An approach for open multivariate analysis of integrated clinical and environmental exposures data

Karamarie Fecho[a,*,1], **Perry Haaland**[b], **Ashok Krishnamurthy**[a,c], **Bo Lan**[d], **Stephen A. Ramsey**[e], **Patrick L. Schmitt**[a], **Priya Sharma**[a], **Meghamala Sinha**[e], **Hao Xu**[a]

[a]Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[b]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[c]Department of Computer Science, University of North Carolina, Chapel Hill, NC, USA

[d]UNC Highway Safety Research Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[e]Oregon State University, Corvallis, OR, USA

Abstract

The Integrated Clinical and Environmental Exposures Service (ICEES) provides regulatory-compliant open access to sensitive patient data that have been integrated with public exposures data. ICEES was designed initially to support dynamic cohort creation and bivariate contingency tests. The objective of the present study was to develop an open approach to support multivariate analyses using existing ICEES functionalities and abiding by all regulatory constraints. We first developed an open approach for generating a multivariate table that maintains contingencies between clinical and environmental variables using programmatic calls to the open ICEES application programming interface. We then applied the approach to data on a large cohort (N = 22,365) of patients with asthma or related conditions and generated an eight-feature table. Due to regulatory constraints, data loss was incurred with the incorporation of each successive feature variable, from a starting sample size of N = 22,365 to a final sample size of N = 4,556 (20.4%), but data loss was < 10% until the addition of the final two feature variables. We then applied a generalized linear model to the subsequent dataset and focused on the impact of seven select feature variables on asthma exacerbations, defined as annual emergency department or inpatient visits for respiratory issues. We identified five feature variables—sex, race, obesity, prednisone, and airborne particulate exposure—as significant predictors of asthma exacerbations. We discuss the advantages and disadvantages of ICEES open multivariate analysis and conclude that, despite limitations, ICEES can provide a valuable resource for open multivariate analysis and can serve as

*Corresponding author. University of North Carolina at Chapel Hill, RENCI, 100 Europa Drive, Suite 540, Chapel Hill, NC, 27517, USA. kfecho@copperlineprofessionalsolutions.com (K. Fecho).
[1]Authors are listed alphabetically.

an exemplar for regulatory-compliant informatic solutions to open patient data, with capabilities to explore the impact of environmental exposures on health outcomes.

Keywords

Open science; Open clinical data; Generalized linear model; Asthma; Environmental exposures; Environmental health

1.   Introduction

Interest in open access to and sharing of electronic health record (EHR) data has been growing in recent years, both among the medical research community and patient advocacy groups. The benefits of such an effort are perhaps best highlighted by the current coronavirus pandemic and the need to rapidly initiate research into the virus and its impacts on health, as well as share data and findings and develop a global response to this unprecedented health crisis. Large-scale Initiatives such as the Columbia Open Health Data (COHD) [1] and the Medical Information Mart for Intensive Care (MIMIC) [2] are advancing efforts to reduce regulatory and institutional barriers surrounding access to EHR data, with the common goal of promoting research, while preserving patient privacy and maintaining institutional assurances. However, further efforts are required to truly leverage EHR data and apply the data to promote global human health and well-being.

As part of the Biomedical Data Translator ('Translator') program [3, 4], funded by the National Center for Advancing Translational Sciences, we have developed a novel, regulatory-compliant, disease-agnostic framework and approach for openly exposing and exploring EHR data that have been integrated at the patient level with a variety of public exposures data: the Integrated Clinical and Environmental Exposures Service (ICEES). ICEES is accessible to anyone on the internet via an application programming interface (API). We have validated ICEES by replicating published research on asthma and related common pulmonary diseases [5-8]. We have extended ICEES to expose multi-institutional data on patients within UNC Health who are also participants within the Environmental Polymorphisms Registry at the National Institute of Environmental Health Sciences [7]. Moreover, as part of the Translator program, we have used ICEES to conduct a multi-institutional study, free of regulatory constraints, over the course of a five-day 'hackathon' [10]. We also have developed a tool for visualizing and exploring ICEES as a 'knowledge graph' of interconnected nodes [11].

While ICEES has demonstrated technical validity and scientific application, the service remains subject to federal and institutional constraints that, while necessary, limit the available functionalities. ICEES currently supports the ability to dynamically define cohorts and explore bivariate relationships between feature variables such as diagnoses, medications, and airborne pollutant exposures. Herein, we describe the development of a novel open approach that supports multivariate analysis using existing regulatory-compliant ICEES functionalities, while maintaining all federal and institutional regulations and preserving patient privacy. We apply the open multivariate approach to a driving use case on asthma,

using a generalized linear model (GLM) to predict asthma exacerbations. Finally, we discuss the advantages and disadvantages of using ICEES for multivariate analysis.

