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

Sara J. Deakyne Davies[1] Robert W. Grundmeier[2] Diego A. Campos[3] Katie L. Hayes[4] Jamie Bell[5] Evaline A. Alessandrini[6] Lalit Bajaj[7] James M. Chamberlain[4] Marc H. Gorelick[8] Rene Enriquez[5] T. Charles Casper[5] Beth Scheid[6] Marlena Kittick[9] J. Michael Dean[5] Elizabeth R. Alpern[10] and the Pediatric Emergency Care Applied Research Network

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

Appl Clin Inform 2018;9:366–376.

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).

[7] Department of Pediatrics, University of Colorado and Children's Hospital Colorado, Aurora, Colorado, United States
[8] Department of Pediatrics, Medical College of Wisconsin and Children's Hospital of Wisconsin, Milwaukee Wisconsin, United States
[9] Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States
[10] Department 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

DOI https://doi.org/10.1055/s-0038-1651496.
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.[1–3] 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

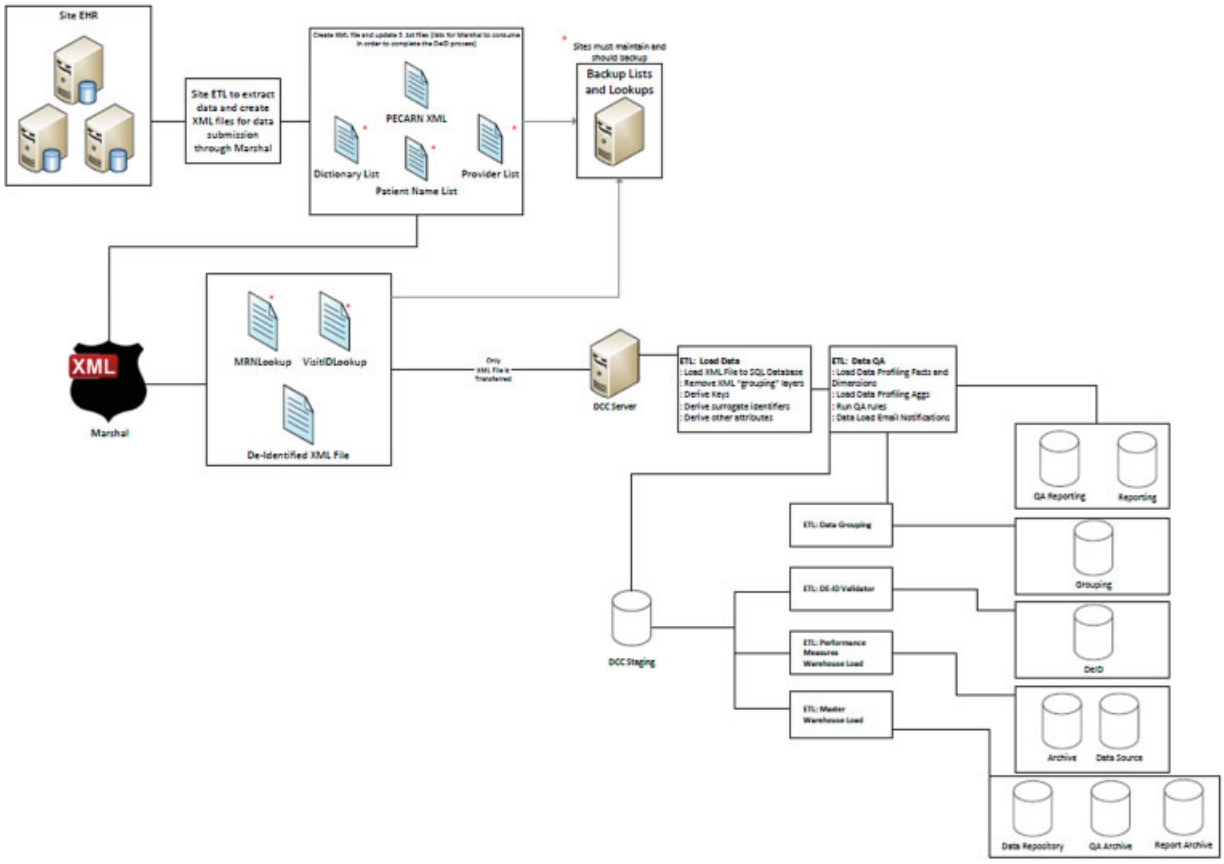

Fig. 1 Registry workflow diagram.

