A distributed geospatial approach to describe community characteristics for multisite studies

Patrick H. Ryan,1,2 Cole Brokamp,1,2 Jeff Blossom,3 Nathan Lothrop,4,5 Rachel L. Miller,6 Paloma I. Beamer,4,5 Cynthia M. Visness,7 Antonella Zanobetti,8 Howard Andrews,9 Leonard B. Bacharier,10 Tina Hartert,11 Christine C. Johnson,12 Dennis Ownby,13 Robert F. Lemanske,14 Heike Gibson,8 Weeber Requia,8 Brent Coul,8 Edward M. Zoratti,12 Anne L. Wright,14,15 Fernando D. Martinez,4,15 Christine M. Seroogy,14 James E. Gern,14 and Diane R. Gold,8,16 on behalf of the CREW Consortium

Abstract

Understanding place-based contributors to health requires geographically and culturally diverse study populations, but sharing location data is a significant challenge to multisite studies. Here, we describe a standardized and reproducible method to perform geospatial analyses for multisite studies. Using census tract-level information, we created software for geocoding and geospatial data linkage that was distributed to a consortium of birth cohorts located throughout the USA. Individual sites performed geospatial linkages and returned tract-level information for 8810 children to a central site for analyses. Our generalizable approach demonstrates the feasibility of geospatial analyses across study sites to promote collaborative translational research.

Introduction

Maintaining patient privacy is a common challenge faced by researchers seeking to understand the relationship of place and health [1–4]. This issue can be especially problematic in multisite studies due to study protocols and confidentiality concerns that limit the sharing of geographic data. Existing approaches to geospatial analyses in multisite studies include the use of a central site to conduct analyses or the application of spatial techniques (e.g. changing geographic coordinates to protect confidentiality, i.e. “geomasking”) to protect patient privacy [4]. However, the former method is limited by data use agreements (DUAs) and the latter may result in exposure misclassification when geomasked locations result in spatial misalignment.

Alternatively, study sites may perform geocoding and geospatial linkages independently before removing identifiable information for joint analyses. This decentralized approach, however, faces challenges in reproducibility and standardization due to geocoding methods, geographic information software (GIS), and expertise that varies across study sites. Here, we describe the application of a novel method to perform reproducible, standardized, and confidential geospatial analyses for multisite studies. Our approach extends a previously developed Decentralized Geomarker Assessment for Multi-site Studies (DeGAUSS) containerization platform to perform geocoding and extraction of polygon feature geospatial data over multiple time periods and large geographic areas [5]. As an example case, we use our approach to ascertaining US census tract-level information for participants enrolled in the Children’s Respiratory and
Environmental Workgroup (CREW), a network of 12 birth cohorts each studying the development of allergy and asthma in childhood [6].

Materials and Methods

Study Population

Our approach was motivated by the CREW consortium, a network of birth cohorts recruited from 1980 to 2020. Information regarding participating cohorts, including eligibility criteria, study recruitment, and other methods, are published [6] and in Supplementary Table 1. Due to its large sample size and geographic distribution, CREW provides a unique platform to examine environmental and community factors that contribute to the disproportionately higher burden of asthma-related morbidity and mortality among disadvantaged communities [7–10]. A CREW data sharing protocol and DUA were approved by the local IRB for each cohort. However, the DUA allowed only limited datasets to be shared and prohibited the distribution of identifying information, including addresses or geocodes.

Distributed Geospatial Analysis

We extended our DeGAUSS software to enable all CREW sites, including those with limited geospatial expertise, to derive spatio-temporal US census tract-level information for their participants at birth. A key advantage to DeGAUSS is the use of a software containerization platform to wrap necessary software, system dependencies, and geospatial data in a stand-alone package that will work the same regardless of its host environment [11]. Previously, we created DeGUASS and applied this tool in the Electronic Medical Records and Genomics (eMERGE) network in a proof-of-concept study [5]. Whereas our prior DeGAUSS software included only a geocoder and code to link geospatial coordinates to nearby roadways and one census tract variable, the CREW consortium required significantly expanded geographical and temporal data to be included for analyses. Therefore, a new custom DeGAUSS container containing decennial US Census data (described below), census tract polygon boundary files for the 1980, 1990, 2000, and 2010 census, and R code (R Foundation for Statistical Computing: Vienna, Austria; 2014) was created to merge census tract-level data to the geocoded locations of CREW birth addresses. Additional details regarding DeGAUSS and the CREW container are provided in the supplementary materials and online [12,13].