2. Materials and methods

All study procedures were approved by the Institutional Review Board at the University of North Carolina at Chapel Hill (protocol #16-2978).

2.1. Technical overview

2.1.1. Open multivariate approach—ICEES is equipped with regulatory-compliant analytic capabilities that allow users to dynamically create cohorts and generate bivariate contingency tables with corresponding Chi Square statistics, probabilities, and frequencies. Motivated by our desire to develop more sophisticated multivariate analytic capabilities, here we describe the development and application of an open approach to conduct multivariate analysis using the functionalities that are currently available in ICEES. The approach leverages the dynamic cohort creation capability in such a way as to maintain feature contingencies and generate a multivariate table, while remaining compliant with all federal and institutional regulations.

2.1.2. ICEES integrated feature tables—Key to the design of ICEES is what we've termed 'ICEES integrated feature tables'. The tables are designed as one-year 'study periods', in which each row represents an individual patient and each column header represents a feature variable. The tables contain integrated data on clinical data elements derived from patient EHRs and exposures data derived from a variety of public sources (e.g., United States Environmental Protection Agency airborne particulate exposures, US Department of Transportation roadway exposures, US Census Bureau American Community Survey socioeconomic exposures, North Carolina Department of Environmental Quality concentrated animal farming operations exposures and landfill exposures). The integration step is achieved using a complex custom software pipeline [8] and requires protected health information (PHI; i.e., geocodes and dates). As such, this step is conducted under a protocol that must be approved by an Institutional Review Board. After integration, however, all PHI elements are removed from the data according to §164.514(b) of the Health Insurance Portability and Accountability Act [12]. The data are then exposed via an open ICEES API [13] that adheres to the Translator Application Programming Interface (TRAPI) standards [14].

2.1.3. Generation of multivariate ICEES integrated feature tables—We developed the multivariate approach in the context of a driving application use case, in which we asked if there is a relationship between asthma exacerbations and the following demographic features, clinical characteristics, and environmental exposures: sex, race, prescription for prednisone, diagnosis of obesity, residential proximity to a major roadway or highway, residential density, and exposure to airborne particulates. These variables were selected on the basis of published studies, including our prior work [6,8,9], which identified these variables as known or suspected to be related to asthma exacerbations. We focused on an existing ICEES cohort of UNC Health patients with asthma or related conditions (see

Ref. [6] for details), and we considered the number of annual emergency department (ED) or inpatient visits for respiratory issues as the primary outcome measure and indicator of asthma exacerbations. We examined asthma exacerbations in year 2010, as this was the first year of data available for the cohort.

The seven features or independent variables and the primary outcome metric or dependent variable are defined and enumerated in Table 1.

While ICEES supports functionalities to examine the bivariate relationship between ED/inpatient visits for respiratory issues and each of the feature variables of interest, it does not directly support the application of multivariate statistical or machine learning models to examine relationships and interactions across multiple feature variables. To apply multivariate models, an eight-feature table was required, with each row representing an individual patient and each column header representing a distinct feature variable, with contingencies maintained across feature variables. To achieve this, we applied the ICEES dynamic cohort creation functionality and used nested bivariate contingencies to generate the requisite multivariate feature table. A visual overview of the approach is provided in Fig. 1.

Specifically, we first selected the asthma cohort, table type (patient or visit), table version, and calendar year of interest as the input parameters. We then created separate cohorts for each level of the dependent variable (i.e., ED/inpatient visits for respiratory issues):

> COHORT:0 = patients in asthma cohort with 0 annual ED/inpatient visits for respiratory issues
>
> COHORT:1 = patients in asthma cohort with 1 annual ED/inpatient visit for respiratory issues
>
> COHORT:2 = patients in asthma cohort with 2 annual ED/inpatient visits for respiratory issues
>
> COHORT:N = patients in asthma cohort with N annual ED/inpatient visits for respiratory issues

The boundary of COHORT:N is determined by both the underlying data (i.e., the maximum number of annual ED/inpatient visits reported for a patient in any given year) and the regulatory constraints imposed on the ICEES service, namely, that cohorts with ≤ 10 patients cannot be created, in which case, the service returns an error message indicating that the data do not exist or that the selected cohort consists of ≤ 10 patients. The practical implication of this regulatory constraint for the efforts described here was that a certain amount of data loss was incurred with each step in the process of generating a multivariate table. We quantified the data loss by comparing the size of the sum of each cohort, or the number of rows for each intermediary table, with the size of the overall sample.

The next step in the process for creating a multivariate table was to create a bivariate contingency table for each of the cohorts generated in the first step. In our example use case, we used *Sex2 x Race*. Because the contingencies between feature variables were maintained,

we were then able to create a tri-variate table, with rows transformed to represent $N = 1$ *patient* (Fig. 2).