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

	The Children's Hospital of Philadelphia	Cincinnati Children's Hospital Medical Center		Children's National Medical Center		The Children's Hospital Colorado	
	Base ED	Base ED	Satellite ED	Base ED	Satellite ED	Base ED	Satellite ED
ED visits in PECARN Registry 2012–June 2016	421,622 (20.9%)	409,967 (20.3%)	196,300 (9.7%)	376,517 (18.6%)	157,802 (7.8%)	316,477 (15.7%)	140,776 (7%)
Age							
Median (IQR)	4.9 (1.7,10.8) year	5.6 (1.8,12.5) year	5.6 (2.0,11.4) year	4.8 (1.7,10.7) year	4.4 (1.8,9.5) year	5.6 (1.9,11.7) year	4.8 (1.9,9.7) year
Sex							
Female	199,620 (47.3%)	200,281 (48.9%)	94,879 (48.3%)	177,254 (47.1%)	76,526 (48.5%)	151,354 (47.8%)	66,164 (47.0%)
Male	221,993 (52.7%)	209,672 (51.1%)	101,416 (51.7%)	199,263 (52.9%)	81,276 (51.5%)	165,119 (52.2%)	74,610 (53.0%)
Missing	9 (0%)	14 (0%)	5 (0%)	0 (0%)	0 (0%)	4 (0%)	2 (0%)
Ethnicity							
Hispanic	32,276 (7.7%)	18,907 (4.6%)	20,124 (10.3%)	88,290 (23.4%)	3,519 (2.2%)	142,852 (45.1%)	36,869 (26.2%)
Non-Hispanic	389,346 (92.3%)	391,060 (95.4%)	176,176 (89.7%)	288,227 (76.6%)	154,283 (97.8%)	173,625 (54.9%)	103,907 (73.8%)
Race							
American Indian/ Alaskan Native	259 (0.06%)	255 (0.06%)	170 (0.08%)	414 (0.1%)	37 (0.02%)	1,334 (0.4%)	507 (0.4%)
Asian/Pacific Islander	13,411 (3.19%)	2,632 (0.64%)	2,597 (1.3%)	4,142 (1.1%)	93 (0.06%)	9,473 (3.0%)	2,308 (1.6%)
Black	257,886 (61.2%)	174,845 (42.6%)	24,745 (12.6%)	225,581 (59.9%)	153,787 (97.5%)	52,847 (16.7%)	1,056 (0.8%)
Other	46,347 (11.0%)	43,551 (10.6%)	31,795 (16.2%)	49,335 (13.1%)	1,567 (1.0%)	86,204 (27.2%)	33,179 (23.6%)
White	102,105 (24.2%)	181,269 (44.2%)	132,981 (67.7%)	33,775 (9.0%)	579 (0.4%)	158,877 (50.2%)	90,715 (64.4%)
Unknown	221 (0.05%)	7,191 (1.8%)	3,936 (2.0%)	59,908 (15.9%)	1,316 (0.83%)	7,597 (2.4%)	12,996 (9.2%)
Missing	1,393 (0.33%)	224 (0.05%)	76 (0.04%)	3,362 (0.89%)	423 (0.27%)	145 (0.05%)	15 (0.01%)
Payer type							
Commercial	135,942 (32.2%)	117,521 (28.7%)	87,639 (44.6%)	80,938 (21.5%)	12,451 (7.9%)	68,587 (21.7%)	72,538 (51.5%)
Governmental	268,621 (63.7%)	274,338 (66.9%)	102,052 (52.0%)	242,552 (64.4%)	122,520 (77.6%)	229,803 (72.6%)	64,267 (45.7%)
Self Pay	17,047 (4.0%)	18,092 (4.4%)	6,607 (3.4%)	18,286 (4.9%)	7,559 (4.8%)	17,877 (5.6%)	3,900 (2.8%)
Other	9 (0%)	16 (0%)	0 (0%)	0 (0%)	0 (0%)	209 (0.07%)	71 (0.05%)
Missing	3 (0%)	0 (0%)	2 (0%)	34,741 (9.2%)	15,272 (9.7%)	1 (0%)	0 (0%)
ED disposition							
Admitted	69,378 (16.5%)	60,835 (14.8%)	12,963 (6.6%)	46,836 (12.4%)	3,429 (2.2%)	36,900 (11.7%)	2,496 (1.77%)

