Fecho *et al.*
BMC Medical Informatics and Decision Making (2025) 25:24
https://doi.org/10.1186/s12911-024-02815-6

BMC Medical Informatics and Decision Making

SOFTWARE

Open Access

FHIR PIT: a geospatial and spatiotemporal data integration pipeline to support subject-level clinical research

Karamarie Fecho[1,2*†], Juan J. Garcia[3†], Hong Yi[1†], Griffin Roupe[1,3] and Ashok Krishnamurthy[1,3]

Abstract

Background Environmental exposures such as airborne pollutant exposures and socio-economic indicators are increasingly recognized as important to consider when conducting clinical research using electronic health record (EHR) data or other sources of clinical data such as survey data. While numerous public sources of geospatial and spatiotemporal data are available to support such research, the data are challenging to work with due to inconsistencies in file formats and spatiotemporal resolutions, computational challenges with large file sizes, and a lack of tools for patient- or subject-level data integration.

Results We developed FHIR PIT (HL7® Fast Healthcare Interoperability Resources Patient data Integration Tool) as an open-source, modular, data-integration software pipeline that consumes EHR data in FHIR® format and integrates the data at the level of the patient or subject with environmental exposures data of varying spatiotemporal resolutions and file formats. We applied FHIR PIT to generate "integrated feature tables" containing patient- or subject-level EHR data integrated with environmental exposures data on two cohorts: one on patients with asthma and related common pulmonary disorders; and a second on patients with primary ciliary dyskinesia and related rare pulmonary disorders. The data were then exposed via the open Integrated Clinical and Environmental Exposures Service, which was then queried to explore relationships between exposures to two representative airborne pollutants (particulate matter and ozone) and annual emergency department or inpatient visits for respiratory issues. We found that hospitalizations for respiratory issues were more common among patients exposed to relatively high levels of particulate matter and ozone and were higher overall among patients with primary ciliary dyskinesia than among patients with asthma.

Conclusions Our manuscript describes a major release of FHIR PIT v1.0 and includes a technical demonstration use case and a clinical application on the use of FHIR PIT to support research on environmental exposures and health outcomes related to asthma and primary ciliary dyskinesia. For application of the tool to common data models (CDMs) other than FHIR, we offer open-source conversion tools to map from the PCORnet, i2b2, and OMOP CDMs to FHIR.

Keywords Asthma, Primary ciliary dyskinesia, HL7® FHIR®, Data integration, Environmental exposures, Airborne pollutant exposures, Socioeconomic exposures, Hospital visits

†Karamarie Fecho, Juan J. Garcia and Hong Yi contributed equally to the work described in the manuscript and are listed in alphabetical order.

*Correspondence:
Karamarie Fecho
kfecho@copperlineprofessionalsolutions.com
Full list of author information is available at the end of the article

Background

Environmental exposures are increasingly recognized as important to consider when conducting research in numerous scientific fields, including medicine [30, 31]. While many public sources of geospatial and spatiotemporal data are available to support research on social and environmental determinants of health and disease, the data can be challenging to work with [4]. For instance, the datasets are often of varying spatial and temporal resolutions and formats, with differing representations for geocodes and time stamps. The exposures source data may be spatially represented as raster data (e.g., gridded air pollution predictions), area data or polygons (e.g., US Census tracts), line data (e.g., roadway data), or point data (e.g., landfill data). The data also may be temporally represented as mean hourly exposure, mean six-hour exposure, maximum 24-h exposure, median five-year exposure, or another temporal representation. This variability complicates data integration and harmonization across input sources. In addition, the datasets are often in different file formats (e.g., SHP vs CSV), and they can be quite large, especially for granular data (e.g., hourly estimates), which presents computational challenges.

The challenges specific to the integration of environmental exposures data with patient data have been well studied [3, 4, 16]. In addition to the challenges noted above, one must consider the need to protect patient privacy and the choice of common data model (CDM) used to represent electronic health record (EHR) data. Few open-source, production-quality, software repositories and tools are available to support at-scale geospatial and spatiotemporal data integration at the level of the individual patient or subject, although several solutions or partial solutions have been put forward. For instance, DeGauss provides a containerized privacy-aware solution that provides patient-level geocoding, e.g., patient primary residential address, and calculates prespecified exposure estimates for several types of exposures (e.g., air pollution, social deprivation, temperature) [1]. The Center for Disease Control and Prevention's National Environmental Public Health Tracking Network Data Explorer offers a web-based catalog and search tool for accessing public health monitoring data and exposures data within the continental United States, with downloads available in tabular format [25]. GeoQuery is a web-based application to access global geospatial estimates (e.g., climate, wealth, land cover), with user-defined spatial and temporal aggregations and downloadable tabular output [15]. NASA's Socioeconomic Data and Applications Center offers a web-based application and visualization tool for exploring global socioeconomic and environmental exposures data (e.g., global poverty, water security, air quality), with subsets of the data available for download in a variety of formats [28].

We developed the HL7® Fast Healthcare Interoperability Resources Patient data Integration Tool (FHIR® PIT) as an open-source data integration pipeline to support patient- or subject-level research on social and environmental determinants of health and disease. FHIR PIT enables the integration of diverse sources of geospatial and spatiotemporal data, including EHR data, clinical study data, and environmental exposures data such as airborne pollutant exposures and socio-economic indicators. As an open-source tool to support geospatial and spatiotemporal data integration at the level of the patient or subject, FHIR PIT thus fills an existing research gap and software need that, with the exception of DeGauss, have not been addressed by other resources.