The next step was to create cohorts for each combination of the three feature variables.

COHORT:100 = patients in asthma cohort with 0 annual ED/inpatient visits for respiratory issues + male sex + African American

COHORT:101 = patients in asthma cohort with 1 annual ED/inpatient visit for respiratory issues + male sex + African American

COHORT:102 = patients in asthma cohort with 2 annual ED/inpatient visits for respiratory issues + male sex + African American

COHORT:N = patients in asthma cohort with N annual ED/inpatient visits for respiratory issues + X sex + X race

For each of the new cohorts, a second bivariate contingency table was generated. In our example, the association was for *Prednisone x ObesityDx*. The cohort creation and bivariate contingency table steps were then repeated for *MaxDailyPM2.5Exposure_StudyMax x RoadwayDistanceExposure2*. As we were interested in an odd number of independent variables, the final step that we applied to the data was to invoke the ICEES univariate functionality (also called feature-rich cohort discovery) to examine frequencies for *EstResidentialDensity*. Upon completion of this step, we then were able to generate an eight-feature multivariate table, with each row representing an individual patient (see Results). For interpretation purposes, and to minimize data loss, we categorized the dependent variable, *TotalEDInpatientVisits*, as 0, 1, … 9+.

2.1.4. Application of multivariate GLM—We developed a GLM algorithm using R to predict TotalEDInpatientVisits (i.e., the outcome or dependent variable) using the seven independent feature variables extracted in the ICEES multivariate table. The variable TotalEDInpatientVisits represents counts of visits over a one-year study period, and the distribution of counts was over-dispersed, or skewed to the right, as expected (i.e., few patients have frequent ED visits or hospital admissions for respiratory issues in any given year). To account for overdispersion, we fit a negative binomial model to the data [15]. The theoretical equation for the negative binomial distribution expressed as a mass probability function was

$$P(X = x \mid r, p) = \frac{\Gamma(x + r)}{x!\,\Gamma(r)}\, p^r (1 - p)^x,$$

where the negative binomial is arising as a distribution of the number of failures X before the rth success in independent trials, with success probability p in each trial (consequently, r 0 and 0 p 1).

To ensure that the negative binomial model accurately fit the data and accounted for overdispersion, we formally tested for overdispersion and its significance using the DHARMa package in R for Residual Diagnostics for Hierarchical (Multi-Level/Mixed) Regression Models [16]. This package applies a simulation-based approach to create readily-

interpretable scaled (quantile) residuals for generalized linear mixed models, including negative binomial GLM. The DHARMa model revealed that the true value of the ratio of the deviance to the degrees of freedom was 0.89 (dispersion parameter), with a P value of $< 2.2e^{-16}$, indicating that the negative binomial model would sufficiently address the overdispersion issue.

Because the frequencies for certain variables (e.g., RoadwayDistanceExposure2) were not evenly distributed across bins, we applied the Synthetic Minority Oversampling Technique (SMOTE) [17] to account for imbalances in the data. The SMOTE approach augments the minority class in order to balance the data such that model performance accounts for cells with otherwise low frequencies. Prior to adopting the SMOTE approach, we considered several other estimation techniques to address data imbalance, including random sampling. There are two main approaches to perform random sampling, both of which we considered: (i) oversampling, or replication of samples from the minority class; and (ii) undersampling, or elimination of samples from the majority class. Because random sampling introduces bias and is considered naïve, with no assumptions regarding the data [18], we chose the SMOTE approach instead.

Satisfied with our approach, we applied GLM to examine both main effects and two- and three-way interactions. We then applied an analysis of variance (ANOVA) to the obtained GLM results. We set α at 0.05.

3. Results

3.1. Eight-feature multivariate table and estimated data loss

We applied the ICEES open multivariate approach to generate an eight-feature multivariate table designed to support our driving application use case on the effects of select demographic variables, socioeconomic exposures, and airborne pollutant exposures on asthma exacerbations (Fig. 3).

We then quantified the data loss that occurs with each step in the process by comparing the size of each recreated cohort with that of the initial cohort (Table 2). Data loss was incurred after the fourth independent variable was incorporated and increased to 57.1% with the incorporation of the seventh and final independent variable.

3.2. Application use case results

We applied a GLM algorithm to the resultant multivariate table and asked the following specific use-case question: are sex, race, prescription for prednisone, diagnosis of obesity, residential proximity to a major roadway or highway, residential density, and/or exposure to airborne particulates predictive, either independently or by interaction, of asthma exacerbations?

