De Novo exposomic geospatial assembly of chronic disease regions with machine learning & network analysis

Andrew Deonarine,[a,b,c,*] Ayushi Batwara,[a,d] Roy Wada,[a] Puneet Sharma,[a] Joseph Loscalzo,[e,f] Bisola Ojikutu,[a,e,f] and Kathryn Hall[a,e,f,g]

[a]Boston Public Health Commission, 1010 Massachusetts Avenue, 6th Floor, Boston, MA 02118, USA
[b]School of Population and Public Health, University of British Columbia, 2206 East Mall, Vancouver, BC V6T 1Z3, Canada
[c]Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, NY 10029, USA
[d]University of California, Berkeley, 110 Sproul Hall #5800, Berkeley, CA 94720-5800, USA
[e]Harvard Medical School, New Research Building, 77 Avenue Louis Pasteur, Room 630M, Boston, MA 02115, USA
[f]Brigham and Women's Hospital, Department of Medicine, 75 Francis Street, Boston, MA 02115, USA
[g]New York Academy of Medicine, 1216 5th Ave, New York, NY 10029, USA

eBioMedicine
2025;112: 105575
Published Online 31 January 2025
https://doi.org/10.1016/j.ebiom.2025.105575

Summary

Background Determining spatial relationships between diseases and the exposome is limited by available methodologies. aPEER (algorithm for Projection of Exposome and Epidemiological Relationships) uses machine learning (ML) and network analysis to find spatial relationships between diseases and the exposome in the United States.

Methods Using aPEER we examined the relationship between 12 chronic diseases and 186 pollutants. PCA, K-means clustering, and map projection produced clusters of counties derived from pollutants, and the Jaccard correlation between these clusters with chronic disease geography (defined as groups of counties with high chronic disease prevalence rates) was calculated. Disease-pollution correlation matrices were used together with network analysis to identify the strongest disease-pollution relationships. Results were compared to *LISA*, Moran's *I*, univariate, elastic net, and random forest regression.

Findings aPEER produced 68,820 human interpretable maps with distinct pollution-derived regions, and acetaldehyde/benzo(a)pyrene was found to be strongly associated with hypertension ($J = 0.5316$, $p = 3.89 \times 10^{-208}$), stroke ($J = 0.4517$, $p = 1.15 \times 10^{-127}$), and diabetes mellitus ($J = 0.4425$, $p = 2.34 \times 10^{-127}$); formaldehyde/glycol ethers with COPD ($J = 0.4545$, $p = 8.27 \times 10^{-131}$); and acetaldehyde/formaldehyde with stroke mortality ($J = 0.4445$, $p = 4.28 \times 10^{-125}$). Methanol, acetaldehyde, and formaldehyde formed distinct regions in the southeast United States (which correlated with both the Stroke and Diabetes Belts) which were strongly associated with multiple chronic diseases. Pollutants predicted chronic disease geography with similar or superior areas under the curve compared to SDOH and preventive healthcare models (determined with random forest and elastic net methods). Conventional geospatial analysis methods did not identify these geospatial relationships, highlighting aPEER's utility.

Interpretation aPEER identified a pollution-defined geographical region associated with chronic disease, highlighting the role of aPEER in epidemiological and geospatial analysis, and exposomics in understanding chronic disease geography.

Funding This work was primarily funded by the BPHC, NHLBI (R03 HL157890) and the CDC, and this work was funded in part by grants from the NIH (U01 HG007691, R01 HL155107, and HL166137), the American Heart Association (AHA24MERIT1185447), and the EU (HorizonHealth 2021 101057619) to JL.

Keywords: Machine learning; Exposome; Chronic disease; Spatial relationships; Stroke belt; Diabetes belt

Introduction

Several diseases follow consistent geographical patterns: stroke, one of the leading causes of mortality in the United States, is geographically associated with a region in the southeast of the country known as the Stroke Belt,[1–4] while a similar region with increased diabetes rates (the Diabetes Belt) has also been defined.[5] There is growing evidence that air pollution, in particular, ozone

*Corresponding author. Boston Public Health Commission, 1010 Massachusetts Avenue, 6th Floor, Boston, MA 02118, USA.
E-mail address: andrew.deonarine@mssm.edu (A. Deonarine).

Research in context

Evidence before this study

Many chronic diseases, such as diabetes and stroke mortality, have well defined geographical distributions in the United States. While the reason for these distributions have been actively investigated for decades, limited studies have examined the role of the exposome. To assess the current scientific literature available, we completed a structured review in Medline, Google Scholar, and PubMed for any publications in English up to June 24, 2024 using the search terms "stroke", "cerebral infarction", "isch(a)emic stroke", "intracerebral h(a)emorrage", "h(a)emorrhagic stroke", or "subarachnoid h(a)emorrage", "diabetes" AND "Stroke Belt", "Stroke Region", "Diabetes Belt", "Diabetes Region", or "Disease Belt". Although there were multiple studies examining the role of genetics and poverty with relation to the geographical distribution of diseases, few examined the exposome and machine learning/artificial intelligence.

Added value of this study

In this study a machine learning algorithm was developed which modelled geospatial relationships between chronic disease rates for 3141 counties and county-level pollution measures in the United States. aPEER uses unsupervised machine-learning to assemble geographical locations (like counties) together into clusters (geographical regions) based on pollution measures, and then compares these geographical regions to chronic disease geography (regions with high rates of chronic disease prevalence). It then finds the best match between pollutants and diseases by calculating the Jaccard correlation coefficient between the sets of counties that comprise the pollution regions and counties in the chronic disease regions. aPEER detected significant relationships between pollutants and several cardiometabolic conditions (using Jaccard correlation coefficient, acetaldehyde/benzo(a)pyrene was found to be strongly associated with hypertension ($J = 0.5316$, $p = 3.89 \times 10^{-208}$), stroke ($J = 0.4517$, $p = 1.15 \times 10^{-127}$), and diabetes mellitus ($J = 0.4425$, $p = 2.34 \times 10^{-127}$); formaldehyde/glycol ethers with COPD ($J = 0.4545$, $p = 8.27 \times 10^{-131}$); and acetaldehyde/formaldehyde with stroke mortality ($J = 0.4445$, $p = 4.28 \times 10^{-125}$). Using just pollution measures, aPEER consistently identified a region in the southeast United States which correlated closely with both the Stroke and Diabetes Belts, and matched the distribution of multiple cardiometabolic diseases. It was possible to predict the geographical distribution of high chronic disease rates using elastic net and random forest regressions from pollution indicators with similar or superior accuracy (determined by receiver operator curves) compared to preventive healthcare or social determinants of health models.

Implications of all the available evidence

It was possible to predict hypertension, COPD, stroke mortality, diabetes, and stroke rates from pollution indicators with comparable or superior accuracy compared to conventional models, and readily identify a region of increased pollution in the United States that closely matched the Stroke and Diabetes Belts using machine learning methods. These results highlight the utility of machine learning in exploring and analysing spatial data, and the importance of pollution in predicting the geographical variation of disease, with implications for cardiometabolic disease pathogenesis and management.

and particulate matter (PM2.5), can also influence the incidence of stroke,[6] asthma,[7] and diabetes.[8] To date, examination of geographical disease distributions by population-level variables is limited by the use of conventional statistical techniques, i.e., choropleths,[9,10] local indicators of spatial association (LISA),[11] and the spatial autocorrelation statistic Moran's *I*.[12] These techniques cannot systematically examine the geographical associations between chronic disease, population level variables and high-dimensional indicators including airborne chemicals and water pollutants.[13,14] Machine learning and network analysis methodologies coupled with the availability of large chronic disease, demographic, and environmental exposure data[15] have created an opportunity to investigate more complex spatial relationships between diseases and pollutants.

