Published in final edited form as:
Med Care. 2013 August ; 51(8 0 3): S22–S29. doi:10.1097/MLR.0b013e31829b1e2c.

Data quality assessment for comparative effectiveness research in distributed data networks

Jeffrey Brown, PhD,
Department of Population Medicine Harvard Medical School and Harvard Pilgrim Health Care Institute 133 Brookline Ave 6th Floor, Boston, MA 02215, USA

Michael Kahn, MD, PhD, and
Section of Pediatric Epidemiology, Department of Pediatrics Colorado Clinical and Translational Sciences Institute and CCTSI Biomedical Informatics University of Colorado Denver Denver, CO USA

Sengwee Toh, ScD
Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute Boston, MA, USA

Abstract

Background—Electronic health information routinely collected during healthcare delivery and reimbursement can help address the need for evidence about the real-world effectiveness, safety, and quality of medical care. Often, distributed networks that combine information from multiple sources are needed to generate this real-world evidence.

Objective—We provide a set of field-tested best practices and a set of recommendations for data quality checking for comparative effectiveness research (CER) in distributed data networks.

Methods—Explore the **requirements for** data quality checking and describe data quality approaches undertaken by several existing multi-site networks.

Results—There are no established standards regarding how to evaluate the quality of electronic health data for CER within distributed networks. Data checks of increasing complexity are often employed, ranging from consistency with syntactic rules to evaluation of semantics and consistency within and across sites. Temporal trends within and across sites are widely used, as are checks of each data refresh or update. Rates of specific events and exposures by age group, sex, and month are also common.

Discussion—Secondary use of electronic health data for CER holds promise but is complex, especially in distributed data networks that incorporate periodic data refreshes. The viability of a learning health system is dependent on a robust understanding of the quality, validity, and optimal

Corresponding author: Jeffrey Brown, PhD Department of Population Medicine Harvard Medical School and Harvard Pilgrim Health Care Institute 133 Brookline Ave 6th Floor, Boston, MA 02215, USA jeff_brown@hphc.org Tel: +1 617 509 9986 Fax: +1 617 859 8112.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

secondary uses of routinely collected electronic health data within distributed health data networks. Robust data quality checking can strengthen confidence in findings based on distributed data network.

Keywords

Comparative effectiveness research; distributed research network; data quality

INTRODUCTION

Electronic health records and other information routinely collected during healthcare delivery and reimbursement can help address the critical need for evidence about the real-world effectiveness, safety, and quality of medical care.[1-7] Consequently, state and federal agencies, private payers, and others are seeking to **use** these data to generate timely and actionable evidence and realize the benefits of a learning health system.[8-10]

Even the largest **individual** data resources are insufficient to investigate interventions with limited use, evaluate rare conditions, identify rare outcomes, and to enable timely decision-making.[1] Combining data from diverse clinical settings also strengthens the generalizability of findings. Therefore, combining multiple data sources is often necessary. Using multiple data resources for comparative effectiveness research (CER) requires efficient mechanisms to access the data while respecting the regulatory, legal, proprietary, and privacy implications of use. In principle, a distributed network or a centralized database could support multi-site CER. The U.S. Food and Drug Administration (FDA)[4, 11, 12] and the Office of the National Coordinator for Health Information Technology,[13] among others, support using distributed data networks that allow data partners to maintain physical control over their data, while permitting authorized users to query the data.[11, 12, 14-19]

In distributed networks data partners maintain control of their data and its uses. Distributed networks typically rely on a common data model that enables queries to be executed identically by all data partners.[20-24] Distributed networks address many of the security, proprietary, legal, and privacy concerns common to multi-site research.[19, 25, 26] An added benefit is that the data are held by those who are best able to consult on proper use and interpretation. Both single-use[27, 28] and multi-use networks are possible. Multi-use networks such as the HMO Research Network (HMORN),[29, 30] the CDC-sponsored Vaccine Safety Datalink,[31-33] and the FDA-sponsored Mini-Sentinel[12] build and maintain data and administrative infrastructure to support multiple studies. **M**ulti-purpose networks often refresh their data as new information becomes available.

Unfortunately, idiosyncratic data quality issues in multi-site environments can arise from variation in data capture, differences and changes in medical coding terminologies and coding practices, local business-rules regarding data adjudication, clinical workflows, and delivery system differences such as formularies, provider contracts, and payment contracts.[34, 35] Without a strategy to identify and address cross-site and temporal data quality issues, the evidence generated by CER using such networks is subject to validity concerns. Therefore, a framework for data quality checking is needed to help researchers identify and resolve these issues. In this paper we **1)** review the **pragmatic data quality** framework developed by Kahn *et al*. (2012),[36] **2)** assess several existing data **quality** checking approaches, **3)** provide several work**ed** examples of data quality checking, and **4)** recommend multi-site data checking approaches.