A flow diagram depicting the distributed approach to geospatial analyses for the CREW consortium is provided in Figure 1. The DeGAUSS container image was created by C.B. at a single location and distributed to each cohort. The DeGAUSS software required cohort users to provide an input.csv file containing the geocoded coordinates of their participants’ birth record address. Site end users also specified the census year to assign the appropriate tract boundary file and census data based on participants’ year of birth. The output data file from the DeGAUSS container contained the original input data, including geocoded locations, and appended census tract information including the census tract FIPS code in which the birth record address was located and census variables. Site end users manually removed identifying information, including the geographic coordinates and census tract FIPS code, prior to returning the de-identified dataset to a central coordinating center.

US Census/American Community Survey Data

Longitudinal US Census data and boundary files for the years 1980, 1990, 2000, and 2010 were downloaded for the entire USA from Social Explorer (www.socialexplorer.com, New York City, NY: Social Explorer 2017; accessed 12/17–1/18) by year, requested variables, and tract level of geography. For 1980, certain census variables were only available from the National Historic GIS data service (Minneapolis, MN: NHGIS; accessed 11/17–1/18) and were updated to reflect variable calculation definitions used in later censuses. As the 2010 census did not include information regarding median household income, median gross rent, or median housing values, these data were downloaded from the 2008–2012 American Community Survey (ACS, https://www.census.gov/programs-surveys/acs/data.html; accessed 12/17–1/18). Additional information regarding the census and ACS variables downloaded and included in the DeGAUSS container for linkage to birth record addresses is available in the supplementary material.

Statistical Analyses

After sharing de-identified data with a central site, descriptive statistics and box-and-whisker plots for all census tract variables were calculated for the combined CREW consortium and for each cohort individually. Comparison of census tract-level to self-reported race, ethnicity, and household income was conducted by plotting the distribution of each census tract-level variable according to self-reported variable. Self-reported household income was compared to census tract median household income using each cohorts’ income categories as collected by questionnaire. Additional details regarding self-reported race, ethnicity, and income information are available in the supplementary material.

Results

All cohorts (n = 12) completed the distributed analysis and returned de-identified data to the coordinating center. Collectively, 8997 participants were enrolled in CREW cohorts, and 98% (n = 8810) of these had birth record addresses geocoded with sufficient precision for linkage to a census tract.

A summary of population, race, ethnicity, and income data for the census tracts in which participants resided at birth is provided in Table 1. CREW participants resided in both low and high population density regions as reflected in the average tract population density (persons per km²) that ranged from 148 for Wisconsin Infant Study Cohort (WISC) participants in Wisconsin to 45,772 for CCCEH participants in New York City (Table 1). Overall, CREW participants lived in tracts that were 67% White, 23% Black, 2% Asian, and 9% Other race. There was, however, variability in tract racial distribution both within and across cohorts (Table 1); eight cohorts enrolled participants from tracts where more than 75% of the population was White, while participants enrolled in URECA, WHEALS, and CCCEH resided in census tracts with populations having a greater proportion of Black or Other race. Most cohorts enrolled participants from tracts having a Hispanic population less than 10%, though the mean Hispanic population in tracts of CCCEH, IIS, TCRS, and URECA (Boston, MA and New York, NY sites) participants ranged from 21 to 57%. The overall mean of households in poverty was 15% in census tracts where CREW participants resided at birth but as shown in Figure 2, there was significant variability within and across cohorts. Additional information on census tract-level...
The distribution of tract-level race data (%White, %Black, %Asian, %Other) according to participants self-reported race is presented in Figure 3. Overall, participants who reported being White or Asian race lived in census tracts with majority White populations, whereas participants who reported Black race resided in census tracts with greater variability in shares of White and Black populations (Figure 3). Additional comparisons of individual-level to neighborhood-level income and ethnicity are provided in the supplementary material.

