Journal of the American Medical Informatics Association, 26(10), 2019, 1064–1073
doi: 10.1093/jamia/ocz042
Advance Access Publication Date: 26 April 2019
Research and Applications

Research and Applications

A novel approach for exposing and sharing clinical data: the Translator Integrated Clinical and Environmental Exposures Service

Karamarie Fecho,[1] Emily Pfaff,[2] Hao Xu,[1] James Champion,[2] Steve Cox,[1] Lisa Stillwell,[1] David B. Peden,[2,3,4] Chris Bizon,[1] Ashok Krishnamurthy,[1,2] Alexander Tropsha,[1,5] and Stanley C. Ahalt[1,2]

[1]Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, [2]North Carolina Translational and Clinical Sciences Institute, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, [3]Division of Allergy, Immunology and Rheumatology, Center for Environmental Medicine, Asthma & Lung Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, [4]Department of Pediatrics, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, and [5]School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Corresponding Author: Karamarie Fecho, PhD, Renaissance Computing Institute, 100 Europa Drive, Suite 540, Chapel Hill, North Carolina, USA 27517 (kfecho@copperlineprofessionalsolutions.com)

Received 25 October 2018; Revised 12 March 2019; Editorial Decision 14 March 2019; Accepted 25 March 2019

ABSTRACT

Objective: This study aimed to develop a novel, regulatory-compliant approach for openly exposing integrated clinical and environmental exposures data: the Integrated Clinical and Environmental Exposures Service (ICEES).

Materials and Methods: The driving clinical use case for research and development of ICEES was asthma, which is a common disease influenced by hundreds of genes and a plethora of environmental exposures, including exposures to airborne pollutants. We developed a pipeline for integrating clinical data on patients with asthma-like conditions with data on environmental exposures derived from multiple public data sources. The data were integrated at the patient and visit level and used to create de-identified, binned, "integrated feature tables," which were then placed behind an OpenAPI.

Results: Our preliminary evaluation results demonstrate a relationship between exposure to high levels of particulate matter $\leq$2.5 μm in diameter ($PM_{2.5}$) and the frequency of emergency department or inpatient visits for respiratory issues. For example, 16.73% of patients with average daily exposure to $PM_{2.5}$ >9.62 $\mu g/m^3$ experienced 2 or more emergency department or inpatient visits for respiratory issues in year 2010 compared with 7.93% of patients with lower exposures (n = 23 093).

Discussion: The results validated our overall approach for openly exposing and sharing integrated clinical and environmental exposures data. We plan to iteratively refine and expand ICEES by including additional years of data, feature variables, and disease cohorts.

Conclusions: We believe that ICEES will serve as a regulatory-compliant model and approach for promoting open access to and sharing of integrated clinical and environmental exposures data.

Key words: open science, patient privacy, regulatory compliance, clinical data, environmental data, data integration, data harmonization, semantic harmonization

INTRODUCTION

The ability to access and share clinical data is critical to accelerate and advance clinical and translational research. Without access to clinical data, inferences and insights gleaned from basic-science research cannot be validated in humans, and likewise, inferences and insights gleaned from clinical research and observation cannot be rigorously investigated in animal studies or in vitro assays. Equally important as access to clinical data is the adequate protection of human subjects.[1] Indeed, multiple regulations surround the use of human subjects in clinical research, enacted in part due to historical abuses.[2]

Yet, the many regulations surrounding clinical research, coupled with cultural sensitivities and technical constraints, often impede vital access to clinical data and decelerate clinical and translational research. First, clinical research, whether involving direct interaction/intervention with human subjects or data from electronic health record (EHR) systems, justifiably requires review by an Institutional Review Board (IRB). However, the approval process, while absolutely essential, can often be slow and cumbersome (see Revised Common Rule in Supplementary Appendix). Second, institutions are often wary to share clinical data due to concerns surrounding Health Insurance Portability and Accountability Act (HIPAA) restrictions and respect for patient privacy, even when clinical data are fully de-identified according to HIPAA. Indeed, this concern has been the topic of much debate and discussion among healthcare institutions, business leaders, and law professionals.[3] A third, less recognized concern is the potential sensitivity of healthcare providers and institutions.[4] For instance, clinical datasets, even de-identified ones, often contain data that can be used to determine rates of procedures (eg, total hip arthroplasty) and adverse events (eg, perioperative hemorrhage) and identify providers (eg, full names) and administrative practices (eg, health insurance payments). Healthcare providers and institutions are understandably reluctant to share such data, as the data can be used in ways that have legal implications and financial consequences. While exceptions apply,[5–8] and several healthcare organizations have launched large-scale open-data initiatives, such as Columbia Open Health Data (COHD)[9] and MIMIC (Medical Information Mart for Intensive Care),[10] the reluctance of healthcare providers and institutions to openly share clinical data remains widespread.

The National Center for Advancing Translational Sciences, a center within the National Institutes of Health, launched the Biomedical Data Translator program in October 2016 in an effort to overcome challenges in the application of the many biomedical datasets that are available today to clinical and translational science and the practice of medicine.[11,12] The goal of the initial feasibility assessment is "to design and prototype a 'Translator' system capable of integrating existing biomedical datasets and 'translating' those data into insights that can accelerate translational research, support clinical care, and leverage clinical expertise to drive research innovations."[12] Central to the program's vision is the ability to openly access and integrate datasets, including clinical datasets and datasets on environmental exposures. This capability compounds issues related to clinical data access because the integration of clinical data with data on environmental exposures requires careful consideration of space and time in order to determine patient- and visit-level exposures. This, in turn, requires access to Protected Health Information (PHI), specifically geocodes and dates, to accurately estimate exposures at any given geographic location and time for any individual patient. We have identified open access to integrated clinical and environmental exposures data as one of the main challenges that must be addressed in order to fully realize the vision of the Translator system.

Objective

As part of the Translator program, we have developed a novel approach and service for exposing integrated clinical and environmental exposures data, which we termed the Integrated Clinical and Environmental Exposures Service (ICEES). ICEES was designed as a disease-agnostic approach to overcome the regulatory, cultural, and technical challenges that hinder efforts to openly share integrated clinical and environmental exposures data. Herein, we describe the development and preliminary evaluation of ICEES.

MATERIALS AND METHODS

Aims

The primary aim of our research and development efforts was to offer open regulatory-compliant access to clinical data on patients in the Carolina Data Warehouse for Health (CDWH), which is the clinical data warehouse for all patients in the UNC Health Care System. A secondary aim was to integrate clinical datasets with external datasets derived from several public databases, including datasets on chemical exposures (eg, airborne pollutants) and socioeconomic exposures (eg, estimated household income), and openly expose the integrated data.