(*Continued*)

Table 1 (*Continued*)

	The Children's Hospital of Philadelphia	Cincinnati Children's Hospital Medical Center		Children's National Medical Center		The Children's Hospital Colorado	
	Base ED	Base ED	Satellite ED	Base ED	Satellite ED	Base ED	Satellite ED
Died	73 (0.02%)	104 (0.03%)	6 (0%)	189 (0.05%)	101 (0.06%)	63 (0.02%)	0 (0%)
Discharged	331,082 (78.5%)	333,789 (81.4%)	176,701 (90.0%)	320,300 (85.1%)	150,835 (95.6%)	255,965 (80.9%)	132,833 (94.4%)
Transferred	3,159 (0.75%)	4,645 (1.13%)	4,084 (2.1%)	1,238 (0.33%)	120 (0.08%)	6,085 (1.92%)	2,430 (1.73%)
Observation unit	9,677 (2.3%)	0 (0%)	0 (0%)	3,953 (1.0%)	347 (0.22%)	6,781 (2.14%)	1,943 (1.38%)
Other	8,185 (1.94%)	10,594 (2.58%)	2,546 (1.30%)	3,441 (0.91%)	2,859 (1.81%)	10,683 (3.38%)	1,074 (0.76%)
Missing	68 (0.02%)	0 (0%)	0 (0%)	560 (0.15%)	111 (0.08%)	0 (0%)	0 (0%)
EHR vendor	Epic	Epic	Epic	Cerner	Cerner	Epic	Epic

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

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 reception 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 supplementary 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.

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

A

Top Data Quality Points					
			Dataset ID	1395	1392
			Submission Period	201601	201602
Rule Description	Rule Fail	Rule Warning	Rule Pass	DQ Flag	DQ Flag
Is the total number of visit records received in the data set within +/- 5% of the average number of visits per month? *(expected range: 5960-10517)*	No, greater than +/- 10% of average number of visits per month.	No, but within +/- 10% of average number of visits per month.	Yes, number of visits records received is with expected range	Pass	Pass
Do all visit records in the data set have an EDDoorYearMonth within the submission period?	No, visits outside submission >=1% of total visits records in data set	No, visits outside submission period > 0 and <1% of total visits records in data set	Yes, all records have EDDoorYearMonth within submission period	Pass	Pass
Is the data set free of duplicate records according to MedRecNbr + EDDoorDate + EDDoorTime?	No, duplicates >=1% of total visits records in data set	No, duplicates >0 and <1% of total visits records in data set	Yes, the data set is free of duplicates	Pass	Pass
Is the total number of records used for evaluations within the expected range? *(expected range: 5960-10517)*	No, greater than +/- 5% of expected range	No, but within +/- 5% of expected range	Yes, number of visits records received is with expected range	Pass	Pass
Is the data set free of patients with more than 10 re-occurring visits?	No, there are more than 1% of patients with reoccuring visits	No, dataset has patients with 5 re-occurring visits	Yes, data set is free of patients >=5 re-occurring visits	Pass	Pass
Is the number of distinct patients with re-occurring visits in the data set between 5% and 10% of total number of distinct patients *(using only records used for evaluations)* ?	No, number of re-occurring visits = 0 or greater than +/- 5% points of expected range (5-10%)	No, number of re-occurring visits within +/- 5%point of expected range (5-10%)	Yes, number of re-occurring visits between 5-10% of distinct patients in dataset.	Pass	Pass
Is the data set free of visits where date seq not BirthDate, EDDoorDate, EDDepartDateTime ?	No, data set has >5% of records used for evaluations with dates preceding BirthDate	No, data set has >0 and <5% of records used for evaluations with dates preceding BirthDate	Yes, all populated BirthDate values occur prior to other date values	Pass	Pass
Is the data set free of visits where any date preceeds BirthDate?	No, data set has >5% of records used for evaluations with dates preceding BirthDate	No, data set has >0 and <5% of visits used for evaluations with dates preceding BirthDate	Yes, all populated BirthDate values occur prior to other date values	Pass	Pass
Do greater than 80% of visits have at least one ICD-9 or ICD-10 diagnosis code?	No, less than 80 of visits have a diagnosis code.	<no soft fail>	Yes, greater than 80% of visits have at least one diagnosis code.	Pass	Pass
Do greater than 95% of visits where patient not LWBS have a ProviderRoleID 1-6 *(Attending, Fellow, Resident, Nurse Practitioner, Physician Assistant or Other Treatment Initiators)* ?	No, less than 95% of visits where patient not LWBS have a ProviderRoleID 1-6?	<no soft fail>	Yes at least 95% of visits where patient not LWBS have a ProviderRoleID 1-6?	Pass	Pass
Do greater than 85% of visits where patient was admitted have at least one ICD-9 or ICD-10 diagnosis code?	No, less than 85% of visits where patient was admitted have a diagnosis code.	At least 85% but less than 95% of visits where patient was admitted have a diagnosis code.	Yes, greater than 85% of visits where patient was admitted have at least one diagnosis code.	Pass	Pass