Discussion

Environmental exposures and the community in which they occur are significant causes of human disease, including asthma [7,14–17]. Disentangling the environmental and social exposures that contribute to health disparities necessitates geographically and culturally diverse studies. The methods described here make the characterization of community characteristics in a confidential yet standardized and reproducible manner more feasible for researchers and policy-makers. Importantly, our method is generalizable to additional types of geographic data, including polygon and point data, allowing other studies to customize and incorporate geocoding and geospatial analyses into their approach.

Our distributed geospatial approach offers some important advantages compared to existing methods, including reduced exposure misclassification, maintaining participant confidentiality, and reducing the need for geospatial expertise at each study site [1]. One existing and alternative method to maintain subject confidentiality is the alteration of participants’ geographic coordinates. Referred to as “geomasking” or “jittering,” this approach involves either a random shift in the location of subjects or a systematic transformation of the locations known only to the researchers [4]. However, this method may introduce errors or biases introduced due to the displacement of the participants’ actual location, particularly for analyses requiring an exact location. For example, the amount of geomasking required to make geographic datasets de-identified according to HIPAA standards may result in incorrect census tracts being linked to individual subjects resulting in exposure misclassification. An alternative approach to incorporating geospatial information into multisite studies is to obtain IRB approval and DUAs to share subjects’ identifiable information with a central site for geocoding and analysis. Challenges with this strategy include hesitation on the part of institutions to share identifiable information (e.g. addresses). Performing geospatial linkages at individual sites is another approach but may produce non-standardized and non-reproducible results due to the use of varying geocoding platforms, software, and dataset.

