Drug Saf (2013) 36 (Suppl 1):S49–S58
DOI 10.1007/s40264-013-0098-7

ORIGINAL RESEARCH ARTICLE

Managing Data Quality for a Drug Safety Surveillance System

Abraham G. Hartzema · Christian G. Reich · Patrick B. Ryan · Paul E. Stang · David Madigan · Emily Welebob · J. Marc Overhage

Abstract

Objective The objective of this study is to present a data quality assurance program for disparate data sources loaded into a Common Data Model, highlight data quality issues identified and resolutions implemented.

Background The Observational Medical Outcomes Partnership is conducting methodological research to develop a system to monitor drug safety. Standard processes and tools are needed to ensure continuous data quality across a network of disparate databases, and to ensure that procedures used to extract-transform-load (ETL) processes maintain data integrity. Currently, there is no consensus or standard approach to evaluate the quality of the source data, or ETL procedures.

Methods We propose a framework for a comprehensive process to ensure data quality throughout the steps used to process and analyze the data. The approach used to manage data anomalies includes: (1) characterization of data sources; (2) detection of data anomalies; (3) determining the cause of data anomalies; and (4) remediation.

Findings Data anomalies included incomplete raw dataset: no race or year of birth recorded. Implausible data: year of birth exceeding current year, observation period end date precedes start date, suspicious data frequencies and proportions outside normal range. Examples of errors found in the ETL process were zip codes incorrectly loaded, drug quantities rounded, drug exposure length incorrectly calculated, and condition length incorrectly programmed.

Conclusions Complete and reliable observational data are difficult to obtain, data quality assurance processes need to be continuous as data is regularly updated; consequently, processes to assess data quality should be ongoing and transparent.

The OMOP research used data from Truven Health Analytics (formerly the Health Business of Thomson Reuters), and includes MarketScan® Research Databases, represented with MarketScan Lab Supplemental (MSLR, 1.2 m persons), MarketScan Medicare Supplemental Beneficiaries (MDCR, 4.6 m persons), MarketScan Multi-State Medicaid (MDCD, 10.8 m persons), MarketScan Commercial Claims and Encounters (CCAE, 46.5 m persons). Data also provided by Quintiles® Practice Research Database (formerly General Electric's Electronic Health Record, 11.2 m persons) database. GE is an electronic health record database while the other four databases contain administrative claims data.

A. G. Hartzema (✉)
College of Pharmacy, University of Florida,
Gainesville, FL, USA
e-mail: Hartzema@ufl.edu

C. G. Reich
AstraZeneca, Waltham, MA, USA

P. B. Ryan · P. E. Stang
Janssen Research and Development LLC, Titusville, NJ, USA

D. Madigan
Department of Statistics, Columbia University,
New York, NY, USA

J. M. Overhage
Siemens Health Services, Malvern, PA, USA

A. G. Hartzema · C. G. Reich · P. B. Ryan ·
P. E. Stang · D. Madigan · E. Welebob · J. M. Overhage
Observational Medical Outcomes Partnership, Foundation
for the National Institutes of Health, Bethesda, MD, USA

1 Background

Substantial research efforts are being directed toward the development of systems to perform sequential surveillance of observational databases for evidence of drug adverse events [1–3]. Currently, however, there is little consensus or guidance for researchers about quality assurance standards necessary for such a surveillance system. Some researchers have studied data validation from the perspective of source record verification to confirm that the electronic health record (EHR) is consistent with what happened to the patient [4–9]. Others have studied validation from the quality perspective within data transformation in software lifecycle processes [10–12]. However, there is paucity of literature that encompassed all aspects of data quality throughout the data management continuum [13–16].

Valuable insights can be obtained from the experiences of the regulated healthcare industry. In clinical product development, quality assurance is defined as “all those planned and systematic actions that are established to ensure that the trial is performed and the data are generated, documented (recorded), and reported in compliance with Good Clinical Practice (GCP) and the applicable regulatory requirement(s) [17].” “Quality control should be applied to each stage of data handling to ensure that all data are reliable and have been processed correctly [17].” Although a drug outcome surveillance system will use data that have already been collected, it is essential that quality assurance practices are implemented throughout the processes of data transformation and analysis, and that findings related to data quality are transparently reported. Such practices will provide greater confidence to stakeholders about the reliability of analysis results and the utility of the database in contributing to drug safety evidence.

The Observational Medical Outcomes Partnership (OMOP) is a public-private partnership conducting methodological research to inform the appropriate use of observational data for active surveillance of medical outcomes. Details about the rationale and design of OMOP have been published elsewhere [1]. The OMOP team conducts research across multiple, disparate, observational databases. Five raw data sets are housed centrally and form the core of the research laboratory. These include MarketScan® Lab Supplemental (MSLR, 1.2m persons), MarketScan® Medicare Supplemental Beneficiaries (MDCR, 4.6m persons), MarketScan® Multi-State Medicaid (MDCD, 10.8m persons), MarketScan® Commercial Claims and Encounters (CCAE, 46.5m persons), and the General Electric Centricity™ (GE, 11.2m persons) database. GE is an electronic health record (EHR) database; the other four databases contain administrative claims data. Additional data sets containing aggregate summary data are housed by partner organizations in a distributed network.

Because the credibility of research rests on the quality of data and validation of the analyses, the OMOP research team has devoted significant effort to developing processes and procedures to monitor and improve data quality, and to validate the various steps involved in preparing and analyzing the data. This paper provides an overview of those processes and procedures, including examples of their use.