While effects of pollution can be relatively small for some conditions, the ubiquity of this exposure elevates the absolute risk at the population level to that of traditional risk factors.[6] The exposome (defined by Wild et al. as the complete set of life-course exposures an individual will encounter[13,16]) encompasses pollutants that might impact an individual's health. Understanding the regional links between the exposome (ex. air pollution measures) and the prevalence of different chronic diseases could promote informed and targeted interventions and policies to mitigate risk in exposed populations.[17] Munzel et al. identified the important role that the exposome plays in several diseases, and emphasised that the co-location of pollutants in the pathogenesis of disease, as well as the role of machine-learning in understanding the exposome needs further investigation,[18] while Fang et al. noted that the high-dimensional data used to quantify the exposome requires reliable statistical analysis methods.[19]

Here, we present a computational pipeline called aPEER (**a**lgorithm for **P**rojection of **E**xposome and **E**pidemiological **R**elationships) which uses unsupervised machine-learning to assemble geographical locations (like counties) together into multiple groups or clusters (geographical regions) based on pollution measures. Once these groups of counties have been

created, they are then compared to groups of counties with high rates of chronic disease prevalence (these groups are referred to as chronic disease geography). It then finds the best match between pollutants and diseases by calculating the Jaccard correlation coefficient between the sets of counties that comprise the pollution regions and counties in the chronic disease regions (Fig. 1). Using a combination of principal component analysis (PCA), K-means clustering, geographical projection, correlation and network analysis (using the Jaccard correlation coefficient[20]) to quantify the correlation between groups of geographical subregions (counties), we identified pollutants in the exposome which were strongly geospatially associated with a disease. Disease-pollution relationships were then validated based on their ability to predict the geographical distribution of regions with high rates of chronic disease prevalence using elastic net and random-forest models. aPEER identified geospatial relationships between multiple chronic diseases and key pollutants that were strongly predictive of chronic disease prevalence. These findings underscore the importance of understanding the potential impact of the exposome on chronic disease prevalence.

Methods

Data sources

The database generated for this study consisted of 226 indicators for 3141 counties (the complete set of indicators from Centre for Disease Control (CDC) PLACES, Environmental Protection Agency's (EPA) EJSCREEN, and EPA AirToxScreen databases) integrated into a dataframe in Python (version 3.9) using Pandas (version 1.3.4).

Health-related indicators for 3141 US counties including rates of chronic disease, participation in preventive services, and risk factors were extracted from the Behavioural Risk Factor Surveillance System (BRFSS) and available through the 2023 CDC PLACES database[21] (Supplementary Table S1). From these datasets we identified 11 disease and health-related measures for analysis (based on the leading contributors to disability-adjusted life years (DALYs) in the United States[22]), specifically, arthritis, asthma, chronic obstructive pulmonary disease (COPD), cancer, coronary heart disease, depression, diabetes, hypertension, obesity, renal disease, and stroke county-level disease prevalence. Stroke mortality prevalence data for ages 35 or older was downloaded from the CDC Stroke Death Rates database (between 2017 and 2019).[23] High disease prevalence or high stroke-mortality counties were defined as having age-adjusted county-level prevalence rates ≥70th percentile.

Pollution data for 9 pollution indicators along with seven social determinants of health (SDOH)/health equity census-tract level measures was extracted from the Environmental Protection Agency (EPA) Environmental Justice (EJSCREEN) 2021 database,[24] together with 177 chemical ambient air concentrations from the EPA's 2018 AirToxScreen database[25] reported at the census block group level (in $\mu g/m^3$), and calculated at the county level by population-weighting the census block group level exposures and then calculating the sum for each county from the blocks. This resulted in annual average pollution levels for each county calculated by county FIPS. Together, the EJSCREEN and AirToxScreen measures resulted in 186 pollution measures examined in this study. Geographical boundary information for counties, in the form of GeoJSON, were obtained from the US Census TIGER database.[26] The 9 EJSCREEN pollution indicators included particulate matter 2.5, ozone, traffic proximity, lead paint exposure, superfund proximity, RMP facility proximity, hazardous waste proximity, underground storage tank exposure, and wastewater discharge exposure (Supplementary Table S1). All models were adjusted to include county-level percent minority, percent linguistic isolation, and percent unemployed rates.

The year of pollution exposure was selected to precede the year when chronic disease rates were reported.

Finding disease–pollution associations with aPEER

The aPEER (**a**lgorithm for **P**rojection of **E**xposome and **E**pidemiology **R**elationships) computational pipeline was developed to find geographical associations between chronic diseases and the exposome (Fig. 1). In a conventional supervised machine-learning model, a dependent variable y (such as chronic disease prevalence at the county level) would be predicted from a set of independent variables x_1, x_2, ...x_n (corresponding to county-level exposome/pollution measurements). However, this conventional approach is limited by its inability to capture geospatial relationships. To address this geospatial modelling gap, aPEER uses a different approach utilising unsupervised machine-learning. First, sets of counties derived from pairs of county-level pollution measures were defined using PCA, K-means clustering, and map projection. These maps (sets of counties) were derived from pollution pairs (we used pairs of pollutants, because at least 2 variables are required for PCA and K-means clustering) which comprise a set of variables $x_{Methanol\text{-}acetaldehyde}$, $x_{Methanol\text{-}formaldehyde}$, ...$x_n$. Then, using county-level disease prevalences for stroke, hypertension, and other chronic conditions, groups of counties with prevalence rates ≥ 70% (referred to as chronic disease geography) were defined as y_{Stroke}, $y_{Hypertension}$, ...y_n. We then measured the correlation between pollution regions $\{x_{Methanol\text{-}acetaldehyde}, x_{Methanol\text{-}formaldehyde}, ...x_n\}$ and chronic disease geographies $\{y_{Stroke}, y_{Hypertension}, ...y_n\}$ using the Jaccard correlation coefficient J (which measures the number of common counties between the set of high-prevalence disease counties and pollution counties), with the