We described a proof-of-concept prototype version of FHIR PIT in a previous publication [32]. Herein, we provide a detailed description of our FHIR PIT v1.0 major release, with support for new types of exposures data, platform-independent execution and testing within a Docker container, sample input data, a tutorial, and extensive documentation. We highlight both a technical demonstration use case and a clinical application use case. We also describe our future plans to further enhance FHIR PIT and extend its applicability and reach across scientific fields.

Implementation

We developed the FHIR PIT data integration pipeline as a collection of modular steps that compose a transformation pipeline to process and link heterogeneous geospatial and spatiotemporal data sources. A detailed overview of the FHIR PIT data integration and implementation pipeline is provided in Figure S1; a detailed overview of the transformation steps used to generate the final output of FHIR PIT, the "integrated feature tables", is provided in Figure S2 (see "Additional file 1" within the "Supplementary materials").

The transformation pipeline is configurable through the Dhall configuration language [6]. The user configures the steps to run, along with the corresponding input/output folders. Additionally, steps can be reused or skipped for faster execution. New transformation steps can be added to the transformation pipeline as needed.

Currently, FHIR PIT supports the integration of several types of data derived from both private and public sources, including EHR data, socioeconomic exposures data, airborne pollutant exposures data, roadway/highway exposures data, data on exposure to concentrated animal feeding operations (CAFO) data, and landfill exposures data (Table 1). Of note, we are in the process

of capturing and cleaning additional sources of socioeconomic and environmental exposures data for incorporation into the FHIR PIT pipeline, including data on public school exposures from the National Center for Education Statistics. We note that the frequency of releases/updates for the FHIR PIT input data is controlled by the data owners and varies. In addition, apart from the patient data, the socioeconomic and environmental exposures data are mostly available as downloads in various formats (e.g., SHP, CSV). However, a subset of cleaned and processed data is accessible via Application Programming Interfaces (APIs) that we have created: an Airborne Pollutant Exposures API; Roadway/Highway Exposures API; and a Socio-economic Exposures API [22, 29].

Of note, we derived FHIR files from UNC Health using a second open-source software pipeline termed Clinical Asset Mapping Program for FHIR (CAMP FHIR; [23]. CAMP FHIR converts data from CDMs (i.e., i2b2, PCORnet, and OMOP) to FHIR files as a meta-CDM. FHIR PIT then consumes the CAMP FHIR output and integrates the data at the patient level with a variety of public sources of socioeconomic and environmental exposures data. Like FHIR PIT, CAMP FHIR is openly available [2]. Thus, for potential FHIR PIT adopters whose healthcare systems have not adopted FHIR as their CDM, CAMP FHIR can be used to facilitate the conversion of their CDM data to FHIR files as the key input file for FHIR PIT.

FHIR PIT supports multiple environmental exposure models. For instance, our airborne pollutant exposures are modeled as raster estimates based on predictive models derived from point sensors and temporally represented as average daily and maximum daily exposures (e.g., average daily exposure to particulate matter $\leq$ 2.5-microns in diameter [$PM_{2.5}$] in μg per meter-cubed). Our exposures data on roadways, landfills, and concentrated animal feeding operations are modeled as point estimates in meters from a specified latitude/longitude (e.g., primary home residence) to the nearest major highway or roadway. Users can incorporate their preferred environmental exposure sources and models into the FHIR PIT pipeline as new sources or integration modules.

New input data of any file format or size can be added to the modular FHIR PIT pipeline. In addition, the transformation steps join FHIR patient records with environmental exposure estimates using patient geocodes (i.e., primary residence) and date(s) of healthcare visit as the spatiotemporal links. At the end of each linkage step, FHIR PIT saves the linked tables to an intermediate folder to support rapid re-execution, should any given step fail to complete (e.g., lack of disk space). To support multiple healthcare visits, FHIR PIT vectorizes FHIR patient records by grouping daily FHIR domains (e.g., Condition, Laboratory, Medication, Procedures) and counting them. The features associated with the patient are then grouped by user-defined study periods (e.g., yearly) and aggregated with user-defined statistics (e.g., counts, mean, median, first measurement, last measurement). A YAML configuration file (i.e., "icees_features.yaml") is used to relabel FHIR domain codes and features from environmental exposure sources into human-readable features. This manually generated file can be configured by users to meet their specific research needs and desired FHIR domain codes. (Please see "Data availability and requirements" and "Availability and implementation" for additional documentation, including the FHIR PIT GitHub repository URL.)

An optional join step follows the integration steps and allows users to join other patient- or subject-level datasets. For example, we have leveraged the join step to incorporate data from a survey-based study on participants who are also UNC Health patients. This allowed us to create a rich dataset containing integrated FHIR patient records, participant survey data, and a variety of environmental exposures data. We successfully applied the integrated dataset to examine the impact of environmental exposures on health outcomes, as defined by both FHIR records and survey data [10].

An optional binning step allows users to bin or categorize variables to facilitate downstream applications and, in some cases, adhere to regulatory restrictions. For instance, we bin environmental exposure estimates to protect patient privacy. The binning step is required by our institution, which considers exposure estimates "secondary Protected Health Information (PHI)" due to the fact that the estimates are derived using PHI (i.e., geocodes and time stamps). The step is necessary, given that many of our application use cases involve exposing the deidentified FHIR PIT–generated integrated data in semi-aggregated form using an open-source web service termed the Integrated Clinical and Environmental Exposures Service (ICEES; [11], also see "Results—Clinical application use case").