2 Data Management Continuum

The upper panel of Fig. 1 depicts a data management continuum as it applies to an active system for monitoring drug safety from observational data. The continuum begins at the point of a healthcare encounter and proceeds through the capture of raw data, transformation and mapping of the data to a common data model (CDM), and the application of analytic methods. Each step in the continuum is associated with unique needs regarding data quality management.

2.1 Data Capture

The data capture process varies with healthcare provider and data source, and reflects activities from the point of the patient encounter to its ultimate representation in a database. Two primary sources of patient information data of potential value for an active surveillance system are EHRs and administrative claims. Electronic health records reflect data captured at the point of care to support clinical care. Administrative claims databases typically capture structured and coded data elements representing the healthcare encounter, which providers of healthcare services (e.g., physicians, pharmacies, hospitals, and laboratories) submit to receive reimbursement [18].

2.2 Transforming Data into a Common Data Model

Data quality issues can be compounded when analyses are performed on multiple databases that have different data capture procedures, linkage schemes, data structures, formats, and coding systems. To minimize these problems but maintain the ability to analyze multiple sources of data, OMOP transforms all data to a common data model [19, 20]. The OMOP CDM [21] is a person-centric relational model with domains inclusive of demographics, observation periods, drug exposure, condition occurrence, procedures, visits, and clinical observations. The CDM builds on previous data models developed to support observational research [22–24]. The process of extracting data from one source, transforming it if necessary and loading into a new

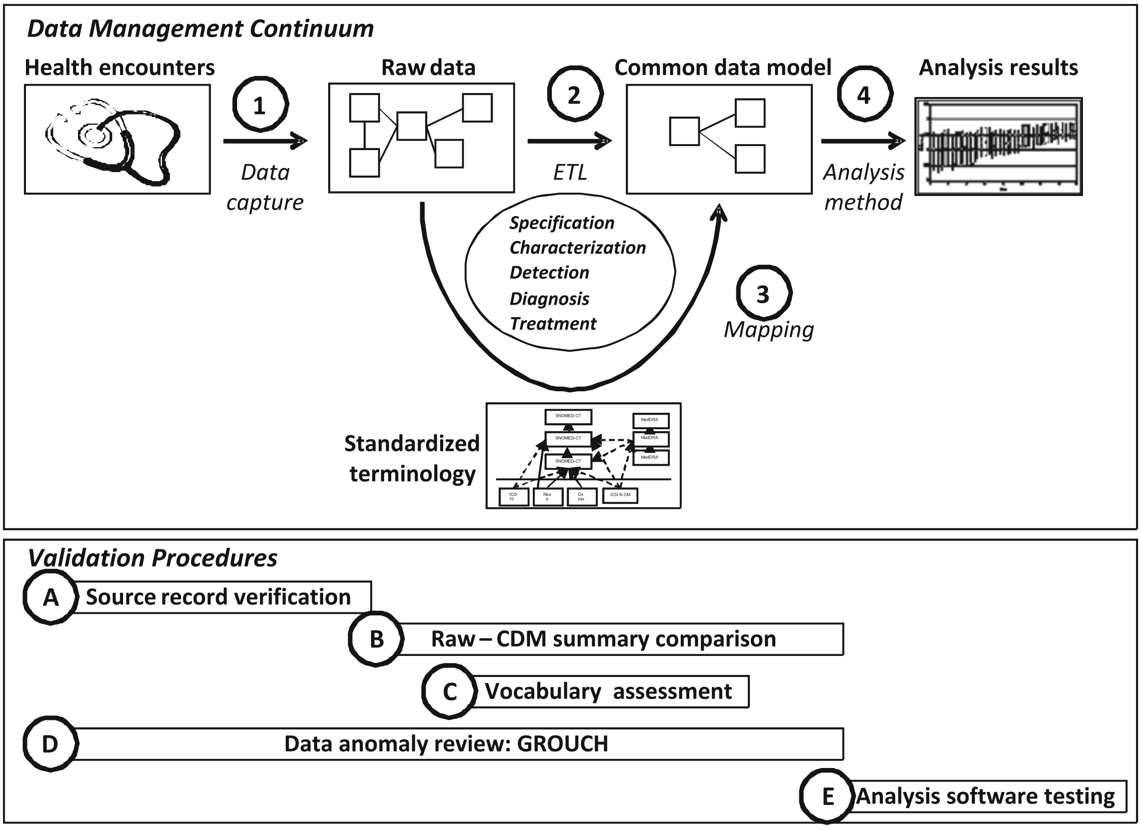


Fig. 1 Data management continuum and corresponding validation procedures developed by OMOP

database structure such as the CDM is referred to as an 'extract-transform-load' (ETL) [25] process. ETL processes can be complex, and significant operational issues can occur at any point. As with any software or application development, creating and executing ETL processes can lead to quality issues including software bugs, lack of documentation, portability, and insufficient or incorrect specifications. Quality issues specific to the ETL process can include incorrect transformation of data from one coding system to another and incorrect derivation of calculated values.