Driving clinical use case

For development and preliminary evaluation of ICEES, our driving clinical use case was asthma, which is a common disease influenced by hundreds of genetic variants and numerous environmental exposures, including exposures to airborne pollutants. Thus, the cohort we selected included all patients in the CDWH with an "asthma-like" condition (see Supplementary Appendix), including patients with (1) a diagnostic code of asthma who were prescribed or administered medications that are typically used to treat asthma, (2) a diagnostic code for a respiratory condition other than asthma who were prescribed or administered medications that are typically used to treat asthma, and (3) a diagnostic code for a respiratory condition other than asthma who were prescribed tests or procedures that are typically used to diagnose asthma. The cohort was identified using a single SQL query of the CDWH. We intentionally selected a broad group of patients, as one of the goals of the Translator program is to move beyond traditional diagnosis of disease to a data-driven reclassification of patients on the basis of shared biomolecular and clinical traits. We note that our approach is by no means restricted to patients with asthma-like conditions; in fact, our approach was developed as a disease-agnostic method for exploring environmental influences on virtually any disease.

Our driving use-case question was the following: is exposure to high levels of particulate matter $\leq$2.5 μm in diameter ($PM_{2.5}$) associated with responsiveness to treatment? For our purposes, we defined high levels of $PM_{2.5}$ on a scale of 1 (low) to 5 (high), whereby the categories represent bins of estimated patient-level exposures. Responsiveness to treatment was defined by the number of emergency department (ED) or inpatient visits for respiratory issues over a 1-year period (<2 per year [responder] vs 2 or more per year [nonres-

ponder]). Additional considerations on the treatment of feature variables are provided in the Design section.

Of importance, a relationship between $PM_{2.5}$ exposure and responsiveness to treatment or asthma exacerbation has been well established in the literature.[13–17] As such, the intent of our driving use-case question was to validate our datasets, software code, and the overall design of ICEES by demonstrating a relationship between $PM_{2.5}$ and asthma. This was critical, as prior studies were conducted as either traditional clinical research studies, with direct measurement of individual-level exposures and health outcomes under an approved IRB protocol, or population-based epidemiological studies using secondary data sources and correlational analyses. In contrast, our approach required the development of a complex analytic pipeline for integrating exposure estimates derived from multiple public data sources with individual-level EHR data using PHI and then openly exposing the results in such a way as to not reveal the integrated individual-level data per se, but still demonstrate individual-level relationships between $PM_{2.5}$ and asthma. Additional challenges are described in the Discussion section.

Design

To achieve our aims in an open regulatory-compliant manner, we developed an approach for creating what we've termed "integrated feature tables" (Figure 1). The approach begins with an IRB-approved protocol and a secure compute environment. For initial development of ICEES, we integrated clinical datasets with a variety of public datasets on environmental exposures, including: airborne exposures data, estimated using the Community Multiscale Air Quality Modeling System[18]; roadway exposures data (ie, an indirect measure of exposure to airborne pollutants), estimated using data derived from the U.S. Department of Transportation; and socioeconomic exposures data (eg, poverty) derived from the U.S. Census Bureau's American Community Survey. IRB approval was required for this step because data integration necessitates the use of patient geocodes (ie, primary residence), date and time stamps, and patient identifiers. However, after the integration step, the data were de-identified according to §164.514(b) of HIPAA ("Safe Harbor" method for patient de-identification of medical records), thus reclassifying the sensitivity of the project and reducing, but not eliminating, regulatory concerns. While the research use of de-identified clinical data is typically classified as exempt from certain federal regulations, in our case, this was not true because our institution views research involving data elements that are generated using PHI as identifiable and subject to IRB review. For instance, the $PM_{2.5}$ exposure estimates were captured as hourly estimates and integrated with patient data using geocodes and date and time stamps. The patient identifiers, geocodes, and date and time stamps were retained after this step to allow for additional integration steps to incorporate estimated exposures derived from different data sources. Thus, the integration procedure was subject to IRB review. However, our institution determined that downstream use of the ICEES service is exempt from IRB review because users are provided with completely de-identified data.

To further address privacy and security concerns related to the integrated feature tables, several additional safeguards were embedded in the design of ICEES. First, after data integration, the feature variables were recoded or binned to further reduce the risk of unintentional leakage of sensitive data (Table 1). The binning strategy

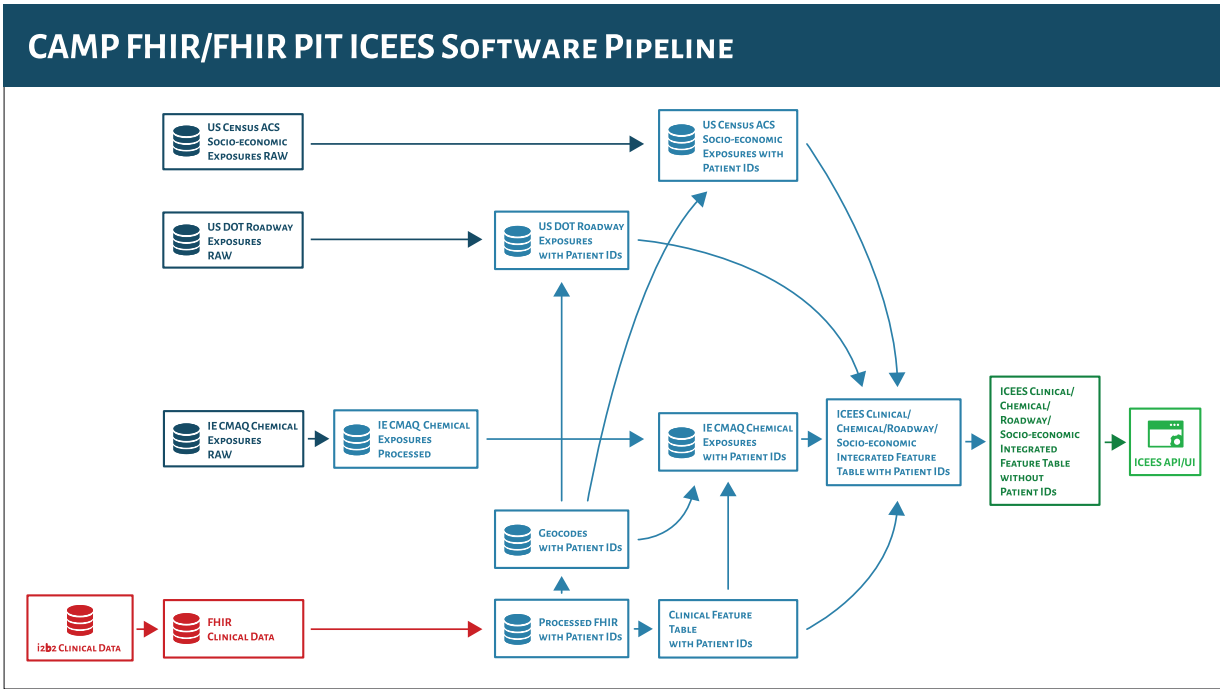

Figure 1. Overview of the process through which the Integrated Clinical and Environmental Exposures Service (ICEES) integrated feature tables are created, including the data sources used to generate the tables. Arrows indicate data processing/integration steps; red denotes private clinical data on patients in the Carolina Data Warehouse for Health; dark blue denotes public data on environmental exposures; light blue denotes stages of processing and integration of clinical data and environmental exposures data, conducted under an Institutional Review Board–approved protocol and within a secure environment; dark green denotes raw, de-identified ICEES integrated feature tables; light green denotes the ICEES OpenAPI and user interface. ACS: American Community Service; API: application programming interface; CAMP FHIR: Clinical Asset Mapping Program for Fast Healthcare Interoperability Resource; DOT: Department of Transportation; FHIR: HL7 Fast Healthcare Interoperability Resource; ID: identifier; IE CMAQ: Institute for the Environment Community Multiscale Air Quality Modeling System; FHIR PIT: FHIR Patient data Integration Tool; UI: user interface.