We found significant main effects of *Race, Prednisone, ObesityDx, MaxDailyPM2.5Exposure_StudyMax,* and *Sex2* (Table 3). Several two- and three-way interactions also were significant. Among two-way interactions, *Prednisone* showed significant interactions with *Sex2, ObesityDx, MaxDailyPM2.5Exposure_StudyMax,* and

RoadwayDistance Exposure2. A significant *Sex2* x *ObesityDx* effect also was apparent. Among three-way interactions (data not shown), *ObesityDx* x *Sex2* x *Race*, *ObesityDx* x *Sex2* x *Prednisone*, and *ObesityDx* x *Race* x *Prednisone* were significant. Higher-level interactions were not significant.

4. Discussion

We demonstrated the ability to programmatically use existing regulatory-compliant ICEES functionalities (i.e., dynamic cohort creation and bivariate contingencies) to generate a multivariate integrated feature table. Importantly, we developed and applied a GLM model to the resultant multivariate table and identified five feature variables—*Prednisone, Race, ObesityDx, Sex*2, *MaxDailyPM2.5Exposure_StudyMax*—as significant predictors of *TotalEDInpatientVisits.*

Importantly, our application findings are in agreement with the published literature. For instance, prednisone is commonly prescribed for the treatment of acute asthma exacerbations in patients who are non-responsive to first-line treatments such as inhaled albuterol [19]. Female sex, obesity, and African American race have previously been identified as variables that contribute to asthma exacerbations. For example, Greenblatt et al. [20] found that female sex and obesity (among other variables) significantly increased the odds of asthma exacerbations. Our prior work [8] and that of others [21] have found a significant association between African American race and increased risk of asthma exacerbations. Finally, exposure to airborne particulate matter is a well-established risk factor for asthma and asthma exacerbations. For example, a study by Requia et al. [22] found a significant association between a two-year increase of 10 μg/m^3 PM$_{2.5}$ in 117 regions in Canada and increased risk in the incidence of asthma. Mirabelli et al. [23] likewise found a significant association between exposure to PM$_{2.5}$ and risk of asthma. We also have demonstrated an association between exposure to high levels of PM$_{2.5}$ and asthma exacerbations [6]. Exposure to major roadways or highways is often used as a proxy for airborne particulate exposure. Indeed, several groups have demonstrated an increase in asthma exacerbations among patients residing in close proximity to a major roadway or highway [24,25]. We did not identify major roadway/highway exposure as a significant predictor of asthma exacerbations. As our patient population is largely rural (unpublished observation), we speculate that exposure to major roadways or highways may not be of primary relevance to asthma exacerbations.

While we have validated the ICEES open multivariate approach described here, several considerations are worthy of discussion. First, ICEES multivariate tables must be created in the context of a driving use case question, with a dependent variable identified and defined as the starting point for the overall approach. While this is not a limitation *per se*, it is a consideration that users should take into account.

Second, while the ICEES multivariate analytic approach is openly available, the ICEES service itself is subject to regulatory constraints that limit the amount of data that can be accessed and the types of analyses that can be performed. Specifically, cohorts   10 patients cannot be created. The impact of this constraint is that a certain amount of data loss will

be incurred whenever a cohort is created that has less than or equal to 10 patients. We are developing a theoretical framework to estimate data loss with different combinations of variables. For instance, suppose the most favorable case, namely, that each feature has only two values and that patients are divided equally among the possible values. Let there be k features in the query. Then, the ultimate cohort, say $C(k-1)$, must have at least 10 subjects to be included in the query output. The penultimate cohort, say $C(k-2)$, must have >=4*10 patients as a minimum. Therefore, the root cohort, say $C(1)$, must have >= $(4^{(k-2)})*10$ patients as a minimum. If there are eight features, then $|C(1)| >= (4^6)*10 = 4096*10 = 40{,}960$ patients are required at minimum to ensure that there is no data loss under the simplest, most favorable assumptions above. We plan to develop a technical approach for presenting this information to users so that they can apply the multivariate approach in an informed manner.

Third, and related to the above consideration, the choice and order of variables influence the open multivariate approach and the final sample size. For instance, a variable that has many missing values or multiple levels will by definition decrease the final sample size. In some cases, the impact of this limitation can be quite large. For example, ICEES currently exposes data on genomic variants, but the data are available for only a minor subset of patients, and so incorporating genomic features from ICEES into a multivariate analysis is not realistic.

Fourth, the data loss that is inherent in the ICEES open multivariate approach may impact model quality. Consider that the final table has one row for every combination of the selected features. The frequency returned for any row in the table for which any previous cohort in the query was less than or equal to 10 will be returned as zero, regardless of the true value, but we know that the true value cannot be greater than 10. If low-frequency rows are randomly distributed across the selected features, then we could assume that the query process may reduce the precision of the model results, but it would not introduce bias. In contrast, if low-frequency rows are not randomly distributed across the selected features, then we may introduce bias into our models, which will systematically affect the accuracy of model results and may lead to spurious conclusions. We are exploring approaches to anticipate and minimize bias.

