Using the CER Hub to ensure data quality in a multi-institution smoking cessation study

Kari L Walker,[1] Olga Kirillova,[1] Suzanne E Gillespie,[1] David Hsiao,[2] Valentyna Pishchalenko,[2] Akshatha Kalsanka Pai,[3] Jon E Puro,[4] Robert Plumley,[5] Rustam Kudyakov,[6] Weiming Hu,[1] Art Allisany,[1] MaryAnn McBurnie,[1] Stephen E Kurtz,[7] Brian L Hazlehurst[1]

For numbered affiliations see end of article.

Correspondence to
Dr Brian Hazlehurst, Kaiser Permanente Northwest, Center for Health Research, 3800 N. Interstate Avenue, Portland, OR 97227, USA;
Brian.hazlehurst@kpchr.org

Received 2 January 2014
Revised 1 June 2014
Accepted 17 June 2014
Published Online First
3 July 2014

To cite: Walker KL, Kirillova O, Gillespie SE, *et al. J Am Med Inform Assoc* 2014;**21**:1129–1135.

ABSTRACT

Comparative effectiveness research (CER) studies involving multiple institutions with diverse electronic health records (EHRs) depend on high quality data. To ensure uniformity of data derived from different EHR systems and implementations, the CER Hub informatics platform developed a quality assurance (QA) process using tools and data formats available through the CER Hub. The QA process, implemented here in a study of smoking cessation services in primary care, used the 'emrAdapter' tool programmed with a set of quality checks to query large samples of primary care encounter records extracted in accord with the CER Hub common data framework. The tool, deployed to each study site, generated error reports indicating data problems to be fixed locally and aggregate data sharable with the central site for quality review. Across the CER Hub network of six health systems, data completeness and correctness issues were prevalent in the first iteration and were considerably improved after three iterations of the QA process. A common issue encountered was incomplete mapping of local EHR data values to those defined by the common data framework. A highly automated and distributed QA process helped to ensure the correctness and completeness of patient care data extracted from EHRs for a multi-institution CER study in smoking cessation.

INTRODUCTION

Comparative effectiveness research (CER) generates evidence on the effectiveness, benefits, and harms of treatments, with the objective of improving health care.[1–7] The expanding use of electronic health record (EHR) systems in care delivery operations results in substantial amounts of clinical data, making possible the measurement and assessment of many aspects of health care.[8 9] Clinical data is recorded in both structured (coded) and unstructured (non-coded and narrative text) fields in EHRs, providing a rich source for understanding details of health status, behaviors, care processes, and outcomes. However, these data are not directly amenable to CER analysis because of their heterogeneous nature, which results from variations in clinical practice, EHR implementations, and organizational priorities affecting the capture and representation of clinical information.[10] For studies involving multiple institutions, CER requires not only coordinated and scalable methods for extracting, aggregating, and analyzing complex clinical data, but also methods to verify that data quality is suitable for the research task.[11]

The CER Hub is a web-based platform for conducting multi-institutional research studies using comprehensive EHR data. The CER Hub platform employs informatics standards for data representation and provides tools for generating research quality data from the entire clinical record, including both structured (coded) and unstructured (text) data fields. The CER Hub enables standardized access to the entire electronic clinical record and accommodates data processing across multiple organizations irrespective of the EHR implementation. In this paper, we focus on quality assurance (QA) activities performed to ensure that data extracted from each study site's EHR data warehouse is reliable and effective in supporting study needs. The work reported in this paper was overseen by institutional review boards from each of the institutions involved, as part of their approval for conducting the CER Hub smoking cessation study.

A study using the CER Hub begins with a protocol describing the research questions and analysis methods. The study protocol also specifies the population and the data that will be used in the analyses. A cornerstone of the CER Hub is the proviso that source data containing protected health information (PHI) must remain at each study site under local governance. To satisfy this requirement while aggregating comprehensive and high quality EHR data from multiple participating study sites, software to extract data meeting study goals is configured using collaborative tools hosted centrally on the CER Hub and then distributed to the CER Hub data providers, allowing all patient and encounter level data to remain at each study site. Only shareable limited datasets specific to the study are returned to the CER Hub central site for analyses. In the case of the QA process, sharing is in the form of aggregate statistical reports returned to the CER Hub central site for QA assessments, leading to feedback that creates refinements in data extraction methods and (sometimes) study designs.

CER Hub common data framework

A key element of the CER Hub platform is use of a common semantics and syntax for EHR data, achieved by a series of transformation steps that we call the common data framework. In particular, data standardization within CER Hub is achieved in three steps:

1. Data are extracted directly from the EHR data warehouse into a structure common to all data providers called the Clinical Research Document (CRD). The CRD schema is based on a research-centric model of the EHR as a transactional repository for documentation of patient contacts and care delivered.
2. Field-based rules are applied to these CRD documents (using the 'emrAdapter' tool described below) to normalize field values using controlled vocabularies (see table 1).
3. Transformation rules are applied (by a second instance of emrAdapter tool) to generate an HL7 Clinical Document

Table 1 Fields of the clinical research document schema (CRDS) which are constrained by controlled vocabularies