2.3 Mapping Across Terminologies

One major component of the ETL process used by OMOP to construct the CDM was the application of a standardized vocabulary of medical conditions. Like other aspects of the ETL process, mapping of medical conditions to a standardized vocabulary was performed to preserve consistent structure and semantic interoperability, and to facilitate analyses across multiple data sources. This mapping is becoming more important as observational data sources become more diverse, expanding beyond US health-plan–based administrative claims databases to EHRs, integrated delivery systems, and international sources. The raw databases in the OMOP network coded medical conditions using 5 different terminologies or coding schemes: International Classification of Diseases, Ninth Revision—Clinical Modification (ICD-9-CM), Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT®), Medical Dictionary for Regulatory Activities (MedDRA®), National Health Service (NHS) Read Codes, and Oxford Medical Information System (OXMIS). Furthermore, drugs can be coded in several different systems including the National Drug Code (NDC), Generic Product Identifier (GPI), Multum, and Multilex drug identifiers, and they can also be recorded as procedures using Current Procedural Terminology (CPT-4), Healthcare Common Procedure Coding System (HCPCS), and ICD-9 procedure codes. MedDRA is the international medical terminology developed under the auspices of the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH).

As part of the ETL process, the codes used in each raw database were identified and mapped to corresponding standard concepts adopted for the CDM. To perform this task, the OMOP team established source-to-concept mappings between many of the common terminologies used to code medical conditions, drugs, or other relevant data domains.

Ideally, each code (e.g., medical condition) in the raw data source can be mapped to exactly one concept in the target source that is exactly equivalent in meaning and granularity (breadth). In reality, two major problems arise: (1) some source codes may not map to any target concept (some terminologies have a broader coverage of concepts than others, leaving codes without an equivalent translation), and (2) the source code requires more than one target concept (e.g., the meaning of a concept in one vocabulary can be represented only using several concepts in another terminology).

Ambiguous mappings that are not one-to-one translations between concepts create problems during the ETL process. If there is no obvious equivalent target concept, the ETL process cannot assign one resulting in data loss. If more than one concept is required to reflect the meaning of the source, the ETL might have to write more than one record, creating a potential for bias or loss of information in the distribution of target concepts. In addition, mapping of concepts is based on the judgment of mapping specialists, who have to assess the meaning of source and target concepts to determine their equivalence. If this assessment is erroneous, a faulty mapping occurs, in which the CDM database does not accurately represent the raw data source.

2.4 Applying Analysis Methods

A primary advantage of using a CDM across a large-scale surveillance system is the ability to develop standardized analysis procedures that can be systematically applied across all participating data sources. Once an analytic procedure is implemented, tested, and validated in one CDM-compliant data source, the procedure should be ready for use in all other CDM-compliant sources. Analytic procedures are computer programs generated using software engineering practices, including software quality assurance. However, for some of the analytic procedures, it is nontrivial to test whether a certain algorithm produces the desired output because of the complexity of the algorithm, its computationally intensive nature, or the fact that the algorithm uses nondeterministic features (e.g., simulation using random number generation).

3 Establishing and Maintaining Data Quality

An overarching principle of data quality management is that the practices used to manipulate data and to manage data quality are documented in enough detail to allow replication. OMOP has specification documents for the ETL process, CDM, vocabulary and vocabulary mapping, and each analysis method. These specifications can be found on the OMOP website. The ETL specification document describes steps taken to convert the raw data source into the CDM and the processes employed to ensure the integrity of the data conversion.

A comprehensive program of data quality management requires validation of the various processes by which the data are recorded, transformed, and analyzed. In medical product manufacturing, FDA guidance defines validation as "establishing documented evidence which provides a high degree of assurance that a specific process will consistently produce a product meeting its pre-determined specifications and quality attributes [26]." In another guidance document for medical products involving software, the FDA considers software validation to be "confirmation by examination and provision of objective evidence that software specifications conform to user needs and intended uses, and that the particular requirements implemented through software can be consistently fulfilled [27]." In each document, emphasis is placed on fully documented specifications and evidence confirming that the specifications have been satisfied.

4 Validation Procedures

4.1 Source Record Verification

For the data management continuum shown in Fig. 1 and used by OMOP, several validation procedures are proposed. Source record verification is performed to determine how well events coded in a given database reflect the actual healthcare encounter. Verification studies attempt to evaluate the performance characteristics of an outcome definition, often in terms of positive predictive value (the proportion of coded events in the database that are confirmed to be true upon review of the source medical record) and sensitivity (the proportion of events confirmed from source records that were identified from the database definition). If the database consists of EHR data, it may be the primary clinical documentation, in which case source record verification has little meaning. Although 'up-coding' may have occurred (where the code for a procedure reimbursed at a higher rate is substituted for the code of diagnostic procedure reimbursed at a lower rate), third party payers are well aware of this possibility and may reimburse at a lower level.

The Clinical Practice Research Database (CPRD), a prominent research database based on clinical data contributed by UK General practitioners, has two levels of diagnostic data level assurance. Practices that contribute data to the CPRD need to meet certain quality standards in capturing and recording the data which undergo extensive validity checks by CPRD before they become part of the network. A second approach to data quality for analysis is the substantial body of work that has been undertaken in the course of research to establish the diagnostic validity of the data for pharmacoepidemiological analysis. Earlier research by Jick et al. and Kahn et al. [28, 29] testifies to this. More recently, a literature review by Herrett et al. [30] identifies 212 publications on the validations of 183 different diagnosis in the CPRD reporting a mean confirmed diagnosis rate of 89 %.