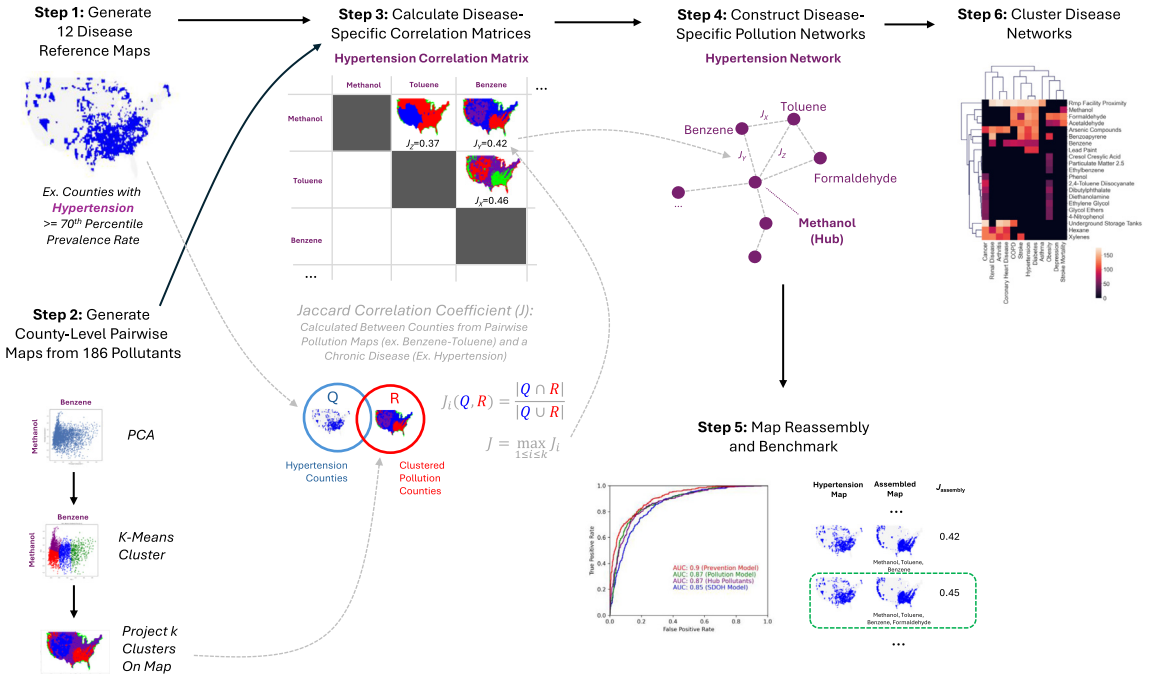

Fig. 1: The 6-Step aPEER workflow. Step 1: Generate reference maps of chronic disease prevalence and stroke mortality (≥70th percentile). Step 2: Clusters derived from principal component analysis (PCA) and *k*-means clustering of 186 pollutants projected on a US map. Step 3: Compare disease and pollution maps using Jaccard correlation coefficient (J). Step 4: Network analysis used to prioritize strongest relationships and identify key disease-related pollutants. Step 5: Findings benchmarked by examining how closely geographical distribution of key pollutants resembles disease maps. Prediction of disease prevalence by pollutants compared to known predictors like risk factors and SDOH. Step 6: Examine relationships among disease-pollution hubs using hierarchical clustering analysis.

highest J values identifying the strongest geographically-defined pollution-disease relationships (resulting in Jaccard values $J_{Stroke\text{-}(Acetaldehyde\text{-}Benzo(a)pyrene)}$, $J_{Hypertension\text{-}(Acetaldehyde\text{-}Benzo(a)pyrene)}$, etc).

In Step 1, we generated 12 reference maps of chronic disease prevalence and stroke mortality by selecting counties with chronic disease rates ≥70th percentile (we selected 70% as an illustrative example, but also completed analyses at the 60th, 80th and 90th percentiles as part of a sensitivity analysis described below, see Supplementary Methods Step 1 for full details). In Steps 2–4, we sought to find the subset of 186 possible pollutants whose geographical distribution best matched each disease reference map. To derive these pollution-disease relationships, we performed binary space decomposition of the 186 pollutants into pairs, and calculated pollution derived clusters of counties (Step 2, see Supplementary Methods Step 2 for full details) using principal component analysis (PCA), K-means clustering, and map projection for each pair of pollution indicators. We then calculated correlation matrices between these pollution-derived clusters and reference maps using the Jaccard correlation coefficient J (illustrated in Step 3 in Fig. 1, which measured the correlation between the set of counties in pollution clusters and set of counties with high disease prevalence for one of the 12 chronic diseases (these counties are depicted in a map of the United States); see Supplementary Methods Step 3 for full details).

We calculated several thousand disease-pollution Jaccard correlation coefficients between the sets of counties derived from pairs of pollutants (such as methanol-formaldehyde), and individual chronic diseases (such as hypertension). To identify the most significant individual pollutants related to a chronic disease, we constructed a pollution network for each chronic disease (i.e., 12 separate pollution networks for hypertension, diabetes, etc) where each node is a pollutant, and each edge is a Jaccard correlation index calculated between a given chronic disease and pair of pollutants (see Fig. 1 Step 3 and 4). The hubs in such a network indicate a pollutant with several statistically significant associations to a given chronic disease (Step 4, see Supplementary Methods Step 4).

The key pollutants for each chronic disease from Steps 2–4 were validated in Step 5: first, key pollutants (hubs) were used to "assemble" counties into pollution clusters, and pollution-disease pair with the highest J correlation coefficients were ranked. Next, we assessed the ability of pollutants to predict the presence of

counties with high disease rates using elastic net and random forest regression, and compared model performance to preventive healthcare and SDOH models. We also compared aPEER's performance to known geospatial analysis methods Moran's *I* and LISA, as well as a baseline elastic net regression model predicting county-level chronic disease rates from pollutants (see Supplementary Methods Step 5). Finally, we examined the relationship among disease-pollution hubs using hierarchical clustering analysis of the hubs identified from the networks (Step 6, see Supplementary Methods Step 6). We also examined how measurement bias and uncertainty in disease prevalence and pollution measures might affect Jaccard correlation coefficient *J* values using a sensitivity analysis in which counties with large confidence intervals were removed and J values recalculated (Supplementary Methods Step 8). All analyses and results were presented by following the MI-CLAIM checklist.[27]

Ethics

Ethics approval was not required for this investigation because the data was publicly available, as no individual-level or re-identifiable data was used.

Code and data availability

A Python implementation of aPEER can be downloaded from: https://github.com/andrewdeo7283/apeer.

Role of funders

The funders did not play any role in the study design, collection, analysis, or interpretation of data, in writing the report, or the decision to submit the paper for publication. KTH is funded by NHLBI R03 HL157890, and AD is funded by the CDC.

Results

In Table 1, the descriptive statistics of the county-level data used in this study are described, highlighting measures from EPA's EJSCREEN and CDC PLACES data. We were able to identify 3141 counties (out of a possible 3144 counties as of 2022) with complete sets of disease and pollution data (with no missingness), representing almost 100% of the geographical area and counties of the United States. The average chronic disease rates calculated from the county-level estimates were nearly identical to the national-level rates reported for these diseases across the United States. We then identified counties with chronic disease rates ≥70% (resulting in about 950 counties for each disease, the resulting maps for each disease are depicted in Supplementary Fig. S1). In Supplementary Table S1, baseline univariate and multivariate elastic net regressions are presented, with the top beta coefficients listed for the 12 chronic disease measures. The highest multivariate beta coefficients include those for carbon tetrachloride in the obesity model ($\beta = 1199.37$, $p = 2.49 \times 10^{-17}$) and in the arthritis model ($\beta = 724.12$, $p = 4.59 \times 10^{-7}$) and that for formaldehyde with stroke mortality ($\beta = 565.81$, $p = 2.77 \times 10^{-7}$).

In Fig. 2, example correlation matrices depicting the geospatial relationships between pairs of pollutants and hypertension (Fig. 2a) and of pollutants and stroke mortality (Fig. 2b) are illustrated (hypertension and stroke mortality were later found to be among the highest disease-pollution associations determined by aPEER using map assembly). Acetaldehyde and formaldehyde had many of the highest associations, with the highest correlations found with pollution pairs acetaldehyde-benzo(a)pyrene ($J = 0.5315$, $p << 0.01$), formaldehyde-diesel PM ($J = 0.5307$, $p << 0.01$), and acetaldehyde-1,3-butadiene ($J = 0.5274$, $p << 0.01$), while a similar pattern of strong associations was found with acetaldehyde and formaldehyde and stroke mortality (Fig. 2b), with the highest associations being benzo(a) pyrene-acetaldehyde ($J = 0.4579$, $p << 0.01$) and formaldehyde-benzene $J = 0.4578$, $p << 0.01$). The correlation matrices for the remaining chronic disease indicators are presented in Supplementary Fig. S2.