CRDS sections	Total number of variables	Variables mapped to controlled vocabularies	Controlled vocabularies
Patient detail	8	Gender	M=Male, F=Female, UN=Unknown or other
		RacePrimary RaceSecondary Ethnicity	CDC defined race and ethnicity coding system adopted by HL7 and Health Information Technology Standards Panel (HITSP) and supported by PHIN Vocabulary Access and Distribution System (PHIN VADS).
Encounter detail	6	EncounterType	1=outpatient—scheduled care, 2=outpatient—emergency department care, 3=outpatient—urgent care/unscheduled or same day visit, 4=outpatient—ancillary, 5=outpatient—observation, 6=inpatient, 7=lab, 8=pharmacy, 9=telephone, 10=assist, 11=e-mail, 12=other
		ServiceDepartment	NUCC Health Care Provider Taxonomy Code Set: v11.0, 1/1/11
Providers	3	ProviderType	1=attending physician, 2=physician trainee, 3=other
		ProviderDept	NUCC Health Care Provider Taxonomy Code Set: v11.0, 1/1/11
Payers	3	InsCoverage	1=Commercial/Private, 2=Veterans Affairs, 3=CHAMPUS 4=Medicare Traditional, 5=Medicare Managed Care Plan, 6=Medicaid Traditional, 7=Medicaid Managed Care Plan, 8=Managed Care HMO/pre-paid, 9=Managed Care PPO/fee-for-service, 10=Managed Care capitated, 11=Self pay/charity, 12=Workers Compensation, 13=Other
Visit diagnosis	5	DiagCode	(see DiagCodingSystem)
		DiagCodingSystem	ICD9CM, SNOMEDCT, Other (give name)
		DiagOrder	1=primary, 2=secondary
Problems	6	ProbCode	(see ProbCodingSystem)
		ProbCodingSystem	ICD9CM, SNOMEDCT, Other (give name)
		ProblemStatus	Active, inactive, chronic, intermittent, recurrent, rule out, ruled out, resolved
Medications	25	MedCode	(see MedCodingSystem)
		MedCodingSystem	NDC, RXNORM, Other (give name)
		MedEventType	Order, Discontinue order, dispense, administer, medication review taking, medication review discontinued, administrative cancellation
		StrengthUnits	1=MEQ/MG, 2=MEQ/ML, 3=MG/ACTUAT, 4=MG, 5=MG/ML, 6=ML, 7=PNU/ML, 8=UNT/MG, 9=UNT/ML
		DoseUnits	1=tablets, 2=capsules, 3=vials, 4=packs, 5=ML, 6=MG, 7=ACTUAT
		Route	FDA route of administration NCI thesaurus OID: 2.16.840.1.113883.3.26.1.1 NCI concept code for route of administration: C38114
		FreqUnits	Seconds, minutes, hours, days, weeks, months
Tobacco	6	TobaccoStatus	1=current, 2=former, 3=never, 4=unknown
		Tobacco Type	1=cigarettes, 2=pipes, 3=cigars, 4=smokeless, 5=unknown
		SmokingPacksPerDay	0=0 packs a day(non-smoker would have this value), 1=1/2 pack a day, 2=1 pack a day,=2 packs a day, 4=3 packs a day, 5=4 packs a day, 6=5 or more packs per day, 7=Unknown
Immunizations	5	Code	(see CodingSystem)
		CodingSystem	CPT, RXNORM, other (give name)
Reasons	4	CodingSystem	Other (give name)
Allergies	7	Code	(see CodingSystem)
		CodingSystem	CPT, RXNORM, other (give name)
		Severity	High, medium, low
		Status	Active, prior history, no longer active
Health maintenance alerts	5	Code	(see CodingSystem)
		CodingSystem	Other (give name)
		ResolutionCode	Done, deferred, pt refused, cancelled/NA
Procedures	5	Code	(see CodingSystem)
		CodingSystem	CPT, SNOMEDCT, LOINC, other (give name)
		Status	Cancelled, held, aborted, active, completed
Referrals	8	Code	(see CodingSystem)
		CodingSystem	CPT, SNOMEDCT, LOINC, other (give name)
		Status	Cancelled, held, aborted, active, completed
		MedSpecialtyCode	NUCC Health Care Provider Taxonomy Code Set: v11.0, 1/1/11
Progress notes	6	Code	(see CodingSystem)
		CodingSystem	Other (give name)
		NoteStatus	Addendum, signed, retracted
Patient instructions	5	Code	(see CodingSystem)
		CodingSystem	Other (give name)

Architecture (CDA[12 13]) document used in subsequent data processing to generate study data.

Creation of CRD files

The CRD schema (CRDS) defines an encounter-based view of a patient's health and care (see the first step of the CER Hub common data framework) capturing *most* of the transactional data generated in the EHR for each patient contact, including vital signs, medications, procedures ordered or performed, diagnoses, progress notes, and after-visit summaries in XML (eXtensible Markup Language) file format. Additional data generated by the health system between patient contacts, such as returned lab results, medication dispenses, and delayed updates to the patient's problems, are captured in the same format. Since the CRDS closely resembles how data are captured in an EHR, it is relatively simple for CER Hub data providers to extract encounter records in this format.

