The Pediatric Emergency Care Applied Research Network Registry: A Multicenter Electronic Health Record Registry of Pediatric Emergency Care

Sara J. Deakyne Davies1 Robert W. Grundmeier2 Diego A. Campos3 Katie L. Hayes4 Jamie Bell5 Evaline A. Alessandrini6 Lalit Bajaj7 James M. Chamberlain4 Marc H. Gorelick8 Rene Enriquez5 T. Charles Casper5 Beth Scheid6 Marlena Kittick9 J. Michael Dean5 Elizabeth R. Alpern10 and the Pediatric Emergency Care Applied Research Network

1Department of Research Informatics, Children’s Hospital Colorado, Aurora, Colorado, United States
2Department of Pediatrics and Department of Biomedical and Health Informatics, Children’s Hospital of Philadelphia, Philadelphia, Pennsylvania, United States
3Department of Biomedical and Health Informatics, Children’s Hospital of Philadelphia, Philadelphia, Pennsylvania, United States
4Department of Pediatrics, Children’s National Medical Center, Washington, District of Columbia, United States
5Department of Pediatrics, University of Utah, Salt Lake City, Utah, United States
6Department of Pediatrics, Cincinnati Children’s Hospital Medical Center, Cincinnati, Ohio, United States

Address for correspondence Elizabeth R. Alpern, MD, MSCE, Department of Pediatrics, Northwestern University Feinberg School of Medicine, Ann & Robert H. Lurie Children’s Hospital, 225 East Chicago Avenue, Box 62, Chicago, IL 60611, United States (e-mail: ealpern@luriechildrens.org).

7Department of Pediatrics, University of Colorado and Children’s Hospital Colorado, Aurora, Colorado, United States
8Department of Pediatrics, Medical College of Wisconsin and Children’s Hospital of Wisconsin, Milwaukee Wisconsin, United States
9Department of Pediatrics, Children’s Hospital of Philadelphia, Philadelphia, Pennsylvania, United States
10Department of Pediatrics, Northwestern University Feinberg School of Medicine, Ann & Robert H. Lurie Children’s Hospital of Chicago, Chicago, Illinois, United States


Abstract

Background Electronic health record (EHR)-based registries allow for robust data to be derived directly from the patient clinical record and can provide important information about processes of care delivery and patient health outcomes.

Methods A data dictionary, and subsequent data model, were developed describing EHR data sources to include all processes of care within the emergency department (ED). ED visit data were deidentified and XML files were created and submitted to a central data coordinating center for inclusion in the registry. Automated data quality control occurred prior to submission through an application created for this project. Data quality reports were created for manual data quality review.

Results The Pediatric Emergency Care Applied Research Network (PECARN) Registry, representing four hospital systems and seven EDs, demonstrates that ED data from disparate health systems and EHR vendors can be harmonized for use in a single registry with a common data model. The current PECARN Registry represents data from 2,019,461 pediatric ED visits, 894,503 distinct patients, more than 12.5 million narrative reports, and 12,469,754 laboratory tests and continues to accrue data monthly.

Conclusion The Registry is a robust harmonized clinical registry that includes data from diverse patients, sites, and EHR vendors derived via data extraction, deidentification, and secure submission to a central data coordinating center. The data provided may be used for benchmarking, clinical quality improvement, and comparative effectiveness research.

Keywords
- electronic health records and systems
- registries
- pediatrics
- emergency medicine
- quality
- data cleaning and deidentification

Received November 21, 2017
Accepted after revision March 29, 2018

Copyright © 2018 Schattauer
ISSN 1869-0327.
Background and Significance

Electronic health records (EHRs) collect and store substantial amounts of patient data with the potential to significantly improve health care delivery, health outcomes, and clinical research. There are clear national priorities to promote the development and use of EHRs with the necessary functionality to improve patient care, increase efficiency, and support performance measurement.\(^1\)\(^2\) EHR-based registries, compilations of data derived directly from the EHR on a cohort of patients, can provide important information about both the processes of care delivery and patient health outcomes without manual data abstraction, and can reduce reliance on less accurate nonclinical (administrative) data for registry activities.\(^4\)\(^7\) An EHR registry also can include free-text information (e.g., provider notes), which can be a valuable source of information not available elsewhere.\(^4\)\(^8\)\(^9\)