In Fig. 3, we sought to model the dependent variable *y*, a binary variable that indicated a county had a high chronic disease prevalence (greater than or equal to the 70% percentile of prevalence rates among counties in the United States, designated a value of 1), or not (designated a value of 0) from a combination of independent (x) variables, which included pollution measures, preventive healthcare delivery rates, or socioeconomic values. The networks for hypertension (Fig. 3a) and stroke mortality (Fig. 3b), together with elastic net and random forest models predicting disease geography from the hub pollutants are presented. Methanol, acetaldehyde, and formaldehyde were identified as hubs in both hypertension (8 hubs) and stroke mortality (3 hubs). Validating the pollution hubs using elastic net and random forest models revealed very specific patterns in the area under the curve (AUCs), with the prevention model performing the best in elastic net models for hypertension (AUC = 0.9) and stroke mortality (AUC = 0.8), while the pollution model (consisting of all 186 pollutants) performed best (AUC = 0.93 and AUC = 0.87) for hypertension and stroke mortality with random forest models (Fig. 3a). The hub pollutant models consistently outperformed the SDOH models irrespective of method for both hypertension and stroke mortality (AUC = 0.79–0.9). In general, the pollution model outperformed all other models when predicting the geographical distribution of the other chronic diseases, especially with random forest models (see Supplementary Fig. S3), with the highest AUC noted for depression (AUC = 0.94 (random forest), AUC = 0.81 (elastic net)) followed by hypertension (AUC = 0.93 (random forest), AUC = 0.87 (elastic net)). In all elastic net and random forest models, we found statistically

	Mean	Std. Deviation	60th Percentile	N	70th Percentile	N	80th Percentile	N	90th Percentile	N
EPA EJSCREEN demographics (%)										
Minority	23.75	20.22	22.84	1257	30.91	943	40.256	629	54.78	315
Low income	35.58	9.97	37.65	1257	40.84	943	43.833	629	48.30	315
Less than HS education	13.14	6.32	13.50	1257	15.58	943	18.161	629	21.37	315
Linguistic isolation	1.88	3.20	1.21	1257	1.70	943	2.558	629	4.59	315
Under 5 yrs	5.82	1.26	5.99	1257	6.27	943	6.590	629	7.12	315
Over 64 yrs	18.79	4.66	19.44	1257	20.54	943	22.017	629	24.71	315
Unemployed	5.36	2.73	5.57	1257	6.19	943	7.036	629	8.43	315
EPA EJSCREEN pollution measures (ug/m3)										
Lead paint	0.29	0.15	0.31	1257	0.37	943	0.43	629	0.50	315
Diesel particulate matter	0.11	0.08	0.11	1257	0.13	943	0.15	629	0.20	315
Air toxic cancer risk	21.23	11.21	23.49	1257	30.00	958	30.00	788	30.00	315
Air toxic respiratory index	0.27	0.15	0.30	1440	0.33	943	0.40	666	0.43	315
Traffic proximity	38.26	157.16	0.00	3141	0.00	3141	13.44	629	103.57	315
Wastewater discharge	0.08	1.35	0.00	3141	0.00	3141	0.00	3141	0.00	315
Superfund proximity	0.06	0.10	0.03	1257	0.04	943	0.08	629	0.16	315
RMP facility proximity	0.52	0.55	0.48	1257	0.65	943	0.89	629	1.30	315
Hazardous waste proximity	0.44	1.00	0.26	1257	0.40	943	0.61	629	1.05	315
Ozone	38.32	12.68	42.02	1257	43.24	943	44.39	629	47.67	315
Particulate matter 2.5	7.13	2.57	8.18	1257	8.58	943	8.91	629	9.30	315
Underground storage tanks	1.52	2.04	1.30	1257	1.78	943	2.38	629	3.38	315
CDC PLACES Health Measures (Population Prevalence (%))										
Arthritis	29.28	4.69	30.5	1257	31.7	962	33.2	635	35.2	321
Hypertension	36.91	6.47	38.1	1273	39.8	954	42.1	634	45	317
Cancer	7.55	1.16	7.8	1380	8.1	1024	8.5	631	9	328
Asthma	9.95	0.92	10.1	1337	10.4	983	10.7	679	11.2	318
Coronary Heart Disease	8.20	1.58	8.6	1328	9.1	945	9.5	656	10.1	346
COPD	8.64	2.24	9	1309	9.6	989	10.4	664	11.6	331
Depression	21.09	3.15	21.9	1270	22.8	951	23.9	635	25.4	324
Diabetes	12.73	2.62	13.1	1265	13.9	947	14.9	629	16.2	322
Renal disease	3.51	0.59	3.6	1399	3.8	972	4	667	4.3	338
Obesity	35.97	4.67	37.4	1271	38.4	972	39.6	650	41.2	325
Stroke	3.87	0.86	4	1290	4.2	1034	4.5	706	5	350
CDC stroke data (rate)										
Stroke mortality	39.32	8.85	40.6	1267	43.1	949	46	633	50.3	316

Table 1: **County-level descriptive statistics for chronic disease and healthcare, pollution indicators and demographic data for 3141 counties used in this study (N = number of counties equal to or above a county-level percentile cutoff).**

significant differences between the AUCs from the hub pollutant compared to the SDOH and prevention models, but did not find differences between the hub pollutant and pollution models (DeLong's $p << 0.01$).

The calibration curves for the various elastic net and random forest models are presented in Supplementary Fig. S4, with higher accuracy in general found for random forest models.

After identifying pollution hubs, we compared them to the β coefficients from elastic net and important features derived from random forest (Fig. 3c). We found that formaldehyde, acetaldehyde, and methanol were consistently highly predictive of hypertension and stroke mortality. The ordering of formaldehyde, acetaldehyde, and methanol was very similar between the aPEER pollution hubs and the beta coefficients derived by elastic net (Supplementary Fig. S3b, and calibration curves for elastic net and random forest regression are depicted in Supplementary Fig. S5). Formaldehyde, acetaldehyde, and methanol were also found to be among the top 10 pollution hubs for COPD, depression, diabetes, and stroke (see Supplementary Fig. S6).

We then determined the strongest disease-pollution associations by assembling pollution maps from pairs of pollutants and ranking the associations by Jaccard correlation coefficient (Fig. 4a), and clustered aPEER pollution hub degree, pollution-related elastic net β coefficients, and random forest pollution-associated