Rationale for the QA process

Large-scale, multi-institution studies using complex EHR data generate a number of data-quality concerns, primarily the validity of the source data assembled. In a CER Hub study this is compounded by the magnitude of the data that can be included. The main issues addressed by the QA process include verification of proper data compilation (per CRD definitions), as well as identification of data completeness and correctness within a study site and detection of possible data semantics discrepancies across sites. Proper CRD generation is ensured by verifying that generated XML files satisfy the CRD schema—specifically, required fields are always populated, repeatable fields are populated appropriately, and empty fields are not mistakenly created. Additional verification ensures that the appropriate data fields (eg, medication dispense) hold valid values (eg, an NDC code is included) and are contained in the correctly typed encounter document (eg, pharmacy encounter, in the case of medication dispense) when the CRD files are created. However, because studies often involve multiple sites, identification of inconsistencies within a study site (eg, vital signs information found in pharmacy or telephone encounters) as well as uniformity in data meaning across sites can also be investigated (eg, we learned that 'completed' and 'signed' are equivalent concepts for Progress Notes Status within our network of study sites). Although each project's topic area will create a focus for detailed study-specific data quality investigations, the overall design objective is a generalizable and highly automated QA process applicable for any CER Hub project to identify discrepant data and assure completeness.

Using emrAdapter to generate data for the QA process

The emrAdapter tool, used in steps #2 and #3 of the CER Hub common data framework, when programmed with QA checks provides the mechanism for generating QA data. The emrAdapter is a general purpose tool for traversing XML documents and selectively applying transformation rules (written in Java) to data in each field of each document[14] (http://www.cerhub.org/web/cerhubpublic/resource-center). In order to construct a single QA process available for all CER Hub studies, the programmable interface of the emrAdapter tool was employed to develop a set of quality checks developed centrally (called the 'QA program') and then run at each study site. The QA program was designed to generate a detailed error report to remain at sites that identified problems to be fixed locally, as well as a summary report containing aggregate data to be shared with the central site (figures 1 and 2). An important task accomplished by the QA program was to ensure that inclusion of codes from standardized terminologies, as defined by the CER Hub common data framework, occurred where required (figure 3). We applied the QA process to data records at step #2 of the CER Hub common data framework (ie, while they were still in CRD format and not yet translated to CDA documents) because this is where we will learn the most about data quality with the least amount of effort.

QA process design

As noted by Kahn *et al*,[15] the key to ensuring validity of data quality across multiple sites is to take a comprehensive approach. Thus the CRD field definitions, and not particular study outcomes, drove our QA process design. Such a strategy would not only assure the data suitable for the study at hand, but also lead to opportunities to improve our common data framework and implementation of it as data extraction vehicle across the network of study sites. To test the completeness and correctness of all fields in the CRD schema, five types of computational processes were developed and used to implement a set of quality checks: (1) frequencies, (2) missing data checks (referred to as 'IsNull checks'), (3) crosstabs, (4) quartile checks, and (5) date comparison checks. The entire CRD was first reviewed by the QA process designers, field by field, to determine appropriate quality checks to perform on each variable. For example, the type of variable (character verses numeric) determined whether a frequency or range check was to be implemented. If the field was a required field, at the minimum a missing check was implemented. If there was a need for a logical comparison with another variable then a cross-tab was implemented (eg, checking patient gender against provider specialty to verify the majority of patients seen by an OBGYN were female). The definition of each type, along with examples of how specific checks were implemented, is summarized in table 2.

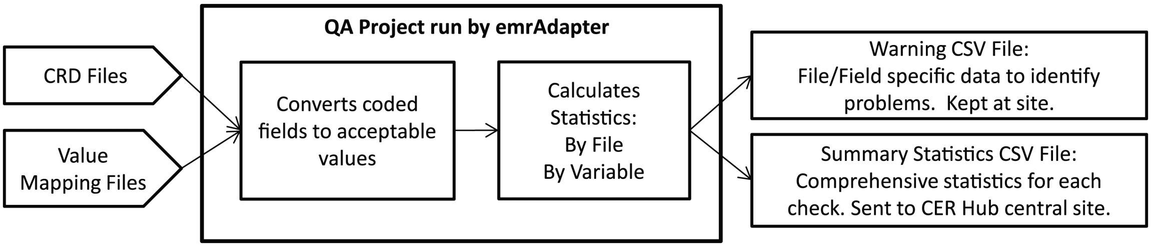
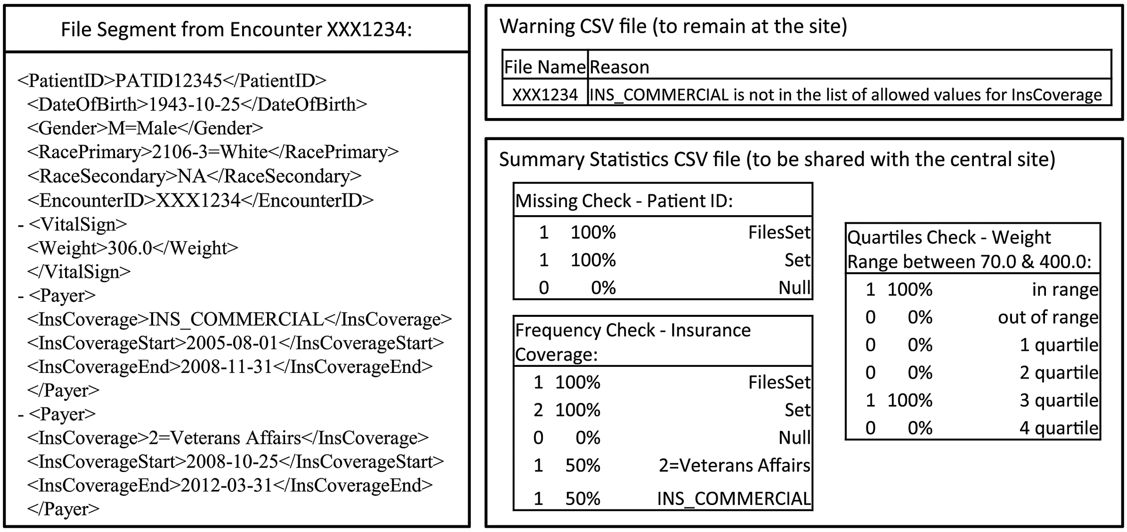