KAHN'S DATA VALIDITY CONCEPTUAL FRAMEWORK

Kahn *et al*. (2012) proposed a **conceptual** framework for assessing the quality of electronic health data that includes five key data concepts, defined below (adapted from Kahn 2012):

- Attribute domain constraints: focus on data value anomalies for individual variables, including distributions, units, and missingness. These checks identify values and distributions inconsistent with expectations (*e.g*., a high proportion of individuals over 120 years old).
- Relational integrity rules: compare elements from one data table to related elements in another data table (*e.g*., every person identifier in the pharmacy table must have a record in the demographic table, **but not necessarily in the enrollment table**).
- Historical data rules: temporal relationships and trend visualizations to identify data gaps, unusual patterns, and dependencies across multiple data values and variables (*e.g*., utilization trends can identify shifts in data capture).
- State-dependent objects rules: extends temporal data assessment to include logical consistency (*e.g*., a series of prenatal ultrasounds should precede a pregnancy outcome).
- Attribute dependency rules: examine conditional dependencies based on knowledge of a clinical scenario (*e.g*., women should not have a diagnosis of prostate cancer).

These data checking domains are commonly found in multi-site data quality checking approaches. However, , data quality checking in distributed data networks has not been well described in the literature, and is generally more complex due to privacy and proprietary concerns and variation in data capture.

DATA QUALITY CHECKING APPROACHES USED BY SELECTED DISTRIBUTED DATA NETWORKS

There are no guidelines to determine whether electronic health data are “valid” or of high “quality” nor any consensus on how to define “valid” or “quality”. Hall *et al* (2012)[37] set out guidelines for good pharmacoepidemiologic practice for database selection and use, and included several recommendations for single-site and multi-site studies.[37] They provide suggestions for data checking, including assessment of the completeness and accuracy of key study variables, check of external validity (*e.g*., is the rate of some metric consistent with external estimates), logic and plausibility checks (*e.g*., assessment of age ranges, missingness, and clinical plausibility), and trending assessments.[37]

Several networks **have published** their data checking procedure, **with varying levels of detail**. The HMORN conducts annual data quality checks of their distributed database – the HMORN Virtual Data Warehouse (VDW). They assess compliance with the common data model, evaluate summary statistics for continuous measures, proportions for categorical variables, missingness frequencies, and trends.[38] Output is reviewed by data area workgroups that consist of HMORN analysts and investigators. HMORN sites must get a "passing" grade before their data is considered acceptable for inclusion in the VDW. The specific metrics needed to earn a "passing" grade are determined by the workgroup.

The Observational Medical Outcomes Partnership (OMOP) has several tools to assess data quality.[39, 40] The **OMOP** approach is similar to the HMORN with respect to review of categorical and continuous variables across an entire database and without focus on a specific study topic. OMOP has a standardized approach to representing the results,[39] and flagging potential problems.[40]

The FDA's Mini-Sentinel project describes four levels of data checking.[41, 42] Level 1 checks review consistency with the Mini-Sentinel Common Data Model data dictionary, focusing on variable names, lengths, formats, and values. For example, acceptable values for the variable SEX are 'M', 'F', 'A', and 'U' so any other value will generate an error flag. Level 2 checks focus on completeness and integrity between variables within a table, or variables between tables. For example, every person identifier in the pharmacy dispensing table must have a record in the demographic file. Level 1 and Level 2 checks apply attribute domain constraints and relational integrity rules to generate binary flags. Level 3 checks – similar to historical data rules - focus on patterns, trends, and cross-variable relationships, but not necessarily relationships that can be characterized as true or false. For example, a graph of monthly pharmacy dispensings or a table of the annual distribution of inpatient and outpatient encounters can identify unusual patterns or cross-site variability. Potential issues are flagged via visual inspection of trends and review of expected within and cross-site consistency. Finally, Level 4 - similar to attribute dependency rules - data checks focus on specific clinical scenarios for which there is a reasonable *a priori* expectation for the findings. Examples include trends in dispensing rates of specific medications by age and sex, vaccinations by age and sex, and the frequency of specific procedures by age and sex.

PRACTICAL SUGESTIONS FOR ASSESSING DATA VALIDITY IN A DISTRIBUTED DATA NETWORK

The approaches described above are consistent in focus and approach. Based on the work reviewed above and our experience operating several multi-site networks, we present a series of practical approaches to help identify data issues and ensure data validity in multi-site multi-purpose distributed networks. Our recommendations have been extensively "field tested" across thousands of data extractions and hundreds of studies. While we continue to find unusual data quality issues, the proposed methods comprise the core data quality checking procedures used in multiple active national networks. We also present considerations uniquely related to data security, privacy, and proprietary issues. Data quality checking for specific studies can use many of the same approaches, but additional study-specific checks are recommended.