Objective

The Pediatric Emergency Care Applied Research Network (PECARN) is the first federally funded network for research on pediatric emergencies.\(^10\)\(^11\) We undertook, within PECARN, the creation of an EHR-based registry with the goal to transcend the limitations of existing administrative databases by collecting emergency department (ED) care data derived directly from the EHR. In this article, we describe the creation of the PECARN Registry, which is part of a larger project funded by the Agency for Healthcare Research and Quality (AHRQ) to use EHR-derived data to provide benchmarked, stakeholder-endorsed, quality metrics as well as audit and feedback to emergency providers on their care derived from these metrics. The PECARN Registry was constructed across multiple sites, two EHR vendors on various software versions, varying operating systems and configurations, and differing clinical workflows.

Methods

To fully achieve the aims of the study and to allow for the measurement of both predetermined and potential quality metrics and research questions related to ED care, we defined the scope of the Registry to include all processes of care within the ED (assessment, treatment, disposition). We established a working group of epidemiologists, quality improvement scientists, clinicians, informaticians, data analysts, biostatisticians, and research coordinators to define discrete and free-text variables. We developed a data dictionary describing the desired data elements, EHR data sources, and whether data elements were recorded multiple times within an ED visit (e.g., vital signs) (Supplementary Material, available in the online version).

We considered established data models that existed at the time this project was started such as Informatics for Integrating Biology and the Bedside (i2b2), Pediatric Health Information System Plus (PHIS +), and Observational Medical Outcomes Partnership (OMOP).\(^12\)\(^14\) These data models were originally constructed with specific uses such as cohort identification for research projects, comparative effectiveness research, and pharmacovigilance activities. Unfortunately, none of these models were able to adequately capture the structure of health care data from ED settings. Although these models had demonstrated flexibility, there remained challenges with representing the timestamp information necessary to understand the sequence and timing of health care events for our anticipated emergency care quality metrics. For example, due to the entity-attribute-value model for observations and measurements in i2b2 and OMOP, it was cumbersome to connect the time a radiology order was placed to the time the study was actually performed, the time images became available for viewing, and subsequently the time the imaging study result report became available. Furthermore, none of these models included locations for storing timestamp data, outside of those that were inherent to their model. However, similar to these models, when available in the source systems, we used common terminologies for coding our data, such as International Classification of Diseases (ICD)-9/10, Current Procedural Terminology (CPT), and Logical Observation Identifiers Names and Codes (LOINC). We also included site-specific codes and names of medications, laboratories, procedures, etc. so that we could access the data that was not in a standard terminology, without adding additional work to the sites to manually map these values. Manually mapping site-specific codes to standard terminologies is a time-intensive activity that requires terminology experts and local expertise for each domain. For efficiency, we instead opted to use expert consensus within the study group to develop categories for mapping values where standard terminologies were not available or were not adequately granular for the anticipated uses in the source systems. For example, relevant categories were determined and applied for mode of arrival and disposition (Supplementary Material, available in the online version). We also used a Web-based application that facilitates linking values for certain data elements to standard reference values related to various performance measures in the report cards.

Narrative data, including clinical notes, radiology reports and impressions, microbiology results, and other laboratory results, were included to facilitate natural language processing (NLP) methods for both planned and future quality improvement and research needs. This required the deidentification of free-text. We collaborated with stakeholders at each site and the data coordinating center (DCC) to establish the overall architecture, source-to-target mapping, deidentification methodology, and data flow (Fig. 1).

Data Extract, Transform, and Load

A significant challenge to establishing a registry based on the EHR is the variability in EHRs themselves. Different vendors, and even different software versions from the same vendor, may have different variable definitions or data table structures. EHR customizations to support the local clinical workflow lead to additional inconsistencies. The PECARN Registry involves four health systems and seven EDs (Table 1). Five of the sites (within three different health systems) used the same EHR vendor, so a common script was used for the extract, transform, and load (ETL) of their data, with
programmed exceptions for site-specific differences. This script was developed by combining expertise at all the sites. Each site provided fragments of code, some used for previous projects, which were then assembled into a single script. Each iteration of the script was run locally at the sites, data were validated, and changes were made to improve the accuracy of the extracted data. Although there was a core script that was used at all sites, differences across sites were reflected in branches of code, or through the use of regular expression pattern matching techniques to standardize disparate values across the sites. Structured query language (SQL) code was packaged using the Scala programming language (https://www.scala-lang.org), and the open source DataExpress toolkit (http://github.com/chop-dbhi/dataexpress) to create an Extensible Markup Language (XML) file based on an XML Schema Definition (XSD) file created for the data model (Supplementary Material, available in the online version). The open source PostgreSQL database (https://www.postgresql.org) was also utilized as a staging database for applying needed mapping and data transforms before creating the XML files. Using the same database system allowed common SQL to be used and minimized variation in syntax which may occur if using a different relational database management (RDBM) system (e.g., Microsoft Access or MySQL). Developing a common ETL script for these five sites was an iterative process that required manual review of the data and mappings to ensure the data were correct, complete, and mapped appropriately.