CRD Files
Value Mapping Files
QA Project run by emrAdapter
Converts coded fields to acceptable values
Calculates Statistics: By File By Variable
Warning CSV File: File/Field specific data to identify problems. Kept at site.
Summary Statistics CSV File: Comprehensive statistics for each check. Sent to CER Hub central site.

Figure 1 Use of the emrAdapter tool and components of the comparative effectiveness research (CER) Hub common data framework for the quality assurance (QA) process: reading in clinical research document (CRD) files and site defined value mapping files, converting discrete fields as defined in the lookup tables, calculating and producing report files.



File Segment from Encounter XXX1234:

```
<PatientID>PATID12345</PatientID>
 <DateOfBirth>1943-10-25</DateOfBirth>
 <Gender>M=Male</Gender>
 <RacePrimary>2106-3=White</RacePrimary>
 <RaceSecondary>NA</RaceSecondary>
 <EncounterID>XXX1234</EncounterID>
- <VitalSign>
 <Weight>306.0</Weight>
 </VitalSign>
- <Payer>
 <InsCoverage>INS_COMMERCIAL</InsCoverage>
 <InsCoverageStart>2005-08-01</InsCoverageStart>
 <InsCoverageEnd>2008-11-31</InsCoverageEnd>
 </Payer>
- <Payer>
 <InsCoverage>2=Veterans Affairs</InsCoverage>
 <InsCoverageStart>2008-10-25</InsCoverageStart>
 <InsCoverageEnd>2012-03-31</InsCoverageEnd>
 </Payer>
```

Warning CSV file (to remain at the site)

File Name	Reason
XXX1234	INS_COMMERCIAL is not in the list of allowed values for InsCoverage

Summary Statistics CSV file (to be shared with the central site)

Missing Check - Patient ID:

1	100%	FilesSet
1	100%	Set
0	0%	Null

Frequency Check - Insurance Coverage:

1	100%	FilesSet
2	100%	Set
0	0%	Null
1	50%	2=Veterans Affairs
1	50%	INS_COMMERCIAL

Quartiles Check - Weight Range between 70.0 & 400.0:

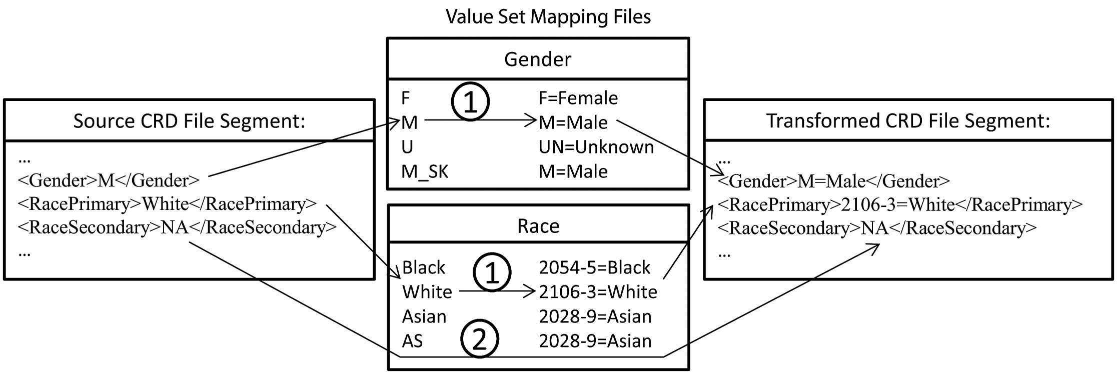
1	100%	in range
0	0%	out of range
0	0%	1 quartile
0	0%	2 quartile
1	100%	3 quartile
0	0%	4 quartile

Figure 2 Example of a segment from a clinical research document file with associated samples of each report type produced during the quality assurance process. Summary statistics vary by computational process depending on objective. Beyond frequencies of the variable's values reviewed, counts indicating breadth (total number of files variable populated, 'filesset') and trend (count of variable populated, encompassing repeating fields, 'set') are produced.