Review Adherence to the Common Data Model

Many multi-site distributed networks employ a common data model to facilitate cross-site analysis. Once the common data model is determined and sites conduct an extract-transform-load (ETL) procedure to populate the model, the ETL process must be checked against the data model. These checks **compare** the extracted data against the data model data dictionary to ensure syntactic consistency and adherence to the data model. There are at least three basic consistency checks: the first evaluates simple syntactic consistency, the second focuses on the table structure, and the third targets expected relationships between tables.

Syntactic correctness refers to whether the transformed variable names, values, length, and format meet the data model specifications. For example, a patient identifier variable could have four checks -- it must be non-missing, be a character format, be left-justified, and be called PATID. The variable ‘sex’ might have five checks – must be a character format, have a length of 1, be non-missing, have values of “F”, “M”, or “U”, and be called SEX.

Consistency with coding terminologies should be evaluated. For example, Healthcare Common Procedure Coding System (HCPCS) codes should be 5 digits, National Drug Codes (NDCs) should not contain any letters, and International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes should always begin with a number, an “E”, or a “V”. For variables that use standardized coding terminologies, all values contained in a data set can be checked against a reference library of all valid values for that terminology. Finally, data models may specify that certain variables include or exclude decimals, dashes, and modifiers and data check should verify that those rules were followed.

Data models are often organized into specific tables that can be linked by person or encounter identifiers. Each table typically has a definition for each row. **For example, a** demographic table **can be** defined as 1 row per unique person and include information such as date of birth, sex, and race – values that typically do not change over time. However, if the demographic table includes a time-varying characteristic such as marital status or zip code, the table could include multiple rows per person. Regardless, the **row** definition must be verified against expectations.

Finally, expected cross-table relationships should be checked for consistency with the data model. These checks rely on identifying relationships that should always be true based on the data model and verifying that the data meet those expectations. For example, in some networks every patient in the pharmacy file should have a corresponding row in the demographic file, and every patient in the utilization file should be in the enrollment file. Violation of these basic data model rules should generate errors for review. **Table 1** illustrates the output of a cross-site, cross-table data check **in the Mini-Sentinel program**. Sites 5, 14, and 17 have low match rates between the enrollment and demographic files. This **finding** is a result of those sites provid**ing** care for non-members. This can occur in integrated delivery systems that operate medical facilities, but should not be observed for insurance companies that do not operate medical facilities. This example highlights that interpretation of differences may require information about data sources that exists outside of the data set and that **observed** differences may not be indicative of poor data quality.

Review Each Data Domain

Data models are commonly divided into domains such as enrollment, demographics, medication dispensing, prescribing, medical utilization, laboratory results, and vital signs. Examples include the HMORN VDW, OMOP, the Mini-Sentinel Common Data Model, the Vaccine Safety Datalink, and Electronic medical record Support for Public health (ESP).[43] Models that combine information across domains into a single table (*e.g*., star schemas)[44] can be stratified by domain to help simplify comparisons.

Independent of the data model, we suggest **extending** data domain checks by evaluating 1) frequency and proportions for categorical variables; 2) distributions **and extreme values** for continuous variables; 3) missingness; 4) “out-of-range” values (as defined by the data curator); 5) expected relationships between variables within the domain; 6) normalized rates (e.g., per person, per member, or per person per month), and 7) temporal trends (weekly, monthly, quarterly, or annual). Trends **in counts or proportions** should be presented overall and by site and can be stratified by age group, sex, or other relevant characteristic. The **seven** basic checks listed above provide a view of the entire database over time and should be implemented across all data domains. These checks should be modified to investigate a specific cohort of patients or topic. We present a series of examples by domain for guidance.

Membership/Enrollment—

Key temporal components such as the start and stop dates of enrollment stratified by the type of coverage and demographic characteristics should be assessed. Suggested checks include:

- Enrollment start dates should precede enrollment end dates
- Maximum enrollment length should not exceed the maximum observation length
- Distribution of enrollment length (mean, median, inter-quartile range, 1st, 5th, 95th, and 99th percentiles)
- Enrollment periods per member
- Proportion of enrollment periods by benefit type (drug coverage, medical coverage, etc)
- Monthly membership, overall and by coverage type

Figure 1 illustrates monthly enrollment changes at two sites in 2010 versus 2011, showing a large drop in January 2010 for 1 site and large drop for both sites in January 2011. These changes can be verified with the sites to ensure that the patterns are expected, and the pattern can help inform researchers using multi-year longitudinal cohorts.

Demographics—

Checks of categorical and continuous variables, including missingness and “outof-range” values should be assessed. Sites can review age and sex distributions to verify consistency with expectations.