Two other sites (within the same health system) shared a different EHR vendor, so a single script was developed for the ETL from these sites. Data were transferred using six discrete reports created from the EHR’s proprietary data extraction tools. These were automatically exported daily and were loaded temporarily into a MS Access (Version 2010, Microsoft Inc., Redmond, Washington, United States) database using Visual Basic, then were exported to the SQL Server (Version 2012, Microsoft Inc.), from which the XML file was created.

For all sites, the data extraction processes produced an XML file that contained all the data for the project. The DCC developed an application called “Marshal” that validated the XML files against the XSD file previously defined for this project. This application also performed the deidentification and data submission tasks described below.

Deidentification

Although all sites had institutional review board approval with Health Insurance Portability and Accountability Act (HIPAA) waivers to send limited quantities of fully identified data to the DCC and the DCC has the appropriate controls in place, such as meeting the requirements for the National Institute of Standards and Technology framework, we sought to further decrease risk by deidentifying clinical documentation prior to transmission to the DCC. We used the De-ID software developed by PhysioNet to replace names with deidentified tags and to shift dates found in narrative elements.15,16 De-ID uses lists of known identifiers (e.g., names) and patterns to
## Table 1 Patient population and EHR vendor by site for January 2012–June 2016

<table>
<thead>
<tr>
<th></th>
<th>The Children's Hospital of Philadelphia</th>
<th>Cincinnati Children's Hospital Medical Center</th>
<th>Children's National Medical Center</th>
<th>The Children's Hospital Colorado</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base ED</td>
<td>Base ED</td>
<td>Satellite ED</td>
<td>Base ED</td>
</tr>
<tr>
<td><strong>ED visits in PECARN Registry 2012–June 2016</strong></td>
<td>421,622 (20.9%)</td>
<td>409,967 (20.3%)</td>
<td>196,300 (9.7%)</td>
<td>376,517 (18.6%)</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td>Median (IQR) year</td>
<td>4.9 (1.7, 10.8)</td>
<td>5.6 (1.8, 12.5) year</td>
<td>5.6 (2.0, 11.4) year</td>
</tr>
<tr>
<td></td>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>199,620 (47.3%)</td>
<td>200,281 (48.9%)</td>
<td>94,879 (48.3%)</td>
<td>177,254 (47.1%)</td>
</tr>
<tr>
<td>Male</td>
<td>221,993 (52.7%)</td>
<td>209,672 (51.1%)</td>
<td>101,416 (51.7%)</td>
<td>199,263 (52.9%)</td>
</tr>
<tr>
<td>Missing</td>
<td>9 (0%)</td>
<td>14 (0%)</td>
<td>5 (0%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>32,276 (7.7%)</td>
<td>18,907 (4.6%)</td>
<td>20,124 (10.3%)</td>
<td>88,290 (23.4%)</td>
</tr>
<tr>
<td>Non-Hispanic</td>
<td>389,346 (92.3%)</td>
<td>391,060 (95.4%)</td>
<td>176,176 (89.7%)</td>
<td>288,227 (76.6%)</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Indian/Alaskan Native</td>
<td>259 (0.06%)</td>
<td>255 (0.06%)</td>
<td>170 (0.08%)</td>
<td>414 (0.1%)</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>13,411 (3.19%)</td>
<td>2,632 (0.64%)</td>
<td>2,597 (1.3%)</td>
<td>4,142 (1.1%)</td>
</tr>
<tr>
<td>Black</td>
<td>257,886 (61.2%)</td>
<td>174,845 (42.6%)</td>
<td>24,745 (12.6%)</td>
<td>225,581 (59.9%)</td>
</tr>
<tr>
<td>Other</td>
<td>46,347 (11.0%)</td>
<td>43,551 (10.6%)</td>
<td>31,795 (16.2%)</td>
<td>49,335 (13.1%)</td>
</tr>
<tr>
<td>White</td>
<td>102,105 (24.2%)</td>
<td>181,269 (44.2%)</td>
<td>132,981 (67.7%)</td>
<td>33,775 (9.0%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>221 (0.05%)</td>
<td>7,191 (1.8%)</td>
<td>3,936 (2.0%)</td>
<td>59,908 (15.9%)</td>
</tr>
<tr>
<td>Missing</td>
<td>1,393 (0.33%)</td>
<td>224 (0.05%)</td>
<td>76 (0.04%)</td>
<td>3,362 (0.89%)</td>
</tr>
<tr>
<td><strong>Payer type</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commercial</td>
<td>135,942 (32.2%)</td>
<td>117,521 (28.7%)</td>
<td>87,639 (44.6%)</td>
<td>80,938 (21.5%)</td>
</tr>
<tr>
<td>Governmental</td>
<td>268,621 (63.7%)</td>
<td>274,338 (66.9%)</td>
<td>102,052 (52.0%)</td>
<td>242,552 (64.4%)</td>
</tr>
<tr>
<td>Self Pay</td>
<td>17,047 (4.0%)</td>
<td>18,092 (4.4%)</td>
<td>6,607 (3.4%)</td>
<td>18,286 (4.9%)</td>
</tr>
<tr>
<td>Other</td>
<td>9 (0%)</td>
<td>16 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Missing</td>
<td>3 (0%)</td>
<td>0 (0%)</td>
<td>2 (0%)</td>
<td>34,741 (9.2%)</td>
</tr>
<tr>
<td><strong>ED disposition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admitted</td>
<td>69,378 (16.5%)</td>
<td>60,835 (14.8%)</td>
<td>12,963 (6.6%)</td>
<td>46,836 (12.4%)</td>
</tr>
</tbody>
</table>