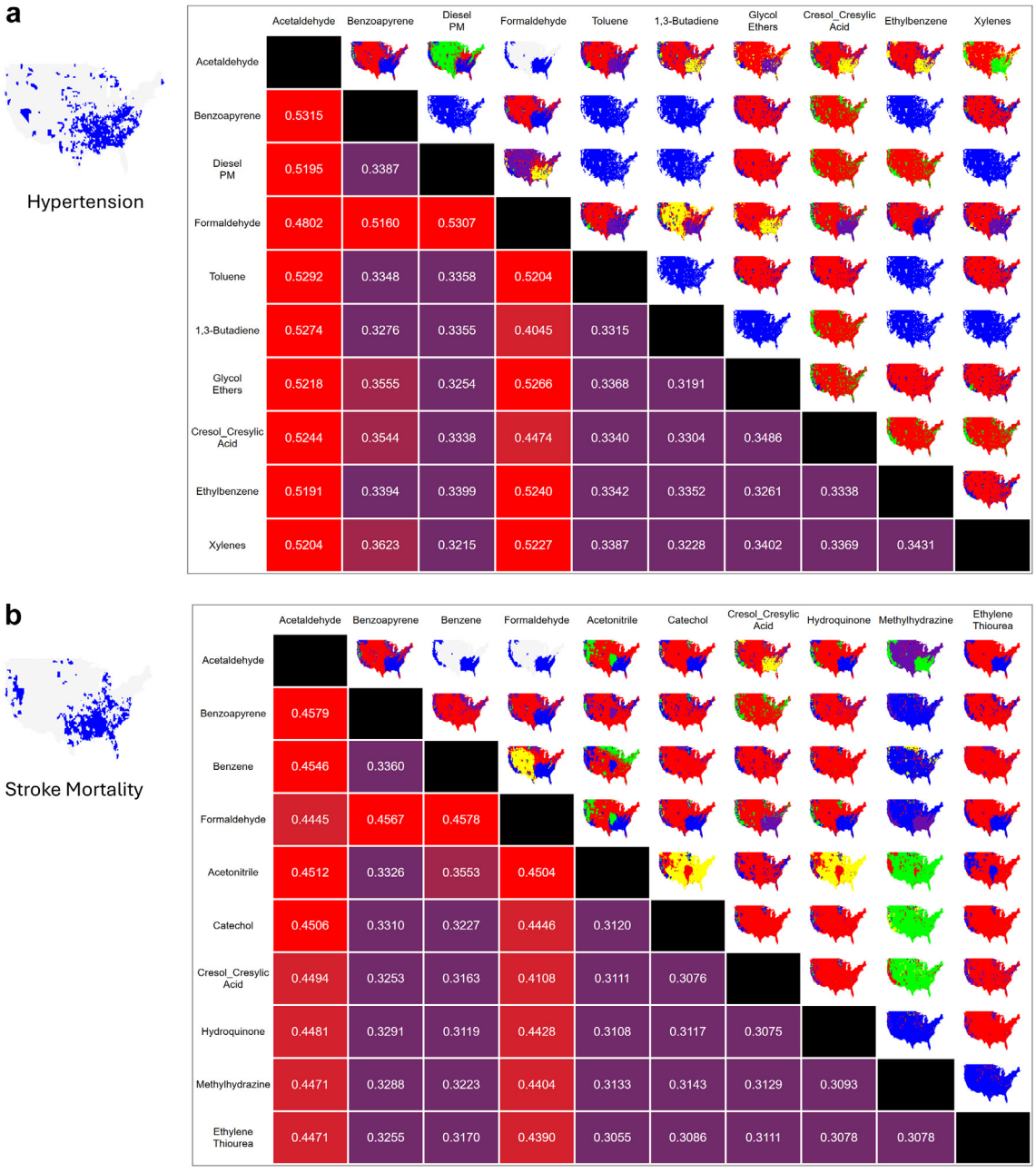

Fig. 2: The Jaccard correlation coefficient values (p << Bonferroni-adjusted threshold 0.001) between the top 10 pollutants (ranked by Jaccard J) for (a) hypertension and (b) stroke mortality (map colors are arbitrary).

feature importance values (Fig. 4b). Four out of the five top assembled map-chronic disease relationships were related to cardiometabolic conditions, including the acetaldehyde-benzo(a)pyrene pollution pair for hypertension ($J = 0.5316$, $p = 3.89 \times 10^{-208}$), formaldehyde-glycol ether for COPD ($J = 0.4545$, $p = 8.27 \times 10^{-131}$), acetaldehyde-benzo(a)pyrene for stroke ($J = 0.4517$, $p = 1.15 \times 10^{-127}$), acetaldehyde-formaldehyde for stroke mortality ($J = 0.4445$, $p = 4.28 \times 10^{-125}$), and acetaldehyde-benzo(a)pyrene for diabetes ($J = 0.4425$, $p = 2.34 \times 10^{-127}$) (Fig. 4a). In Fig. 4b–a consistent pattern of formaldehyde, acetaldehyde, and methanol clustering together is apparent; these pollutants also clustered together using β coefficients from elastic net. These relationships were partly noted after clustering feature importance from random forest analysis, with

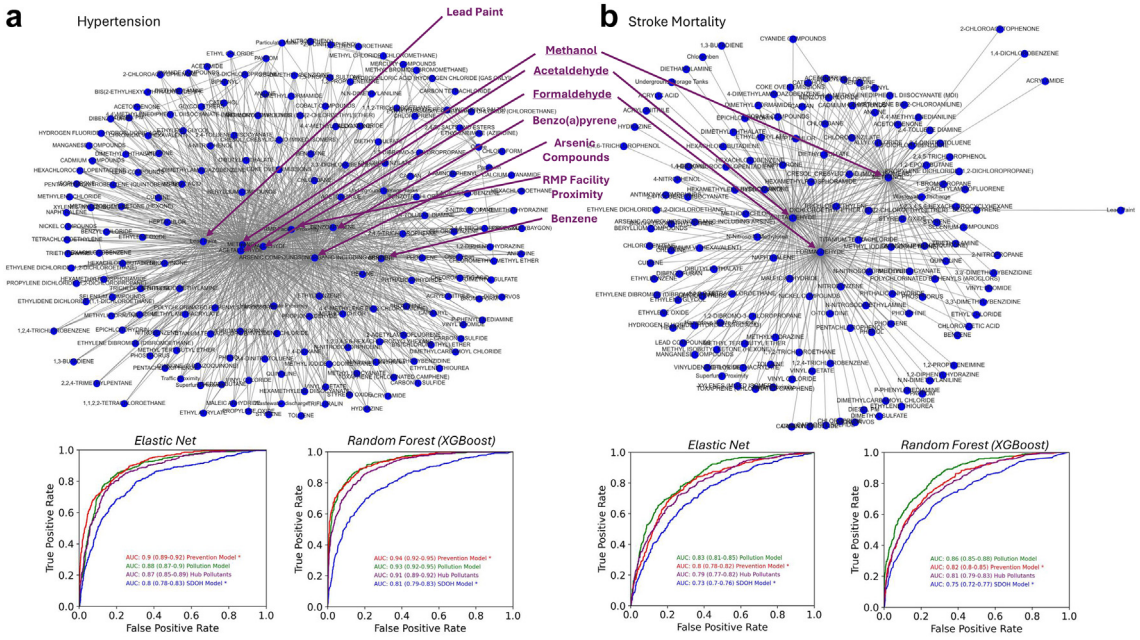

Fig. 3: (a and b) Pollution networks for (a) hypertension and (b) stroke mortality, with elastic net and random forest models predicting the geographical distribution of a given disease using pollution hubs compared to SDOH, prevention, and pollution (area under the curve (AUC) values presented with DeLong's confidence intervals, and "*" indicates if a model AUC is statistically different compared to the AUC for the "Hub Pollutants" model). (c) A comparison of the pollutants identified as being highly predictive of hypertension and stroke mortality using aPEER (pollution hubs), elastic net (β coefficients), and random forest models (importance). Three pollutants (formaldehyde, methanol, and acetaldehyde) consistently appeared irrespective of the analysis method employed (highlighted in red).

acetaldehyde clustering separately from methanol and formaldehyde (the full list of disease-pollution associations is presented in Supplementary Fig. S7). To assess the robustness of the pollution-disease associations noted with aPEER, we completed a sensitivity analysis by varying the chronic disease cutoff, using the ≥60th, ≥70th, ≥80th, and ≥90th percentiles (Supplementary Fig. S8). Clustering the results from aPEER and elastic net results mostly showed acetaldehyde, formaldehyde, and methanol grouping together at the ≥70th percentile cutoff, while less consistent results were noted with random forest regression, suggesting that aPEER and elastic net may be methodologically similar.