Regardless of the limitations, we believe that the ICEES open multivariate approach provides a unique, regulatory-compliant service, with broad application. We are now comparing GLM model robustness and results with the API output versus the underlying data. We are also developing additional multivariate models such as random forest and causal inference. Finally, we are expanding the service to support additional use cases, including primarily ciliary dyskinesia and other rare respiratory disorders, drug-induced liver injury, coronavirus infection, and rare disease phenotypes.

Acknowledgments

Funding support

References

[1]. 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. [PubMed: 30480666]

[2]. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035. [PubMed: 27219127]

[3]. The Biomedical Data Translator Consortium. The Biomedical Data Translator program: conception, culture, and community. Clin Transl Sci 2019;12(2):86–90. [PubMed: 30412337]

[4]. The Biomedical Data Translator Consortium. Toward a universal biomedical data translator. Clin Transl Sci 2019;12(2):91–4. [PubMed: 30412340]

[5]. Ahalt SC, Chute CG, Fecho K, * Glusman G, Hadlock J, Solbrig H, Overby-Taylor C, Pfaff E, Ta C, Tatonetti N, Weng C. The NCATS Biomedical Data Translator Consortium. Clinical data: sources and types, regulatory constraints, applications. Clin Transl Sci 2019;12(4):329–33. 10.1111/cts.12638 [PubMed: 31074176] * Authors are listed alphabetically; K. Fecho served as lead author.

[6]. Fecho K, Pfaff E, Xu H, Champion J, Cox S, Stillwell L, Bizon C, Peden D, Krishnamurthy A, Tropsha A, Ahalt SC. A novel approach for exposing and sharing clinical data: the Translator Integrated Clinical and Environmental Exposures Service. J Am Med Inf Assoc 2019;26(10):1064–73. 10.1093/jamia/ocz042.

[7]. Pfaff ER, Champion J, Bradford RL, Clark M, Xu H, Fecho K, Krishnamurthy A, Cox S, Chute CG, Overby Taylor C, Ahalt S. Fast Healthcare Interoperability Resources (FHIR) as a meta model to integrate common data models: development of a tool and quantitative validation study. JMIR Med Inform 2019;7(4):e15199. 10.2196/15199. [PubMed: 31621639]

[8]. Xu H, Cox S, Stillwell L, Pfaff E, Champion J, Ahalt SC, Fecho K. FHIR PIT: an open software application for spatiotemporal integration of clinical data and environmental exposures data. BMC Med Inf Decis Making 2020;20. 10.21203/rs.2.19633/v1. article 53.

[9]. Fecho K*, Garantziotis S, Krishnamurthy A, Pfaff E, Schmitt C, Schurman S, Shuptrine S, Xu H, Ahalt S. Open integrated analysis of multi-institutional data using ICEES. Peer-reviewed poster, AMIA 2021 Virtual Informatics Summit. * Apart from the first/lead and last/senior author, all authors are listed alphabetically.

[10]. Fecho K, * Arunachalum S, Champion J, Chute CG, Gersing K, Glusman G, Hadlock J, Lee J, Pfaff E, Robinson M, Sid E, Ta C, Xu H, Zhu R, Zhu Q. Peden DB, and the Biomedical Data Translator Consortium. Sex, obesity, diabetes, and exposure to particulate matter: scientific insights revealed by analysis of open clinical data sources during a five day hackathon. J Biomed Inf 2019;100:103325. 10.1016/j.jbi.2019.103325 [Special Communication], * Apart from the first/lead and last/senior author, all other authors are listed in alphabetical order.

[11]. Cox S, Ahalt SC, Balhoff J, Bizon C, Fecho K, Kebede Y, Morton K, Tropsha A, Wang P, Xu H. TranQL: an interactive query and visualization environment for federated knowledge graphs. JMIR Med Inform 2020;8(11):e17964. 10.2196/17964. [PubMed: 33226347]

[12]. Health insurance portability and accountability Act, HIPAA, 'Safe Harbor' method for patient de-identification of medical records. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.

[13]. ICEES and related resources. ICEES OpenAPI: https://icees.renci.org:16340/apidocs/. ICEESGitHubrepository: https://github.com/NCATS-Tangerine/icees-api/tree/master/; CAMP FHIR GitHub respository: https://github.com/NCTraCSIDSci/camp-fhir; FHIR PIT GitHub repository: https://github.com/xu-hao/FHIR-PIT.