Table 1. ICEES feature variables: name, description, and binning strategy

Feature Variable[a]	Description and Binning Strategy
CohortID	*Cohort ID# & input variables used to define cohort*
PatientID	*Randomly assigned patient identifier*
StudyPeriod	*Years 2010, 2011, 2012, 2013, 2015*
AgeStudyStart	*Age at "study" start date: calculated from birth date and binned as 0…2, 3…17, 18…34, 35…50, 51…69, 70–89 years*[b]
Sex	*Male, Female, Unknown*
Race	*Caucasian, African American, Asian, Native Hawaiian/Pacific Islander, American/Alaskan Native, Other*
Ethnicity	*Hispanic (1=Yes, 0=No)*
ObesityICD	*ICD code*[c] *for obesity anytime over "study" period (1=Yes, 0=No)*
ObesityBMI	*BMI ≥ 30 anytime over "study" period (1=Yes, 0=No)*
Diagnosis: AsthmaDx, CroupDx, ReactiveAirwayDx, CoughDx, PneumoniaDx	*One or more diagnoses of select respiratory disorders (defined by high-level ICD categories) over "study" period; 1=Yes, 0= No for each diagnosis)*
EstResidentialDensity	*US Census Bureau ACS 2012-2016 estimated total population [block group], binned according to US Census Bureau definitions (1=rural [0, 2500), 2=urban cluster [2500, 50000), 3=urbanized area [50000, inf))*
EstResidentialDensity25Plus	*US Census Bureau ACS 2012-2016 estimated total population aged 25 years or older [block group], binned as quintiles (1, 2, 3, 4, 5)*
EstProbabilityNonHispWhite	*US Census Bureau ACS 2012-2016 estimated proportion of persons who are non-Hispanic white [block group], binned as quartiles (1, 2, 3, 4)*
EstProbabilityHouseholdNonHispWhite	*US Census Bureau ACS 2012-2016 estimated proportion of households that are non-Hispanic white [block group], binned as quartiles (1, 2, 3, 4)*
EstProbabilityHighSchoolMaxEducation	*US Census Bureau ACS 2012-2016 estimated proportion persons aged 25 years or older with a HS diploma or less at their highest level of schooling [block group], binned as quartiles (1, 2, 3, 4)*
EstProbabilityNoAuto	*US Census Bureau ACS 2012-2016 estimated proportion of households without an automobile [block group], binned as quartiles (1, 2, 3, 4)*
EstProbabilityNoHealthIns	*US Census Bureau ACS 2012-2016 estimated proportion of persons without health insurance [block group], binned as quartiles (1, 2, 3, 4)*
EstProbabilityESL	*US Census Bureau ACS 2012-2016 estimated proportion of persons 5 years or older who sometimes speak a language other than English at home [block group], binned as quartiles (1, 2, 3, 4)*
EstHouseholdIncome	*US Census Bureau ACS 2012-2016 estimated median household income [block group], binned as quintiles (1, 2, 3, 4, 5)*
MajorRoadwayHighwayExposure	*US Census TIGERline distance in meters from household to nearest major road/highway (1 = 0-49, 2 = 50-99, 3 = 100-199, 4 = 200-299, 5 = 300-499, 6 = >=500 meters)*
RoadwayDistanceExposure	*US DOT distance in meters from household to nearest roadway (1 = 0-49, 2 = 50-99, 3 = 100-199, 4 = 200-299, 5 = 300-499, 6 = >=500 meters)*
RoadwayType	*UNC DOT roadway classification (eg, major highway)*
RoadwayAADT	*US DOT annual average daily traffic estimate*
RoadwaySpeedLimit	*US DOT roadway speed limit*
RoadwayLanes	*UNC DOT roadway number of lanes*
AvgDailyPM2.5Exposure	*UNC IE average estimated average daily PM2.5 exposure over "study" period, binned by values (1, 2, 3, 4, 5)*
MaxDailyPM2.5Exposure	*UNC IE average estimated maximum daily PM2.5 exposure over "study" period, binned by values (1, 2, 3, 4, 5)*
AvgDailyOzoneExposure	*UNC IE average estimated average daily ozone exposure over "study" period, binned by values (1, 2, 3, 4, 5)*
MaxDailyOzoneExposure	*UNC IE average estimated maximum daily ozone exposure over "study" period, binned by values (1, 2, 3, 4, 5)*
Medications Prescribed or Administered: Prednisone, Fluticasone, Mometasone, Budesonide, Beclomethasone, Ciclesonide, Flunisolide, Albuterol, Metaproterenol, Diphenhydramine, Fexofenadine, Cetirizine, Ipratropium, Salmeterol, Arformoterol, Formoterol, Indacaterol, Theophylline, Omalizumab, Mepolizumab	*One or more prescriptions/administrations of medication over "study" period (1=Yes, 0=No for each medication)*
TotalEDInpatientVisits	*Total number ED/inpatient visits for respiratory issue(s) (defined by same ICD codes used to pull patients with asthma-like conditions) over "study" period (0, 1, 2, 3, …)*

ACS: American Community Survey; BMI: body mass index; DOT: Department of Transportation; ED: emergency department; ICD: International Classification of Diseases; ICEES: Integrated Clinical and Environmental Exposures Service; IE: Institute for the Environment; PM25: particulate matter ≤2.5 μm in diameter; UNC: University of North Carolina.

[a]The feature variables listed in the table are those for the patient-level tables, which include data on each patient for each year of available data (ie, data on individual patients are represented as rows in the table). Similar feature variables are available for the visit-level tables, although the variables are sometimes treated differently. For example, $PM_{2.5}$ exposures are expressed in relation to the 24-hour and 2-week period before visits, not in relation to the 1-year "study" period, as was done for the patient-level tables. Additional feature variables (eg, laboratory measures) are available for select years. Further information can be accessed via the ICEES OpenAPI.

[b]Maximum age is 89 years, per Health Insurance Portability and Accountability Act regulations.