The final step in the FHIR PIT data integration pipeline involves the deidentification of the FHIR PIT output file by removing all PHI per the Safe Harbor Method of the Health Insurance Portability and Accountability Act (HIPAA; [26]). This deidentification step allows the resultant FHIR PIT integrated feature tables containing patient-level data derived from both EHR data and environmental exposures data sources to be openly shared for research purposes as HIPAA-compliant datasets, albeit in compliance with institutional constraints. For instance, our institution restricts the use of the deidentified FHIR PIT integrated feature tables to those persons who have

Table 1 FHIR PIT data sources

Data Type	Data Format	Data Source	Availability	Frequency of release/update	URL
EHR data (e.g., patient demographics, diagnoses, drug prescriptions or administrations)	FHIR	UNC Health	Private	Dependent on application use case needs	N/A
Socioeconomic exposures data (e.g., household income, access to health insurance, access to automobile)[a]	Area (polygon), CSV	US Census Bureau American Community Survey	Public	5-year releases	https://www.census.gov/programs-surveys/acs/data.html
Airborne pollutant exposures (e.g., particulate matter, ozone)[a]	Raster (grid), CSV	US Environmental Protection Agency	Public	Contingent on federal release process and policy, typically lags behind calendar year by one year or more	https://www.epa.gov/cmaq
Major roadway/highway exposures[a]	Line, SHP	US Department of Transportation, Federal Highway Administration, Highway Performance Monitoring System	Public	Contingent on federal release process and policy, typically lags behind calendar year by one year or more	https://www.fhwa.dot.gov/policyinformation/hpms.cfm
Major roadway/highway exposures[a]	Line, SHP	US Census Bureau, Topologically Integrated Geographic Encoding and Referencing (TIGER)/Line Roadway System	Public	Contingent on federal release process and policy, typically lags behind calendar year by one year or more	https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html
CAFO exposures	Point, CSV	NC Department of Environmental Quality	Public	Annual	https://deq.nc.gov/cafo-map
Landfill exposures	Point, CSV	NC Department of Environmental Quality	Public	Annual	https://www.nconemap.gov/datasets/ncdenr::active-permitted-landfills/about
Public school exposures[b]	Point, CSV	National Center for Education Statistics	Public	Annual	https://nces.ed.gov/

Abbreviations: *CAFO* concentrated animal feeding operations, *CSV* comma separated values, *EHR* electronic health record, *SHP* shapefile

[a] We have made a subset of the processed environmental exposures data available to users via application programming interfaces (APIs) for programmatic access: Socio-economic Exposures Service; Airborne Pollutant Exposures Service; Roadway Exposures Service [22]. These services have been called successfully by external users and applied to address research questions [29]

[b] Not yet included in the FHIR PIT data integration pipeline

been approved by the Institutional Review Board (IRB) as study personnel, but the data are freely accessible in a semi-aggregated form by the open ICEES service [11].

Note that because all steps in the FHIR PIT integration pipeline are modular, they can be reused or skipped. For example, if a user's institution does not require that exposure estimates are binned, then that user may skip the binning step when running the software pipeline. The same is true with the optional join step and even the final deidentification step, which users may wish to skip if their data does not leave a secure environment. Thus, FHIR PIT can be tailored to meet a given user's data integration and downstream analysis needs.

Security requirements

FHIR PIT relies on geocodes and time stamps for integration of clinical and environmental exposures data. These data elements are considered PHI under HIPAA and thus the datasets that FHIR PIT integrates fall under federal and institutional regulations, including approval by an IRB, adherence to HIPAA, and institutional regulations related to data storage and movement. For our work, we store all data and run the FHIR PIT software pipeline within a PHI/HIPAA/Tier 3 secure enclave and under an IRB-approved protocol. After the final deidentification step, the FHIR PIT output files or integrated feature tables are no longer under federal or HIPAA regulations; however, the data remain under institutional regulations. At our institution, this means that the deidentified integrated feature tables must remain on a secure, password-protected server, accessible only by IRB-approved study personnel, although the data can be shared in semi-aggregated form via the open ICEES service [11].

Results

Technical demonstration use case

For users interested in adopting FHIR PIT for their geospatial and spatiotemporal data integration needs, the FHIR PIT GitHub repository contains extensive documentation, including: a detailed README file; example code snippets; sample input data; data transformation details; and specific run instructions (see "Data availability and requirements" and "Availability and implementation" for URLs). The sample input datasets include dummy data on "patient" FHIR files formatted in JSON, as shown in Fig. 1 for an example record from a "Condition" table. This example record shows that a patient had an encounter or healthcare visit on November 06, 2019, in which the patient was diagnosed with "acute viral pharyngitis (disorder)".

The sample environmental exposures datasets include files of varying format and geospatial and spatiotemporal resolution. An example CSV snippet for estimated concentrations of $PM_{2.5}$ (denoted pm2.5) and ozone is shown in Table 2.

This example shows the average daily estimated concentration of $PM_{2.5}$ and the maximum daily 8-h estimated concentration of ozone, as well as standard deviations, for -70.2386 latitude and 41.6696 longitude on January 1, 2, and 3, 2010. Additional airborne pollutant exposures such as carbon monoxide and formaldehyde are also available for select years.

The sample socio-economic exposure datasets are provided in CSV format and include estimates for median household income, health insurance, access to an automobile, and other metrics. The sample roadway exposure datasets are provided in SHP format and can be applied to calculate nearest distance. The sample landfill and CAFO datasets are provided as point estimates.

After installing and running FHIR PIT, a user will be able to generate a sample integrated feature table, a snippet of which is shown in Table 3. in CSV format.

This example shows dummy data on three "patient" records (indexed as 0, 1, 2) for the year 2010. The first record or row is for a male 17 years-of-age, with an average daily $PM_{2.5}$ exposure of 22.135 $\mu g/m^3$ and a maximum daily ozone exposure of 67.853 ppbV across a one-year study period. The patient's primary residence is located 462.221 m from a major roadway or highway. The patient did not have any emergency department (ED) or inpatient visits, diagnoses of asthma, or prescriptions/administrations of prednisone over the study period.