[14]. Translator application programming interface (TRAPI). https://github.com/NCATSTranslator/ReasonerAPI/issues.

[15]. Lehman RR, Archer KJ. Penalized negative binomial models for modeling an overdispersed count outcome with a high-dimensional predictor space: application predicting micronuclei frequency. PloS One 2019;14(1):e0209923. 10.1371/journal.pone.0209923. [PubMed: 30620740]

[16]. Hartig F. DHARMa: residual Diagnostics for hierarchical (Multi-Level/Mixed) regression Models. R package version 0.4.3. July 7, 2021. https://CRAN.R-project.org/package=DHARMa.

[17]. Brownlee J. SMOTE for imbalanced classification with Python. January Machine Learning Mastery 2020;17. Updated January 5, 2021, https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.

[18]. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from imbalanced data sets. Switzerland AG: Springer Nature; 2018. 10.1007/978-3-319-98074-4.

[19]. Alangari AA. Corticosteroids in the treatment of acute asthma. Ann Thorac Med 2014;9(4):187–92. [PubMed: 25276236]

[20]. Greenblatt RE, Zhao EJ, Henrickson SE, Apter AJ, Hubbard RA, Himes BE. Factors associated with exacerbations among adults with asthma according to electronic health record data. Asthma Res Pract 2019;5:1. 10.1186/s40733-019-0048-7. eCollection 2019, https://www.ncbi.nlm.nih.gov/pubmed/30680222. [PubMed: 30680222]

[21]. Keet CA, McCormack MC, Pollack CE, Peng RD, McGowan E, Matsui EC. Neighborhood poverty, urban residence, race/ethnicity, and asthma: rethinking the inner-city asthma epidemic. J Allergy Clin Immunol 2015;135(4):655–62. 10.1016/j.jaci.2014.11.022. [PubMed: 25617226]

[22]. Requia WJ, Adams MD, Koutrakis P. Association of $PM_{2.5}$ with diabetes, asthma, and high blood pressure incidence in Canada: a spatiotemporal analysis of the impacts of the energy generation and fuel sales. Sci Total Environ 2017;584:1077–83. 585. [PubMed: 28169030]

[23]. 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. 10.1289/EHP92. [PubMed: 27385358]

[24]. Perez L, Lurmann F, Wilson J, Pastor M, Brandt SJ, Künzli N, McConnell R. Nearroadway pollution and childhood asthma: implications for developing "win-win" compact urban development and clean vehicle strategies. Environ Health Perspect 2012 Nov;120(11):1619–26. 10.1289/ehp.1104785. Epub 2012 Sep. 24. [PubMed: 23008270]

[25]. Schurman SH, Bravo MA, Innes CL, Jackson WB, McGrath JA, Miranda ML, et al. Toll-like receptor 4 pathway polymorphisms interact with pollution to influence asthma diagnosis and severity. Sci Rep 2018:12713. 10.1038/s41598-018-30865-0. [PubMed: 30140039]
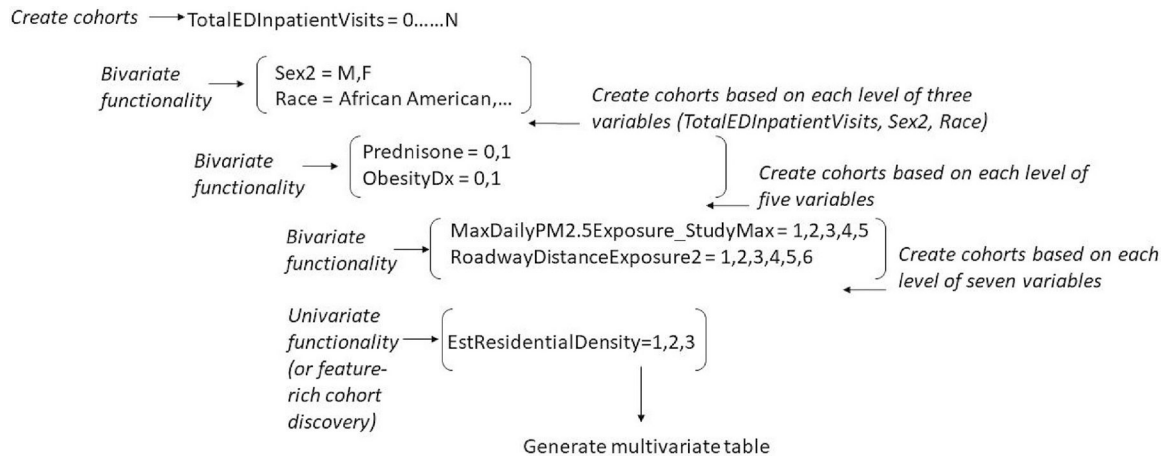
Fig. 1.
High-level overview of process for generating an ICEES multivariate table by application of dynamic cohort creation and nested bivariate contingencies. Levels or bins for each variable are defined in source documentation available from the OpenAPI and also accessible as an ICEES OpenAPI endpoint. See Table 1 for the feature variable definitions and enumeration used in this study.

