant questions through inference. To evaluate this, we computed the effects of interventions on features by modifying the network to simulate interventions. First, we removed undirected edges. We then learned the parameters of our learned causal DAG, given the network structure and the data. Next, we constructed a mutilated network to simulate a perfect intervention by setting a target node to a particular value. Finally, we tested the effects of three interventions on TotalEDInmpatientVisits: Obesity = 1 (all patients forced to be obese); Prednisone = 1 (all patients forced to be using prednisone); and Sex = Male (all patients forced to be male). The expectations, based on the causal inference network developed under section "Causal network analysis", were that interventions on obesity and prednisone would have direct effects on the number of TotalEDInpatientVisits, whereas an intervention on sex would not have direct effects. We note that while the interventions on obesity and prednisone are feasible, we recognize that an intervention on sex is not; however, we included sex as a test of the causal model and our assumptions, not its realistic implementation.

Results

Feature importance

In our feature importance analysis using a random forest model, we found that Prednisone, Race, Obesity,

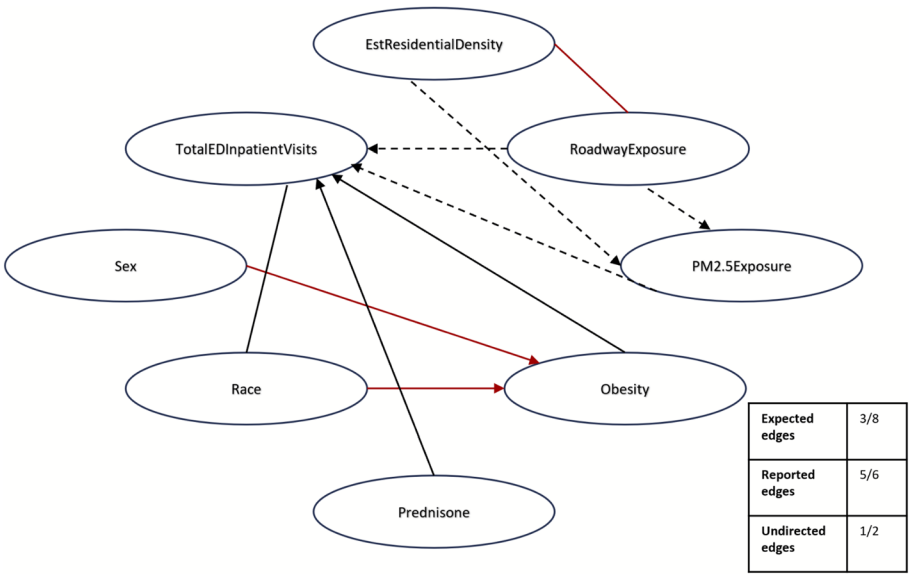

Fig. 3 Inferred causal graph. Solid black lines represent inferred expected edges based on subject matter expertise combined with published literature (true positives), dashed lines represent missed expected edges (false negatives), and red lines represent unexpected edges, meaning not expected based on subject matter expertise or the published literature (false positives)

RoadwayExposure, and PM2.5Exposure were the main contributing factors to asthma attacks (Fig. 2).

Causal analysis

Having completed the random forest analysis, we then conducted an independent causal analysis. First, we applied a PC algorithm to the ICEES multivariate feature table using the same eight feature variables used for the random forest analysis. In Fig. 3, we show the inferred casual graph. Expected relationships between features based on subject matter expertise and published literature are represented in black lines (solid and dashed, respectively). There were eight such expected edges, which we used to measure the structure learning accuracy of the causal algorithm. Solid black lines represent expected edges (true positives) that were reported via the PC algorithm, while dashed lines represent edges that were expected but missed (false negatives). Newly found relationships inferred by the PC algorithm, that were not expected, are represented in red (false positive). We note that there were a few undirected edges detected, for which the algorithm was not able to determine directionality.

Three of eight expected edges as determined by subject matter expertise were inferred; two out of three additional edges expected edges as reported in the literature were inferred (see section "Causal network analysis" for details). The expected directed edge from Race → TotalEDInpatientVisits was missed.