B

Visit/Patient Counts	*Dataset ID* / *Submission Period*	1395 / 201601		1392 / 201602	
Number of Visit Records in XML submission		8,596		8,725	
Less: number of records with EDDoorYYYYMM outside of submission period[1]		0	0.0%	0	0.0%
Less: number of duplicate records[1]		0	0.0%	0	0.0%
Less: visits where patients has more than 10 re-occurring visits[1]		0	0.0%	0	0.0%
Equals: number of visit records used for evaluation[4]		**8,596**	**100.0%**	**8,725**	**100.0%**
Less: number of visits where patient left without being seen[2]		**121**	**1.41%**	**185**	**2.12%**
Equals: number of visit records used for evaluation less patients who left without being seen[5]		**8,475**	**98.59%**	**8,540**	**97.88%**
Number of distinct patients in data submission[3]		**7,902**		**8,046**	
Number of patients with re-occurring visits[3]		**621**	**7.86%**	**611**	**7.59%**

Other Data Quality Points	*Dataset ID* / *Submission Period*	1395 / 201601		1392 / 201602	
Number of visits with dates not in sequence of BirthDate - EDDoorDateTime - EDDischDateTime[2]		0	0.0%	0	0.0%
Number of visits that have any date preceding BirthDate[2]		0	0.0%	0	0.0%
Number of visits where patient has at least one ICD-9 or ICD-10 diagnosis code reported[2]		8,442	98.21%	8,492	97.33%
Number of visits where patient was admitted AND has at least one ICD-9 or ICD-10 diagnosis code reported[4]		1,478	100.0%	1,424	100.0%
Number of visits with at least one ProviderRoleID 1-6 where patient did not leave without being seen[2]		8,435	98.13%	8,486	97.26%
Number of visits where patient was seen by an Attending and patient did not leave without being seen[2]		8,023	93.33%	8,023	91.95%
Number of visits where patient was seen by an Attending, Physician Assistant or Nurse Practitioner and patient did not leave without being seen[2]		8,434	98.12%	8,480	97.19%

1 - Percentages are out of number of visit records in XML submission.
2 - Percentages are out of number of records used for evaluation.
3 - Percentage is out of number of distinct patients.
4 - Percentages are out of number of records used for evaluation where EDDisposition = admitted.
5 - Percentages are out of number of records used for evaluation less patients who left without being seen.
6 - Sparkline reflects % for all rows unless % is not expressed, then reflects N.

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.

C

Denominator: Vital	Group N = 8,571 (100%)			Group N = 8,695 (100%)		
	Total	Pctg	% Bar	Total	Pctg	% Bar
VitalDate	8,571	100%		8,695	100%	
VitalTime	8,571	100%		8,695	100%	
SystolicBP²	6,078	71%		6,369	73%	
DiastolicBP²	6,078	71%		6,369	73%	
HeartRate²	8,543	100%		8,667	100%	
O2Sat	5,268	61%		5,160	59%	
Supplemental_O2	2,144	25%		2,082	24%	
RespiratoryRate²	8,517	99%		8,653	100%	
TempC²	8,515	99%		8,635	99%	
TempRouteID	8,224	96%		8,377	96%	
WeightEstimated	8,312	97%		8,388	96%	
WeightKg²	8,312	97%		8,388	96%	

Denominator: EDMeds	Group N = 5,264 (61%)			Group N = 5,464 (63%)		
	Total	Pctg	% Bar	Total	Pctg	% Bar
MedCode	5,264	100%		5,464	100%	
MedName	5,264	100%		5,464	100%	
MedOrderedDate	5,260	100%		5,460	100%	
MedOrderedTime	5,260	100%		5,460	100%	
MedOrderedDose	5,097	97%		5,286	97%	
MedOrderedDoseUnits	5,097	97%		5,286	97%	
MedOrderedRoute	5,258	100%		5,458	100%	
MedOrderedSchedule	5,260	100%		5,460	100%	
MedAdministeredDate	5,264	100%		5,464	100%	
MedAdministeredTime	5,264	100%		5,464	100%	
MedAdministeredDose	5,127	97%		5,322	97%	

Fig. 2 (*Continued*)

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.[17] 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.