Our DeGAUSS method overcomes these limitations because the geospatial data and software are developed at a central site, ensuring that all individual study sites run the same software on identically constructed datasets. Importantly, our method also accounts for spatial, temporal, and informational changes in census tract boundaries and census data over the study period.
<table>
<thead>
<tr>
<th>Cohort</th>
<th>Participants with deocoded birth address</th>
<th>Census year</th>
<th>Population (mean, SD)</th>
<th>Population density (/km²)</th>
<th>% White</th>
<th>% Black</th>
<th>% Asian</th>
<th>% Other</th>
<th>% Hispanic</th>
<th>Median household income (2012 USD)</th>
<th>% Below poverty level</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAS</td>
<td>825</td>
<td>1990</td>
<td>1457 (970)</td>
<td>93 (13)</td>
<td>4 (13)</td>
<td>2 (2)</td>
<td>1 (1)</td>
<td>1 (1)</td>
<td>1 (1)</td>
<td>73,449 (26,708)</td>
<td>6 (5)</td>
</tr>
<tr>
<td>CCAAPS</td>
<td>762</td>
<td>2000</td>
<td>1346 (938)</td>
<td>79 (28)</td>
<td>17 (28)</td>
<td>1 (2)</td>
<td>2 (1)</td>
<td>1 (1)</td>
<td>1 (1)</td>
<td>64,895 (27,632)</td>
<td>10 (15)</td>
</tr>
<tr>
<td>CCCEH</td>
<td>724</td>
<td>2000</td>
<td>45,772 (1956)</td>
<td>16 (9)</td>
<td>41 (28)</td>
<td>2 (2)</td>
<td>42 (21)</td>
<td>57 (27)</td>
<td>1 (1)</td>
<td>31,171 (7308)</td>
<td>34 (9)</td>
</tr>
<tr>
<td>COAST</td>
<td>280</td>
<td>2000</td>
<td>877 (882)</td>
<td>89 (11)</td>
<td>4 (5)</td>
<td>3 (6)</td>
<td>4 (3)</td>
<td>4 (1)</td>
<td>7 (1)</td>
<td>71,090 (20,443)</td>
<td>5 (6)</td>
</tr>
<tr>
<td>EHAAS</td>
<td>499</td>
<td>1990/2000</td>
<td>3944 (3936)</td>
<td>83 (23)</td>
<td>9 (19)</td>
<td>5 (4)</td>
<td>3 (6)</td>
<td>5 (7)</td>
<td>7 (7)</td>
<td>86,218 (33,980)</td>
<td>8 (8)</td>
</tr>
<tr>
<td>IIS</td>
<td>481</td>
<td>2000</td>
<td>1380 (880)</td>
<td>79 (12)</td>
<td>4 (2)</td>
<td>3 (1)</td>
<td>15 (11)</td>
<td>23 (17)</td>
<td>2 (1)</td>
<td>57,774 (23,415)</td>
<td>9 (8)</td>
</tr>
<tr>
<td>INSPIRE</td>
<td>1921</td>
<td>2010</td>
<td>628 (692)</td>
<td>73 (25)</td>
<td>20 (23)</td>
<td>2 (2)</td>
<td>6 (5)</td>
<td>7 (7)</td>
<td>4 (3)</td>
<td>47,606 (20,237)</td>
<td>16 (12)</td>
</tr>
<tr>
<td>MAAP</td>
<td>141</td>
<td>2010</td>
<td>1165 (700)</td>
<td>82 (19)</td>
<td>11 (17)</td>
<td>4 (5)</td>
<td>3 (2)</td>
<td>4 (3)</td>
<td>3 (1)</td>
<td>69,042 (27,253)</td>
<td>9 (7)</td>
</tr>
<tr>
<td>TCRS</td>
<td>1142</td>
<td>1980</td>
<td>1172 (813)</td>
<td>85 (14)</td>
<td>3 (4)</td>
<td>1 (1)</td>
<td>10 (12)</td>
<td>21 (21)</td>
<td>2 (1)</td>
<td>51,535 (17,849)</td>
<td>12 (8)</td>
</tr>
<tr>
<td>URECA - Baltimore</td>
<td>163</td>
<td>2000/2010</td>
<td>6480 (3468)</td>
<td>16 (22)</td>
<td>79 (26)</td>
<td>1 (2)</td>
<td>4 (5)</td>
<td>3 (7)</td>
<td>3 (7)</td>
<td>33,202 (12,999)</td>
<td>28 (13)</td>
</tr>
<tr>
<td>URECA - Boston</td>
<td>143</td>
<td>2000/2010</td>
<td>8244 (4544)</td>
<td>30 (26)</td>
<td>46 (29)</td>
<td>5 (6)</td>
<td>20 (11)</td>
<td>25 (17)</td>
<td>4 (3)</td>
<td>44,680 (20,190)</td>
<td>23 (11)</td>
</tr>
<tr>
<td>URECA - New York</td>
<td>118</td>
<td>2000/2010</td>
<td>33,137 (14,409)</td>
<td>26 (15)</td>
<td>41 (19)</td>
<td>3 (3)</td>
<td>30 (13)</td>
<td>51 (20)</td>
<td>2 (3)</td>
<td>30,967 (17,645)</td>
<td>31 (12)</td>
</tr>
<tr>
<td>URECA - St. Louis</td>
<td>178</td>
<td>2000/2010</td>
<td>2375 (1292)</td>
<td>21 (25)</td>
<td>76 (27)</td>
<td>1 (2)</td>
<td>3 (8)</td>
<td>2 (3)</td>
<td>1 (1)</td>
<td>31,467 (10,297)</td>
<td>28 (11)</td>
</tr>
<tr>
<td>WHEALS</td>
<td>1248</td>
<td>2000/2010</td>
<td>2376 (1196)</td>
<td>41 (38)</td>
<td>53 (41)</td>
<td>2 (4)</td>
<td>5 (6)</td>
<td>4 (10)</td>
<td>4 (1)</td>
<td>49,092 (24,995)</td>
<td>20 (15)</td>
</tr>
<tr>
<td>WISC</td>
<td>185</td>
<td>2010</td>
<td>148 (301)</td>
<td>96 (3)</td>
<td>1 (1)</td>
<td>1 (1)</td>
<td>3 (3)</td>
<td>3 (3)</td>
<td>3 (3)</td>
<td>50,871 (9775)</td>
<td>9 (5)</td>
</tr>
</tbody>
</table>