Effects of Intervention

Having learned a causal network from the data, we then used it to answer relevant questions by making inferences. To evaluate the network, we tested the effects of three simulated interventions on TotalEDInpatientVisits. Specifically, to substantiate the causal relationships identified section "Causal analysis", we tested the effects of interventions based on the following expected claims:

- Claim (a). Obesity should have a direct effect on TotalEDInpatientVisits. Hence, conducting an intervention on the node "Obesity" (i.e., forcing all patients to be obese) should produce a direct change (increase or decrease, accordingly) in the probability distribution of TotalEDInpatientVisits.
- Claim (b). Prednisone should have a direct effect on TotalEDInpatientVisits. Hence, conducting an intervention on the node "Prednisone" (i.e., focusing all patients to be using prednisone) should produce a direct change (increase or decrease, accordingly) in the probability distribution of TotalEDInpatientVisits.
- Claim (c). Sex should not have a direct effect on TotalEDInpatientVisits, as our causal network identified only an indirect effect of sex by way of a direct effect on obesity. Hence, conducting an intervention on the node "Sex" (i.e., forcing all patients to be male) should not produce a direct change (increase or decrease, accordingly) in the probability distribution of TotalEDInpatientVisits.

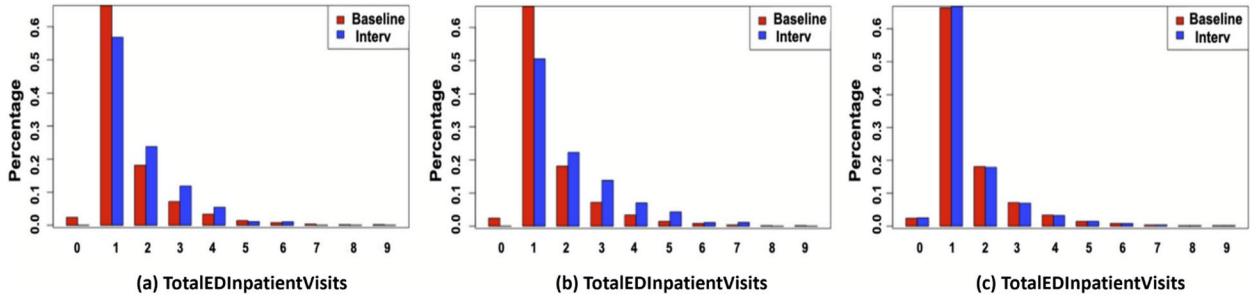

Fig. 4 The change in the mean number (% increase) of TotalEDInpatientVisits after each intervention: **a** 0.5681 to 0.6642 mean number of visits (9.62% increase) for Obesity; **b** 0.5681 to 0.7271 mean number of visits (15.90% increase) for Prednisone; and **c** 0.5681 to 0.5722 mean number of visits (0.42% increase) for Sex. Interv = intervention

We conducted these three simulated interventions on our learned causal network. To test Claim (a), we created a mutilated network by fixing the state of Obesity to 1, which means we forced Obesity to be present. For Claim (b), we fixed the state of Prednisone to be 1, meaning that we forced prednisone use to be present. For Claim (c), we fixed the state of Sex to be Male. Next, we compared the changes in the probability distribution of TotalEDInpatientVisits before and after these three *ad hoc* simulated interventions to confirm the expected causal influences (Fig. 4). The change in the probability distribution for TotalEDInpatientVisits for interventions (a) and (b) shifted to the right with each intervention due to their causal relationships to the outcome: 0.5681 to 0.6642 mean number of visits (9.62% increase) for obesity (Fig. 4a); 0.5681 to 0.7271 mean number of visits (15.90% increase) for prednisone (Fig. 4b). For intervention (c), the change in the probability distribution before and after the intervention was negligible (Fig. 4c): 0.5681 to 0.5722 mean number of visits (0.42% increase) for sex. Thus, intervening on obesity and prednisone caused a shift to the right in the number of annual ED or inpatient visits for respiratory diseases, as expected, given that our causal model showed direct effects of each variable on the outcome. In contrast, intervening on Sex had a negligible effect on the probability distribution of TotalEDInpatientVisits, also as expected, given that our causal model showed only an indirect effect of sex on the outcome by way of obesity.