We compared the findings from aPEER with Moran's *I* and LISA, and found that neither of these methods identified statistically significant geographical patterns in pollution or selected diseases (Supplementary Figs. S9 and S10), indicating that aPEER may be more robust when detecting geospatial patterns in pollution data. We also examined if the disease-pollution relationships identified by aPEER were confounded by population levels, but no significant relationships were noted with selected diseases and pollutants (Supplementary Fig. S11). Additionally, we examined if aPEER was identifying the similarities between disease and pollution distributions, but no distributional similarities were apparent (Supplementary Fig. S12). The pollutants identified by aPEER as important (hubs) were not identified in the original baseline elastic net model, highlighting the limitations of the baseline model. In Supplementary Fig. S13, we determined how uncertainty in the county-level chronic disease prevalence measures and pollution measures could affect the Jaccard correlation coefficient. Using confidence interval widths as a measure of uncertainty, we calculated the density distributions of selected chronic disease and pollutant measure confidence widths (Supplementary Fig. S13a and b). We then performed a sensitivity analysis to determine how the Jaccard correlation coefficient *J* changed with uncertainty. For each pollutant and disease, we removed county outliers with confidence interval widths above the 99th, 97.5th, and 95th percentiles, and recalculated sample *J*, with very little change noted (Supplementary Fig. S13c) with variations in uncertainty.

Finally, we completed a preliminary qualitative analysis to determine if aPEER could be potentially used in time series analysis. To do this, we examined how stroke chronic disease geography (i.e., counties selected for high disease prevalence) changed over the years

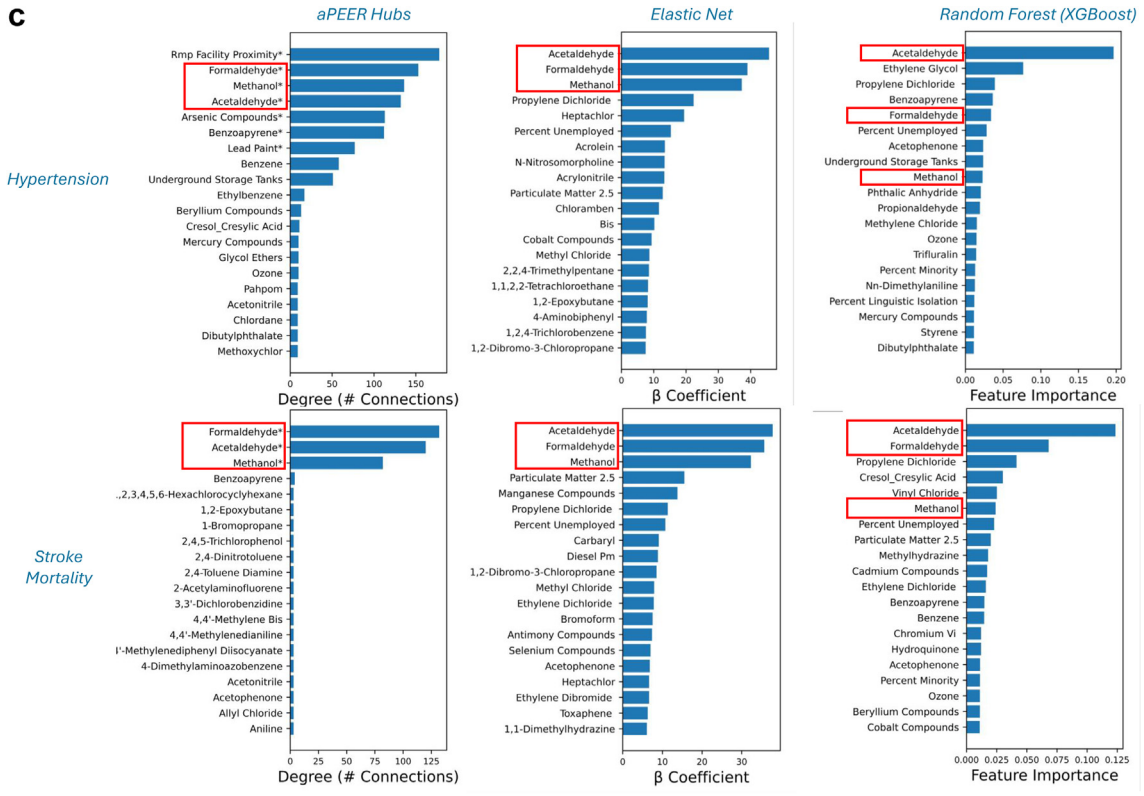

Fig. 3: Continued.

2020–2022, as well as the counties produced from clustering acetaldehyde and benzo(a)pyrene over the years 2017–2019 (Supplementary Fig. S14). Qualitative inspection of the disease and pollution maps revealed that they appeared consistent over time, suggesting that aPEER could be extended using a time-series model.

Discussion

In this investigation, it was possible to identify previously unidentified geospatial disease-pollution relationships using the aPEER algorithm between 12 chronic disease indicators and 186 pollutants, particularly between hypertension, diabetes, stroke mortality, and stroke and the pollutants acetaldehyde, formaldehyde, and methanol (Fig. 4). The associations between acetaldehyde, formaldehyde, and methanol and cardiometabolic diseases identified through correlation matrices (Fig. 2) and network analysis (Fig. 3a and b) were confirmed by elastic net and random forest regression (Fig. 3c), while statistically significant geographical distributions of diseases were not noted using conventional methods such as Moran's *I*, LISA, or benchmark univariate/elastic net regression models. These associations were also persistent even when we performed a sensitivity analysis varying the cutoffs from the 60th–90th percentile, with consistent clustering of acetaldehyde, formaldehyde, and methanol prominently noted at the 70th percentile in the aPEER and elastic net regression analysis (Supplementary Fig. S8), suggesting that at and above this threshold pollutants begin to play a significant role in disease prevalence for several cardiometabolic conditions. Air pollutants were found to be better at predicting cardiometabolic disease than conventional models based on healthcare system measures and the SDOH. The fact that aPEER generated a region from the exposome (especially acetaldehyde, formaldehyde, and methanol) that strongly resembled both the Stroke Belt and Diabetes Belts provides strong evidence for a potential linkage between stroke mortality, hypertension, diabetes, stroke and other cardiometabolic conditions and these pollutants. From the results of this study, three main conclusions can be drawn.

Firstly, aPEER identified a region in the southeast United States defined by hub pollutants which is roughly correlated with both the Stroke Belt and Diabetes Belts (Fig. 4a), and was highly associated with stroke, COPD, diabetes, hypertension, and stroke mortality. Partial explanations for regional variations in chronic diseases focus on risk factors, comorbidities, lifestyle, and SDOH factors together with the impacts of structural and environmental racism.[1,28] As well, aPEER