(Continued)
recognize identifiable information and replaces it with deidentified tags. To preserve time intervals, dates were shifted by a constant amount across all ED visits for an individual child. We modified the De-ID software to remove the patient name and medical record number for all free-text associated with that patient using a search algorithm that uses a medical-record-to-name mapping file. In addition, we used lists of treating providers at an institution, and lists of common male and female names to create a lookup list of terms to remove additional identifiers within the free-text. Marshal ingested the XML file and then fed text to the De-ID PERL program to remove the identifiers, and replaced them with deidentified tags. The exceptions were dates, which were shifted with the interval preserved per patient, the patient identifier and encounter identifier, which were logged in a file kept locally by the study sites, with a map between the local identifier and the study identifier. Note that although the dates are shifted, the time element is preserved, so that the time of day is known. The shifting has a window that is sufficiently narrow, to preserve seasonality. The scrubbed text was then reintegrated back into an XML file via Marshal. Only study-coded identifiers and deidentified tags remained in the final produced XML. Prior to finalizing this process, 30% of records from each site were compared manually between limited fully identified and deidentified versions.

### Data Transfer and Automated Quality Control

After an XML file was deidentified, it was transferred via a secure Web service to the DCC. An ETL process was developed to extract the data from the XML file and load it into a temporary SQL data warehouse at the DCC. As part of the DCC’s ETL process, we developed business rules, such as range, missingness, and element comparison rules to ensure the quality of the data (Fig. 2A). The data were validated against these business rules to detect values that were out-of-range or did not meet the expected pattern prior to being loaded into the data warehouse.

#### Manual Quality Control

After initial creation of the scripts, data were captured for a single day at each site and were uploaded to the DCC to test the ETL process, deidentification, and receipt of data. Data transferred to the DCC were compared against source data in the EHR system by analysts, research coordinators, and site investigators. In addition, comprehensive data quality reports were created using SAS 9.3 (SAS Institute, Cary, North Carolina, United States) to provide an overview of the data and for comparison between sites (Fig. 2A–D). The data quality reports identified unmapped values, missing values, and the distribution of values to aid in detecting problems with the ETL (Fig. 2B–D). Note that rules for missingness for each variable were based on expected coverage of a given variable accounting for both clinical and situational expertise. For example, the threshold for the missing proportion of systolic blood pressure to “alarm” the data quality report (percent highlighted in red in Fig. 2C) is lower (e.g., 20% missingness) than for documentation of supplemenary oxygen, which is not clinically expected to be routinely performed for all patients (proportion of missingness visualized in yellow but percent not highlighted in red).