Medication Use—Medication information typically includes the medication dispensed/prescribed, the days of supply, number of units (*e.g.*, pills, canisters), and the date dispensed/prescribed. If **standard** coding terminologies such as RXnorm or NDCs are used, the recorded values should match **valid values** the terminology dictionary. Suggested medication checks include:

- Frequency of records by days supplied and amount dispensed/prescribed, including identification of unexpected values such as <0, 0, 0-<1, and > 100.
- Records with missing values
- Records that do not match the terminology dictionary
- Records per month
- Users (*i.e.*, individuals with at least one dispensing/prescription) per month
- Dispensings/prescriptions per user and per health plan member per month
- Days supplied per dispensing

Several other dispensing data checks are illustrated in the Supplemental Digital Content (SDC; see SDC Table 1 and SDC Figures 1-3).

Medical Utilization—Medical utilization data often include care setting (*e.g.*, inpatient, ambulatory, emergency department), diagnoses recorded, procedures performed, facility identifier, provider identifier, and service dates. Suggested checks include:

- Encounters per patient (defined as a person with at least one encounter)
- Encounters per member (includ**es** individuals without any encounter)
- Encounters by care setting
- Diagnoses and procedures per encounter
- Procedures by procedure code type (e.g., HCPCS, ICD-9)
- Diagnoses by diagnosis code type (e.g., ICD-9-CM, SNOMED-CT)
- Encounters per patient and per member per month

Figure 2 shows encounters by month for 2 consecutive ETLs for 1 site. The saw-tooth pattern is unexpected; it could be the result of a data capture issue or a change in the ETL logic that was introduced in the most recent data extraction. Discussion with the data partner identified a data extraction error that was corrected. Temporal comparisons of ETLs with**in** a site can help identify missing data (SDC Figure 4 illustrates 2 sequential data extracts, the second of which had a month of missing data).

Clinical Data—Data such as laboratory results and vital signs can be tested like the other domains, with relevant stratifications. Investigation of laboratory test results are complex because results can be idiosyncratic and data capture can vary substantially across sites. Basic clinical expectations should be reviewed for validity such as men weighing more than women and blood pressure increasing by age. Examples of temporal trends include the number of tests per month, tests per member per month, and average values per year. **Figure 3** provides two laboratory data check examples showing individuals with one or more laboratory test results per 1,000 members. The charts show two different patterns, substantial cross-site variation, and within site trend changes. This figure illustrates differences in data capture across sites that would generate additional investigation by researchers using the data for specific purposes, and may influence development of research protocols or selection of sites.

Assess Expected Clinical Relationships

Expected clinical relationships should be assessed. For example, the rate of hip fractures among 60-65 year-old women, and ankle fractures among 18-22 year-old males are metrics that can be compared across sites with the expectation of cross-site consistency. The intent is to identify a clinical condition or event with an expected pattern or expected consistency across sites, and test the hypothesis of similarity. **We also** recommend evaluating relationships that should never occur such as the number of pregnant men and women with prostate cancer.[45-49] Because most electronic data contain these types of errors, a data set with no “never” relationships is itself suspect and it is up to the researcher to investigate **with the data partner** the process used to implement the cleaning.

Additional Considerations

Data checking in a distributed multi-site environment requires the transfer of information from the sites to a coordinating center that acts as data curator. Transfer of information to any external entity raises security, privacy, and proprietary concerns.

Additionally, data partners interpret privacy regulations differently, but most prohibit transfer of patient-level information or tables with low cell counts without approval of their Institutional Review Board and/or privacy officer. Data checking output can avoid patient privacy issues by transferring only stratified count information. Proprietary concerns can introduce **additional** barriers. Examples include dispensings by NDC or generic name (this can allow identification of preferred pharmacy vendors or the formulary status of products), **per-member per-month counts,** and counts of members by diagnosis code, procedure code, or other clinical codes as this can expose implied quality measures. Finally, some partners may object to the sheer volume of information often required for comprehensive data checking, requiring a balancing of partner needs with those of the multi-site network data curator.

Study-specific data checks also should be performed. Study-specific checks should **investigate** the exposure, outcome, and covariates of interest **in detail**. Assessment of metrics such as days supplied per dispensing, dispensings per user, and total days of exposure per user can help identify cross-site variability. These checks can identify data quality issues not observed by the network data checking process.

DISCUSSION

Assessing data validity in multi-site distributed data networks is complex. There are no clear rules for what constitutes a “valid” data resource or even which metrics to use to assess data validity. Data validity **approaches** often rest on the experience of the investigator **and work of data curators**. We **recommend a range** data checks for CER studies to help identify potential data issues and provide assurance to CER stakeholders that routine and appropriate data quality checking was conducted.