[c]The ICD codes for obesity and for the selection of patients with asthma-like conditions can be found in the Supplementary Appendix.

was based on expert consultation, published literature, or data distributions, depending on the feature variable. For example, ages were coded as 0-2, 3-17, 18-34, 35-50, 51-69, or 70-89 years, similar to the approach taken in our previous work.[5,6] For $PM_{2.5}$, hourly estimates were binned using pandas cut; bin values were then reviewed by a subject matter expert and confirmed using published literature on airborne exposures and asthma exacerbations.[16] Medications were coded as 1 if they were administered or prescribed at least one time over the study period, and 0 otherwise. Diagnoses were coded in the same manner. In addition to the use of binning as a security measure, the application programming interface (API) through which the integrated feature tables can ultimately be queried was designed to provide the end user with aggregated counts only, not the underlying patient- or visit-level data.

An additional consideration when integrating clinical data and environmental exposures data is the treatment of space and time. For instance, exposure to $PM_{2.5}$ varies by time of day (eg, daytime vs nighttime), season (eg, spring vs fall), and location (eg, rural vs urban setting).[19] Likewise, spatiotemporal factors influence a patient's clinical profile. For instance, patient characteristics (eg, age), clinical measures (eg, laboratory tests), and clinical outcomes (eg, number of ED visits) depend on reference time frames. Moreover, clinical signs and symptoms (eg, wheezing) are subject to environmental influences (eg, outdoors vs indoors). We accounted for these facets of the data in the design of the integrated feature tables (see Table 1). Specifically, time was accounted for by considering 1-year study periods to provide a frame of reference for features such as age at study start, number of annual ED or inpatient visits for respiratory issues, and airborne exposures over the 1-year study period. Space was accounted for by considering airborne exposures in relation to a patient's primary residence. While we have not yet accounted for patient mobility (eg, school vs home, work vs home), our general approach is designed to accommodate mobility. In fact, we recently obtained nationwide data on public schools, and we are developing an analytic approach to incorporate those data into the integrated feature tables.

RESULTS

User interface

A Flasgger user interface for the ICEES OpenAPI has been developed (Figure 2). Documentation on the service is available and includes a hyperlink from the API to a collection of files in the ICEES GitHub repository; these files provide plain-text documentation on the data sources, feature variables, integrated feature tables, and existing functionalities. ICEES API example queries also are available via a hyperlink from the API to the ICEES GitHub README file. In addition, an ICEES Jupyter notebook has been created to demonstrate existing functionalities and new applications. Scientific and technical team members have been working together to improve the design of the user interface by, for example, including drop-down menus for selection of feature variables and bins. The expectation is for the two teams to work together to improve the user friendliness of ICEES over time, as a certain amount of technical expertise is currently required to access the data. We anticipate that this will be an iterative refinement process, one that will continue as we expand ICEES and engage new users.

Functionalities

We designed ICEES to initially provide four main functionalities and several related functionalities that support varying degrees of scientific discovery: (1) cohort discovery, (2) feature-rich cohort discovery, (3) hypothesis-driven 2 × 2 feature associations, and (4) exploratory 1 × N feature associations. Each functionality is

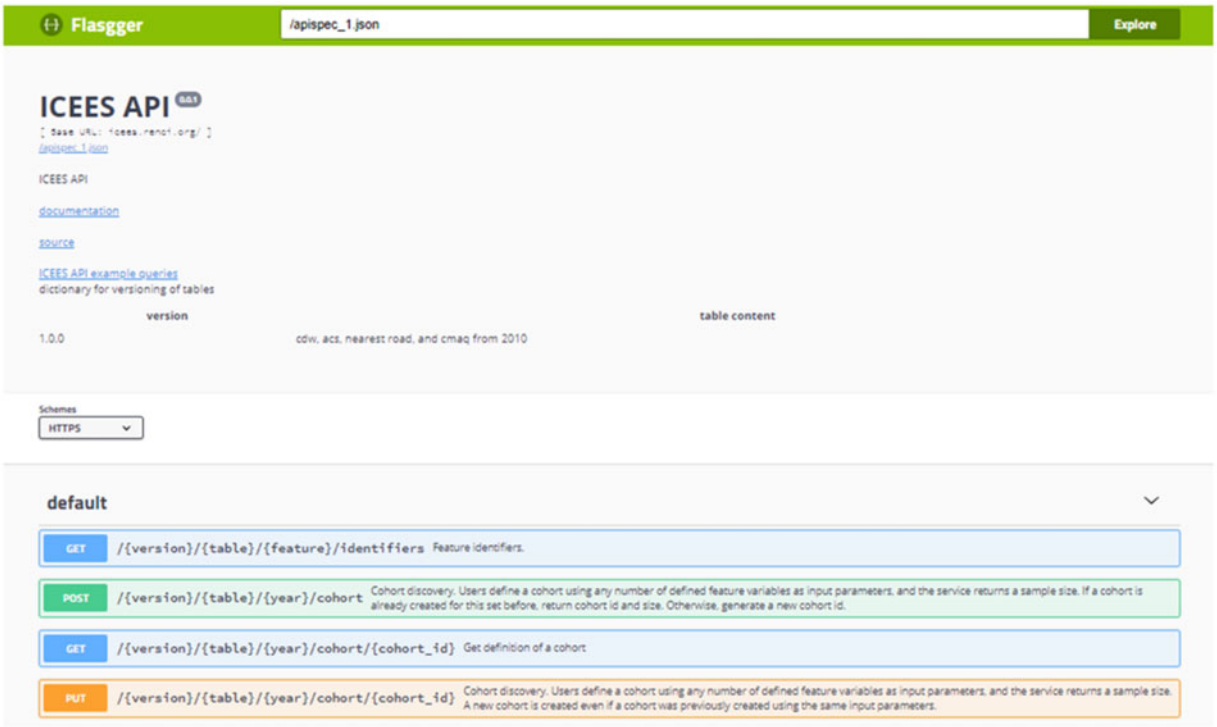

Figure 2. The Integrated Clinical and Environmental Exposures Service application programming interface.

described briefly in the following sections.

Cohort discovery. The first functionality allows for simple cohort discovery and also provides an additional safeguard against accidental leakage of sensitive data. Specifically, a user defines a cohort using any number of feature variables as input parameters, and the service returns a cohort ID# and sample size. Two examples are the following:

Input:

feature variables: {}

version: 1.0.0

table: patient

year: 2010

Output:

"size": 23093,

"cohort_id": "COHORT:22"

Input:

feature variables: {"Race":{"operator":"=","value":"African American"},"Sex":{"operator":"=","value":"M"}}

version: 1.0.0

table: patient

year: 2010

Output:

"size": 2567,

"cohort_id": "COHORT:38"

The output can be interpreted as follows: among all patients with an asthma-like condition in 2010 (cohort 22, n = 23 093), cohort 38 comprises 2567 patients who are men and African American.

Note that if a user-defined set of input parameters yields a cohort of ≤10 patients, then the system automatically returns an error message of: "Input features invalid or cohort ≤10 patients. Please try again." This safeguard was instituted per institutional guidelines and at the request of the CDWH Oversight Committee. An additional safeguard that is built into all four functionalities is that Data Use Agreement (DUA)–like terms and conditions are included as part of the response to the query (see Supplementary Appendix). While we acknowledge that these terms and conditions will be difficult to enforce, we believe that they will serve as a deterrent against inappropriate use of the data. Moreover, they include information on attribution, including funding support for the project, which will help to ensure proper acknowledgement of the service. As a final security feature, service requests that are submitted at a rate of >10 per second are blocked to prevent against malicious attacks on the system.