We then proceeded with monthly block submissions until the entire previous calendar year was submitted. After each monthly submission, data quality reports were reviewed for out-of-range values, missingness, unmapped values, categorical anomalies, and temporal trends. Members of the study team reviewed the EHR source data at the sites to identify and correct any issues and scripts were refined as needed. After the initial year was submitted, we began monthly data submissions, which occur 4 weeks after the completion of the calendar month to ensure that the majority of laboratory results, billing data, and hospital discharge information are available for extraction.

---

**Table 1 (Continued)**

<table>
<thead>
<tr>
<th></th>
<th>The Children’s Hospital of Philadelphia</th>
<th>Cincinnati Children’s Hospital Medical Center</th>
<th>Children’s National Medical Center</th>
<th>The Children’s Hospital Colorado</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base ED</td>
<td>Base ED</td>
<td>Satellite ED</td>
<td>Base ED</td>
</tr>
<tr>
<td>Died</td>
<td>73 (0.02%)</td>
<td>104 (0.03%)</td>
<td>6 (0%)</td>
<td>189 (0.05%)</td>
</tr>
<tr>
<td>Discharged</td>
<td>331,082 (78.5%)</td>
<td>333,789 (81.4%)</td>
<td>176,701 (90.0%)</td>
<td>320,300 (85.1%)</td>
</tr>
<tr>
<td>Transferred</td>
<td>3,159 (0.75%)</td>
<td>4,645 (1.13%)</td>
<td>4,084 (2.1%)</td>
<td>1,238 (0.33%)</td>
</tr>
<tr>
<td>Observation unit</td>
<td>9,677 (2.3%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>3,953 (1.0%)</td>
</tr>
<tr>
<td>Other</td>
<td>8,185 (1.94%)</td>
<td>10,594 (2.58%)</td>
<td>2,546 (1.30%)</td>
<td>3,441 (0.91%)</td>
</tr>
<tr>
<td>Missing</td>
<td>68 (0.02%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>560 (0.15%)</td>
</tr>
<tr>
<td>EHR vendor</td>
<td>Epic</td>
<td>Epic</td>
<td>Cerner</td>
<td>Cerner</td>
</tr>
</tbody>
</table>

Abbreviations: ED, emergency department; EHR, electronic health record; IQR, interquartile range; PECARN, Pediatric Emergency Care Applied Research Network.
Results

The PECARN Registry currently includes a total of 176 distinct variables including demographics, encounter characteristics, timestamps, vital signs, clinical scores, clinical care orders, results, medications, coded diagnoses and procedures, and free-text narratives from the entire ED encounter (Supplementary Material, available in the online version). Many of the 176 variables occur multiple times for each patient, such as heart rate being taken multiple times throughout the ED encounter, and all occurrences are captured in the registry with an associated timestamp. The PECARN Registry basic element characteristics are in ▶ Table 2 with over 2,019,461 visits and 894,503 unique patients represented in the study period. The database includes over 12.5 million narrative reports, more than 4.45 million heart rates documented, and 12,469,754 laboratory tests (▶ Table 2).

Monthly reports are provided to more than 490 clinicians on their individual performance on 20 quality metrics and to more than 50 site managers/physician leaders on overall site performance. Visits are attributed to providers, and metrics from the visit are included in their reports if the provider was ascribed to a patient at any time during the patient’s ED care. These reports are provided ~45 days after the completion of each month.

Some data elements had large numbers of response options that differed at each site. For example, the “mode of arrival” often contained the specific name of a local emergency medical service (EMS) agency. To simplify the mapping

---

**Fig. 2** (A) Data quality report: business rules examples. (B) Data quality report: submission summary and overview. (C) Data quality report: element completeness example. (D) Data quality report: element distribution example.
challenge and to anticipate new values that could appear over time, we used regular expression pattern matching to map local values to an allowed value for the registry (e.g., any arrival mode containing the abbreviation “BLS”—indicating basic life support—was mapped to the registry code for “EMS ground”).