One important note when running FHIR PIT is that the application may be memory-intensive, depending on the size of the input data. For example, for the sample data available in the GitHub repository, the application requires 4 GB RAM to run through all of the data processing and integration steps when running Docker natively. (Docker Desktop on Windows/Mac may incur additional memory overhead.)

Clinical application use case

Here, we validate FHIR PIT and the generated integrated feature tables that support ICEES by way of a direct comparison between an ICEES cohort of patients with asthma and related common pulmonary disorders and a second ICEES cohort of patients with primary ciliary dyskinesia (PCD) and related rare pulmonary disorders (hereafter referred to as ICEES asthma and ICEES PCD cohorts, respectively). The ICEES asthma cohort was described previously [11]. In brief, the cohort included all patients in our clinical data warehouse with an "asthma-like" condition, 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

```
{
  "resourceType": "Bundle",
  "entry": [
    {
      "resource": {
        "resourceType": "Condition",
        "id": "00634035-5b5e-402b-bfa4-81e375eae861",
        "clinicalStatus": {
          "coding": [
            {
              "system": "http://terminology.hl7.org/CodeSystem/condition-clinical",
              "code": "resolved"
            }
          ]
        },
        "verificationStatus": {
          "coding": [
            {
              "system": "http://terminology.hl7.org/CodeSystem/condition-ver-status",
              "code": "confirmed"
            }
          ]
        },
        "code": {
          "coding": [
            {
              "system": "http://snomed.info/sct",
              "code": "195662009",
              "display": "Acute viral pharyngitis (disorder)"
            }
          ],
          "text": "Acute viral pharyngitis (disorder)"
        },
        "subject": {
          "reference": "Patient/d8a00854-f768-4195-a025-9af34b2e0e9f"
        },
        "encounter": {
          "reference": "Encounter/18326582-7a65-4519-8c46-5bcc514a6b74"
        },
        "onsetDateTime": "2019-11-06T01:08:11+00:00",
        "abatementDateTime": "2019-11-19T01:08:11+00:00",
        "recordedDate": "2019-11-06T01:08:11+00:00",
        "meta": {
          "tag": [
            {
              "system": "https://smarthealthit.org/tags",
              "code": "synthea-5-2019"
            }
          ]
        }
      }
    }
  ]
}
```


Fig. 1 Example JSON record from a FHIR "Condition" table

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 ICEES PCD cohort also was described previously [9]. Briefly, the cohort included all patients in our clinical data warehouse with a definitive diagnostic code for cystic fibrosis (CF) (ICD-E84) or a possible diagnosis of CF, idiopathic bronchiectasis (IB), or PCD, using broad, expert-defined inclusion criteria. The inclusion of a definitive diagnosis of IB or PCD as a criterion was not possible, given that a diagnostic code does not exist for IB or PCD.

For both cohorts, FHIR PIT–generated integrated feature tables were used to support the ICEES instances. We focused on one primary outcome measure, annual emergency department (ED) or inpatient visits for respiratory issues, and two representative environmental exposures,

airborne $PM_{2.5}$ and ozone exposures, during the year 2016, which was the most recent year for which exposures variables were available for both cohorts.

We found that for both ICEES cohorts, the percentage of patients with ED/inpatient visits for respiratory issues was higher among those exposed to relatively high levels of $PM_{2.5}$ or ozone than among those exposed to relatively low levels of $PM_{2.5}$ or ozone (Fig. 2). Overall, annual ED/inpatient visits for respiratory issues were more frequent among patients in the ICEES PCD cohort than among patients in the ICEES asthma cohort.

Thus, for two separate patient cohorts, one comprising patients with asthma or another common pulmonary disorder, and the other comprising patients with PCD or another rare pulmonary disorder, we were able to run FHIR PIT to generate integrated feature tables, load the tables behind the ICEES APIs, and run queries to demonstrate significant relationships between exposure to PM2.5 or ozone and an increase in respiratory issues requiring an ED or inpatient hospital visit.

Discussion

FHIR PIT was developed to fill a biomedical informatics need for a tool that supports the integration of geospatial and spatiotemporal datasets at the patient or subject level. We demonstrated the ability of FHIR PIT to generate two sets of integrated feature tables derived from an ICEES asthma cohort and an ICEES PCD cohort. We further showed that the data can be loaded behind an ICEES API and queried to demonstrate meaningful

Table 2 Example snippet from airborne pollutant exposures data, sourced from the US Environmental Protection Agency

Date	FIPS	Longitude	Latitude	pm25_daily_average	pm25_daily_average_stderr	ozone_daily_8hour_maximum	ozone_daily_8hour_maximum_stderr
1/1/2010	25001012101	-70.2386	41.6696	16.569	7.454	32.678	6.7798
1/2/2010	25001012101	-70.2386	41.6696	5.553	3.4195	35.938	7.7447
1/3/2010	25001012101	-70.2386	41.6696	5.775	2.7642	28.291	7.0392

Table 3 Example snippet of the patient-level FHIR PIT integrated feature table created from the FHIR PIT PreprocCSVTable transformation step showing data elements from both the EHR (year, Sex, AgeStudyStart, TotalEDInpatientVisits, AsthmaDx, Prednisone) and environmental exposures data sources (AvgDailyPM2.5Exposure, MaxDailyOzoneExposure, RoadwayDistance)

year	index	Sex	AgeStudyStart	AvgDailyPM2.5Exposure	MaxDailyOzoneExposure	RoadwayDistance	TotalEDInpatientVisits	AsthmaDx	Prednisone
2010	0	Male	17	22.13492346	67.85280609	462.2205866	0	0	0
2010	1	Male	20	22.71102758	94.79951477	252.796139	0	0	0
2010	2	Male	17	10.52038211	28.94073296	301.8245516	0	0	0