Step #2 of the CER Hub common data framework involves transforming data that populate discrete fields of CRD documents to values of standard vocabularies. As mentioned above, a value-mapping process was implemented by a programmed instance of the emrAdapter to achieve this data transformation for each site. Each site creates and maintains lookup tables as files, called 'Value Set Mapping Files', which define how site-specific EHR values map to values of the approved standard vocabularies of the CER Hub common data framework (figure 3). The emrAdapter reads these mapping definition files and applies text string replacements for all identified input values. The 'QA report file', one of two files generated by the QA program, highlights all invalid values resulting from this mapping process. These warnings are produced in a file that could contain PHI, and thus are not shared outside the local study site; instead, the local programmer uses them to learn details about what aspects of the data sample are out of compliance. A second file generated by the QA program, the 'QA result file', provides summary information on results of the checks applied to the sample and is shared with the central study site for quality improvement opportunities. The QA Process workflow is shown in detail in figure 4.

QA process implementation

The QA process was applied in a CER Hub study that assessed delivery of smoking cessation services in primary care of six participating health systems. In this case, samples of 10 000 primary care encounters per year, all from the study period January 1, 2006 to June 20, 2012, along with all associated pharmacy and 'Other' encounters, were extracted and used for QA at each site (table 3). Encounter records of type 'Other' include study-relevant data that are not linked to a patient visit but are typically generated between patient visits by the health system (eg, a problem list update generated sometime after the visit). Multiple rounds of the QA process were implemented to cover the entire time period in question and to allow for quality improvement intervention. The first round was restricted to January 1, 2006 to December 31, 2010; by the second round the remaining files from January 1, 2011 to June 30, 2012 had been reviewed; and a final round encompassed the entire study period of January 1, 2006 to June 30, 2012. For each round, a random sampling strategy was used to ensure the capture of encounters representative of data needs specified by the study protocol. It is worth noting that the iterative nature of



Source CRD File Segment:

```
...
<Gender>M</Gender>
<RacePrimary>White</RacePrimary>
<RaceSecondary>NA</RaceSecondary>
...
```

Value Set Mapping Files

Gender

F	F=Female
M	M=Male
U	UN=Unknown
M_SK	M=Male

Race

Black	2054-5=Black
White	2106-3=White
Asian	2028-9=Asian
AS	2028-9=Asian

Transformed CRD File Segment:

```
...
<Gender>M=Male</Gender>
<RacePrimary>2106-3=White</RacePrimary>
<RaceSecondary>NA</RaceSecondary>
...
```

Figure 3 Field value mapping process: text string matches are used to identify if the original value entry is present in the look up table, if so, then the assigned value from the look up table is used when creating the output file (1). If the original entry does not match any of the values in the lookup table, no replacements occur (2).

Table 2 Descriptions of quality checks designed and programmed for the quality assurance process

Type	Number implemented	Description	Example of use as quality check
Frequency	45	Counts (and percent of total) of values appearing for a given variable	▶ To determine if the distributions of discrete fields, like encounter types, gender, or insurance coverage, were as expected ▶ Verified all coded entries were mapped correctly in accord with Value Set Mapping Files
IsNull	35	Missing check, which produced a count and percent of the number of times a null entry was found for a given variable	▶ Helped to determine if a site had not populated any required fields, like Date of Birth or Encounter ID ▶ Highlighted which sites populated the optional data fields, that is, Immunization data
Crosstabs	39	Crosstabs compared values of two variables and counted associated pairs of values found in the data	Utilized to determine logic concerns, that is: ▶ Department compared to gender would identify men visiting obstetricians or ▶ Determine how many times the Referral Start Date was populated when Referral End Date was not or vice versa
DateCompare	6	This check depended on a predefined inequality expression between two date fields and produced count and percent of entries that satisfied the comparison	Most checks of this type were verifying end dates did not occur before start dates, that is: ▶ Insurance Start Date≤Insurance End Date or ▶ Encounter Start Date≤Death Date
Quartiles	30	Count and percent of continuous variables by quartiles of a predefined acceptable interval or identified a value as out of range	▶ Vital signs data: height, pulse, systolic BP, etc. ▶ Spirometry data: pre Fev1Fvc, post Fev1Fvc, peak flow, etc.

identifying and correcting data in this process requires a strong collaborative relationship between members of the study sites and the CER Hub QA staff.[15]

RESULTS

For the smoking cessation study reported here, a total of 155 logic checks were created and deployed in three rounds of the QA process (tables 2 and 3). Staff at the CER Hub central site reviewed result files generated by the QA program on each round of the QA process. Summary reports were generated by these staff and returned to each study site, including results from each quality check as well as an outline summarizing all data concerns for the study site meeting the needs of the overall study (see figure 4 for workflow of the QA process). Over three rounds of applying the process, the rates of successful quality checks increased from a starting point of 70–83% across sites to 100% correct for all study sites (table 3). The Date Comparison check type was the most successful, with a pass rate of 100% across all six study sites in its first deployment, which occurred in round 2 (see table 3). First-round summary reports indicated that half of the study sites had the EncounterType variable correctly populated 0–51% of the time. However, by the second round, all EncounterType values were verified correct. The problem with EncounterType demonstrates one example of the most common quality issue identified: incorrect or incomplete mapping of local EHR data values to the approved (controlled) values defined by the CER Hub common data framework. Overall, the quality check success rate for these prescribed mappings was 48–68% in round 1, but improved to 100% correct by the end of round 3.