Discussion

In this paper, we demonstrated the ability to use the ICEES OpenAPI to answer important questions about causal relationships between factors affecting asthma attacks. We focused on a large cohort of patients with asthma or related conditions and a dataset that included data derived from EHRs and a variety of public sources of environmental exposures data. We selected eight feature variables for our analyses; namely, sex, race, obesity, prednisone use, airborne particulate matter exposure, major roadway/highway exposure, residential density, and annual number of ED or inpatient hospital visits for respiratory issues. The racial categories were self-reported, as defined according to our hospital's EHR system, and included to capture potential racial disparities for further investigation. We applied a random forest algorithm and identified prednisone, race, and obesity as significant predictors of annual ED or inpatient visits for respiratory issues, followed by residential distance from a major roadway/highway, airborne particulate exposure, and sex. We then applied an independent causal inference model to the data, using the same feature variables, and found that prednisone and obesity were causally related to annual ED or inpatient visits, and sex and race were found to be indirectly related to annual ED or inpatient visits via a causal relationship to obesity. To validate our causal model, we then performed simulated interventions based on our causal network. Specifically, we tested the effects of "forcing" all patients to be obese, using prednisone, and of the male sex. As expected, we found that forcing all patients to be obese or using prednisone had a direct effect on annual ED or inpatient visits, whereas forcing all patients to be male did not have a direct effect. The results of our interventions, while carrying an undefined degree of statistical uncertainty, generally support our causal network analysis. Indeed, one of the strengths of causal analysis modeling, unlike predictive modeling, is that it minimizes the influence of confounding. Nonetheless, confounding remains a consideration due to factors that were unaccounted for such as physician prescribing practices regarding the use of prednisone.

Our results are largely consistent with previously published literature. For instance, prednisone, which is commonly prescribed for patients who are non-responsive to first-line treatments such as inhaled albuterol [18], has been identified as a factor associated with asthma

exacerbations and ED or inpatient visits for respiratory issues [19]. Female sex, obesity, and Black African American race have previously been identified as factors that contribute to asthma attacks [20]. In another work by our group [10] and others [21], obesity and sex were found to be highly related to asthma attacks. Several other works [9, 22] have additionally found a significant association between Black African American race and increased risk of asthma attacks. Exposure to major roadways or highways has also been found to be a risk factor for asthma. Several studies [23, 24] have demonstrated an increase in asthma attacks among patients residing in close proximity to a major roadway or highway. Our findings on the relationship between roadway exposures and asthma exacerbations have been inconsistent, with evidence to support [20] and negate [25] a relationship.

One factor that we expected to find in our model as causally related to asthma attacks, but did not, is exposure to airborne particulate matter. Exposure to airborne particulate matter is a well-established trigger for asthma attacks [8, 9, 19, 20, 25, 26]. The failure to detect a causal relationship between exposure to airborne particulate matter and asthma attacks likely reflects the imbalance in the distribution of patients across bins. Indeed, we are actively refining both our exposure models and our binning strategy. For instance, instead of using a Python algorithm to bin the airborne pollutant exposures, we are considering a binning strategy based on subject matter expertise alone.

Conclusions

EHR data, while being a rich data source for important clinical information, are mostly observational and generally challenging to access due to regulatory constraints. Performing real-world interventions are not only costly, but even impractical, given the need to integrate large data sources across various domains. Causal inference provides an excellent tool to simulate clinical interventions and answer questions about the effects of medical and healthcare interventions. In this study, we used the regulatory-compliant open ICEES service to generate a multivariate feature table and apply a causal inference model, as well as conduct simulated interventions, to explore the influence of key demographic factors and environmental exposures on asthma attacks. Our results were largely consistent with expectations based on subject matter expert opinion and the published literature. As part of our future studies, we are expanding our causal inference model to include additional features and additional years of data in order to reflect the underlying causal relationships at a larger scale, while supporting additional use cases, including a cohort of patients with primarily ciliary dyskinesia or another rare respiratory disorder.