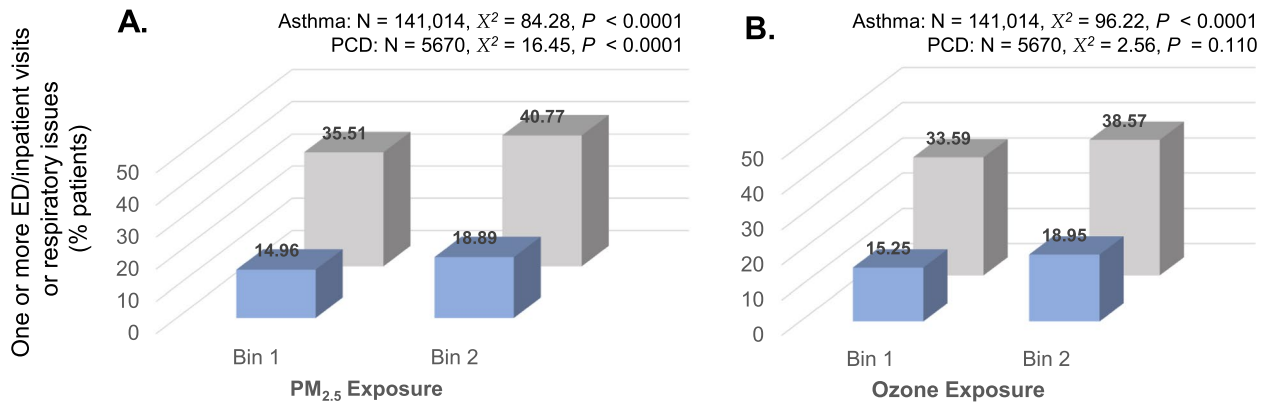

Fig. 2 Association between ED or inpatient visits for respiratory issues and exposure to particulate matter or ozone. **A** Exposure to $PM_{2.5}$ ($\mu g/m^3$). Asthma: Bin: 1 = (3.273, 7.801] $\mu g/m^3$; Bin 2 = (7.810, 10.830]. PCD: Bin 1 = (5.510, 8.741]; Bin 2 = (8.741, 10.891]. **B** Exposure to ozone (ppbV). Asthma: Bin 1 = (27.798, 38.998], Bin 2 = (38.998, 46.453]. PCD: Bin 1 = (27.854, 38.234]; Bin 2 = (38.324, 45.141]

relationships between environmental exposures to representative airborne pollutants (PM2.5 and ozone) and ED or inpatient hospital visits for respiratory issues.

Key features of FHIR PIT

The FHIR PIT data integration pipeline includes a number of attractive features. For instance, FHIR PIT is modular in design, which facilitates the incorporation of new FHIR elements, geospatial and spatiotemporal data sources, and environmental exposure models. For instance, users can define the FHIR elements of interest in the "icees_features.yaml" file to meet their specific research needs. Likewise, the environmental exposure models can be tailored to meet a user's need (e.g., MaxDailyPM2.5Exposure vs AvgDailyPM2.4Exposure). Indeed, the steps in the integration pipeline can be reused or updated as needed. In addition, FHIR PIT supports the integration of geospatial and spatiotemporal datasets of varying spatial and temporal resolution and representation, and users can incorporate their preferred exposure models and datasets. For example, we recently acquired data on public school exposures from the National Center for Education Statistics. We are now developing models to consider patient mobility among school-aged children by examining differences in their environmental exposures at home versus school and then determining if those differences have health impacts. For example, one might find that asthma exacerbations are more common during the school day versus at home or during the academic calendar year versus the summer months among children whose school is located near a major roadway or highway. Although FHIR PIT currently assumes a patient or subject resides in a single geolocation across a defined study period, the software can be extended to account for potential changes in geolocation and daily mobility, for example, home versus school exposure or home versus work exposure. The FHIR PIT data integration pipeline also is configurable using the Dhall configuration language. For example, the data integration pipeline can be reconfigured for partial (re)execution in order to reduce compute time. Similarly, FHIR domain codes or data source features can be added and/or relabeled in a configurable way. Moreover, FHIR PIT output can be deidentified per HIPAA Safe Harbor, thus making the data ready for public release. Lastly, FHIR PIT can be easily configured to include join steps for the incorporation of additional sources of identifiable patient or subject data. Likewise, FHIR PIT can be configured to exclude unneeded steps, e.g., the binning step.

Software considerations

While FHIR PIT contains several attractive features, the software also has limitations and considerations that should be considered. First, the FHIR PIT code is dependent on a YAML configuration file ("icees_features.yaml"), which is manually generated to support new use cases and is used to define and enumerate feature variables and support mappings between feature variables and FHIR elements. Second, codebase updates require proficiency in multiple tools and software languages, including Python, Dhall (configuration language), Apache Spark (cluster-computing framework engine), Scala, sbt (build tool for Scala and JAVA), and YAML (data serialization language). To address this limitation, we recently Dockerized FHIR PIT to encapsulate the sample input files and the software and platform dependencies and facilitate execution and deployment. Third, the application can be memory-intensive, depending on the size of the input data. As noted under "Technical demonstration use case", with the sample data available in the GitHub repository, the application requires 4 GB RAM to run through all of the data processing and integration steps when running Docker natively. In addition, FHIR PIT code annotation and technical documentation are in continual editing for completeness and may not meet the needs of all external users, in which case, we encourage users to submit an issue to the GitHub repository or contact us directly for troubleshooting and/or feature requests (see "Availability and implementation"). Despite these considerations and the general complexity of the software pipeline, we note that two external users (GR and SN) have successfully run FHIR PIT using the sample data we provided in the GitHub repository.