The performance characteristics of data recorded in a database for the patient's actual status vary by the outcome studied. Some diagnostic codes have demonstrated adequate performance [5–9, 31–33], while other codes have lower performance [25, 34–36]. However, source record verification can be context specific, and performance measures may vary between patient populations, across sites of care, across data sources, and over time as medical practice patterns evolve. Therefore, caution should be used when attempting to generalize from past studies. For example, many publications have described source record verification of diagnostic codes for acute liver injury [37], hemorrhage [38–40], acute myocardial infarction [32, 41], and acute renal failure [42, 43], but the performance of these codes varies considerably. It is important to reinforce that drug safety surveillance is a secondary use of clinical data. As such, the performance of a surveillance system will be influenced by the data capture practices used in the clinic (electronic data capture versus manual capture), and by the accuracy with which the recorded data represent the patient's physiological state, diagnosis made and procedures performed.

Source record verification requires access to person-level information, such as patient medical records. As a result, some data sources (e.g., large data aggregators) considered for use in surveillance cannot readily conduct primary source record verification as protection of personal health information makes this extremely difficult. Furthermore, primary source record verification is an extremely time-consuming and resource-intensive process that becomes increasingly challenging with the need for large sample sizes necessary in safety evaluations across multiple databases. Often, it is only the presumed positive cases based on a single outcome definition that are verified, while not as robust as a full source record verification that would also identify cases not found.

4.2 Validation of ETL: Raw-CDM Summary Comparison

There are several approaches to ensure the fidelity of the ETL process. One is to validate each transformation procedure through a set of test cases (the classic software validation approach). Another is to compare the resulting data to the raw data. Since these two data sets are not directly comparable (CDM data have undergone transformation), one solution is to compare summary statistics about the relevant domains and their relationships. These statistics could include the number of persons, demographic distributions, number of conditions, number of drugs, and length of data capture over time. Since the CDM data are represented in a standard format, the relevant summary statistics can be generated using a standardized tool. The same summary statistics need to be derived from the original datasets. If the two sets of summary statistics match, there is a high level of assurance that the data transformation was performed correctly. Any discordance identified through this process should be documented and reconciled. It is important to note that this validation process focuses only on the fidelity of the transformation and does not provide insight about data quality issues that originated in the raw data or occurred after transformation.

4.3 Validation of Code Mapping: Vocabulary Assessment

Building high-quality mapping tables between terminologies is resource intensive, and because terminologies constantly change, maintaining mapping tables is an ongoing process. Mapping between 2 concepts requires the understanding and judgment of an expert in domains such as medical diagnoses, drugs, and procedures. Therefore, this measure is somewhat subjective, as many clinical terms are ambiguous and medical practice is in constant flux.

A process for performing a rigorous review of mapping quality involves convening a jury of domain experts, such as medical coders, vocabulary developers, and medical informaticists. These experts would, for example, establish criteria for the percentage of source-to-concept mappings that are of sufficient quality to be used in a drug surveillance system. They would then analyze the quality of a representative sample of mapped relationships to determine if those criteria are satisfied. The quality of mapped relationships should be assessed by semantic equivalence (confirm that 2 codes are fully synonymous with each other); relationship precision (confirm that the differences in concept granularity are consistently addressed); relationship accuracy (confirm that the relationships are correct based on current clinical understanding); and interdependence of codes (confirm that relationships between concepts with a given vocabulary are adequately preserved through the mapping to a target vocabulary). Table 1 provides examples of issues in mapping.

4.4 Data Anomaly Review

The data anomaly review involves the detection of suspicious and implausible data elements that warrant further

Table 1 Vocabulary assessment—conditions

Potential for quality issues	Incorrect mapping
	Incomplete mapping
	Semantic mismatch
	Hierarchy mismatch
Quality check SNOMED vs. ICD-9-CM vs. MedDRA	Spot checking
	Comparing record numbers
	Comparing whether DOI-HOI associations can be reproduced in selected methods
Test OMOP HOI—original definition ICD-9-CM codes	Only HOI used that have no additional diagnostic/therapeutic procedure, lab test, radiology test or EKG definition

SNOMED Systematized Nomenclature Of Medicine, *ICD-9-CM* International Classification of Diseases, Ninth Revision, Clinical Modification, *MedDRA* Medical Dictionary for Regulatory Activities, *OMOP* Observational Medical Outcomes Partnership, *HOI* health outcomes of interest

investigation. Unlike the validation procedures discussed before, each of which correlates to a specific data progression step, a comprehensive data anomaly review can identify issues found within the total process of data capture, ETL, and mapping processes.

Specific data quality checks that could be integrated into a data anomaly review process have been proposed. Notably, Hennessy et al. evaluated 4 parameters for data integrity: (1) unexplained variation in the number of prescriptions per month; (2) proportion of Medicaid prescription records for which the NDC corresponded to a record in a commercially available NDC database; (3) the ratio of hospitalizations to population size; and (4) the frequency of miscoding of diagnoses [44]. In addition, the HMO Research Network, a federated data system of multiple health data sources, has conducted a series of examinations of the participating data sources within its Virtual Data Warehouse (VDW) and used several data characterizations to assess patterns across the data network. These include demographic summarizations, such as the proportion of patients in each participating source by age and gender; condition frequency, such as the prevalence of a specific disease over time and within subpopulations of interest (e.g., elderly women); data density over time, such as number of prescriptions dispensed per month; and value distribution, such as the range of days' supply for each prescription or the average blood pressure by age [45–48]. As described below, these targeted checks are incorporated into a comprehensive assessment tool in OMOP.

The approach used by OMOP to manage data anomalies encompasses four specific activities: (1) characterization of data sources; (2) detection of data anomalies; (3) determination of the cause of data anomalies; and (4) remediation. Each of these activities should be implemented to the satisfaction of both the data holder and the central coordinating center before any analytic work proceeds.