Acknowledgements

The authors wish to acknowledge Stanley C. Ahalt, Former Director of the Renaissance Computing Institute and current Dean of the UNC School of Data Science and Society, for support and advice on the work described herein; David B. Peden for his expertise on the asthma use case; Emily R. Pfaff and James Champion for their help with the patient data; and Sarav Arunachalam, Stephen A. Appold, Alejandro Valencia Arias, and Lisa Stillwell for their help with the environmental exposures data. The authors also thank Ms. Marie Rape of the Regulatory Service at the UNC Chapel Hill NC Translational and Clinical Sciences Institute for regulatory guidance (CTSA—UM1TR004406).

Authors' contributions

MS prepared the first draft of the manuscript and performed the research modeling and experimentation. BL, PLS, PS, and HX helped in the data generation and data analysis behind this research. PH and KF were our advisors and guided us throughout the research with their valuable feedback and mentoring. KF also contributed to the writing. AK and SR reviewed the initial draft and provided overall project guidance. MS did the major writing and the experimentation behind this research. BL, PLS, PS and HX helped in the data generation and data analysis behind this research. PH and KF were our advisors and guided us throughout the research with their valuable feedabck and mentoring. AK and SR helped us the final reviewing. KF did the major revision and point by point response of the reviewers.

Funding

This project was funded with awards from the National Center for Advancing Translational Sciences, National Institutes of Health [OT3TR002020, OT2TR003430, UL1TR002489, UL1TR002489-03S4, OT2TR003428].

Availability of data and materials

The ICEES asthma OpenAPI can be accessed at https://icees-asthma.renci.org/apidocs. The associated public GitHub repositories include: https://github.com/ExposuresProvider/icees-api; https://github.com/ExposuresProvider/FHIR-PIT; https://github.com/NCTraCSIDSci/camp-fhir.
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

A waiver of informed consent for research [45 CFR 46.116(d)] and a waiver of HIPAA authorization [45 CFR 164.512(i)(2)(ii)] were granted by the Institutional Review Board at the University of North Carolina at Chapel Hill (protocol 16-2978).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, USA. [2]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [3]Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [4]Department of Computer Science, University of North Carolina, Chapel Hill, NC, USA. [5]UNC Highway Safety Research Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

Received: 27 November 2023 Accepted: 1 January 2025
Published: 15 January 2025

References

1. Pearl J. Causality: models, reasoning, and inference. Economet Theor. 2003;19(675–685):46.
2. Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82:669–710. https://doi.org/10.1093/biomet/82.4.669.