Clinical use case considerations

In the use-case application described herein, we found that hospital visits for respiratory issues were more frequent overall among patients in the ICEES PCD cohort than among patients in the ICEES asthma cohort, which is perhaps not surprising, given that the ICEES PCD cohort is compromised of patients with severe, primarily undiagnosed, respiratory disease, including confirmed CF and suspected CF, PCD, or IB. For both the ICEES asthma cohort and the ICEES PCD cohort, we found an increase in hospital visits for respiratory issues among patients exposed to relatively high levels of airborne pollutants when compared to those exposed to relatively low levels, thus replicating our prior findings and those of other groups on asthma [8, 11, 17, 18, 24] and extending them to derive new insights into PCD.

One important consideration to note when applying FHIR PIT to support clinical use cases such as those supported by ICEES relates to our binning approach for the airborne pollutant exposures, which is a FHRI PIT step that is required by our institution. For instance, we've found that the binning approach that we currently use

for airborne pollutant exposures, i.e., pandas.cut [19], frequently results in small cell sizes for the extremes of the patient distributions, meaning patients with very low exposures or very high exposures. We often have to collapse contiguous bins at the extreme distributions in order to adjust for the small cell sizes. Indeed, the approach we took for the current work followed that of our prior work [8], in which we collapsed the lowest and highest bins. However, in the use-case application reported here, we additionally collapsed the second lowest bin to account for a cell size of < 20 patients in the ICEES PCD cohort and allow us to directly compare the ICEES PCD and asthma cohorts. We have applied other approaches, e.g., pandas.qcut [20], but those approaches tend to overfit. We are exploring additional binning approaches, for example, applying a strict upper limit for airborne pollutant exposures as specified by authorities such as the US Environmental Protection Agency [27] or allowing users to specify the bins; however, the former approach removes the granularity, while the latter approach introduces computational and regulatory challenges. We are working with groups such as the Environmental Health Language Collaborative [7] to identify a robust and generalizable solution to this challenge for ICEES. We point out the binning challenge here only to inform potential FHIR PIT adopters. We emphasize, however, that FHIR PIT can be adapted to support whatever binning approach works best for users and their datasets, or the binning step can be skipped entirely, if desired.

Future directions

In considering further improvements to FHIR PIT, we are currently working to: enhance unit and functional testing; extend the data integration pipeline to support new sources of environmental exposures data; improve our binning approach for airborne pollutant exposures; and apply FHIR PIT to new clinical use cases such as post-acute sequelae of SARS-CoV-2 infection and other diseases. As open-source software, we believe that our efforts on FHIR PIT will fill a critical need in patient- or subject-level geospatial and spatiotemporal data integration and the harmonization of geospatial and spatiotemporal datasets across diverse use-case applications.

Conclusion

FHIR PIT was developed to fill a need for a geospatial and spatiotemporal tool to support patient- or subject-level integration of clinical data such as EHR data and environmental exposures data such as airborne pollutant exposures. Since the initial demonstration use case and publication of the prototype [32], the FHIR PIT v1.0 code and modular integration steps have matured to support a number of use cases and multiple sources of environmental exposures data. In addition, FHIR PIT output contributes to the open-source "knowledge graph"–based Biomedical Data Translator system [12].

The FHIR PIT software code has matured significantly since the prototype deployment and the initial validation test. The FHIR PIT GitHub repository contains extensive documentation to support its adoption by independent users, including: a detailed README file; example code snippets; sample input data; data transformation details and specific run instructions; and a Docker file. The sample input data includes sample FHIR files that contain dummy "patient" datasets and sample environmental exposures datasets. We have demonstrated that the code, documentation, and sample files included in the FHIR PIT GitHub repository are sufficient to support independent use of FHIR PIT by external users. For application of the tool to CDMs other than FHIR, we offer open-source conversion tools to map from the PCORnet, i2b2, and OMOP CDMs to FHIR. Specifically, pcornet-to-fhir [33] maps from the PCORnet CDM to FHIR, and CAMP FHIR [2, 23] maps from the PCORnet, i2b2, and OMOP CDMs to FHIR.

Data availability and requirements

Project name: FHIR PIT (v1.0 release: https://github.com/ExposuresProvider/FHIR-PIT/releases/tag/v1.0).

Project home page: https://github.com/ExposuresProvider/FHIR-PIT.

Operating system(s): platform independent, Linux preferred.

Programming language: Python; Dhall; Apache Spark; Scala; sbt; YAML.

Other requirements: sufficient RAM (application requires 4 GB RAM to run through all of the data processing and integration steps using the sample input).

License: MIT License, ©2024 Renaissance Computing Institute, https://github.com/ExposuresProvider/FHIR-PIT/blob/master/LICENSE.

Any restrictions to use by non-academics: none.

Abbreviations

CAFO	Concentrated animal feeding operation
CF	Cystic fibrosis
CSV	Comma separate values
ED	Emergency department
EHR	Electronic health record
FHIR PIT	HL7 Fast Healthcare Interoperability Resources Patient data Integration Tool
HIPAA	Health Insurance Portability and Accountability Act
IB	Idiopathic bronchiectasis
ICEES	Integrated Clinical and Environmental Exposures Service
IRB	Institutional review board
PCD	Primary ciliary dyskinesia
PHI	Protected health information
$PM_{2.5}$	Particulate matter ≤ 2.5-microns in diameter
SHP	Shapefile

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02815-6.