D

Denominator: Visits	Value Description	Group N = 8,596 (100%) Total	Pctg	% Bar	Group N = 8,725 (100%) Total	Pctg	% Bar
SexID	-2 No Data	0	0.0%		0	0.0%	
	-1 Value Not Mapped	0	0.0%		0	0.0%	
	F Female	4,128	48.02%		4,275	49.0%	
	M Male	4,468	51.98%		4,450	51.0%	
	U Stated Unknown	0	0.0%		0	0.0%	
TriageCategoryID	-2 No Data	4	0.05%		6	0.07%	
	-1 Data Not Mapped	0	0.0%		0	0.0%	
	1 ESI1	69	0.8%		72	0.83%	
	2 ESI2	1,713	19.93%		1,711	19.61%	
	3 ESI3	2,539	29.54%		2,522	28.91%	
	4 ESI4	3,318	38.6%		3,456	39.61%	
	5 ESI5	953	11.09%		958	10.98%	
	6 Other	0	0.0%		0	0.0%	
ArrivalModeID	-2 No Data	3,539	41.17%		2,750	31.52%	
	-1 Data Not Mapped	0	0.0%		0	0.0%	
	1 EMS Air	0	0.0%		1	0.01%	
	2 EMS Ground	203	2.36%		190	2.18%	
	3 Non-EMS/Walking	4,545	52.87%		5,325	61.03%	
	4 Other	309	3.6%		459	5.26%	
	5 Stated Unknown	0	0.0%		0	0.0%	

Denominator: Vital	Value Description	Group N = 8,571 (100%) Total	Pctg	% Bar	Group N = 8,695 (100%) Total	Pctg	% Bar
HeartRate	Missing HeartRate	28	0.33%		28	0.32%	
	<70	578	6.74%		606	6.97%	
	70-79	1,126	13.14%		1,150	13.23%	
	80-89	1,810	21.12%		1,894	21.78%	
	90-99	1,832	21.37%		1,892	21.76%	
	100-109	2,068	24.13%		2,153	24.76%	
	110-119	1,972	23.01%		2,135	24.55%	
	120-129	2,286	26.67%		2,437	28.03%	
	130-139	1,947	22.72%		2,017	23.2%	
	140-149	1,695	19.78%		1,658	19.07%	
	150-159	1,236	14.42%		1,110	12.77%	
	160-169	921	10.75%		836	9.62%	
	>=170	641	7.48%		542	6.23%	

Fig. 2 (*Continued*)

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

Table 2 PECARN Registry scope and sample data characteristics: January 2012 through June 2016

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

Abbreviations: ED, emergency department; GCS, Glasgow Coma Scale; ICD-9/10 CM, International Classification of Diseases-9/-10 Clinical Modification; PECARN, Pediatric Emergency Care Applied Research Network.

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

including sites with two different EHR vendors (Epic and Cerner), which account for a large share of the national ED EHR market.[26] 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.[18,19] 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 Health Resources and Services Administration (HRSA), the Maternal and Child Health Bureau (MCHB), and the Emergency Medical Services for Children (EMSC) Network Development Demonstration Program under cooperative agreements U03MC00008, U03MC00001, U03MC00003, U03 MC00006, U03MC00007, U03MC22684, and U03MC22685.

Conflict of Interest

None.