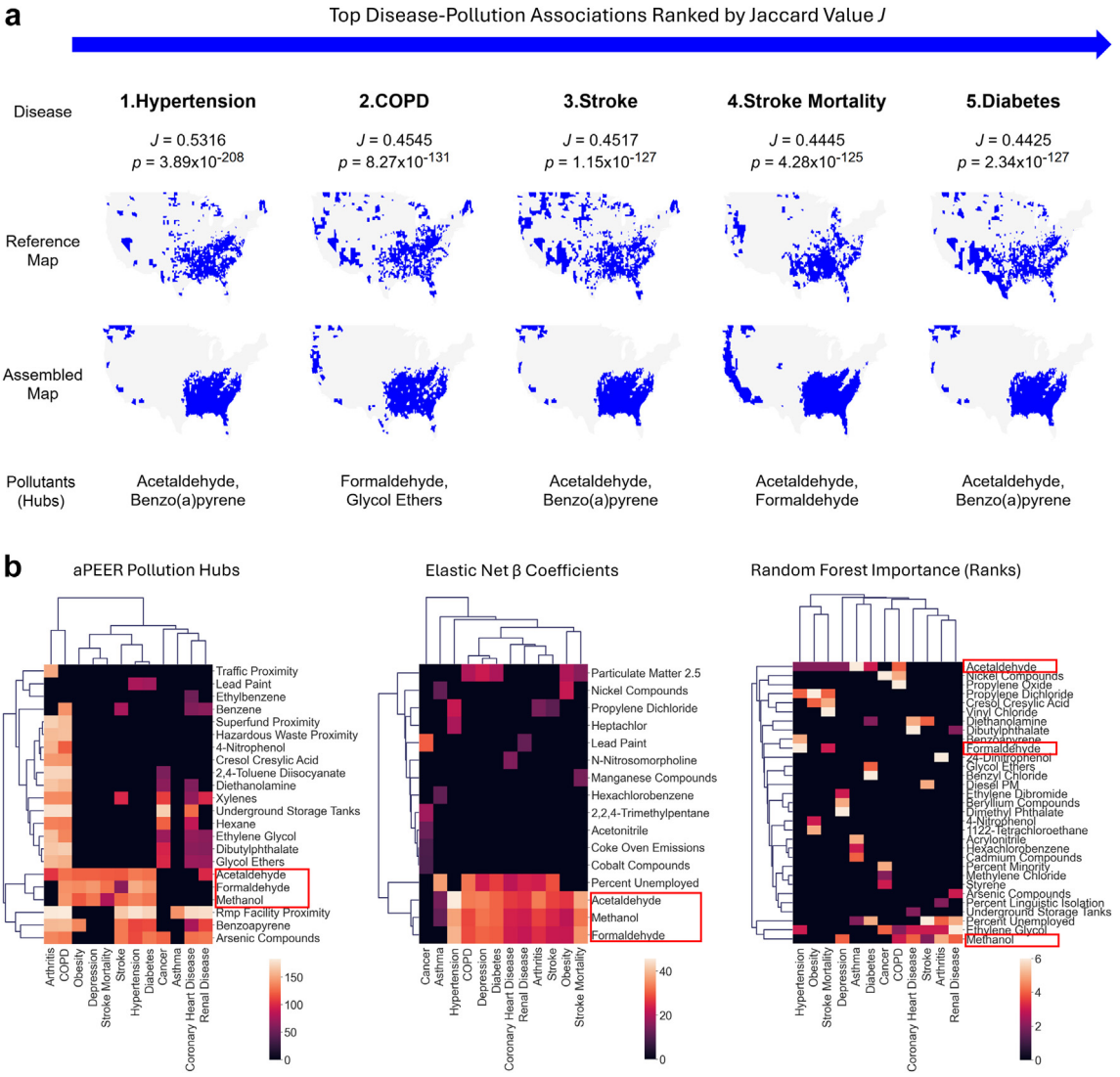

Fig. 4: (a) The top 5 disease-pollution associations derived from assembled pollution maps (70th percentile) ranked by Jaccard correlation coefficients, and (b) clustered heatmap of aPEER pollution hubs, elastic net β coefficients, and random forest importance features showing methanol, formaldehyde, and acetaldehyde closely clustered together, and strongly associated with multiple cardiometabolic diseases (highlighted in red).

identified regions on the west coast that had high rates of stroke mortality, which are not identified using the conventional Stroke Belt definition. Air pollution measures such as diesel particulate matter and PM2.5 are known to contribute to inflammation and stroke, and this may be one of the major pathways through which air pollution results in an increase in stroke rates. Many of the air pollutants identified by aPEER such as formaldehyde and acetaldehyde have documented relationships with chronic diseases such as cardiovascular, respiratory, and cancer-related conditions, and have been previously reviewed[29]). For instance, formaldehyde has been associated with stroke mortality, and hypertension,[30] but there have been fewer studies characterising these relationships in the United States. It is possible that these pollutants directly contribute to the pathogenesis of cardiometabolic diseases. Another pathway may be indirect, where air pollutants contribute to risk factors for stroke and diabetes. More recently, an investigation found an association between organic aerosols and the Stroke Belt,[31] and recapitulated very similar results to those found in this investigation. Importantly, in contrast to the observational associations observed by Pye et al.,[31] our analysis uniquely demonstrated that it is possible to assemble the Stroke Belt from hub pollutants (Fig. 4a), and that these pollutants perform nearly the same or better than established preventive services and SDOH reference models.

The ability of aPEER to produce explainable, human-interpretable maps from simple pollution combinations partly addresses the explainable artificial intelligence (XAI) problem of other machine learning techniques such as elastic net regression and random forest regressions, which rely on "black-box" coefficient optimization and creation of abstract decision trees, respectively. In the large correlation matrix of clustered maps there are several unique maps with distinct geographical distributions, and aPEER could be used to better understand how climate change, pollution, and features such as geographical elevation (which may be associated with some of the clusters) are correlated with disease distribution.

Secondly, previous investigations have identified PM2.5,[32] ozone,[33] and selenium (deficiency)[1] as being associated with increased Stroke Belt stroke rates, but few significant environmental predictions or associations have been otherwise noted. Additionally, previous studies focused on SDOH/equity factors (and in particular the African American population) and the possible cultural and genetic causes of increased stroke; by contrast, this investigation identified modifiable environmental factors that comprise the exposome, in particular air pollution, that might further explain this risk. Our results may indicate that issues such as environmental racism and exposure to specific compounds should be prioritised for investigation and intervention not only for stroke mortality, but also for diabetes, COPD, hypertension, and other chronic diseases with high AUCs (see Supplementary Fig. S4a and b).

Thirdly, this investigation highlights the role of unsupervised machine learning in analysing geographical information and finding associations between different indicators. By combining dimensionality reduction, clustering, and regression analysis for validation, it was possible to detect associations between pollution indicators and chronic diseases that would not normally be detectable. For instance, using an elastic net regression model to predict chronic disease rates from 186 different pollutants identified different pollutants compared to aPEER (except for stroke mortality, where formaldehyde and methanol were found to be significant). This observation may partly explain why pollution indicators have not been extensively studied previously for different chronic diseases. For example, Ji et al.[33] used a combination of machine learning and multilevel modelling to analyse environmental and SDOH associations with stroke, and identified ozone as having a strong association. While the relationship with ozone was replicated in our analysis, it did not appear to be the strongest relationship. This difference in outcome may be partly due to the data employed by Ji et al., namely CDC 500 Cities data, which is a subset of the CDC PLACES data used here.

Exploring and discovering relationships between multiple diseases and the exposome was not possible using conventional methods such as baseline elastic net regression, LISA, or Moran's *I*, highlighting aPEER's utility as a geospatial analysis tool. aPEER is not limited to pollution data, and can be extended to include other exposomic or geospatial data, and the ecological association of those data with the geographical distribution of other health indicators. In addition, aPEER produces clearly demarcated cluster boundaries, which reduces the need for arbitrary thresholds that sometimes are used to identify geographical regions. Hence, aPEER could be used as a general epidemiological tool to investigate ecological geospatial relationships between different geospatial measurements (such as pollution and disease rates) at different geographical resolutions (such as counties, census-tracts, zip codes, census blocks, and precincts). This method could be further enhanced through the incorporation of satellite imagery to understand better how the built environment could enhance the prediction of disease rates; in this vein, we are investigating whether different correlations (such as an area-weighted Jaccard correlation coefficient or tetrachoric correlation) and different clustering methods (such as generating pairwise disease maps and using 1-dimensional clustering algorithms) would yield better results.

Limitations of this investigation include the ecological nature of the data and relationships examined: although different geographical resolutions were used and were found to be concordant, these relationships should be confirmed using individual-level diagnosis of different chronic diseases and exposures to air pollution and other pollution indicators. Additionally, this modelling work was completed in the United States, and generalizability to other countries is yet to be determined. Finally, another major limitation of this work is that the analysis needs to be verified using other sources of pollution data at different levels of spatial resolution.