CREW, Children’s Respiratory and Environmental Workgroup; CAS, Childhood Allergy and Asthma Study; COAST, Childhood Origins of Asthma Study; CCAAPS, Cincinnati Childhood Allergy and Air Pollution Study; CCCEH, Columbia Center for Children’s Environmental Health; EHAAS, Epidemiology of Home Allergens and Asthma Study; INSPIRE, Infant Susceptibility to Pulmonary Infections and Asthma Following RSV Exposure; IIS, Infant Immune Study; MAAP, Microbes, Allergy, Asthma, and Pets; TCRS, Tucson Children’s Respiratory Study; URECA, Urban Environment and Childhood Asthma; WHEALS, Wayne County Health, Environment, Allergy, and Asthma Longitudinal Study; WISC, Wisconsin Infant Study Cohort.
Fig. 2. Percentage* of household poverty in census tracts where CREW participants resided at birth. CAS, Childhood Allergy and Asthma Study; COAST, Childhood Origins of Asthma Study; CCAAPS, Cincinnati Childhood Allergy and Air Pollution Study; CCCEH, Columbia Center for Children's Environmental Health; EHAAS, Epidemiology of Home Allergens and Asthma Study; INSPIRE, Infant Susceptibility to Pulmonary Infections and Asthma Following RSV Exposure; IIS, Infant Immune Study; MAAP, Microbes, Allergy, Asthma, and Pets; TCRS, Tucson Children’s Respiratory Study; URECA, Urban Environment and Childhood Asthma (BA, Baltimore; BO, Boston; NY, New York; SL, St. Louis); WHEALS, Wayne County Health, Environment, Allergy, and Asthma Longitudinal Study; WiSC, Wisconsin Infant Study Cohort.

Fig. 3. Comparison of census tract race to participants' self-reported race.
Participants in CREW cohorts were born over nearly four decades, and, as detailed in the accompanying supplement, linking geocoded addresses with the appropriate census tract at the time of participant birth was critical to our approach. By including census tract boundaries and accompanying data from 1980, 1990, 2000, and 2010 in our DeGAUSS software, we were able to link study participants to the appropriate census tract and data for their date of birth. Of note, the DeGAUSS platform is flexible and amenable to additional geographic data and analyses, including derivation of nearby greenspace, distance to roadways and other locations, estimating gridded air pollution exposures, area-based indices of neighborhood deprivation, and others. Specific details and example uses are available on the DeGAUSS website [12,13].

As part of the NIH Environmental influences on Child Health Outcomes (ECHO) program, the objective of CREW is to understand the etiology of childhood asthma and determine environmental and genetic contributors. As such, CREW offers a unique opportunity to examine the significantly higher rates of asthma prevalence, hospitalization, and morbidity among children residing in households and neighborhoods with lower SES, as compared to White children residing in higher socioeconomic status (SES) households and communities [18–20]. Multiple influences contribute to these disparities including disproportionately higher exposure to air pollution, poor housing, limited access to care, and other adverse physical and toxicant contributors, and inherited factors may increase individuals’ sensitivity to these [21–25]. However, environmental exposures and inherited factors alone do not fully explain the observed health disparities. For that reason, hardships linked to poverty, including discrimination, stress, family turmoil, violence, instability, and others have been posited to play an important role in the social patterning of disease [26]. These social determinants of health may be separate from access to medical care and are important drivers not only in asthma morbidity, but also in a wide range of both adult and pediatric health outcomes, including mortality [27,28]. The importance of socioeconomic factors is highlighted by observations that disparities in health outcomes across strata of SES remain present within racial/ethnic groups [28].

Our methods and results should be considered, however, in the context of some limitations. Census tract boundaries alone may not accurately describe individuals’ neighborhood experience and the use of decennial census data rather than ACS data resulted in reduced temporal precision but allowed us to increase our historic reach to 1980. There are also limitations to the US Census data including a lack of specificity in certain variables such as ethnicity. For example, Hispanic participants in the TCRS and IIS in Tucson, AZ, likely differ in ancestry from Hispanic participants from cohorts in New York City, NY. Finally, our approach requires expertise at individual institutions to obtain patient or electronic health records.

In conclusion, we demonstrated the use of a distributed approach to conduct geospatial analyses for the 12 CREW cohorts that is also applicable to other multisite studies. Future applications of our method will include additional gridded data including land cover, air pollution models, and meteorological information. Future research to understand the etiology of childhood asthma will incorporate longitudinal residential locations throughout childhood and multilevel analyses of individual and neighborhood environments.