A related feature allows users to provide a descriptive name for a cohort that is discovered. Specifically, users can choose an ICEES cohort identifier plus a descriptive name for a defined cohort. If a descriptive name has been used previously, the following error message will be returned: "Name is already taken. Please choose another name." Users also can retrieve an ICEES cohort identifier from a descriptive name.

Feature-rich cohort discovery. The second functionality is an expansion of the first one and provides feature-rich cohort discovery. Users input a predefined cohort identification number, and the service returns all feature variables associated with that cohort. The feature variables are presented as counts of patients per bin. Note that cell sizes with ≤10 counts are returned as such, thus providing yet another safeguard against data leakage. Moreover, missing data points add extra "noise" to sample sizes, meaning that sample sizes may not be identical across feature variables. This limitation is inherent when working with any dataset, but it adds an additional safeguard, as it makes it more challenging for users to identify cohorts with ≤10 patients.

Input:

version: 1.0.0

table: patient

year: 2010

cohort_id: COHORT:22

Output:

feature	count
AgeStudyStart = 0-2	2013 8.72%
AgeStudyStart = 3-17	3759 16.28%
AgeStudyStart = 18-34	3542 15.34%
AgeStudyStart = 35-50	4910 21.26%
AgeStudyStart = 51-69	6416 27.78%
AgeStudyStart = 70+	2453 10.62%

...

The previous list demonstrates the second functionality of ICEES. For the selected input parameters, the service returns the age distribution of patients in cohort 22 (all patients in 2010, n = 23 093). Tables for the remaining feature variables follow the previous list and are not shown here. For convenience, the user has the option of viewing and downloading the results in either text/tabular view or JSON format.

Two additional ICEES functionalities are available, as described subsequently and demonstrated in the Preliminary Evaluation Results section.

Hypothesis-driven 2 × 2 feature associations. The third functionality is intended to support hypothesis-driven simple queries. Users input a predefined cohort identification number and two defined clinical feature variables, and the system returns a 2 × 2 feature table with counts of patients per cell and a corresponding chi-square statistic and associated *P* value. Note that we anticipate offering a

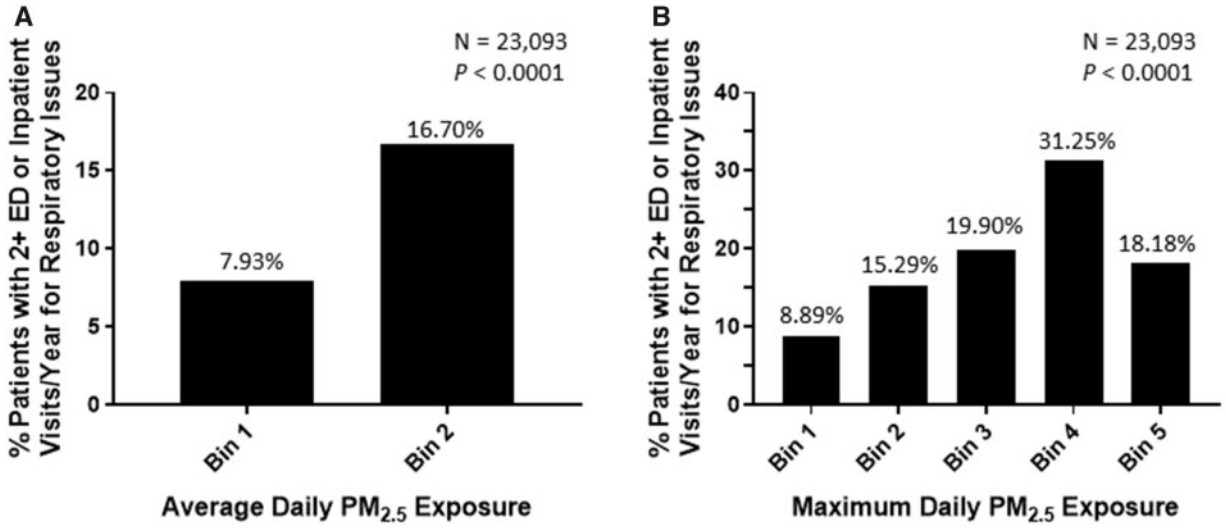

Figure 3. (A) Hypothesis-driven 2 × 2 feature association and (B) exploratory 1 × N feature association demonstrating an association between exposure to particulate matter <=2.5 μm in diameter (PM2.5) and emergency department (ED) and/or inpatient visits for respiratory issues among patients with asthma-like conditions in year 2010 (n = 23 093). Input for the results shown in panel A: {"feature_a":{"TotalEDInpatientVisits":{"operator":"<","value":2}},"feature_b":{"AvgDailyPM2.5Exposure":{"operator":"<","value":3}}}; version: 1.0.0; table: patient; year: 2010; cohort_id: 22. Bins for the results shown in panel A: Bin 1 = 1.58–9.62 μg/m^3; Bin 2 = 9.62–17.33 μg/m^3. Input for the results shown in panel B: {"feature":{"TotalEDInpatientVisits":{"operator":"<","value":2}},"maximum_p_value":0.1}; version: 1.0.0; table: patient; year: 2010; cohort_id: 22. Bins for the results shown in panel B: Bin 1 = 6.77–42.02 μg/m^3; Bin 2 = 42.02–46.21 μg/m^3; Bin 3 = 46.21–47.06 μg/m^3; Bin 4 = 47.06–51.72 μg/m^3; Bin 5 = 51.72–114.94 μg/m^3.

choice of statistics, such as Mutual Information scores and approaches for handling covariates.

Exploratory 1 × N feature associations. The fourth and final functionality supports more exploratory queries. Users input a cohort identification number and a single defined feature variable, and the service returns a 1 × N feature table with counts of patients per cell and corresponding corrected chi-square statistics and associated *P* values. As with the third functionality, we plan to offer a choice of statistics in the future.

Preliminary evaluation results

We have made progress on our driving use-case question and demonstrated a relationship between $PM_{2.5}$ exposure and ED or inpatient visits for respiratory issues (Figure 3). This relationship was established for both average daily $PM_{2.5}$ exposure (Figure 3A) and maximum daily $PM_{2.5}$ exposure (Figure 3B). Results for average daily $PM_{2.5}$ exposure are presented in the context of the third ICEES functionality (hypothesis-driven 2 × 2 feature associations) and demonstrate a greater proportion of patients with two or more annual ED or inpatient visits for respiratory issues among patients with higher average daily $PM_{2.5}$ exposure than among those with lower average daily $PM_{2.5}$ exposure (Figure 3A) ($P < .0001$). Results for maximum daily $PM_{2.5}$ exposure are presented in the context of the fourth ICEES functionality (hypothesis-driven 1 × N feature associations) and demonstrate the dose dependency of the relationship between maximum daily $PM_{2.5}$ exposure and ED or inpatient visits for respiratory issues in that the proportion of patients with two or more annual ED or inpatient visits for respiratory issues increases with increases in maximum daily $PM_{2.5}$ exposure (Figure 3B) ($P < .0001$). While the results presented in Figure 3B are not strictly linear, they clearly trend in the expected direction. Moreover, the sample size for the highest exposure group is small (n = 11), suggesting that the binning strategy may need to be refined, which is something that we are actively exploring (see Validation of Preliminary Findings section).