References

1 Quality Forum. Available at: http://www.qualityforum.org/News_And_Resources/Press_Releases/2008/Medicare_Law_Provision_Will_Make_Quality_Front_and_Center_in_America_s_Efforts_to_Successfully_Reform_Our_Nation_s_Healthcare_System.aspx. Accessed May 27, 2010
2 Committee on Ways and Means United States House of Representatives. Available at: http://waysandmeans.house.gov/media/pdf/111/hitech.pdf. Accessed May 27, 2010
3 United States Department of Health and Human Services. Available at: https://www.hhs.gov/sites/default/files/hhs-it-strategic-plan-final-fy2017-2020.pdf. Accessed March 11, 2018
4 Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013;51(08, Suppl 3):S30–S37
5 Roane TE, Patel V, Hardin H, Knoblich M. Discrepancies identified with the use of prescription claims and diagnostic billing data following a comprehensive medication review. J Manag Care Pharm 2014;20(02):165–173
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
7 Heintzman J, Bailey SR, Hoopes MJ, et al. Agreement of Medicaid claims and electronic health records for assessing preventive care quality among adults. J Am Med Inform Assoc 2014;21(04):720–724
8 Friedman DJ. Assessing the potential of national strategies for electronic health records for population health monitoring and research. Vital Health Stat 2 2006;(143):1–83
9 Devine EB, Capurro D, van Eaton E, et al. Preparing electronic clinical data for quality improvement and comparative effectiveness research: the SCOAP CERTAIN Automation and Validation Project. EGEMS (Wash DC) 2013;1(01):1025
10 Pediatric Emergency Care Applied Research Network. Rationale, development, and first steps. Pediatr Emerg Care 2003;19(03):185–193
11 Pediatric Emergency Care Applied Research Network. Rationale, development, and first steps. Acad Emerg Med 2003;10(06):661–668
12 Informatics for integrating biology and the bedside. Available at: http://www.i2b2.org. Accessed November 08, 2017
13 Observational Medical Outcomes Partnership. Available at: http://omop.org/. Accessed November 08, 2017
14 Narus SP, Srivastava R, Gouripeddi R, et al. Federating clinical data from six pediatric hospitals: process and initial results from the PHIS+ Consortium. AMIA Annu Symp Proc 2011;2011:994–1003
15 Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):E215–E220
16 Neamatullah I, Douglass MM, Lehman LW, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32
17 Health Information Technology: Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology, 2014 Edition; Revisions to the Permanent Certification Program for Health Information Technology. Available at: https://www.federalregister.gov/documents/2012/09/04/2012-20982/health-information-technology-standards-implementation-specifications-and-certification-criteria-for#p-33. Accessed March 05, 2018
18 Grundmeier RW, Masino AJ, Casper TC, et al; Pediatric Emergency Care Applied Research Network. Identification of long bone fractures in radiology reports using natural language processing to support healthcare quality improvement. Appl Clin Inform 2016;7(04):1051–1068
19 Goyal MK, Johnson TJ, Chamberlain JM, et al. Pediatric Care Applied Research Network (PECARN). Racial and ethnic differences in antibiotic use for viral illness in emergency departments. Pediatrics 2017;140(04):e20170203
20 Mahmoudi E, Kotsis SV, Chung KC. A review of the use of Medicare claims data in plastic surgery outcomes research. Plast Reconstr Surg Glob Open 2015;3(10):e530
21 Hannan EL, Samadashvili Z, Cozzens K, et al. Appending limited clinical data to an administrative database for acute myocardial infarction patients: the impact on the assessment of hospital quality. Med Care 2016;54(05):538–545
22 Buzzetti R, Salvatore D, Baldo E, et al. An overview of international literature from cystic fibrosis registries: 1. Mortality and survival studies in cystic fibrosis. J Cyst Fibros 2009;8(04):229–237
23 Crandall W, Kappelman MD, Colletti RB, et al. ImproveCareNow: the development of a pediatric inflammatory bowel disease improvement network. Inflamm Bowel Dis 2011;17(01):450–457
24 Salvatore D, Buzzetti R, Baldo E, et al. An overview of international literature from cystic fibrosis registries 2. Neonatal screening and nutrition/growth. J Cyst Fibros 2010;9(02):75–83
25 Charlifue S, Tate D, Biering-Sorensen F, et al. Harmonization of databases: a step for advancing the knowledge about spinal cord injury. Arch Phys Med Rehabil 2016;97(10):1805–1818
26 Gregg H. 50 Things to Know About Epic, Cerner, MEDITECH, McKesson, athenahealth and Other Major EHR Vendors; 2014. Available at: http://www.beckershospitalreview.com/healthcare-information-technology/50-things-to-know-about-epic-cerner-meditech-mckesson-athenahealth-and-other-major-ehr-vendors.html. Accessed June 23, 2016