Characterization of each data source must be fully transparent to capture and communicate local knowledge of the source data to a central coordinating center, provide context for comparing sources across the network, and interpret results of drug safety studies. To address this requirement, the OMOP Research Team developed the Observational Source Characteristics Analysis Report (OSCAR) [49], which provides a comprehensive set of descriptive analyses of key CDM tables and fields. OSCAR produces descriptive statistics about the overall population in a given data source.

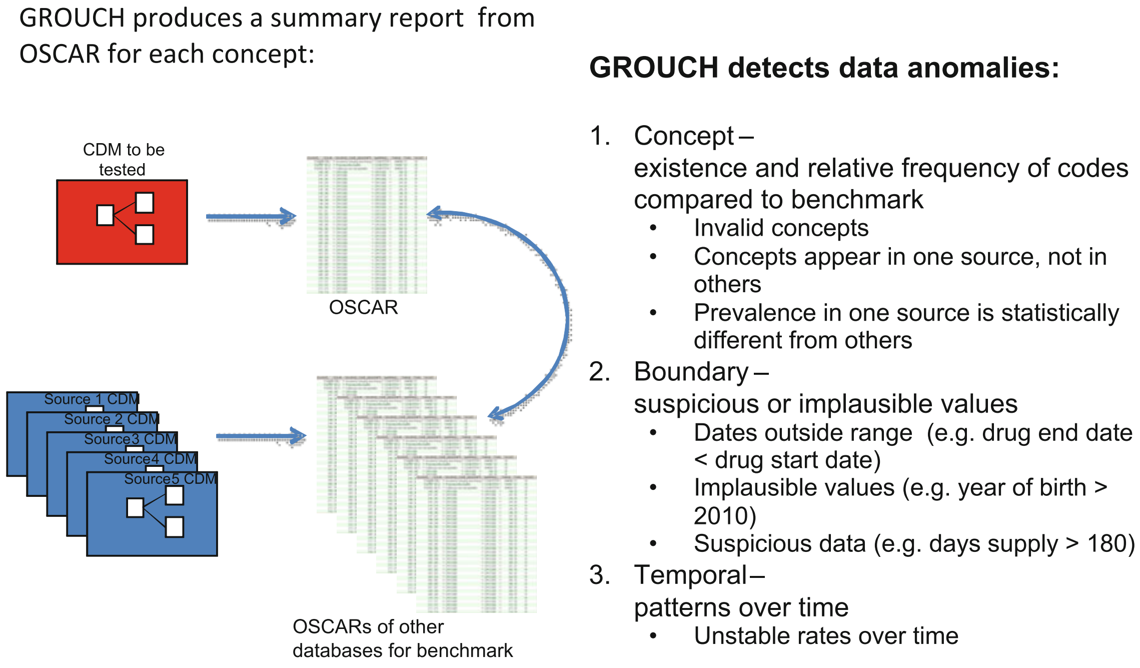
Another tool, complementing OSCAR, is the Natural History Analysis (NATHAN) [50], that characterizes specific subpopulations of interest. Data summaries produced by NATHAN can identify unusual patterns in defined subpopulations that may not have been observed when summarizing the overall database. OSCAR and NATHAN produce summaries of population-level distributions and large subpopulations, statistics that represent small subpopulations or populations with low cell counts are not included (e.g., those over 80 years of age); and dates are abstracted as month and year of service to minimize concerns about exchange of identifiable information.

Several different types of data anomalies may warrant review (Table 2). Data anomalies can occur at different levels of the data: within a data source at the individual record level (a drug exposure record is missing a start date); within a data source at the population level (the number of males far exceeds the number of females); within a data source over time (the number of laboratory tests per person in one year is much higher than in the previous year); and across data sources at the population level (the prevalence of a specific drug is significantly lower compared with other sources). Detection of data anomalies requires comparison to other data sources, and responsibility for detection is shared between data providers and the coordinating center.

In the OMOP environment, once data are robustly characterized, the summaries produced by OSCAR and NATHAN serve as a starting point for the detection of anomalies. In order to detect potential data quality issues, the OMOP team created a tool, Generalized Review of OSCAR Unified Checking (GROUCH) [51], which produces a summary report for each data source warning of implausible and suspicious data observed in the OSCAR summary. GROUCH identifies potential issues across all OMOP CDM tables, including potential concerns with all drug exposures and all conditions. See Fig. 2 for GROUCH operations. This allows for data quality review of specific

Table 2 Types of data anomalies

Types of data anomalies	Examples
Incomplete: missing values or entire fields	• A dataset does not capture ‘race’ • A dataset has no record of ‘year of birth’
Implausible: highly unlikely or practically impossible	• A person has a ‘year of birth’ that exceeds the current year • A drug is recorded with an invalid concept from the standardized vocabulary • An observation period has an end date that is before the start date
Suspicious: somewhat unlikely and justifying a closer look in case they reveal data issues	• The proportion of males is significantly higher than in other sources • The frequency of condition records in a given month is substantially lower than the prior or subsequent months • A source has many exposures to a given drug compared with other data sources • A source does not have patients with a given procedure that all other sources have records for



GROUCH produces a summary report from OSCAR for each concept:

GROUCH detects data anomalies:

1. Concept – existence and relative frequency of codes compared to benchmark
   - Invalid concepts
   - Concepts appear in one source, not in others
   - Prevalence in one source is statistically different from others
2. Boundary – suspicious or implausible values
   - Dates outside range (e.g. drug end date < drug start date)
   - Implausible values (e.g. year of birth > 2010)
   - Suspicious data (e.g. days supply > 180)
3. Temporal– patterns over time
   - Unstable rates over time

OSCAR = Observational Source Characteristics Analysis Report; **GROUCH** = Generalized Review of OSCAR Unified Checking; **CDM** = common data model

Fig. 2 Data anomaly review—GROUCH

drugs or conditions (including population-level prevalence of the health outcomes of interest and gender-stratified rates, such as males with pregnancy and females with prostate cancer). The full list of data quality checks performed in GROUCH and the description of the 35 different types of data checks are available on the OMOP website.