Validation of preliminary findings

We validated the ICEES service and use-case results in several ways. First, before binning, the quality of the raw integrated data was assessed by examining distributions for errors and integration issues (eg, missing variables, incomplete variables, improperly formatted variables, linkage errors). The raw integrated data also were randomly sampled in order to examine data for individual patients or visits. Second, an independent investigator validated the binning cut points for the exposures data and replicated the API output shown in Figure 3 using the raw integrated data. We are currently conducting an additional study with more complex use-case questions and alternative binning methods to systematically determine if there are substantial differences between binning methods and between binned data and raw integrated data.

Privacy and security safeguards

As discussed previously, ICEES was designed to include a number of built-in safeguards or security features designed to protect against the accidental leakage of sensitive data, while also enabling open access to integrated clinical and environmental exposures data. Specifically:

1. ICEES tables are de-identified after the clinical data are integrated with environmental exposures data;
2. Feature variables are binned or recoded;
3. Only aggregated counts of patients or visits are returned to users;
4. HIPAA-defined PHI is excluded from the integrated feature tables;
5. An error message is returned if the input feature variables identify a cohort of ≤10 patients;

6. Users are not explicitly informed as to why the input feature variables are invalid under scenario (5);
7. Bins with ≤10 counts are returned as such;
8. Missing data points add "noise" to sample sizes;
9. Service requests are restricted to ≤10 per second;
10. Text for DUA-like terms and conditions are returned in the output for service calls;
11. ICEES tables are housed on a secure server located at the Renaissance Computing Institute; and
12. ICEES tables are stored on an encrypted hard drive while at rest and encrypted via SSL while in motion.

DISCUSSION

As part of the Translator program, we have developed a regulatory-compliant framework and approach for openly exposing and sharing clinical data that have been integrated with data on environmental exposures: ICEES. We have provided preliminary evidence for the validity and utility of ICEES by replicating a known relationship between exposure to $PM_{2.5}$ and responsiveness to treatment among patients with asthma-like conditions, our driving demonstration use case.

We emphasize that the ability to reproduce prior findings using ICEES is not a trivial task, but rather is fraught with complexities and potential pitfalls. For instance, the $PM_{2.5}$ exposure estimates are based on sensor readings within relatively large geographical grids and derived using a complex analytic model that attempts to account for factors such as temperature, humidity, and altitude. Moreover, all exposure estimates are in reference to a patient's primary residence, which presumably is a home address, but because EHR systems are intended to support administrative tasks, not research, the primary residence may instead represent a billing address. In addition, the binning strategy that was used to expose the exposure estimates may render the "signal" lost in the noise. Indeed, while we have based the binning strategy on a combination of expert consultation, published literature, and logical inferences based on data distributions, the development of appropriate binning strategies for the feature variables has proven to be challenging and something we are systematically investigating in a different study.

In addition to the binning strategy, we continue to critically assess and evaluate other aspects of ICEES. For instance, our preliminary evaluation was based on one year of data. Having demonstrated proof of concept, we are now in the process of integrating and exposing multiple years of data on ~160 000 patients total. We also have begun analyzing visit-level data, as opposed to the patient-level data shown in Figure 3. The visit-level data will allow us to directly examine airborne pollutant exposures over the 2-week period immediately preceding ED or inpatient visits for respiratory issues.[16] Finally, we have begun exploring additional feature variables that are currently captured in ICEES. For example, we plan to determine whether we can replicate the established association between ozone exposure and asthma exacerbations.[20–24] We also plan to establish whether we can reproduce published literature demonstrating an association between roadway exposure and asthma exacerbations.[25–27] Additionally, we will determine whether we can use ICEES to demonstrate a relationship between asthma and established socioeconomic risk factors such as race, ethnicity, poverty, household educational attainment, and English language proficiency.[28–34]

We note that several healthcare organizations have launched large-scale open-data initiatives, including COHD[9] and MIMIC.[10] Our results complement and extend these other initiatives by providing capabilities beyond those available through COHD or MIMIC. For instance, COHD is restricted to clinical co-occurrence data, whereas ICEES is designed as a patient-level or visit-level N x N array of feature variables. MIMIC provides access to data on critical care only and access is restricted to persons with a researcher designation who have completed human-subjects training and obtained a fully executed DUA. In contrast, ICEES is available to anyone and accessible via an OpenAPI. Of significance, ICEES offers access to data beyond that available through COHD, MIMIC, or any EHR system; namely, environmental exposures data that have been integrated with clinical data at the patient and visit level. To the best of our knowledge, no other healthcare organization offers such data.

The current design of the integrated feature tables is geared toward patients with asthma-like conditions. However, we emphasize that ICEES serves as a disease-agnostic, scalable, open framework and approach for exploring myriad other diseases and conditions that are sensitive to environmental exposures. In fact, we are developing new clinical use cases and associated ICEES tables to address diseases and conditions that are sensitive to environmental exposures. For instance, recent findings suggest an association between exposure to airborne pollutants and risk of diseases and conditions as diverse as cardiovascular disease,[35] diabetes,[36] dementia,[37] and premature death.[18] We also wish to integrate new data sources that appear to have a major influence on asthma and other diseases, such as proximity to concentrated animal farming operations.[38]

We acknowledge that new use cases and data sources will require considerable subject matter expertise. However, the general framework and approach is completely scalable and can be adapted to virtually any use case. Moreover, components of ICEES (eg, demographic data, medications) are invariant and can be reused for new use cases; other components (eg, laboratory values, environmental exposures) are variant and will require initial preprocessing, but will then serve as invariant components for reuse in subsequent use cases. In addition, the API was designed to be generalizable and can be readily adapted for new use cases and data types. Thus, we expect the scalability of ICEES to improve iteratively over time as new applications are realized and implemented.

ICEES was designed and implemented with numerous safeguards and security features that are intended to minimize the risk of intentional or inadvertent leakage of sensitive patient data, while also permitting access to clinical data that have been integrated at the patient and visit level with environmental exposures data derived from multiple sources. These safeguards were informed by the CDWH Oversight Committee and required an honest, back-and-forth dialogue among scientific, technical, and regulatory teams at our institution. We greatly appreciate the willingness of the committee and team members to engage in the dialogue necessary to develop the service in such a way as to accommodate the vision of the Translator program and meet increasing societal demands for open access to patient data,[39] while also making every effort to ensure for the privacy and security of data on individual patients. We expect to remain engaged with the committee as we continue to develop ICEES. For instance, new cohorts will require regulatory review by the IRB and the CDWH Oversight Committee. Thus, we anticipate an ongoing partnership with our institution's regulatory teams. We encourage readers to learn from our example.