TotalEDInpatientVisits	Sex2	Race	frequency
totaledinpatientvisits_1	sex2_male	race_native_hawaiian/pacific_islander	1
totaledinpatientvisits_1	sex2_male	race_caucasian	3782
totaledinpatientvisits_1	sex2_male	race_african_american	1657
totaledinpatientvisits_1	sex2_male	race_asian	89
totaledinpatientvisits_1	sex2_male	race_unknown	195
totaledinpatientvisits_1	sex2_male	race_american/alaskan_native	61
totaledinpatientvisits_1	sex2_male	race_other	786
totaledinpatientvisits_1	sex2_female	race_native_hawaiian/pacific_islander	4
totaledinpatientvisits_1	sex2_female	race_caucasian	4805
totaledinpatientvisits_1	sex2_female	race_african_american	2423
totaledinpatientvisits_1	sex2_female	race_asian	111
totaledinpatientvisits_1	sex2_female	race_unknown	256
totaledinpatientvisits_1	sex2_female	race_american/alaskan_native	70
totaledinpatientvisits_1	sex2_female	race_other	821
totaledinpatientvisits_2	sex2_male	race_native_hawaiian/pacific_islander	1
totaledinpatientvisits_2	sex2_male	race_caucasian	1002
totaledinpatientvisits_2	sex2_male	race_african_american	504
totaledinpatientvisits_2	sex2_male	race_asian	19
totaledinpatientvisits_2	sex2_male	race_unknown	27

Next iteration with four feature variables will results in a cohort <=10 patients, thus stopping the loop

Fig. 2.
Example ICEES tri-variate table, with rows in aggregate form representing the number of patients sharing the characteristics defined in each column. Each row can thus be duplicated to represent *N = 1* patient.

	A	B	C	D	E	F	G	H	I
1	TotalEDInpatientVisits	Sex2	Race	Prednisone	ObesityDx	MaxDailyPM2.5Exposure_StudyMax	RoadwayDistanceExposure2	EstResidentialDensity	frequency
2	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_1	estresidentialdensity_1	6
3	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_1	estresidentialdensity_2	3
4	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_1	estresidentialdensity_3	0
5	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_2	estresidentialdensity_1	1
6	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_2	estresidentialdensity_2	1
7	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_2	estresidentialdensity_3	0
8	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_3	estresidentialdensity_1	5
9	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_3	estresidentialdensity_2	2
10	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_3	estresidentialdensity_3	0
11	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_4	estresidentialdensity_1	4
12	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_4	estresidentialdensity_2	1
13	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_4	estresidentialdensity_3	0
14	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_5	estresidentialdensity_1	8
15	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_5	estresidentialdensity_2	0
16	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_5	estresidentialdensity_3	0
17	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_6	estresidentialdensity_1	35
18	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_6	estresidentialdensity_2	17
19	totaledinpatientvisits_0	sex2_male	race_caucasian	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_6	estresidentialdensity_3	0
20	totaledinpatientvisits_0	sex2_male	race_african_am	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_1	estresidentialdensity_1	5
21	totaledinpatientvisits_0	sex2_male	race_african_am	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_1	estresidentialdensity_2	1
22	totaledinpatientvisits_0	sex2_male	race_african_am	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_1	estresidentialdensity_3	0
23	totaledinpatientvisits_0	sex2_male	race_african_am	prednisone_0	obesitydx_0	maxdailypm2.5exposure_studymax_2	roadwaydistanceexposure2_2	estresidentialdensity_1	4

Fig. 3.
Excerpt from ICEES eight-feature multivariate table. The frequency column allows users to generate patient-level rows by, for instance, creating six separate rows for the features defined in row two and assigning a pseudo-identifier to each row.

Table 1

Feature variables used to generate multivariate table.