The data checks recommended are generic by design and do not address all possible data domains or scenarios. The domains included above represent the most common data domains for CER, but since the conceptual frame is the same for all domains, the data checking steps listed here can be applied to other domains such as patient reported outcomes, registry data, and other elements in electronic health records. In addition, multi-site networks (and specific studies) often incorporate data updates or “refreshes” that add newer data to the network. In such instances, all data checking must be repeated to ensure that the updates did not introduce new data quality problems, as illustrated in our examples. Our experience is that data updates using a previously validated **ETL** offers no protection against new data quality problems resulting from unknown or underappreciated changes in the local data resources. In fact, it is not uncommon for data partners in distributed networks to learn of local data changes from the network data checking process.

Data quality checking is typically conducted “behind the scenes”, with results excluded from public reports. We suggest changing this paradigm by including information about the data quality approach and results as part of the standard CER reporting template. At a minimum, investigators should report key data checking metrics for the primary exposure, outcome, and covariate measures. This would help ensure that investigators implement extensive data checks and provide stakeholders with enough information to assess the likelihood that the data sources are appropriate for the study design.

Well-conducted CER requires the appropriate combination of data resources, study design, and statistical analysis. Poor choice of any of the three can result in invalid results. Well-designed data checking activities can identify data issues and determine the suitability of the data source for the study design and statistical method. To fully realize the potential of electronic health data for CER, our understanding of data issues, including methods for identifying them, must advance in parallel with advances in study design and statistical methods. Only when all three legs of the CER stool are solid will CER take its rightful place in evidence generation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: Drs. Toh and Brown were supported by Agency for Healthcare Research and Quality grant R01HS019912 (Scalable PArtnering Network for Comparative Effectiveness Research). Dr. Kahn was supported by Agency for Healthcare Research and Quality grant R01HS019908 (Scalable Architecture for Federated Translational Inquires Network), and NIH/NCATS Colorado CTSI Grant Number TR000154 (Colorado Clinical and Translational Sciences Institute).

REFERENCES

1. Federal Coordinating Council for Comparative Effectiveness Research. Report to the President and Congress. 2009. Available at: http://www.hhs.gov/recovery/programs/cer/cerannualrpt.pdf
2. McClellan M. Drug safety reform at the FDA--pendulum swing or systematic improvement? N Engl J Med. 2007; 356:1700–1702. [PubMed: 17435081]
3. Alina Baciu, KS.; Sheila, P.; Burke, editors. The Future of Drug Safety: Promoting and Protecting the Health of the Public. Institute of Medicine of the National Academies; Washington, D.C.: 2006.
4. Behrman RE, Benner JS, Brown JS, et al. Developing the Sentinel System--a national resource for evidence development. N Engl J Med. 2011; 364:498–499. [PubMed: 21226658]
5. Platt R, Wilson M, Chan KA, et al. The New Sentinel Network - Improving the Evidence of Medical-Product Safety. N Engl J Med. 2009; 361:645–647. [PubMed: 19635947]
6. Strom, B.; Kimmel, SE.; Hennessy, S. Pharmacoepidemiology. Wiley; West Sussex, England: 2012.
7. Buntin MB, Jain SH, Blumenthal D. Health information technology: laying the infrastructure for national health reform. Health Aff (Millwood). 2010; 29:1214–1219. [PubMed: 20530358]
8. Olsen, L. IOM Roundtable on Evidence-Based Medicine. The Learning Healthcare System: Workshop Summary. Washington, DC: 2007. Available at
9. Institute of Medicine. Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary. The National Academies Press; 2011.
10. Young, PL.; Olsen, L.; McGinnis, JM. Value in Health Care: Accounting for Cost, Quality, Safety, Outcomes, and Innovation: Workshop Summary. The National Academies Press; 2010.
11. Brown, JS.; Lane, K.; Moore, K., et al. Defining and Evaluating Possible Database Models to Implement the FDA Sentinel Initiative. U.S. Food and Drug Administration; FDA-2009-N-0192-0005.2009. Available at: http://www.regulations.gov/search/Regs/home.html#documentDetail?R=090000648098c282
12. Platt R, Carnahan RM, Brown JS, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. Pharmacoepidemiol Drug Saf. 2012; 21(Suppl 1):1–8. [PubMed: 22262586]
13. [June, 2012] Standards & Interoperability Framework: Query Health Initiative. 2012. Available at: http://wiki.siframework.org/Query+Health.
14. Brown, JS.; Holmes, J.; Maro, J., et al. Report 1: Design Specifications for Network Prototype and Research Cooperative.;Developing a Distributed Research Network and Cooperative to Conduct Population-based Studies and Safety Surveillance. Agency for Healthcare Research and Quality; Rockville, MD: HHSA29020050033I.January 30 2009. Available at: Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm
15. Maro JC, Platt R, Holmes JH, et al. Design of a National Distributed Health Data Network. Ann Intern Med. 2009; 151(5):341–4. [Epub ahead of print]. [PubMed: 19638403]
16. Moore KM, Duddy A, Braun MM, et al. Potential population-based electronic data sources for rapid pandemic influenza vaccine adverse event detection: a survey of health plans. Pharmacoepidemiol Drug Saf. 2008; 17:1137–1141. [PubMed: 18763248]
17. Toh S, Platt R, Steiner JF, et al. Comparative-effectiveness research in distributed health data networks. Clin Pharmacol Ther. 2011; 90:883–887. [PubMed: 22030567]
18. Yih WK, Kulldorff M, Fireman BH, et al. Active surveillance for adverse events: the experience of the Vaccine Safety Datalink project. Pediatrics. 2011; 127(Suppl 1):S54–64. [PubMed: 21502252]
19. Geiger, H. Decentralizing the analysis of health data. Center for Democracy and Technology; Washington, DC: Mar 22. 2012 Available at: https://www.cdt.org/paper/decentralizing-analysis-health-data
20. Platt R, Davis R, Finkelstein J, et al. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. Pharmacoepidemiol Drug Saf. 2001; 10:373–377. [PubMed: 11802579]
21. Pace WD, Cifuentes M, Valuck RJ, et al. An electronic practice-based network for observational comparative effectiveness research. Ann Intern Med. 2009; 151:338–340. [PubMed: 19638402]