Supplementary Material 1. FHIR PIT data integration pipeline and transformation steps.

Acknowledgements
The authors gratefully acknowledge Hao Xu for leading the design and implementation of the prototype FHIR PIT software code. The authors also wish to acknowledge; Stanley C. Ahalt, Dean of the UNC School of Data Science and Society and former Director of the Renaissance Computing Institute, for his support and advice on the work described herein; Emily Pfaff and James Champion for their efforts on CAMP FHIR; and Sarav Arunachalam, Stephen A. Appold, Alejandro Valencia Arias, Brian Naess, and Lisa Stillwell for their help with the environmental exposures data. Finally, the authors are especially grateful to Sue Nolte for independent testing of the FHIR PIT software, documentation, and sample data and to Michael Knowles, Margaret Leigh, and David B. Peden for serving as clinical subject matter experts on the asthma and PCD use cases.

Availability and implementation
FHIR PIT open-source software code, technical documentation, and sample input data can be found at: https://github.com/ExposuresProvider/FHIR-PIT. See README overview at https://github.com/ExposuresProvider/FHIR-PIT/blob/master/README.md; specific run instructions and data transformation details at https://github.com/ExposuresProvider/FHIR-PIT/blob/master/spark/README.md; and (deidentified) sample input data at https://github.com/ExposuresProvider/FHIR-PIT/tree/master/data/input. FHIR PIT also has been dockerized to encapsulate the software and platform dependencies and facilitate execution and deployment (see https://github.com/ExposuresProvider/FHIR-PIT/blob/master/Dockerfile).
Contact: fhir-pit@renci.org.

Authors' contributions
K.F. prepared the first draft of the manuscript, served as a software tester; contributed to use case development; and conducted the ICEES queries described herein; J.G. and H.Y. co-led FHIR PIT software development; J.G. contributed documentation and sample data to the FHIR PIT GitHub repository; G.R. contributed new environmental exposures data and software testing; H.Y. led ICEES OpenAPI software development; A.K. provided overall project guidance; all authors reviewed the final manuscript draft and approved submission.

Funding
This work was supported by the National Center for Advancing Translational Sciences, National Institutes of Health [award numbers OT3TR002020, OT2TR003430, UL1TR002489, UL1TR002489-03S4].

Declarations

Ethics approval and consent to participate
The work described in this manuscript was approved by the Institutional Review Board at the University of North Carolina at Chapel Hill (studies #16–2978 [asthma] and #21–0099 [PCD]).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [2]Copperline Professional Solutions, LLC, Pittsboro, NC, USA. [3]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

Received: 14 June 2024 Accepted: 9 December 2024
Published: 15 January 2025

References

1. Brokamp C, Wolfe C, Lingren T, Harley J, Ryan P. Decentralized and reproducible geocoding and characterization of community and environmental exposures for multi-site studies. J Am Med Inform Assoc. 2018;25(3):309–14. https://doi.org/10.1093/jamia/ocx128. https://degauss.org.
2. CAMP FHIR (Clinical Asset Mapping Program for HL7 Fast Healthcare Interoperability Resources), CAMP FHIR v1.0.4, released August 14, 2023. GitHub repository. https://github.com/NCTraCSIDSci/camp-fhir.
3. Choirat C, Braun D, Kioumourtzoglou M-A. Data science in environmental health research. Curr Epidemiol Rep. 2019;6:291–9.
4. Clark LP, Zilber D, Schmitt C, et al. A review of geospatial exposure models and approaches for health data integration. J Expo Sci Environ Epidemiol. Published online September 6, 2024. https://doi.org/10.1038/s41370-024-00712-8.
5. Cui Y, Eccles KM, Kwok RK, Joubert BR, Messier KP, Balshaw DM, et al. Integrating multiscale geospatial environmental data into large population health studies: challenges and opportunities. Toxics. 2022;10:403.
6. The Dhall configuration language. https://dhall-lang.org/.
7. Environmental Health Language Collaborative (EHLC). Environmental Health Language Collaborative: Harmonizing Data. Connecting Knowledge. Improving Health. December 11, 2023. https://www.niehs.nih.gov/research/programs/ehlc.
8. Fecho K,* Ahalt SC, Appold S, Arunachalam S, Pfaff E, Stillwell L, Valencia A, Xu H, Peden D. Development and application of an open tool for sharing and analyzing integrated clinical and environmental exposures data: asthma use case. JMIR Form Res. 2022;6(4):e32357; https://doi.org/10.2196/32357.
9. Fecho K,* Ahalt SC, Knowles M, Krishnamurthy A, Leigh M, Morton K, Pfaff E, Wang M, Yi H. Leveraging open electronic health record data and environmental exposures data to derive insights into rare pulmonary disease. Front Artif Intell. 2022;5:918888 (special issue on Biomedical Informatics Applications in Rare Diseases). https://doi.org/10.3389/frai.2022.918888.
10. Fecho K,* Garantziotis S, Krishnamurthy A, Pfaff E, Schmitt C, Schurman S, Shuptrine S, Xu H, Ahalt A. Open integrated analysis of multi-institutional data using ICEES. AMIA 2021 Virtual Annual Informatics Summit, March 2021.
11. 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 Inform Assoc. 2019;26(10):1064–73. https://doi.org/10.1093/jamia/ocz042.
12. Fecho K, Thessen AE, Baranzini SE, et al. Progress toward a universal biomedical data translator [published online ahead of print, 2022 May 25]. Clin Transl Sci. 2022;15(8):1838–47. https://doi.org/10.1111/cts.13301.
13. FHIR PIT input data, 2024: UNC Health EHR data in FHIR format (https://www.unchealthcare.org/); NIEHS Personalized Environment and Genes Study (PEGS) participant data (formerly known as Environmental Polymorphisms Registry) (https://www.niehs.nih.gov/research/clinical/studies/pegs/index.cfm); US Environmental Protection Agency airborne pollutant exposures data (https://www.epa.gov/hesc/rsig-related-downloadable-data-files); US Department of Transportation, Federal Highway Administration, Highway Performance Monitoring System major roadway/highway exposures data (http://www.fhwa.dot.gov/policyinformation/travel_monitoring/pubs/aadt/); US Census Bureau Topologically Integrated Geographic Encoding and Referencing (TIGER)/Line roadway data (http://www.census.gov/geo/maps-data/data/tiger-line.html); US Census Bureau American Community Survey socio-economic exposures data (https://www.census.gov/programs-surveys/acs/data.html); NC Department of Environmental Quality concentrated animal feeding operations (CAFO) exposures data (https://deq.nc.gov/cafo-map); NC Department of Environmental Quality landfill exposures data (https://www.nconemap.gov/datasets/ncdenr::active-permitted-landfills/about); and National Center for Education Statistics (https://nces.ed.gov/datatools/).