Once potential data anomalies are identified, their nature must be determined. Anomalies identified as such could represent legitimate data, true ‘extreme’ values (either at the patient- or population-level, e.g., large numbers of males in the Veterans Affairs data), or could be true data anomalies (e.g. a date of birth in the twenty-first century). The ETL of raw source data into the CDM could introduce systematic errors. Table 3 provides examples of data anomalies identified between the raw data and the CDM and the effect these may have on the analysis. Here the goal should be to detect ETL programming errors, identify systematic errors in the raw data capture, and understand and document other inconsistencies. The responsibility for determining the cause of data anomalies resulting from the raw data rests with the data provider, who has local domain

Table 3 Raw CDM: A summary of comparison results

Type of inconsistency	Effect of DOI or HOI
MarketScan Databases (MSLR, MDCD, MDCR, and CCAE)	
Zip codes 001–009 incorrectly loaded	No effect on HOI or DOI, no method taking geographical region into account
Procedure drug mapping incorrect, small (%) number of extra procedure drugs	No effect on DOI
Drug quantity rounded, errors in quantity for fractions (like ½ for ointments)	No effect on DOI, no method taking drug quantity into account
GE Database (electronic health record)	
Gender by age calculated based on 2008, not 2009	No effect on methods
Drug exposure length incorrectly programmed, resulting in values deviating in 3.72 % of cases	Small effect on DOI era length
Condition length incorrectly programmed, resulting in values deviating in a small number of cases	Possibly small effect on HOI era length

MSLR MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* General Electric Healthcare, *DOI* drugs of interest, *HOI* health outcomes of interest

expertise and access to patient-level data to investigate suspicious patterns. Developing remedial actions and resolving data anomalies that result from the ETL process are a shared responsibility. Currently, there are no automated tools to facilitate this process; therefore, the cause of data anomalies must be determined on a case-by-case basis, which can consume significant time and resources.

Lastly, once the cause of a data anomaly is determined, the appropriate course of action must be developed. There are 3 options for remediation of flagged data: (1) correct them; (2) delete them; or (3) leave them unchanged. Remediation should be context-specific, and all remedies should be documented so that the analysis process is fully transparent and reproducible. For example, for corrections, the original data values, and the amended data values should be recorded. For deletions, the values or records being removed should be recorded. All data modifications should be documented and implemented thru a standardized routine and should not be manually applied to the data. Decisions to leave data unchanged should be accompanied by a rationale, so that unnecessary duplication of the data quality management process can be avoided if future checks uncover the same issue, and so that investigators are fully aware of outstanding data issues. After remediation, the ETL specification document should be revised to reflect the data manipulations, the ETL code should be updated, and the ETL process should be rerun, resulting in an improved CDM dataset.

4.5 Analysis Software Testing

The issues surrounding testing and validation of analytic methods and the programming algorithms by which they are implemented are beyond the scope of this discussion. They have been included in Fig. 1 for completeness.

5 Ongoing Management of Data Quality

Once a plan for data quality management is in place, validation procedures need to be re-applied periodically because the source databases and the processes that populate them are not static, and components of the ETL procedure may change as medical practices change. Thus, practices used to maintain data quality should be viewed as continuously evolving in response to changing data sources and methods. However, not all validation steps need to be repeated every time a database is refreshed or analyzed. For example, source record verification should be reassessed only when the data capture process has changed. ETL validation should remain constant until the database structure changes, and vocabulary assessment should be repeated when vocabularies change or new codes are introduced in the raw data. The data anomaly review should be updated with each data refresh to determine whether the new data are consistent with the prior assessment.

6 Discussion

Data quality assurance procedures are necessary across the data management continuum to ensure a reliable active surveillance system. Establishing and maintaining data quality for a drug safety surveillance system requires a comprehensive approach that spans the data management continuum. Currently, there is no consensus or standard approach to evaluating the quality of observational healthcare databases for drug outcome research, or for maintaining data quality during data transformation and analysis. We have presented a framework for a multi-component approach designed to establish and maintain data quality for a drug safety surveillance system and introduced several tools to accomplish this task. OMOP tools and processes promote transparency and facilitate shared understanding across the data network and central coordinating center. ETL validation through raw-summary comparison enables complete specification and evaluation of decision rules. These procedures and tools allow one to identify, diagnose and treat issues of potential concern prior to drug safety analyses. This is a responsibility shared with all stakeholders.

7 Conclusions

We developed this quality assurance framework as a consequence of the process that we followed to prepare our data for analysis. An area to explore for further research is to provide a quantitative assessment of the framework across multiple databases to assess the relative merits of each type of data quality issue. The approach offered, along with the data quality assessment tools employed within the OMOP research program, can provide research teams (at a data source or at a central coordinating center) with the foundation for a transparent and robust process to establish and maintain data quality.

