Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Sparse modeling of spatial environmental variables associated with asthma

Timothy S. Chang [a,*], Ronald E. Gangnon [b], C. David Page [c], William R. Buckingham [d], Aman Tandias [e,1], Kelly J. Cowan [f,2], Carrie D. Tomasallo [g], Brian G. Arndt [e], Lawrence P. Hanrahan [e], Theresa W. Guilbert [f,3]

[a] Department of Biostatistics and Medical Informatics, School of Medicine and Public Health, University of Wisconsin, 5795 Medical Sciences Center, 1300 University Ave, Madison, WI 53706, USA
[b] Department of Biostatistics and Medical Informatics, School of Medicine and Public Health, University of Wisconsin, 603 Warf Office Building, 610 Walnut St, Madison, WI 53706, USA
[c] Department of Biostatistics and Medical Informatics, School of Medicine and Public Health, University of Wisconsin, 6743 Medical Sciences Center, 1300 University Ave, Madison, WI 53706, USA
[d] Applied Population Laboratory, Department of Rural Sociology, University of Wisconsin, 308b Agricultural Hall, 1450 Linden Dr, Madison, WI 53706, USA
[e] Department of Family Medicine, School of Medicine and Public Health, University of Wisconsin, 1100 Delaplaine Ct, Madison, WI 53715, USA
[f] Department of Pediatrics, School of Medicine and Public Health, University of Wisconsin, Madison, WI 53706, USA
[g] Division of Public Health, Bureau of Environmental and Occupational Health, Wisconsin Department of Health Services, Room 150, 1 West Wilson Street, Madison, WI 53703, USA

ARTICLE INFO

Article history:
Received 7 February 2014
Accepted 12 December 2014
Available online xxxx

Keywords:
Asthma
Sparsity
Spatial statistics
Environmental variables
Electronic health record

ABSTRACT

Geographically distributed environmental factors influence the burden of diseases such as asthma. Our objective was to identify sparse environmental variables associated with asthma diagnosis gathered from a large electronic health record (EHR) dataset while controlling for spatial variation. An EHR dataset from the University of Wisconsin's Family Medicine, Internal Medicine and Pediatrics Departments was obtained for 199,220 patients aged 5–50 years over a three-year period. Each patient's home address was geocoded to one of 3456 geographic census block groups. Over one thousand block group variables were obtained from a commercial database. We developed a Sparse Spatial Environmental Analysis (SASEA). Using this method, the environmental variables were first dimensionally reduced with sparse principal component analysis. Logistic thin plate regression spline modeling was then used to identify block group variables associated with asthma from sparse principal components. The addresses of patients from the EHR dataset were distributed throughout the majority of Wisconsin's geography. Logistic thin plate regression spline modeling captured spatial variation of asthma. Four sparse principal components identified via model selection consisted of food at home, dog ownership, household size, and disposable income variables. In rural areas, dog ownership and renter occupied housing units from significant sparse principal components were associated with asthma. Our main contribution is the incorporation of sparsity in spatial modeling. SASEA sequentially added sparse principal components to Logistic thin plate regression spline modeling. This method allowed association of geographically distributed environmental factors with asthma using EHR and environmental datasets. SASEA can be applied to other diseases with environmental risk factors.

Abbreviations: UW eHealth–PHINEX, University of Wisconsin Electronic Health Record–Public Health Information Exchange; EHR, electronic health record.

* Corresponding author at: University of Wisconsin, Department of Biostatics and Medical Informatics, 5795 Medical Sciences Center, 1300 University Ave, Madison, WI 53706, USA. Tel.: +1 608 265 5693; fax: +1 608 265 7916.

E-mail addresses: tschang3@uwalumni.com (T.S. Chang), ronald@biostat.wisc.edu (R.E. Gangnon), page@biostat.wisc.edu (C. David Page), wrbuckin@wisc.edu (W.R. Buckingham), aman.tandias@fammed.wisc.edu (A. Tandias), kelly.cowan@vtmednet.org (K.J. Cowan), carrie.tomasallo@dhs.wisconsin.gov (C.D. Tomasallo), brian.arndt@fammed.wisc.edu (B.G. Arndt), larry.hanrahan@fammed.wisc.edu (L.P. Hanrahan), theresa.guilbert@cchmc.org (T.W. Guilbert).

[1] Present address: Division of Public Health, Bureau of Environmental and Occupational Health, Wisconsin Department of Health Services, Room 150, 1 West Wilson Street, Madison, WI 53703, USA.

[2] Present address: Department of Pediatrics, College of Medicine, University of Vermont, 111 Colchester Avenue, Burlington, VT 05401, USA.

[3] Present address: Department of Pediatrics, Cincinnati Children's Hospital Medical Center, MLC 2021, 3333 Burnet Avenue, Cincinnati, OH 45229, USA.

http://dx.doi.org/10.1016/j.jbi.2014.12.005
1532-0464/

1. Introduction

While there is continued interest in associating genes with disease using methods such as genome-wide association studies [1], approximately 23% of disease burden and death can be attributed to environmental factors [2]. It is important to associate diseases with a strong environmental component, including respiratory infections, cardiovascular disease, cerebrovascular disease, and asthma [2], with geographical environmental factors. Methods that consider spatial variation and interpretability of results will increasingly be utilized as clinical, environmental, and geographical datasets become more readily available. Our paper applies sparsity with spatial modeling to study the association of environmental factors and asthma.