DISCUSSION

The CER Hub QA process relies on distributed use of a common data framework enabling automation to be deployed across sites to efficiently apply quality checks to records, and fields within records, to uncover data inconsistencies in large samples of data records. We found that large scale, multi-institution comparisons made possible by automation can

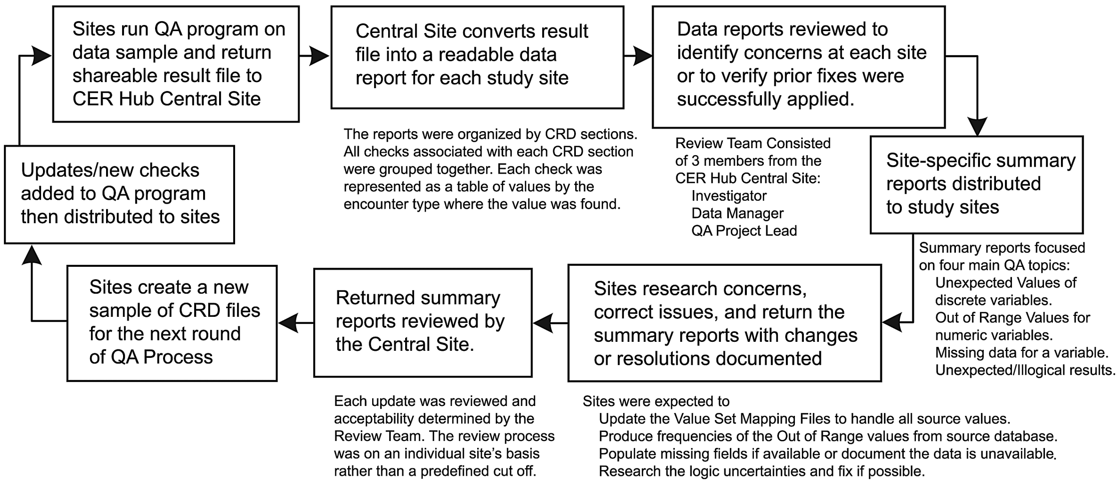

Sites run QA program on data sample and return shareable result file to CER Hub Central Site
Central Site converts result file into a readable data report for each study site
Data reports reviewed to identify concerns at each site or to verify prior fixes were successfully applied.
The reports were organized by CRD sections. All checks associated with each CRD section were grouped together. Each check was represented as a table of values by the encounter type where the value was found.
Review Team Consisted of 3 members from the CER Hub Central Site: Investigator Data Manager QA Project Lead
Site-specific summary reports distributed to study sites
Updates/new checks added to QA program then distributed to sites
Summary reports focused on four main QA topics: Unexpected Values of discrete variables. Out of Range Values for numeric variables. Missing data for a variable. Unexpected/Illogical results.
Sites create a new sample of CRD files for the next round of QA Process
Returned summary reports reviewed by the Central Site.
Sites research concerns, correct issues, and return the summary reports with changes or resolutions documented
Each update was reviewed and acceptability determined by the Review Team. The review process was on an individual site's basis rather than a predefined cut off.
Sites were expected to Update the Value Set Mapping Files to handle all source values. Produce frequencies of the Out of Range values from source database. Populate missing fields if available or document the data is unavailable. Research the logic uncertainties and fix if possible.

Figure 4 Quality assurance (QA) process workflow.

sometimes reveal otherwise hidden inconsistencies in the data. To elaborate on one example from our study, when the 'Secondary Race' patient demographic variable was mapped to its acceptable standardized values, an incorrect mapping to patient race of 'Other' became evident by the distribution of its occurrence across study sites (frequency ranged from 0.29% to 99.65%). Further investigation revealed that three of the six sites were collapsing 'Unknown' secondary race into 'Other', which was incorrect because 'Other' is sometimes what is in fact collected from patients. Once fixed, the value 'Other' for 'Secondary Race' appeared less than 1% of the time across all participating sites.

Ensuring quality in data derived from EHRs for multi-institution CER is a challenge faced by multiple informatics platforms for research that have been developed over the last decade (eg, i2b2/SHRINE,[16] SHARPn,[17] and 11 large American Recovery and Reinvestment Act of 2009 (ARRA) investments in clinical infrastructure for CER recently funded by the Agency for Health Care Research and Quality[18]). Sittig *et al*[19] compare and contrast six such platforms, including CER Hub, but do not include the specifics of data QA methods in their review. Table 4 provides a brief overview of the six projects that they surveyed. Califf[20] describes the new PCORNet, which aims to be a 'network of networks', and includes 11 different clinical data research networks based on a multitude of informatics platforms for aggregating research data across institutions for CER. Holmes *et al*[21] provide a comprehensive literature review of distributed research networks that utilize multi-institution electronic clinical data for research, although their topical focus was primarily policy and data governance issues of the networks, more than the informatics platforms utilized by the networks. One large difference between the CER Hub and some other platforms with more substantial infrastructure investments (eg, Mini-Sentinel and the Vaccine Safety Datalink) is that CER Hub does not maintain a study-independent data warehouse that is curated to improve data quality in advance of study needs. Instead, methods for transforming data to improve quality are developed study by study, and then applied as a composite transform in generation of study-specific datasets.