22. Wagner EH, Greene SM, Hart G, et al. Building a research consortium of large health systems: the Cancer Research Network. J Natl Cancer Inst Monogr. 2005:3–11. [PubMed: 16287880]

23. Go AS, Magid DJ, Wells B, et al. The Cardiovascular Research Network: a new paradigm for cardiovascular quality and outcomes research. Circ Cardiovasc Qual Outcomes. 2008; 1:138–147. [PubMed: 20031802]

24. Robb MA, Racoosin JA, Sherman RE, et al. The US Food and Drug Administration's Sentinel Initiative: expanding the horizons of medical product safety. Pharmacoepidemiol Drug Saf. 2012; 21(Suppl 1):9–11. [PubMed: 22262587]

25. McGraw D, Rosati K, Evans B. A policy framework for public health uses of electronic health data. Pharmacoepidemiol Drug Saf. 2012; 21(Suppl 1):18–22. [PubMed: 22262589]

26. Rosati K. Using electronic health information for pharmacovigilance: the promise and the pitfalls. J Health Life Sci Law. 2009; 2:173–239.

27. Salmon DA, Akhtar A, Mergler MJ, et al. Immunization-safety monitoring systems for the 2009 H1N1 monovalent influenza vaccination program. Pediatrics. 2011; 127(Suppl 1):S78–86. [PubMed: 21502251]

28. Velentgas P, Bohn RL, Brown JS, et al. A distributed research network model for post-marketing safety studies: the Meningococcal Vaccine Study. Pharmacoepidemiol Drug Saf. 2008; 17:1226–1234. [PubMed: 18956428]

29. Andrade, S.; Raebel, M.; Boudreau, D., et al. Chapter 12: Health Maintenance Organizations / Health Plans.. In: Strom, B.; Kimmel, SE.; Hennessy, S., editors. Pharmacoepidemiology. Wiley; West Sussex, England: 2012. p. 163-188.

30. Hornbrook MC, Hart G, Ellis JL, et al. Building a virtual cancer research organization. J Natl Cancer Inst Monogr. 2005:12–25. [PubMed: 16287881]

31. The Centers for Disease Control and Prevention (CDC). [June, 2012] Vaccine Safety Datalink (VSD) Project. Available at: http://www.cdc.gov/vaccinesafety/Activities/VSD.html.

32. Chen RT, Glasser JW, Rhodes PH, et al. Vaccine Safety Datalink project: a new tool for improving vaccine safety monitoring in the United States. The Vaccine Safety Datalink Team. Pediatrics. 1997; 99:765–773. [PubMed: 9164767]

33. Davis RL, Kolczak M, Lewis E, et al. Active surveillance of vaccine safety: a system to detect early signs of adverse events. Epidemiology. 2005; 16:336–341. [PubMed: 15824549]

34. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005; 58:323–337. [PubMed: 15862718]

35. Suissa S, Garbe E. Primer: administrative health databases in observational studies of drug effects--advantages and disadvantages. Nat Clin Pract Rheumatol. 2007; 3:725–732. [PubMed: 18037932]

36. 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):S21–29. [PubMed: 22692254]

37. Hall GC, Sauer B, Bourke A, et al. Guidelines for good database selection and use in pharmacoepidemiology research. Pharmacoepidemiol Drug Saf. 2012; 21:1–10. [PubMed: 22069180]

38. Bauck, A.; Bachman, D,BD.; Riedlinger, K.; Walker, K.; Luke, S.; Bardsley, J.; Donovan, J. [June, 2012] Developing a Consistent Structure for VDW QA checks. 2011. Available at: http://www.hmoresearchnetwork.org/archives/2011/concurrent/A1-02-Bauck.pdf.