In summary, we identified key pollutants associated with multiple chronic diseases, such as stroke, hypertension, COPD and diabetes using aPEER. It was possible to identify pollutants that predicted the geospatial distribution of chronic diseases with higher accuracy than conventional preventive and SDOH factors, highlighting the importance of the exposome in the pathogenesis of multiple chronic diseases, and the role that modifiable environmental exposures play in disease. Future directions include performing a time-series analysis with aPEER across multiple years, analysing smaller regions within the United States (such as cities) to determine if there are smaller chronic disease regions such as Stroke or Diabetes belts, creating chronic disease maps from combinations of diseases, and using different correlation metrics (a geospatially weighted Jaccard correlation, or tetrachoric correlation coefficient) and machine learning models (t-SNE instead of PCA, or DBSCAN instead of K-means clustering).

Contributors

AD, AB, RW, and KH conceptualized the study. AD, AB, RW, JL, and KH created the methodology. AD and AB conducted the investigation. AD and AB visualized the study. PS, BO, and KH acquired the funding. AD and KH completed the project administration. AD and KH supervised the study. AD, RW, and KH wrote the first draft of the manuscript. All authors were involved in reviewing and editing the manuscript. All authors had access to the raw data used in the study. AD and AB have directly accessed and verified the reported underlying data. All authors accepted responsibility to submit the manuscript for publication.

Data sharing statement

Sample code is available on Github (https://github.com/andrewdeo7283/apeer) and Supplementary Figs and Tables are available in Supplementary Materials.

Declaration of interests

We declare no competing interests.

Acknowledgements

The authors would like to thank Dr. Sian Tsuei (Chan School of Public Health, Harvard Medical School) and Dr. Ella Douglas-Durham (Boston Public Health Commission) for helping to review this manuscript, and the Boston Public Health Commission for providing the information technology services to support this work.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2025.105575.

References

1 Merrill PD, Ampah SB, He K, et al. Association between trace elements in the environment and stroke risk: the reasons for geographic and racial differences in stroke (REGARDS) study. *J Trace Elem Med Biol*. 2017;42:45–49.
2 Howard G, Labarthe DR, Hu J, Yoon S, Howard VJ. Regional differences in African Americans' high risk for stroke: the remarkable burden of stroke for Southern African Americans. *Ann Epidemiol*. 2007;17:689–696.
3 Esenwa C, Ilunga Tshiswaka D, Gebregziabher M, Ovbiagele B. Historical slavery and modern-day stroke mortality in the United States stroke belt. *Stroke*. 2018;49:465–469.
4 Lanska DJ, Kuller LH. The geography of stroke mortality in the United States and the concept of a stroke belt. *Stroke*. 1995;26:1145–1149.
5 Myers CA, Slack T, Broyles ST, Heymsfield SB, Church TS, Martin CK. Diabetes prevalence is associated with different community factors in the diabetes belt versus the rest of the United States. *Obesity*. 2017;25:452–459.
6 Lee KK, Miller MR, Shah ASV. Air pollution and stroke. *J Stroke Cerebrovasc Dis*. 2018;20:2–11.
7 Zhang Y, Yin X, Zheng X. The relationship between PM2.5 and the onset and exacerbation of childhood asthma: a short communication. *Front Pediatr*. 2023;11:1191852.
8 Valdez RB, Tabatabai M, Al-Hamdan MZ, et al. Association of diabetes and exposure to fine particulate matter (PM2.5) in the Southeastern United States. *Hyg Environ Health Adv*. 2022;4:100024.
9 Center for Disease Control. PLACES: local data for better health. Center for disease Control. https://experience.arcgis.com/experience/22c7182a162d45788dd52a2362f8ed65. Accessed September 20, 2022.
10 Center for Disease Control. Social determinants of health - United States disease surveillance system. Center for disease Control. https://gis.cdc.gov/grasp/diabetes/diabetesatlas-sdoh.html. Accessed September 20, 2022.
11 Anselin L. Local indicators of spatial association - LISA. *Geogr Anal*. 1995;27:93–115.
12 Moran PAP. Notes on continuous stochastic phenomena. *Biometrika*. 1950;37:17–23.
13 Wild CP. Complementing the genome with an 'exposome': the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev*. 2005;14:1847–1850.
14 Smith MT, Rappaport SM. Building exposure biology centers to put the E into 'G x E' interaction studies. *Environ Health Perspect*. 2009;117:A334–A335.
15 Vrijheid M. The exposome: a new paradigm to study the impact of environment on health. *Thorax*. 2014;69:876–878.
16 Vermeulen R, Schymanski EL, Barabási A-L, Miller GW. The exposome and health: where chemistry meets biology. *Science*. 2020;367:392–396.
17 Kulick ER, Kaufman JD, Sack C. Ambient air pollution and stroke: an updated review. *Stroke*. 2023;54:882–893.
18 Münzel T, Sørensen M, Hahad O, Nieuwenhuijsen M, Daiber A. The contribution of the exposome to the burden of cardiovascular disease. *Nat Rev Cardiol*. 2023;20:651–669.
19 Fang M, Hu L, Chen D, et al. Exposome in human health: utopia or wonderland? *Innovation*. 2021;2:100172.
20 Jaccard P. The distribution of the flora in the alpine zone[1]. *New Phytol*. 1912;11:37–50.
21 Greenlund KJ, Lu H, Wang Y, et al. PLACES: local data for better health. *Prev Chronic Dis*. 2022;19:E31.
22 US Burden of Disease Collaborators, Mokdad AH, Ballestros K, et al. The state of US health, 1990-2016: burden of diseases, injuries, and risk factors among US States. *JAMA*. 2018;319:1444–1472.
23 Center for Disease Control. *Stroke Death rates, total population 35 and older*. Center for Disease Control; 2022. published online Sept 27. https://www.cdc.gov/dhdsp/maps/national_maps/stroke_all.htm. Accessed November 14, 2022
24 Owusu C, Flanagan B, Lavery AM, et al. Developing a granular scale environmental burden index (EBI) for diverse land cover types across the contiguous United States. *Sci Total Environ*. 2022;838:155908.
25 Environmental Protection Agency. *2018 AirToxScreen: assessment results*. Environmental Protection Agency; 2022. published online Aug 16. https://www.epa.gov/AirToxScreen/2018-airtoxscreen-assessment-results. Accessed October 31, 2022
26 Her YG, Yu Z. Mapping the US Census data using the TIGER/line shapefiles. *Environ Data Inf Serv*. 2021;2021. https://doi.org/10.32473/edis-ae557-2021.
27 Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26:1320–1324.
28 Longstreth WT. The REasons for geographic and racial differences in stroke (REGARDS) study and the national institute of neurological disorders and stroke (NINDS). *Stroke*. 2006;37:1147.
29 Sinharoy P, McAllister SL, Vasu M, Gross ER. Environmental aldehyde sources and the health implications of exposure. *Adv Exp Med Biol*. 2019;1193:35–52.
30 Wang S, Han Q, Wei Z, Wang Y, Deng L, Chen M. Formaldehyde causes an increase in blood pressure by activating ACE/AT1R axis. *Toxicology*. 2023;486:153442.
31 Pye HOT, Ward-Caviness CK, Murphy BN, Appel KW, Seltzer KM. Secondary organic aerosol association with cardiorespiratory disease mortality in the United States. *Nat Commun*. 2021;12:7215.
32 Yuan S, Wang J, Jiang Q, et al. Long-term exposure to PM2.5 and stroke: a systematic review and meta-analysis of cohort studies. *Environ Res*. 2019;177:108587.
33 Ji J, Hu L, Liu B, Li Y. Identifying and assessing the impact of key neighborhood-level determinants on geographic variation in stroke: a machine learning and multilevel modeling approach. *BMC Publ Health*. 2020;20:1666.