Kahn *et al*[15] identify the following general methods (types of rules) for assessing data quality within informatics platforms that use EHR data for multi-institution research. They say that rule types to implement QA should ideally provide: (1) validation of values in individual fields, (2) validation of relationships between data entities (including fields, records, datasets), (3) validation of time represented by data entities, (4) life-cycle validation for objects referred to by data (eg, coherence in starts and stops on clinical orders), and (5) validation of other (non-time) properties of real-world objects referred to by data (eg, any aggregate value should equal the sum of the component or atomic level values).

The CER Hub QA process currently addresses some but not all aspects of the Kahn *et al* framework. In particular, the QA

Table 3 Number of clinical research document files processed for each round of quality assurance (QA) by Encounter type

Electronic health record	Site 1 GE centricity	Site 2 Health connect/Epic	Site 3 Health connect/Epic	Site 4 Health connect/Epic	Site 5 Epic	Site 6 VISTA
Total files reviewed	84 359	952 919	313 844	374 818	326 637	338 239
Round 1						
Files reviewed	28 549	196 062	121 347	128 487	115 046	131 400
Outpatient	9888	50 000	50 000	50 000	50 000	50 000
Outpatient—ancillary	–	–	–	–	7449	–
Pharmacy	6873	26 804	51 165	7964	–	37 227
Other	11 788	119 258	20 181	70 523	57 597	44 173
Checks passed (n=145)	70%	73%	83%	77%	72%	83%
Frequency (n=44)	55%	50%	52%	48%	48%	68%
Quartiles (n=30)	77%	53%	87%	70%	67%	70%
IsNull (n=35)	51%	91%	100%	94%	77%	97%
Crosstabs (n=36)	100%	100%	100%	100%	100%	100%
Round 2						
Files reviewed	12 380	60 406	35 475	42 080	56 238	38 386
Outpatient	3100	15 000	15 000	15 000	15 000	15 000
Outpatient—ancillary	–	–	–	–	10 944	–
Pharmacy	3668	7310	14 458	2046	–	10 461
Other	5612	38 096	6015	25 034	30 294	12 925
Checks passed (n=155)	90%	91%	90%	90%	88%	97%
Frequency (n=45)	82%	84%	80%	78%	82%	91%
Quartiles (n=30)	77%	83%	80%	83%	83%	97%
IsNull (n=35)	100%	94%	100%	100%	86%	100%
Crosstabs (n=39)	100%	100%	97%	100%	100%	100%
Date compare (n=6)	100%	100%	100%	100%	100%	100%
*Round 3**						
Files reviewed	43 400	696 451	157 022	204 251	155 353	168 453
Outpatient	14 456	65 000	65 000	65 000	65 000	65 000
Outpatient—ancillary	–	–	–	–	11 966	–
Pharmacy	11 471	58 124	65 714	39 158	–	46 455
Other	17 473	573 327	26 299	100 093	78 387	56 998

*All checks passed after the third round of QA.

Table 4 Six large-scale projects implementing informatics platforms for comparative effectiveness research (CER), as surveyed by Sittig *et al*[19]

Project name	Project description
The Comparative Effectiveness Research Hub (CER Hub)	A web-based platform for implementing multi-institutional studies using the MediClass system for processing comprehensive electronic medical records, including both coded and free-text data elements
Washington Heights/Inwood Informatics Infrastructure for Comparative Effectiveness Research (WICER)	Creating infrastructure to facilitate patient-centered, comprehensive analysis of populations in New York City, NY by leveraging data from existing EHRs, and combining data from institutions representing various health care processes
Scalable PArtnering Network for Comparative Effectiveness Research: Across Lifespan, Conditions, and Settings (SPAN)	Uses its Virtual Data Warehouse (VDW) to provide a standardized, federated data system across 11 partners spread out across the nation
The Partners Research Patient Data Registry (RPDR)	An enterprise data warehouse combined with a multi-faceted user interface (i2b2) that enables clinical research and CER across Partners Healthcare in Boston, MA
The Indiana Network for Patient Care (INPC) Comparative Effectiveness Research Trial of Alzheimer's Disease Drugs (COMET-AD)	Started in 1994 as an experiment in community-wide health information exchange serving five major hospitals in Indianapolis, IN. Using data from hospitals and payers statewide to monitor various health care processes and outcomes
The Surgical Care Outcomes Assessment Program Comparative Effectiveness Research Translation Network (SCOAP-CERTN)	Assessing how well an existing statewide quality assurance and quality improvement registry (ie, the Surgical Care Outcomes Assessment Program) can be leveraged to perform CER

process quality checks reported here addressed their type 1 (providing validation of field values) and type 2 rules (providing validation among data objects generally), and some aspects of type 3 rules (providing some validation of time-based events represented by data entities). However, because our QA program focuses on the encounter record as unit of analysis, we currently have very little capacity to implement type 4 and 5 rules as defined in the Kahn *et al* framework. These deviations from their framework give us new directions to consider for improving our QA process.