39. [June, 2012] Observational Medical Outcomes Partnership.. OSCAR - Observational Source Characteristics Analysis Report (OSCAR) Design Specification and Feasibility Assessment. 2011. Available at: http://omop.fnih.org/OSCAR.

40. [June, 2012] Observational Medical Outcomes Partnership.. Generalized Review of OSCAR Unified Checking. 2011. Available at: http://omop.fnih.org/GROUCH.

41. Curtis LH, Weiner MG, Boudreau DM, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. Pharmacoepidemiol Drug Saf. 2012; 21(Suppl 1):23–31. [PubMed: 22262590]

42. Curtis, LHWM.; Beaulieu, NU.; Rosofsky, R.; Woodworth, TS.; Boudreau, DM.; Cooper, WO.; Daniel, GW.; Nair, VP.; Raebel, MA.; Brown, JS. [June, 2012] Mini-Sentinel Year 1 Common

Data Model - Data Core Activities. 2012. Available at: http://www.mini-sentinel.org/data_activities/details.aspx?ID=128.

43. Klompas M LR, Daniel J, Haney G, Hou X, Campion FX, Kruskal BA, DeMaria A, Platt R. Electronic medical record Support for Public health (ESP): Automated Detection and Reporting of Statutory Notifiable Diseases to Public Health Authorities. Advances in Disease Surveillance. 2007:3.
44. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010; 17:124–130. [PubMed: 20190053]
45. Hennessy S, Leonard CE, Palumbo CM, et al. Quality of Medicaid and Medicare data obtained through Centers for Medicare and Medicaid Services (CMS). Med Care. 2007; 45:1216–1220. [PubMed: 18007173]
46. Hennessy S, Bilker WB, Weber A, et al. Descriptive analyses of the integrity of a US Medicaid claims database. Pharmacoepidemiol Drug Saf. 2003; 12:103–111. [PubMed: 12647699]
47. Brown PJ, Warmington V. Info-tsunami: surviving the storm with data quality probes. Inform Prim Care. 2003; 11:229–233. discussion 234-227. [PubMed: 14980063]
48. Brown PJ, Warmington V. Data quality probes-exploiting and improving the quality of electronic patient record data and patient care. Int J Med Inform. 2002; 68:91–98. [PubMed: 12467794]
49. Brown PJ, Harwood J, Brantigan P. Data quality probes--a synergistic method for quality monitoring of electronic medical record data accuracy and healthcare provision. Stud Health Technol Inform. 2001; 84:1116–1119. [PubMed: 11604902]