3. Pearl J. Causality: models, reasoning, and inference. 2nd ed. New York: Cambridge University Press; 2000. p. 2009.
4. Rizzi DA. Causal reasoning and the diagnostic process. Theoret Med. 1994;15(3):315–33.
5. Rubin D. Estimating causal effects of treatments in randomized and non randomized studies. J Educ Psychol. 1974;66:688–701. https://doi.org/10.1037/h0037350.
6. Robins JM, Greenland S. The role of model selection in causal inference from nonexperimental data. Am J Epidemiol. 1986;123(3):392–402. https://doi.org/10.1093/oxfordjournals.aje.a114254.
7. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37–48.
8. Fecho K, Pfaff E, Xu H, Champion J, Cox S, Stillwell L, Bizon C, Peden D, Krishnamurthy A, Tropsha A, Ahalt SC. A novel approach for exposing and sharing clinical data: the translator integrated clinical and environmental exposures service. J Am Med Inform Assoc. 2019;26(10):1064–73. https://doi.org/10.1093/jamia/ocz042.1.
9. Xu H, Cox S, Stillwell L, Pfaff E, Champion J, Ahalt SC, Fecho K. FHIR PIT: an open software application for spatiotemporal integration of clinical data and environmental exposures data. BMC Med Inform Decis Mak 2020;20:article 53. https://doi.org/10.21203/rs.2.19633/v1.
10. Fecho K, Ahalt S, Arunachalum S, Champion J, Chute CG, Gersing K, Glusman G, Hadlock J, Lee J, Pfaff E, Robinson M, Sid E, Ta C, Xu H, Zhu R, Zhu Q, Peden DB, and The Biomedical Data Translator Consortium. Sex, obesity, diabetes, and exposure to particulate matter among patients with severe asthma: Scientific insights from a comparative analysis of open clinical data sources during a five-day hackathon. J Biomed Inform. 2019;100:103325 [Special Communication]. https://doi.org/10.1016/j.jbi.2019.103325.
11. Kuhn M. Building predictive models in r using the caret package. J Statistical Software. 2008;28(1):1–26.
12. Sinha M. Causal structure learning from experiments and observations. Graduate Thesis 2019. https://ir.library.oregonstate.edu/concern/graduate_thesis_or_dissertations/7h149w16r.
13. Sinha M, Tadepalli P, Ramsey SA. Pooling vs voting: an empirical study of learning causal structures. Conference paper, Association for the Advancement of Artificial Intelligence 2019. https://why19.causalai.net/papers/siha-why19.pdf.
14. Sinha M, Tadepalli P, Ramsey SA. Voting-based integration algorithm improves causal network learning from interventional and observational data: an application to cell signaling network inference. PLoS ONE. 2021;16(2): e0245776.
15. Spirtes P, Glymour C, Scheines R. Causation, prediction, and search Adapt Comput. Machine Learning. Cambridge: MIT Press; 2000.
16. Druzdzel MJ. The role of assumptions in causal discovery. 2009. http://d-scholarship.pitt.edu/6017/1/wupes09.pdf.
17. Pearl J. Graphical models for probabilistic and causal reasoning. In: Quantified representation of uncertainty and imprecision, pages 367–89. Springer, 1998.
18 Alangar AAi. Corticosteroids in the treatment of acute asthma. Annals Thoracic Med. 2014;9(4):187.
19. Fecho K, Ahalt SC, Appold S, Arunachalam S, Pfaff E, Stillwell L, Valencia A, Xu H, Peden D. Development and application of an open tool for sharing and analyzing integrated clinical and environmental exposures data: asthma use case. JMIR Form Res. 2022;6(4): e32357. https://doi.org/10.2196/32357.
20. Lan B, Haaland P, Krishnamurthy A, Peden DB, Schmitt PL, Sharma P, Sinha M, Xu H, Fecho K. Open application of statistical and machine learning models to explore the impact of environmental exposures on health and disease: an asthma use case. Int J Environ Res Public Health. 2021;18(21):11398. https://doi.org/10.3390/ijerph182111398. (**[published as part of a special issue titled "Application of Biostatistical Modelling in Public Health and Epidemiology"]**).
21. Greenblatt RE, Zhao EJ, Henrickson SE, Apter AJ, Hubbard RA, Himes BE. Factors associated with exacerbations among adults with asthma according to electronic health record data. Asthma Research Pract. 2019;5(1):1–11.
22 Keet CA, McCormack MC, Pollack CE, Peng RD, McGowan E, Matsu ECi. Neighborhood poverty, urban residence, race/ethnicity, and asthma: rethinking the inner-city asthma epidemic. J Allergy Clin Immunol. 2015;135(3):655–62.
23. Perez L, Lurmann F, Wilson J, Pastor M, Brandt SJ, Künzli N, McConnell R. Near-roadway pollution and childhood asthma: implications for developing "win–win" compact urban development and clean vehicle strategies. Environ Health Perspect. 2012;120(11):1619–26.
24. Schurman SH, Bravo MA, Innes CL, Jackson WB, McGrath JA, Miranda ML, Garantziotis S. Toll-like receptor 4 pathway polymorphisms interact with pollution to influence asthma diagnosis and severity. Sci Reports. 2018;8(1):1–11.
25. Fecho K, Haaland P, Krishnamurthy A, Lan B, Ramsey S, Schmitt PL, Sharma P, Sinha M, Xu H. An approach for open multivariate analysis of integrated clinical and environmental exposures data. Inform Med Unlocked. 2021;26: 100733. https://doi.org/10.1016/j.imu.2021.100733.
26. Mirabelli MC, Vaidyanathan A, Flanders WD, Qin X, Garbe P. Outdoor PM2.5, ambient air temperature, and asthma symptoms in the past 14 days among adults with active asthma. Environ Health Perspect. 2016;124(12):1882–90.

Publisher's Note