Feature Variable	Variable Definition and Enumeration
TotalEDInpatientVisits	*Total number ED or inpatient visits for respiratory issue(s) over the 'study' period (0, 1, 2, 3, …)*
Sex2	*Male (M, 0), Female (F, 1)*
Race	*Caucasian, African American, Asian, Native Hawaiian/Pacific Islander, American/Alaskan Native, Other*
Prednisone	*One or more prescriptions for prednisone over 'study' period (1 = Yes, 0 = No)*
ObesityDx	*One or more diagnostic codes for obesity over 'study' period (1 = Yes, 0 = No)*
MaxDailyPM2.5Exposure_StudyMax	*US Environmental Protection Agency estimated maximum daily exposure to airborne particulate matter 2.5-μm in diameter ($PM_{2.5}$) over 'study' period, binned using pandas.cut (1, 2, 3, 4, 5)*
RoadwayDistanceE.xposure2	*US Department of Transportation distance in meters from household to nearest roadway (1 = 0-49, 2 = 50-99, 3 = 100-149, 4 = 150-199, 5 = 200-249, 6 = 250 m)*
EstResidentialDensity	*US Census Bureau American Community Survey 2007–2011 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))*

Abbreviations: $PM_{2.5}$ = particulate matter 2.5-μm in diameter.

Table 2

Quantification of data loss with ICEES open multivariate approach.[a]

Feature Variable Added[b]	Total ICEES Rows (N)	Maximum Possible Rows (N)	Missing Rows (N)	Missing Rows/Maximum Possible Rows (%)
Starting sample size	*22365*	*N/A*	*N/A*	*N/A*
Sex2	22365	22365	0	0
Race	22365	22365	0	0
Prednisone	22365	22365	0	0
ObesityDx	22208	22361	153	0.68
MaxDailyPM2.5Exposure_StudyMax	15861	17390	1529	8.79
RoadwayDistanceExposure2	5022	8262	3240	39.2
EstResidentialDensity	4556	10615	6059	57.1

[a] Starting sample size before filtering for patients who were active in the 'study' period (calendar year 2010): N = 163302.

[b] Feature variables were added in the order listed, following the schema shown in Fig. 1.

Table 3

ANOVA results for GLM model with main effects and 2-way interactions.[a]

Main Effect or Interaction	df	Deviance	Residual df	Residual Deviance	P value	Sig
NULL			14936	8796.1		
Sex2	1	16.701	14935	8779.4	4.376e-05	***
Race	5	141.052	14930	8638.3	< 2.2e-16	***
Prednisone	1	153.832	14929	8484.5	< 2.2e-16	***
ObesityDx	1	28.412	14928	8456.1	9.806e-08	***
MaxDailyPM2_5Exposure_StudyMax	2	36.204	14926	8419.9	1.375e-08	***
RoadwayDistanceExposure2	5	9.601	14921	8410.3	0.087363	
EstResidentialDensity	1	0.274	14920	8410	0.600541	
Sex2:Race	5	9.395	14915	8400.6	0.094305	
Sex2:Prednisone	1	0.871	14914	8399.7	0.35066	
Sex2:ObesityDx	1	6.249	14919	8403.4	0.012426	*
Sex2:MaxDailyPM2_5Exposure_StudyMax	2	0.428	14906	8363	0.80749	
Sex2:RoadwayDistanceExposure2	5	2.997	14896	8347.6	0.700431	
Sex2:EstResidentialDensity	1	2.814	14855	8314.8	0.093454	
Race:Prednisone	2	18.555	14912	8381.2	9.351e-05	***
Race:ObesityDx	2	1.35	14909	8373.5	0.509129	
Race:MaxDailyPM2_5Exposure_StudyMax	3	1.589	14903	8361.4	0.661954	
Race:RoadwayDistanceExposure2	25	9.65	14871	8338	0.997493	
Race:EstResidentialDensity	5	1.129	14850	8313.7	0.951543	
Prednisone:ObesityDx	1	10.07	14908	8363.5	0.001507	**
Prednisone:MaxDailyPM2_5Exposure_StudyMax	1	9.507	14902	8351.9	0.002047	
Prednisone:RoadwayDistanceExposure2	5	14.696	14866	8323.3	0.011744	
Prednisone:EstResidentialDensity	1	0.175	14849	8313.5	0.676105	
ObesityDx:MaxDailyPM2_5Exposure_StudyMax	1	1.313	14901	8350.6	0.251903	
ObesityDx:RoadwayDistanceExposure2	5	2	14861	8321.3	0.849104	
ObesityDx:EstResidentialDensity	1	0.658	14848	8312.8	0.417263	
MaxDailyPM2_5Exposure_StudyMax: RoadwayDistanceExposure2	5	3.681	14856	8317.6	0.596173	
MaxDailyPM2_5Exposure_StudyMax: EstResidentialDensity	1	0.115	14847	8312.7	0.734287	
RoadwayDistanceExposure2 EstResidentialDensity	5	0.862	14842	8311.8	0.972891	

Abbreviations: ANOVA = analysis of variance; df = degrees of freedom; GLM = generalized linear model; Sig = significance level (*:0.05, **: 0.01, ***0.001).

[a] Negative binomial model, link: log, dependent variable: *TotalEDInpatientVisits*. Three-way and higher interactions are not included in the table for readability.