Although we strove to include standard terminologies when available, surprisingly, even standard terminologies required as part of meaningful use were not uniformly implemented at sites during the entirety of this project. For example, the transition from ICD-9 to ICD-10 occurred in October 2015 for all participating sites, after the registry had been operating for several years. As meaningful use requirements promoted the use of LOINC terminology, these codes became available in 2016 at three sites, but are still currently unavailable for extraction at four of the sites. Systematized Nomenclature of Medicine (SNOMED) codes, available for diagnosis codes at only some of our sites, were reliant on mapping tables provided by third-party vendors to attach SNOMED codes to ICD-10 diagnosis descriptions. In addition, some codes were vendor-specific (e.g., Generic Product Identifier codes for medications from Medi-Span). Using a Web-based application, we manually reviewed all laboratories and medications monthly to identify those that may be involved with specific performance measures on the quality report cards. For example, we centrally identified all rapid group A Beta hemolytic streptococcal tests used in the cohort definition of a performance measure.

The deidentification process resulted in complete deidentification of medical record number (MRN) and patient name in discrete fields and within free text when written correctly, and 97% deidentification of other identifiers. Alterations of a name, such as nicknames, a misspelling, or missing leading or trailing spaces between the name and an adjacent word have resulted in some missed deidentification. Ongoing quality checks are performed and continuing improvements to the Marshal program are undertaken, such as updates to the graphical user interface, updates to the internal functions of the application to reduce the likelihood of failure due to lack of adequate memory resources, and improved error handling.

Due to the potential presence of remaining identifiers, we treat this as a data warehouse with personal health information (PHI), and maintain the same security controls as used for other PHI. However, data sets derived from this data warehouse that do not contain free text and narrative fields are considered deidentified as all dates have been shifted.
After the deidentification process and incorporation into a central data warehouse, free-text radiology narrative and impression data have been successfully used to perform NLP on identification of long bone fractures for a quality of care performance measure metric.18

Throughout the project, data quality procedures have continued and evolved. For each file submission, data that did not pass validation required an iterative resubmission process by the site until the data met the defined requirements or until an exception was granted after review by the overseeing research team (e.g., exceptions were granted noting that microbiology antibiotic sensitivity data were not available for some sites). Ongoing review and quality assurance checks revealed new data problems that arose after the initiation of our project, leading to an iterative process of resubmission and review, to improve the ETL script or mappings to improve accuracy. Some modifications were due to changes in the workflow or EHR build at a particular site. For example, one site had a change in how pain scores were recorded during the study period, originally documented in two different fields, one for the type of score and the other for the actual value (e.g., type = “BIERI” and Score = “0”), and subsequently documented in a field specific to the score (e.g., BIERI = “0”). Issues were recognized utilizing a combination of ongoing quality report review and manual review of patient records, and the ETL was updated to reflect changes.

Over the time this project has been active, new values were found in the EHRs that needed to be mapped to our terminologies. To avoid making changes to the program every time a new value occurred in the source systems that did not map using our existing regular expressions, we added a delimited file to the site local ETL process, where the site can add additional values to be mapped. This optional text file allows each site to add mapping changes without changing preliminary extracts and updates that are applied. This step is utilized in the PostgreSQL database using a simplified pattern for designating values to particular fields in sections a site may need to alter.

In addition to changes being made related to data quality, some changes were also made to optimize the ETL process. For example, one site experienced monthly run times of up to 5 hours for their ETL scripts. We discovered the site had created numerous additional events for tracking hospital metrics leading to larger tables slowing down the execution time of the script. The script was modified to shorten the run time within the range of other sites (~10 minutes for smaller monthly census sites and ~20 minutes for larger monthly
census sites). Resources needed at each site to run the ETL included an environment available to load the software (PERL, DataExpress, Marshal, Postgres), at least 4 gigabytes of memory, and 4 processors. One site initially had a virtual machine configured with a fixed amount of memory to meet these requirements, but eventually transitioned to a configuration that permitted expandable memory usage due to unexpectedly high memory requirements for the deidentification software at this site. With this new configuration, deidentification completes in just over 20 minutes for a smaller census and just under an hour for a larger census. It is now recommended that each site have at least 8 gigabytes of memory for the deidentification software. The burden of the ETL on local database resources was minimal, and did not require running the ETL during off hours or on a separate database. Each site also required an analyst with access to the EHR database and server or virtual machine who could run the ETL scripts, and troubleshoot any issues that arose. During initial development, the analyst time commitment was ~20 to 30% of full-time equivalent. Once the scripts and processes were finalized, the time commitment decreased to ~2 active hours per month, including time required to run the scripts, confirm all processes completed, pass the XML file through the deidentification program, submit the final XML file, and validate the data.