14. Garcia JJ, Yi H, Fecho K, Krishnamurthy A. FHIR-PIT: Link FHIR records with environmental exposures. Poster presentation at BioIt World 2023, May 16-18, 2023.
15. Goodman S, BenYishay A, Lv Z, Runfola D. GeoQuery: Integrating HPC systems and public web-based geospatial data tools. Comp Geosci. 2019;2019(122):103–12.
16. Hu H, Liu X, Zheng Y, He X, Hart J, James P, et al. Methodological challenges in spatial and contextual exposome-health studies. Crit Rev Environ Sci Technol. 2023;53:827–46.
17. Lan B,* Haaland P, Krishnamurthy A, Peden DB, Schmitt PL, Sharma P, Sinha M, Xu H, Fecho K. Open application of statistical and machine learning models to explore the impact of environmental exposures on health and disease: an asthma use case. Int J Environ Res Public Health 2021;18(21):11398 [published as part of a special issue titled "Application of Biostatistical Modelling in Public Health and Epidemiology"]; https://doi.org/10.3390/ijerph182111398.
18. Mirabelli MC, Vaidyanathan A, Flanders WD, Qin X, Garbe P. Outdoor PM2.5, Ambient air temperature, and asthma symptoms in the past 14 days among adults with active asthma. Environ Health Perspect. 2016;124(12):1882–90. https://doi.org/10.1289/EHP92.
19. NumFOCUS, Inc. pandas.cut. 2024. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.cut.html.
20. NumFOCUS, Inc. pandas.qcut. 2024. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.qcut.html.
21. US Department of Health and Human Services (US DHHS). Health Insurance Privacy and Accountability Act, Safe Harbor Method for De-identification. October 25, 2022. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.
22. RENCI Environmental Exposures APIs (undated). Socioeconomic Exposures Service, https://bdt-social.renci.org/socio_environmental_exposures_api/v1/ui/; Airborne Pollutant Exposures Service, https://bdt-cmaq.renci.org/cmaq_exposures_api/v1/ui/; Roadway Exposures Service, https://bdt-proximity.renci.org/roadway_proximity_api/v1/ui/.
23. 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. https://doi.org/10.2196/15199. https://github.com/NCTraCSIDSci/camp-fhir.
24. Sharma P,* Haaland P, Krishnamurthy A, Lan B, Schmitt PL, Sinha M, Xu H, Fecho K. Evaluating robustness of a generalized linear model when applied to electronic health record data accessed using an openAPI. Health Inform J. 2023;29(2); https://doi.org/10.1177/14604582231170892.
25. US Centers for Disease Control and Prevention (US CDC). (2023). National Environmental Public Health Tracking Network Data Explorer. https://ephtracking.cdc.gov/DataExplorer/.
26. US Department of Health and Human Services, Form Approved OMB# 0990–0379, Expiration Date 8/31/2023. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#safeharborguidance.
27. US Environmental Protection Agency (EPA). National Ambient Air Quality Standards (NAAQS) for PM. March 6, 2024. https://www.epa.gov/pm-pollution/national-ambient-air-quality-standards-naaqs-pm.
28. US NASA Socioeconomic Data and Applications Center (SEDAC). (2023). SEDAC: A data center in NASA's Earth Observing System Data and Information System (EOSDIS)- Hosted by CIESIN at Columbia University. https://sedac.ciesin.columbia.edu.
29. Valencia A, Stillwell L, Appold S, Arunachalam S, Cox S, Xu H, Schmitt CP, Schurman SH, Garantziotis S, Xue W, Ahalt SC, Fecho K. Translator Exposure APIs: open access to data on airborne pollutant exposures, roadway exposures, and socio-environmental exposures and use case application. IJERHP. 2020;17(14):5243. https://doi.org/10.3390/ijerph17145243.
30. Vineis P, Robinson O, Chadeau-Hyam M, Dehghan A, Mudway I, Dagnino S. What is new in the exposome? Environ Int. 2020;143:105887. https://doi.org/10.1016/j.envint.2020.105887.
31. Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev. 2005;14:1847–50.
32. 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 Inform Decis Mak. 2020;20:53. https://doi.org/10.21203/rs.2.19633/v1.
33. Yi H. pcornet-to-fhir mapping tool. 2024. https://github.com/RENCI/tx-pcornet-to-fhir.

Publisher's Note