Acknowledgements The Observational Medical Outcomes Partnership is funded by the Foundation for the National Institutes of Health (FNIH) through generous contributions from the following: Abbott, Amgen Inc., AstraZeneca, Bayer Healthcare Pharmaceuticals, Inc., Biogen Idec, Bristol-Myers Squibb, Eli Lilly & Company, GlaxoSmithKline, Janssen Research and Development, Lundbeck, Inc., Merck & Co., Inc., Novartis Pharmaceuticals Corporation, Pfizer Inc, Pharmaceutical Research Manufacturers of America (PhRMA), Roche, Sanofi-aventis, Schering-Plough Corporation, and Takeda. Drs. Schuemie, Stang, and Ryan are employees of Janssen Research and Development. Dr. Schuemie received a fellowship from the Office of Medical Policy, Center for Drug Evaluation and Research, Food and Drug Administration. Dr. Reich is an employee of AstraZeneca. Drs. Schuemie, Madigan and Hartzema have received funding previously from FNIH. J. Marc Overhage and Emily Welebob have no conflicts of interest to declare.

The authors thank and acknowledge the contributions of the OMOP Distributed Research Partners in phases of this research, who were supported by a grant from FNIH. Assistance with writing and manuscript preparation was provided by Ken Scholz, PhD, with financial support from FNIH.

This article was published in a supplement sponsored by the Foundation for the National Institutes of Health (FNIH). The supplement was guest edited by Stephen J.W. Evans. It was peer reviewed by Olaf H. Klungel who received a small honorarium to cover out-of-pocket expenses. S.J.W.E has received travel funding from the FNIH to travel to the OMOP symposium and received a fee from FNIH for the review of a protocol for OMOP. O.H.K has received funding for the IMI-PROTECT project from the Innovative Medicines Initiative Joint Undertaking (http://www.imi.europa.eu) under Grant Agreement no 115004, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007–2013) and EFPIA companies' in kind contribution.

References

1. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Int Med. 2010;153(9):600–6.
2. Coloma PM, Trifirò G, Schuemie MJ, Gini R, Herings R, Hippisley-Cox J, et al. Electronic healthcare databases for active drug safety surveillance: is there enough leverage? Pharmacoepidemiol Drug Saf. 2012;21(6):611–21.
3. FDA. The Sentinel Initiative: A National Strategy for Monitoring Medical Product Safety. May 2008 [cited 2012 September 15]. http://www.fda.gov/Safety/FDAsSentinelInitiative/ucm089474.htm.
4. Donahue JG, Weiss ST, Goetsch MA, Livingston JM, Greineder DK, Platt R. Assessment of asthma using automated and full-text medical records. J Asthma. 1997;34(4):273–81.
5. Hennessy S, Leonard CE, Freeman CP, Deo R, Newcomb C, Kimmel SE, et al. Validation of diagnostic codes for outpatient-originating sudden cardiac death and ventricular arrhythmia in Medicaid and Medicare claims data. Pharmacoepidemiol Drug Saf. 2010;19(6):555–62.
6. Lee DS, Donovan L, Austin PC, Gong Y, Liu PP, Rouleau JL, et al. Comparison of coding of heart failure and comorbidities in administrative and clinical data for use in outcomes research. Med Care. 2005;43(2):182–8.
7. Miller DR, Oliveria SA, Berlowitz DR, Fincke BG, Stang P, Lillienfeld DE. Angioedema incidence in US veterans initiating angiotensin-converting enzyme inhibitors. Hypertension. 2008;51(6):1624–30.
8. So L, Evans D, Quan H. ICD-10 coding algorithms for defining comorbidities of acute myocardial infarction. BMC Health Serv Res. 2006;6:161.
9. Varas-Lorenzo C, Castellsague J, Stang MR, Tomas L, Aguado J, Perez-Gutthann S. Positive predictive value of ICD-9 codes 410 and 411 in the identification of cases of acute coronary syndromes in the Saskatchewan Hospital automated database. Pharmacoepidemiol Drug Saf. 2008;17(8):842–52.
10. Software Engineering—Product Quality—Part 1: Quality Model. Geneva, Switzerland: International Organization for Standardization; 2001.
11. Kan SH. Metrics and models in software quality engineering. 2nd ed. Boston: Addison-Wesley; 2002.
12. Glass RL. Building quality software. Upper Saddle River: Prentice-Hall; 1992.
13. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. 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.
14. Wang RY, Storey VC, Firth CP. A framework for analysis of data quality research. IEEE Trans Knowl Data Eng. 1995;7(4):623–40.
15. Pipino LL, Lee YW, Wang RY. Data quality assessment. Commun ACM. 2002;45(4):211–8.
16. Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. ACM Comput Surv. 2009;41(3):1–52.
17. Guidance for Industry E6 Good Clinical Practice: Consolidated Guidance. 1996 [cited Oct 5, 2010]. http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm073122.pdf.
18. Hennessy S. Use of health care databases in pharmacoepidemiology. Basic Clin Pharmacol Toxicol. 2006;98(3):311–3.

19. Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. Med Care. 2012;50(Suppl):S60–7.
20. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. JAMIA. 2012;19(1):54–60.
21. OMOP. Common Data Model (version 4); 2012 [cited 2012 November 12]. http://omop.org/CDMvocabV4.
22. Reisinger SJ, Ryan PB, O'Hara DJ, Powell GE, Painter JL, Pattishall EN, et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. J Am Med Inform Assoc. 2010;17(6):652–62.
23. Li L. A conditional sequential sampling procedure for drug safety surveillance. Stat Med. 2009;28(25):3124–38.
24. Informatics for Integrating Biology and the Bedside (i2b2) Software. [cited November 18, 2010]. https://www.i2b2.org.
25. Leonard CE, Haynes K, Localio AR, Hennessy S, Tjia J, Cohen A, et al. Diagnostic E-codes for commonly used, narrow therapeutic index medications poorly predict adverse drug events. J Clin Epidemiol. 2008;61(6):561–71.
26. Guideline on General Principles of Process Validation. 1987 [cited Cot 5, 2010]. http://www.fda.gov/Drugs/GuidanceCompliance RegulatoryInformation/Guidances/ucm124720.htm.
27. General Principles of Software Validation: Guidance for Industry and FDA Staff. 2002 [cited Oct 5, 2010]. http://www.fda.gov/RegulatoryInformation/Guidances/ucm126954.htm.
28. Jick SS, Kaye JA, Vasilakis-Scaramozza C, Garcia Rodriguez LA, Ruigomez A, Meier CR, et al. Validity of the general practice research database. Pharmacotherapy. 2003;23(5):686–9.
29. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. Br J Gen Pract. 2010;60(572):e128–36.
30. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. Br J Clin Pharmacol. 2010;69(1):4–14.
31. Garcia Rodriguez LA, Perez Gutthann S. Use of the UK General Practice Research Database for pharmacoepidemiology. Br J Clin Pharmacol. 1998;45(5):419–25.
32. Pladevall M, Goff DC, Nichaman MZ, Chan F, Ramsey D, Ortiz C, et al. An assessment of the validity of ICD Code 410 to identify hospital admissions for myocardial infarction: the Corpus Christi Heart Project. Int J Epidemiol. 1996;25(5):948–52.
33. Wahl PM, Rodgers K, Schneeweiss S, Gage BF, Butler J, Wilmer C, et al. Validation of claims-based diagnostic and procedure codes for cardiovascular and gastrointestinal serious adverse events in a commercially-insured population. Pharmacoepidemiol Drug Saf. 2010;19(6):596–603.
34. Harrold LR, Saag KG, Yood RA, Mikuls TR, Andrade SE, Fouayzi H, et al. Validity of gout diagnoses in administrative data. Arthritis Rheum. 2007;57(1):103–8.
35. Lewis JD, Schinnar R, Bilker WB, Wang X, Strom BL. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. Pharmacoepidemiol Drug Saf. 2007;16(4):393–401.
36. Strom BL. Data validity issues in using claims data. Pharmacoepidemiol Drug Saf. 2001;10(5):389–92.
37. Jinjuvadia K, Kwan W, Fontana RJ. Searching for a needle in a haystack: use of ICD-9-CM codes in drug-induced liver injury. Am J Gastroenterol. 2007;102(11):2437–43.
38. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. Med Care. 2005;43(5):480–5.
39. Lain SJ, Roberts CL, Hadfield RM, Bell JC, Morris JM. How accurate is the reporting of obstetric haemorrhage in hospital discharge data? A validation study. Aust N Z J Obstet Gynaecol. 2008;48(5):481–4.
40. Lopushinsky SR, Covarrubia KA, Rabeneck L, Austin PC, Urbach DR. Accuracy of administrative health data for the diagnosis of upper gastrointestinal diseases. Surg Endosc. 2007;21(10):1733–7.
41. Austin PC, Daly PA, Tu JV. A multicenter study of the coding accuracy of hospital discharge administrative data for patients admitted to cardiac care units in Ontario. Am Heart J. 2002;144(2):290–6.
42. Liangos O, Wald R, O'Bell JW, Price L, Pereira BJ, Jaber BL. Epidemiology and outcomes of acute renal failure in hospitalized patients: a national survey. Clin J Am Soc Nephrol. 2006;1(1):43–51.
43. Waikar SS, Wald R, Chertow GM, Curhan GC, Winkelmayer WC, Liangos O, et al. Validity of international classification of diseases, ninth revision, clinical modification codes for acute renal failure. J Am Soc Nephrol. 2006;17(6):1688–94.
44. Hennessy S, Leonard CE, Palumbo CM, Newcomb C, Bilker WB. Quality of Medicaid and Medicare data obtained through Centers for Medicare and Medicaid Services (CMS). Med Care. 2007;45(12):1216–20.
45. Butani AL, Sherwood N, Adams K, et al. The VDW vital signs file: strengths, issues and recommendations for the future. Poster presented at the 15th Annual HMO Research Network Conference, Danville; 2009.
46. Hornbrook MC, Hitz P, Pardee R, et al. The VDW demographic and enrollment files: strengths, issues, and recommendations for the Future. Presented at the 15th annual HMO research network conference, Danville; 2009.
47. Moore KM, Cheetham C, Dublin S, et al. VDW pharmacy file: strengths, weaknesses and recommendations. Poster presented at the 15th annual HMO research network conference, Danville; 2009.
48. Saylor G, Ellis JL, Raebel MA, et al. Formalization of the laboratory result content area of the VDW. Poster presented at the 14th Annual HMO research network conference, Minneapolis; 2008.
49. OMOP. Observational Source Characteristics Analysis Report (OSCAR) Design Specification and Feasibility Assessment. 2010 [cited 2012 June 18]. http://omop.org/OSCAR.
50. OMOP. NATHAN—Utility of Natural History Information; 2010 [cited 2012 June 18]. http://omop.org/NATHAN.
51. OMOP Implementation 2011 [cited 2012 December 12]. http://omop.org/OMOPimplementation.