Discussion

The PECARN Registry demonstrates harmonization of ED data from disparate sources for use in a single registry with a common data model. For sites using one EHR vendor, we were able to use a common script, with minor site-specific accommodations, to query the EHR data directly. For sites using another EHR vendor, the process involved passing the data through an intermediary SQL server database after extraction from the EHR using proprietary reports. The common data model included timestamp data for ED events, allowing analysis of timing between events and providing insight into ED workflows.

The PECARN Registry contains EHR data that includes all vital signs, clinical scores, clinical documentation, medication orders and administrations, laboratory and imaging results, events, and orders related to ED care. The data model was flexible enough to allow for all occurrences of each of these ED care clinical elements, with each related to a timestamp, so that we did not need to limit ourselves to only a subset of clinical data in the database. This represents a rich data source for benchmarking, quality improvement (including audit and feedback), and comparative effectiveness research. For example, the PECARN Registry allows adjustment for severity of illness using numerous clinical variables, and even supports the use of NLP to extract information that may reside only in narrative documentation.18–21 These types of data are not available in registries based only on billing or pharmacy data.20,21 Differing from disease-specific registries,22–25 this registry allows the study of any injury or illness presenting to the ED. In contrast with other existing registries and databases that contain only some ED data, the PECARN Registry has a high level of detail for many variables specifically related to the ED course, including triage acuity level, multiple timestamps throughout the course of ED care, clinical scores, as well as final laboratory and radiology results. Furthermore, the PECARN Registry’s incorporation of narrative data increases the utility of the registry and facilitates a more complete picture of the patient and their ED course through the use of NLP to further define patient phenotypes.18

Although the PECARN Registry initially included seven sites, we worked to increase the scope of the endeavor by

<table>
<thead>
<tr>
<th>Data field</th>
<th>Total number populated in data field</th>
<th>Number of ED visits with data field populated</th>
<th>Mean number populated in data field per ED visit</th>
<th>Median number populated in data field per ED visit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distinct patients</td>
<td>894,503</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ED encounters</td>
<td>2,019,461</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ICD-9/ICD-10 CM diagnosis code</td>
<td>5,536,038</td>
<td>1,995,355</td>
<td>2.74</td>
<td>2</td>
</tr>
<tr>
<td>Laboratory result</td>
<td>12,469,754</td>
<td>474,071</td>
<td>6.17</td>
<td>0</td>
</tr>
<tr>
<td>ED medication order</td>
<td>2,637,339</td>
<td>1,198,958</td>
<td>1.31</td>
<td>1</td>
</tr>
<tr>
<td>Discharge medication prescription</td>
<td>1,447,220</td>
<td>890,190</td>
<td>0.72</td>
<td>0</td>
</tr>
<tr>
<td>Radiology exam</td>
<td>714,366</td>
<td>524,384</td>
<td>0.35</td>
<td>0</td>
</tr>
<tr>
<td>Narrative documents</td>
<td>12,666,442</td>
<td>2,003,160</td>
<td>6.27</td>
<td>5</td>
</tr>
<tr>
<td>Heart rate vital sign</td>
<td>4,451,500</td>
<td>1,932,988</td>
<td>2.20</td>
<td>1</td>
</tr>
<tr>
<td>Asthma score</td>
<td>358,202</td>
<td>104,807</td>
<td>0.177</td>
<td>0</td>
</tr>
<tr>
<td>Pain score</td>
<td>4,916,721</td>
<td>1,893,004</td>
<td>2.43</td>
<td>2</td>
</tr>
<tr>
<td>GCS score</td>
<td>239,880</td>
<td>158,170</td>
<td>0.119</td>
<td>0</td>
</tr>
<tr>
<td>Distinct ED care providers</td>
<td>10,812</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

including sites with two different EHR vendors (Epic and Cerner), which account for a large share of the national ED EHR market. In addition, the PECARN Registry includes four main academic EDs and three community satellite EDs, indicating that diversity of sites can be included within a centralized registry. As additional sites join the PECARN Registry, those on the same vendor as the five who shared a script are offered the opportunity to use this script. Future applications of these methods are already at work with a project to better understand and improve sepsis-related care for children. At a subset of PECARN Registry sites, we are expanding our data collection efforts to include available prehospital care (e.g., ambulatory or EMS care) as well as subsequent inpatient care.