Acknowledgments. The Children’s Environment and Respiratory Workgroup (CREW) is a collaboration of the following cohorts, institutions, investigators, and staff (principal investigators are indicated by an asterisk). The authors would particularly like to thank Michael Evans, Andrew Moseby, Zhouwen Liu, Brent Olson, Suzanne Havstad, Soma Datta, Howard Andrews, and Agustin Calatroni for running the DeGAUSS software to link census and address data at their cohort sites.


Infant Immune Study (IIS): Anne L. Wright*, Fernando D. Martinez*, Wayne Morgan, Debra A. Stern, Dean Billheimer, Brian Hallmark, Paloma Beamer, Nathan Lothrop, Lynn Taussig, Heidi Erickson, Marilyn Halonen, Amber Spangenberg, David Spies.

Wisconsin Infant Study Cohort (WISC): Marshfield Clinic Research Institute: Casper Benderixen, Katherine Barnes, Tera Johnson, Terry Foss, Elizabeth Armagost, Vicki Moon, Tammy Koepel, Erin Donnerbauer, Wayne Frome, and Brent Olson.


The Infant Susceptibility to Pulmonary Infections and Asthma Following RSV Exposure (INSPIRE): Vanderbilt University Medical Center, Nashville, TN: Tina V. Hartert MD, MPH**, Andrew Abreo MD+, Nieck Achten

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/cts.2021.7


Administrative Center, University of Wisconsin-Madison: James E. Gern,* Gina Crisafi, Dorothy Fleorkee, and Rick Kelley.


Genetics Core, University of Chicago. Carole Ober*, Daniel Nicolea.

Geospatial Core, Harvard School of Public Health, Boston, MA: Diane Gold*, Heike Gibson, Brent Coull, Antonella Zanobetti, Webber Requia, Joel Schwartz, Itai Kloog, Qian Di, Peter James, Marcia Jimenez Pescador, and Jaime Hart.

Microbiome Core, University of California-San Francisco: Susan Lynch* and Katherine McCauley.

This manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

CREW is funded by HHS/NIH grant SUG3OD023282. Additional support was provided by individual cohorts’ grants/contracts:

Columbia University: P01 ES09600, R01 ES008977, P30 ES09089, R01 ES013163, R827027.

Tucson Children’s Respiratory Study (TCSR): R01HL132523, R01HL56177, K25HL103970.

Infant Immune Study (IIS): R01AI42268, K25HL103970.

Childhood Origins of Asthma Study (COAST): P01 HL070831, U10 HL064305, R01 HL068179.

Urban Environmental and Childhood Asthma Study (URECA): NO1-AI-25496, NO1-AI-25482, HHSN272200900052C, HHSN272201000052I, NCRB/NIH RR00552, M01RR00533, 1 UL1RR025771, M01RR00071, 1UL1RR024156, ULTR001079, 5UL1RR024992-02, NCATS/NIH ULTR00404.

Cincinnati Childhood Allergy and Air Pollution Study (CCAPAS): R01 ES11170, R01 ES019890.

The Epidemiology of Home Allergens and Asthma Study (EHAAS): R01 AI053786.

Wayne County Health, Environment, Allergy and Asthma Longitudinal Study (WHEALS): R01 AI050681, R56 AI050681, R01 AI066774, R21 AI059415, K01 AI070606, R21 A1069271, R21 HL113010, R21 ES022351, P01 AI089473, R21 AI080066, R01 AI10450, R01 HD082147 and the Fund for Henry Ford Health System.

Childhood Allergy Study (CAS): R01 AI024156, R03 HL067427, R01 AI051598, Blue Cross Foundation Johnson, and the Fund for Henry Ford Hospital.

Microbes, Allergy, Asthma and Pets (MAAP): P01 AI089473 and Fund for Henry Ford Hospital.

Infant Susceptibility to Pulmonary Infections and Asthma following RSV Exposure (INSPIRE): NIH/NIAIID U19 AI 095227, NIH/NCATS U11 TR 002243, NIH/NIAIID K24 AI 077930, NIH/NIHLBI R21 HD 087864, NIH/NIHLDI X01 HL 134583.

Wisconsin Infant Study Cohort (WISC): U19 AI04317, NCATS ULTR000427, the charitable donors to the Marshfield Clinic Health System Foundation, and the Upper Midwest Agricultural Safety and Health Center (UMASH) U54 OH010170.

Disclosures. The authors have no conflicts of interest to declare.

References