CONCLUSION

As part of the CER Hub informatics platform for conducting studies using multi-institution EHR data, we developed a general QA process that can be used to monitor and ensure data quality. The QA process employs a QA program that is developed centrally, then distributed to participating CER Hub data providers. This design ensures that sensitive EHR data remain local and assists the local programmer in improving the quality of data extracted for a CER Hub study. The QA program also produces an aggregate report shared with the CER Hub central site that provides capabilities to verify data quality, ensure data completeness, enable quality improvement, and inform study design.

Author affiliations
[1]Kaiser Permanente Northwest, Center for Health Research, Portland, Oregon, USA
[2]Kaiser Permanente Hawaii, Center for Health Research, Honolulu, Hawaii, USA
[3]Emory University, Atlanta, Georgia, USA
[4]OCHIN Inc, Portland, Oregon, USA
[5]VA Puget Sound Health Care System, Seattle, Washington, USA
[6]Baylor Health Care System, Center for Clinical Innovation, Dallas, Texas, USA
[7]Oregon Health & Science University, Portland, Oregon, USA

Contributors KW, SG, MAM, BH, OK designed the QA process. AA, OK, KW, DH, VP, AKP, JP, RP, RK, WH implemented and refined the QA process. KW, SK, BH were involved in the main manuscript writing. All authors reviewed/edited manuscript drafts.

Funding This research, and the CER Hub project (http://www.cerhub.org), is funded by grant R01HS019828 (Hazlehurst, PI) from the Agency for Health Care Research and Quality (AHRQ), US Department of Health and Human Services.

Competing interests None.

Ethics approval The IRBs of six participating health systems.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Institute of Medicine. *Initial national priorities for comparative effectiveness research*. Washington, DC: Institute of Medicine, 2009.
2 Federal Coordinating Council for Comparative Effectiveness Research. *Report to the President and Congress*. Washington, DC: Department of Health and Human Services, 2009. http://www.hhs.gov/recovery/programs/cer/cerannualrpt.pdf
3 AHRQ: what is comparative effectiveness research. http://www.effectivehealthcare.ahrq.gov/index.cfm/what-is-comparative-effectiveness-research1/ (accessed 15 Feb 2011).
4 Iglehart JK. Prioritizing comparative-effectiveness research—IOM recommendations. *N Engl J Med* 2009;361:325–8.
5 Clancy C, Collins FS. Patient-Centered Outcomes Research Institute: the intersection of science and health care. *Sci Transl Med* 2010;2:18.
6 Lauer MS, Collins FS. Using science to improve the nation's health system: NIH's commitment to comparative effectiveness research. *JAMA* 2010;303:2182–3.
7 Selker HP, Strom BL, Ford DE, *et al*. White paper on CTSA consortium role in facilitating comparative effectiveness research: September 23, 2009 CTSA consortium strategic goal committee on comparative effectiveness research. *Clin Transl Sci* 2010;3:29–37.
8 D'Avolio LW, Farwell WR, Fiore LD. Comparative effectiveness research and medical informatics. *Am J Med* 2010;123(12 Suppl 1):e32–7.
9 Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc* 2013;20:117–21.
10 Kokkonen EW, Davis SA, Lin HC, *et al*. Use of electronic medical records differs by specialty and office settings. *J Am Med Inform Assoc* 2013;20:e33–8.
11 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 2013;20:144–51.
12 Dolin RH, Alschuler L, Beebe C, *et al*. The HL7 clinical document architecture. *J Am Med Inform Assoc* 2001;8:552–69.
13 Dolin RH, Alschuler L, Boyer S, *et al*. HL7 clinical document architecture, release 2. *J Am Med Inform Assoc* 2006;13:30–9.
14 Frasier R, Allisany A, Hazlehurst BL. *The EmrAdapter Tool: A General-Purpose* Translator for Electronic Clinical Data. *Proceedings of the AMIA Annual Symposium* 2012:1740. http://proceedings.amia.org/amia-55142-a2012-1.124619/an-353-1.127895
15 Kahn MG, Raebel MA, Glanz JM, *et al*. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Med Care* 2012;50(Suppl):S21–9.
16 Weber GM, Murphy SN, McMurry AJ, *et al*. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *J Am Med Inform Assoc* 2009;16:624–30.
17 Rea S, Pathak J, Savova G, *et al*. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. *J Biomed Inform* 2012;45:763–71.
18 Forum, EDM, "Informatics Tools and Approaches To Facilitate the Use of Electronic Data for CER, PCOR, and QI: Resources Developed by the PROSPECT, DRN, and Enhanced Registry Projects" (2013). *Issue Briefs and Reports*. Paper 11. http://repository.academyhealth.org/edm_briefs/11
19 Sittig DF, Hazlehurst BL, Brown J, *et al*. A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogenous clinical data. *Med Care* 2012;50(Suppl):S49–59.
20 Califf RM. The Patient-Centered Outcomes Research Network: a national infrastructure for comparative effectiveness research. *N C Med J* 2014;75:204–10.
21 Holmes JH, Elliott TE, Brown JS, *et al*. Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature. *J Am Med Inform Assoc* 2014;21:730–6.