Despite our best efforts, there are limitations to our work building this pediatric ED care registry. A known limitation of EHR data is that the data are not collected with the same purpose as prospectively collected research data, so it may include errors and missing values. We strove, however, to contain potential data quality issues by establishing a standardized reviewing process of variables on a monthly basis. We explore missingness and regularly validate that the absence of the data accurately reflects the EHR source. For example, although disposition was missing in ~0.15% of visits at one site, this accurately reflected the EHR of the visits affected. This process identified a deficiency in the clinical workflow related to documentation of disposition leading to a local quality improvement process at that site. Another limitation of this registry is that the inclusions of ED data points was guided by the quality of care performance measures of interest and, therefore, not every single available ED data point was collected. However, limiting our variables to the 176 selected allowed us to ensure that these were well-defined. We prioritized developing a model that captured the most pertinent elements of ED care, while minimizing coding complexity to utilize the data to achieve our aims. A lack of standard terminologies for many data domains also limits the distributable effectiveness and required additional manual steps for identifying values for quality metrics per differing site. However, we were able to use the data before standard terminologies existed in the source systems, and developed methods for investigators to easily review and add new values as needed.

Conclusion

We demonstrated successful implementation of a robust harmonized clinical registry using EHR data across seven sites, within four health systems using two EHR vendors for inclusion in a central ED registry used for quality improvement in the ED setting. Sites using the same EHR shared ETL scripts with some site-specific customizations. When common terminologies were not available in the source system, the data could still be used successfully by utilizing a mapping tool. The clinical data within the registry establishes a rich data source for provider benchmarking, quality improvement, and comparative effectiveness research. The registry is currently supporting numerous quality improvement activities across the participating institutions including completeness of vital signs, appropriate use of systemic steroids for asthma exacerbations, asthma symptom reassessment, and pain management for children with fractures.

Despite the widespread availability of EHR systems, collaborative research and quality improvement activities across multiple sites remain difficult due to challenges with data harmonization. Even when sites use the same EHR software, implementation decisions differ between sites to accommodate complex clinical workflows such as those found in pediatric emergency care settings. Despite these challenges, we successfully harmonized data across multiple pediatric emergency care sites.

Clinical Relevance Statement

To establish a pediatric emergency care EHR-based registry, we centrally developed and locally implemented a pipeline of software tools to facilitate data extraction, deidentification, generation of XML files, and secure submission of data to a central data coordinating center. In the absence of available standard terminologies, we developed software solutions for facilitating the identification and assignment of standardized values for use in our quality metrics and were successful in harmonizing data across sites and EHR vendors. The clinical data within the registry provides a rich data source for benchmarking, quality improvement, and comparative effectiveness research.

Multiple Choice Questions

1. The software toolkit utilized for ETL and to create the XML file at five of the sites is called:
   a. De-ID
   b. Marshal
   c. DataExpress
   d. SAS
   Correct Answer: The correct answer is option c. DataExpress toolkit (https://github.com/chop-dbhi/dataexpress) is the software used to create an Extensible Markup Language (XML) files for the five sites using a common script for this project.

2. Which of the following is a standard terminology used when harmonizing laboratory data?
   a. LOINC
   b. CPT
   c. ICD
   d. RxNorm
   Correct Answer: The correct answer is option a. LOINC stands for Logical Observation Identifiers Names and Codes and is a terminology used to identify measurements, observations, and documents (http://loinc.org).

Note

The information or content and conclusions are those of the authors and should not be construed as the official
position or policy of, nor should any endorsements be inferred by HRSA, HHS, or the U.S. Government.

**Protection of Human and Animal Subjects**

The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects and was approved by the institutional review boards of all study sites and the DCC.

**Funding**

This project work was supported by the Agency for Healthcare Research and Quality (AHRQ) grant R01HS020270. The PECARN infrastructure was supported by the National Institutes of Health (NIH), and the Emergency Medical Services for Children (EMSC) Network Development Program under cooperative agreements U03MC00008, U03MC00001, U03MC00003, U03 MC00006, U03MC00007, U03MC22684, and U03MC22685.

**Conflict of Interest**

None.

**References**

6. Quan H, Parsons GA, Ghali WA. Validity of procedure codes in International Classification of Diseases, 9th revision, clinical modification administrative data. Med Care 2004;42(08):801–809