Last, we assert that ICEES has the potential to serve as a regulatory-compliant technical prototype and pipeline for openly

exposing integrated clinical and environmental exposures data—irrespective of clinical data model. With this in mind, we recently developed an extensible clinical data conversion pipeline that invokes a custom software application, CAMP FHIR (Clinical Asset Mapping Program for Fast Healthcare Interoperability Resource), to transform clinical data from common data models (eg, Informatics for Integrating Biology and the Bedside, National Patient-Centered Clincial Research Network, Observational Medical Outcomes Partnership) into HL7 Fast Healthcare Interoperability Resource (FHIR) files.[40] A subsequent custom software application, FHIR PIT (Patient data Integration Tool), then integrates the clinical data with environmental exposures data from multiple sources before stripping the data of PHI and binning feature variables to create ICEES tables. Of note, FHIR PIT is modular and extensible and can be adapted for virtually any type of data that is of interest to clinical researchers and requires geocodes, dates, and identifiers for integration with EHR data. We expect CAMP FHIR and FHIR PIT to greatly improve the scalability of ICEES and automate the creation of integrated feature tables. CAMP FHIR and FHIR PIT are currently undergoing unit testing and will be released under an open GitHub license.

CONCLUSIONS

As part of the Translator program, we have developed ICEES as a novel service that provides regulatory-compliant open access to clinical data integrated with environmental exposures data at the patient and visit level. Our preliminary evaluation of ICEES has demonstrated the validity of the integrated datasets and our overall approach. We believe that ICEES, through an iterative process of refinement and expansion, will advance the vision of the Translator program and generate new knowledge that will inform and accelerate clinical and translational research. Moreover, we believe that the ICEES framework and approach will serve as a regulatory-compliant model for promoting open sharing of integrated clinical and environmental exposures data.

FUNDING

This work was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, grant numbers OT3TR002020 (Stanley C. Ahalt, PhD, PI) and OT2TR002514 (Alexander Tropsha, PhD, PI). The Carolina Data Warehouse for Health is the clinical data warehouse for data on patients in the UNC Health Care System. The Carolina Data Warehouse for Health is supported by the North Carolina Translational and Clinical Sciences Institute and funded by the National Center for Advancing Translational Sciences (UL1TR002489). John Buse, MD, PhD, PI.

AUTHOR CONTRIBUTIONS

KF led the design and evaluation of the Integrated Clinical and Environmental Exposures Service (ICEES), contributed to the driving use case, conducted the preliminary analysis of ICEES, contributed to the design of FHIR PIT, and prepared the first draft of this manuscript. EP extracted and transformed raw clinical data into the ICEES format, provided regulatory and data-security expertise, and contributed to the design of ICEES and CAMP FHIR. HX contributed to the technical design, implementation, and deployment of the ICEES preprocessing pipeline, back end, and front end. HX also was the primary software developer for the ICEES API back end and front end and currently serves as the primary software developer for the FHIR PIT software application. JC was the primary software developer for the predecessor to FHIR PIT and currently serves as the primary software developer for the CAMP FHIR software application. SC provided technical oversight of the project and contributed to the design of ICEES, CAMP FHIR, and FHIR PIT. LS developed APIs to access the environmental exposures data, managed the environmental exposures data, and contributed software code to address spatial and temporal aspects of the data integration pipeline. DBP led the development of the driving use case, contributed to the design of ICEES, and provided critical feedback on the preliminary results. CB contributed to the design of ICEES and provided critical feedback on the preliminary results. AK and AT provided critical feedback on the design of ICEES and the preliminary results. SCA and KF together conceptualized ICEES. SCA further provided critical feedback on the design of ICEES and the preliminary results. All authors reviewed the manuscript, offered feedback, approved journal submission, and agreed to be accountable for the work.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGEMENTS

The authors acknowledge that this work was inspired by the work of other members of the Biomedical Data Translator Consortium. Specifically, the work of Nicholas Tatonetti, Chunhua Weng, Casey Ta, and others involved with Columbia Open Health Data motivated the conceptualization of the Integrated Clinical and Environmental Exposures Service (ICEES). The authors also wish to acknowledge the oversight and support provided by the Biomedical Data Translator Consortium and the North Carolina Translational and Clinical Sciences Institute, including Ms Marie Rape, who provided a careful review of the regulatory text and terminology. Kelsey Urgo created Figure 1; Finally, Kenneth Morton of CoVar Applied Technologies provided critical input on ICEES; Colin K. Curtis created a Jupyter notebook for ICEES; Vinicious Medeiras Alves and Eugene Muratov of University of North Carolina (UNC) School of Pharmacy independently validated the ICEES use-case results and provided critical feedback; Michael Stealey of the Renaissance Computing Institute contributed expertise on Swagger APIs; Steve Appold of UNC Kenan-Flagler Business School and Ann Moss Joyner and Allan Parnell of Cedar Grove Institute for Sustainable Communities provisioned the U.S. Census Bureau American Community Service socioeconomic exposures data; Sarav Arunchalam of the UNC Institute for the Environment provisioned the Community Multiscale Air Quality Modeling System airborne exposures data; Alex Valencia of the UNC Institute for the Environment provisioned the U.S. Department of Transportation roadway exposures data; and Charles Schmitt of the National Institute of Environmental Health Science contributed expertise on the exposures data. The authors greatly appreciate these contributions.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Emanuel EJ, Wendler D, Grady C. What makes clinical research ethical? *JAMA* 2000; 283 (20): 2701–11.
2. Mandal J, Acharya S, Parija SC. Ethics in human research. *Trop Parasitol* 2011; 1 (1): 2–3.
3. Lubarski B. Re-identification of "anonymized" data. Georgetown Law Technology Review; April 2017. https://www.georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017 Accessed March 12, 2019.