1.1. Asthma risk factors

Asthma is a chronic respiratory disease with variable and recurring symptoms, airflow obstruction, bronchial hyperresponsiveness, and inflammation [3]. Its prevalence rose by 15% in the last 10 years [4]. Based on a Wisconsin Department of Health Services asthma surveillance report, approximately 14% of adults and 10% of children have been diagnosed with asthma in Wisconsin [5]. In 2009, 5300 people were hospitalized and 21,000 went to an emergency department with a principal diagnosis of asthma. Eleven percent of adults with asthma had an emergency department visit and 20% had urgent care visits for symptoms [5].

Asthma onset is associated with multiple, complex factors. While some are non-modifiable such as sex and age [6], many others are associated with the environment and residential location. These include educational attainment, household income, health insurance, smoking, physical activity, and obesity [6]. Medical conditions influenced by the environment and associated with asthma include atopy [7], allergic reactions [8], airway hyperreactivity [9], and airway responsiveness [10]. Over 370 outdoor and indoor environmental factors have been associated with asthma including substances from building materials, cleaning products, personal care products, central heating systems, maintenance, and humidification devices [11].

1.2. Geographical analysis of asthma

Geographic information system (GIS) analyses have been used to study geographic environmental variables associated with asthma. The most studied variable was air pollution [12], which has been measured via passive measurement, direct measurement, proximity to roadways, and traffic carbon emissions. Besides air pollution, asthma was associated with climate differences [13], latitude [14], and socioeconomic status [15]. Socioeconomic status, specifically male employment, was positively associated with asthma in a Southern California study, where access to care and the hygiene hypothesis—the idea that limited exposure to bacterial and viral pathogens during childhood result in a predisposition to allergy [16,17]—were proposed as explanations.

Fewer asthma studies have incorporated local environmental variables aggregated at the level of census tracts or block groups. Census tracts and block groups are geographic areas developed by the United State Census Bureau and contain 1500–8000 and 600–3000 people, respectively. Using census tract data, asthma diagnosis was correlated with houses facing highway intersection [18] and sociodemographic characteristics of race, sex, and education [19]. Fewer studies have used block group level variables. Socioeconomic status was associated with asthma diagnosis using block group level data [15]. Many of these analyses used questionnaire data to determine asthma diagnosis, which may be limited by self-report bias [20]. These analyses involved less than 5700 participants, 10 environmental variables, and census geographic regions from only a portion of a state.

1.3. Environmental variables associated with EHR data

Environmental variables and built environments have been studied using EHR data. For example, nitrogen oxides were tested for association with diseases including asthma diagnoses obtained from EHR datasets in primary care [21]. Body mass index (BMI) calculated from EHR data was positively associated with the number of fast food restaurants near a person's home [22].

Schwartz et al. [23] used an EHR dataset, environmental community-level variables, and multilevel statistical analysis to demonstrate that lower BMI was associated with higher socioeconomic status and areas with more venues for physical activity.

1.4. Spatial statistics to study disease

Spatial statistics offer methods to incorporate geographic location to identify risk factors associated with disease [24]. The spatial statistics utilized in this study included a generalized additive model. Generalized additive models [25] are generalized linear models with predictors that involve a linear sum of smooth functions.

Previous health studies that utilized spatial generalized additive models investigated the association of air pollution and mortality, tuberculosis drug resistance patterns in Peru [26], and geographic distribution of heart disease [27].

Spatial statistics, specifically additive models, have been combined with sparsity. COSSO [28] and SpAM [29] extended the lasso estimator [30] while another approach created a new sparsity-smoothness penalty [31].

1.5. Objective

Our goal was to identify an interpretable set of environmental risk factors of asthma distributed geographically. Other studies have combined environmental variables and EHR data, spatial statistics and disease, and spatial statistics and sparsity. Our main contribution is the addition of sparsity to spatial statistics. As applied to geographically distributed EHR and environmental datasets, we describe this methodology as Sparse Spatial Environmental Analysis (SASEA).

2. Material and methods

2.1. Source of clinical data

Our research group developed the University of Wisconsin Electronic Health Record–Public Health Information Exchange (UW eHealth–PHINEX), an EHR data exchange between University of Wisconsin (UW) Departments of Family Medicine, Internal Medicine, and Pediatrics and the Wisconsin Division of Public Health. Further details have been described previously [32]. Briefly, the database contains clinical care variables such as disease diagnoses, medications, and laboratory test results. Patient home addresses from year 2012 were geocoded to year 2000 block groups, the smallest geographic area the US Census Bureau publishes. Block groups were linked to detailed demographic and environmental data from the ESRI Business Analyst database [33]. The data exchange is a HIPAA Privacy Rule compliant-limited dataset, and the Wisconsin Division of Public Health is blinded to patient/provider specific information. All patient identifiers were removed from the data except birth month and year, ZIP code, and census

Please cite this article in press as: Chang TS et al. Sparse modeling of spatial environmental variables associated with asthma. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.12.005