4. Wacker J, Kolbe M. The challenge of learning from perioperative patient harm. *Trends Anaesthesia Critical Care* 2016; 7–8: 5–10.
5. Fecho K, Lunney AT, Boysen PG, Rock P, Norfleet EA. Postoperative mortality after inpatient surgery: incidence and risk factors. *Ther Clin Risk Manag* 2008; 4 (4): 681–8.
6. Fecho K, Moore CG, Lunney AT, Rock P, Norfleet EA, Boysen PG. Anesthesia-related perioperative adverse events during in-patient and outpatient procedures. *Int J Health Care Qual Assur* 2008; 21 (4): 396–412.
7. Fecho K, Jackson F, Smith F, Overdyk FJ. In-hospital resuscitation: opioids and other factors influencing survival. *Ther Clin Risk Manag* 2009; 5: 961–8.
8. Williams GD, Muffly MK, Mendoza JM, Wixson N, Leong K, Claure RE. Reporting of perioperative adverse events by pediatric anesthesiologists at a tertiary children's hospital: targeted interventions to increase the rate of reporting. *Anesth Analg* 2017; 125 (5): 1515–23.
9. Ta CN, Dumontier M, Hripcsak G, Tatonetti NP, Weng C. Columbia Open Health Data for clinical concept prevalence and co-occurrence from electronic health records. *Sci Data* 2018; 5: 180273.
10. Johnson AEW, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3: 160035.
11. Biomedical Data Translator Consortium. The Biomedical Data Translator program: conception, culture, and community. *Clin Transl Sci* 2019; 12 (2): 91–4.
12. The Biomedical Data Translator Consortium. Toward a universal biomedical data translator. *Clin Transl Sci* 2019; 12 (2): 86–90.
13. Brunekreef B, Janssen NAH, de Hartog J, Harssema H, Knape M, van Vliet P. Air pollution from truck traffic and lung function in children living near motorways. *Epidemiology* 1997; 8 (3): 298–303.
14. Delfino RJ. Epidemiologic evidence for asthma and exposure to air toxics: linkages between occupational, indoor, and community air pollution research. *Environ Health Perspect* 2002; 110 (Suppl 4): 573–89.
15. Brugge D, Durant JL, Rioux C. Near-highway pollutants in motor vehicle exhaust: a review of epidemiologic evidence of cardiac and pulmonary health risks. *Environ Health* 2007; 6: 23.
16. Mirabelli MC, Vaidyanathan A, Flanders WD, Qin X, Garbe P. Outdoor $PM_{2.5}$, ambient air temperature, and asthma symptoms in the past 14 days among adults with active asthma. *Environ Health Perspect* 2016; 124 (12): 1882–90.
17. Salimi F, Morgan G, Rolfe M, *et al*. Long-term exposure to low concentrations of air pollutants and hospitalization for respiratory diseases: a prospective cohort study in Australia. *Environ Int* 2018; 121 (Pt 1): 415–20.
18. Chang SY, Vizuete W, Serre M, *et al*. Finely resolved on-road $PM_{2.5}$ and estimated premature mortality in central North Carolina. *Risk Anal* 2017; 37 (12): 2420–34.
19. Liu Z, Hu B, Wang L, Wu F, Gao W, Wang Y. Seasonal and diurnal variation in particulate matter (PM10 and PM2.5) at an urban site of Beijing: analyses from a 9-year study. *Environ Sci Pollut Res Int* 2015; 22 (1): 627–42.
20. Balmes JR. The role of ozone exposure in the epidemiology of asthma. *Environ Health Perspect* 1993; 101 (Suppl 4): 219–24.
21. Delfino RJ, Coate BD, Zeiger RS, Seltzer JM, Street DH, Koutrakis P. Daily asthma severity in relation to personal ozone exposure and outdoor fungal spores. *Am J Respir Crit Care Med* 1996; 154 (3): 633–41.
22. Lierl MB, Hornung RW. Relationship of outdoor air quality to pediatric asthma exacerbations. *Ann Allergy Asthma Immunol* 2003; 90 (1): 28–33.
23. Schildcrout JS, Sheppard L, Lumley T, Slaughter JC, Koenig JQ, Shapiro GG. Ambient air pollution and asthma exacerbations in children: an eight-city analysis. *Am J Epidemiol* 2006; 164 (6): 505–17.
24. Sheffield E, Zhou J, Shmool JLC, Clougherty JE. Ambient ozone exposure and children's acute asthma in New York City: a case-crossover analysis. *Environ Health* 2015; 14: 25.
25. English P, Neutra R, Scalf R, Sullivan M, Waller L, Zhu L. Examining associations *between* childhood asthma and traffic flow using a geographic information system. *Environ Health Perspect* 1999; 107 (9): 761–7.
26. Venn AJ, Lewis SA, Cooper M, Hubbard R, Britton J. Living near a main road and the risk of wheezing illness in children. *Am J Respir Crit Care Med* 2001; 164 (12): 2177–80.
27. Lin S, Munsie JP, Hwang SA, Fitzgerald E, Cayo MR. Childhood asthma hospitalization and residential exposure to state route traffic. *Environ Res Section A* 2002; 88 (2): 73–81.
28. Litonjua AA, Carey VJ, Weiss ST, Gold DR. Race, socioeconomic factors, and area of residence are associated with asthma prevalence. *Pediatr Pulmonol* 1999; 28 (6): 394–401.
29. Smith LA, Hatcher-Ross JL, Wertheimer R, Kahn RS. Rethinking race/ethnicity, income, and childhood asthma: racial/ethnic disparities concentrated among the very poor. *Public Health Rep* 2005; 120 (2): 109–16.
30. Lara M, Akinbami L, Flores G, Morgenstern H. Heterogeneity of childhood asthma among Hispanic children: Puerto Rican children bear a disproportionate burden. *Pediatrics* 2006; 117 (1): 43–53.
31. Forno E, Celedón JC. Health disparities in asthma. *Am J Respir Crit Care Med* 2012; 185 (10): 1033–5.
32. Keet CA, Matsui ED, McCormack MC, Peng TD. Urban residence, neighborhood poverty, race/ethnicity, and asthma morbidity among children on Medicaid. *J Allergy Clin Immunol* 2017; 140 (3): 822–7.
33. Assari S, Moghani Lankarani M. Poverty status and childhood asthma in white and black families: National Survey of Children's Health. *Healthcare (Basel)* 2018; 6 (2): E62.
34. Montgomery MP, Allen ED, Thomas O, *et al*. Association between pediatric asthma care quality and morbidity and English language proficiency in Ohio. *J Asthma* 2018 May 8. May[Epub ahead of print]
35. Hennig F, Fuks K, Moebus S, *et al*.; Heinz Nixdorf Recall Study Investigative Group. Association between source-specific particulate matter air pollution and hs-CRP: local traffic and industrial emissions. *Environ Health Perspect* 2014; 122 (7): 703–10.
36. Weinmayr G, Hennig F, Fuks K, *et al*.; Heinz Nixdorf Recall Investigator Group. Long-term exposure to fine particulate matter and incidence of type 2 diabetes mellitus in a cohort study: effects of total and traffic-specific air pollution. *Environ Health* 2015; 14: 53.
37. Chen H, Kwong JC, Copes R, *et al*. Living near major roads and the incidence of dementia, Parkinson's disease, and multiple sclerosis: a population-based cohort study. *Lancet* 2017; 389 (10070): 718–26.
38. Kravchenko J, Rhew SH, Akushevich I, Agarwal P, Lyerly HK. Mortality and health outcomes in North Carolina communities located in close proximity to hog concentrated animal feeding operations. *NC Med J* 2018; 79 (5): 278–88.
39. Broes S, Lacombe D, Verlinden M, Huys I. Sharing human samples and patient data: opening Pandora's box. *J Cancer Policy* 2017; 13: 65–9.
40. Pfaff ER, Champion J, Cox S, *et al*. All roads lead to FHIR: an extensible clinical data conversion pipeline. *Paper presented at AMIA 2019, March 25–28*, 2019; San Francisco, CA.