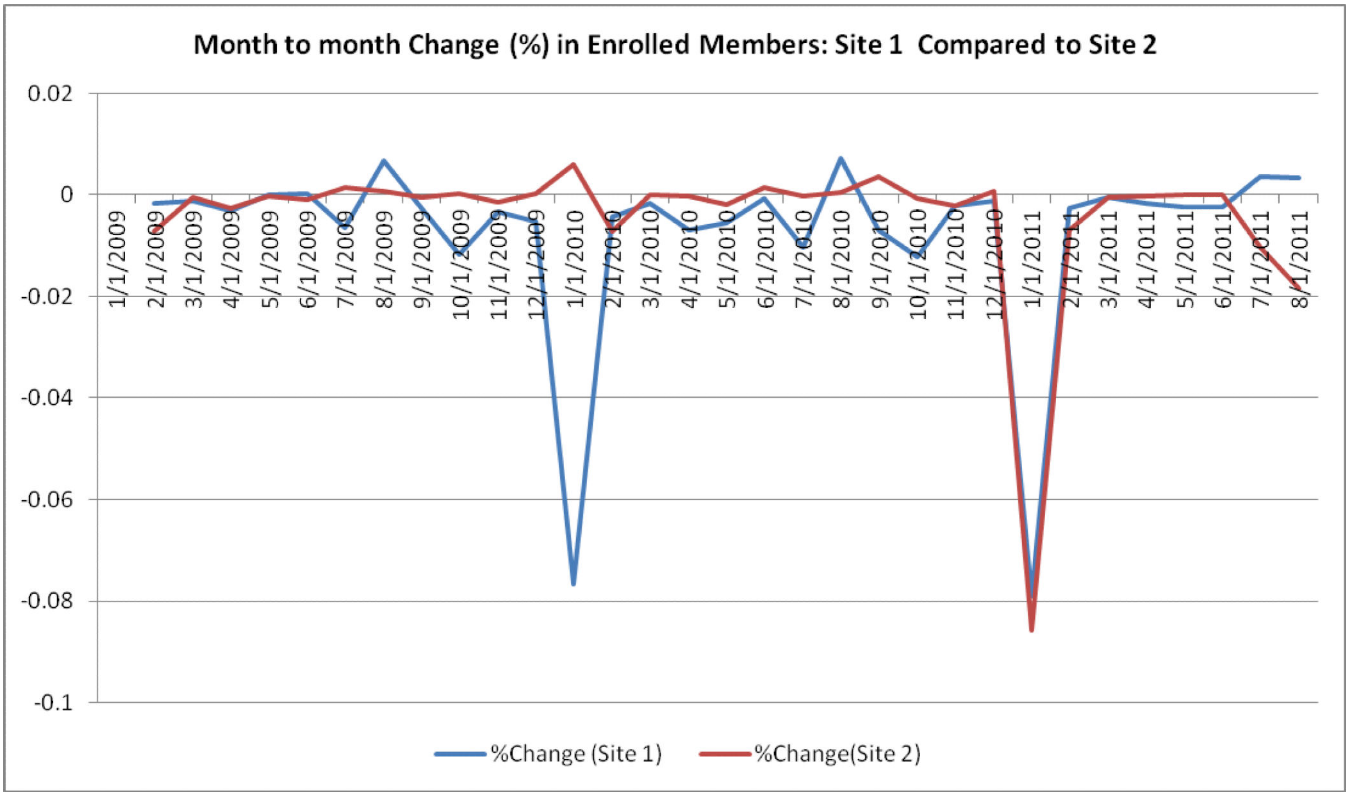

Figure 1.

Comparison of change in monthly enrollment for 2 sites

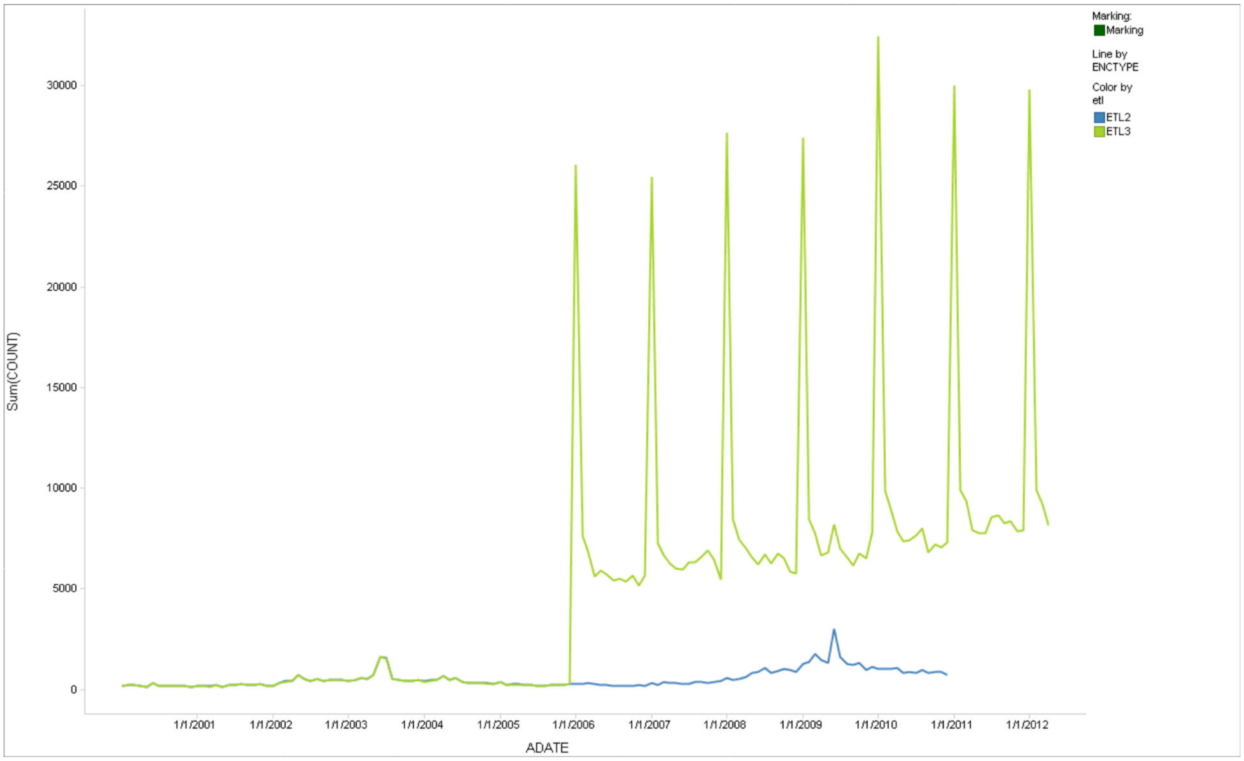

Figure 2.

Comparison of change in monthly encounters across two subsequent data extractions

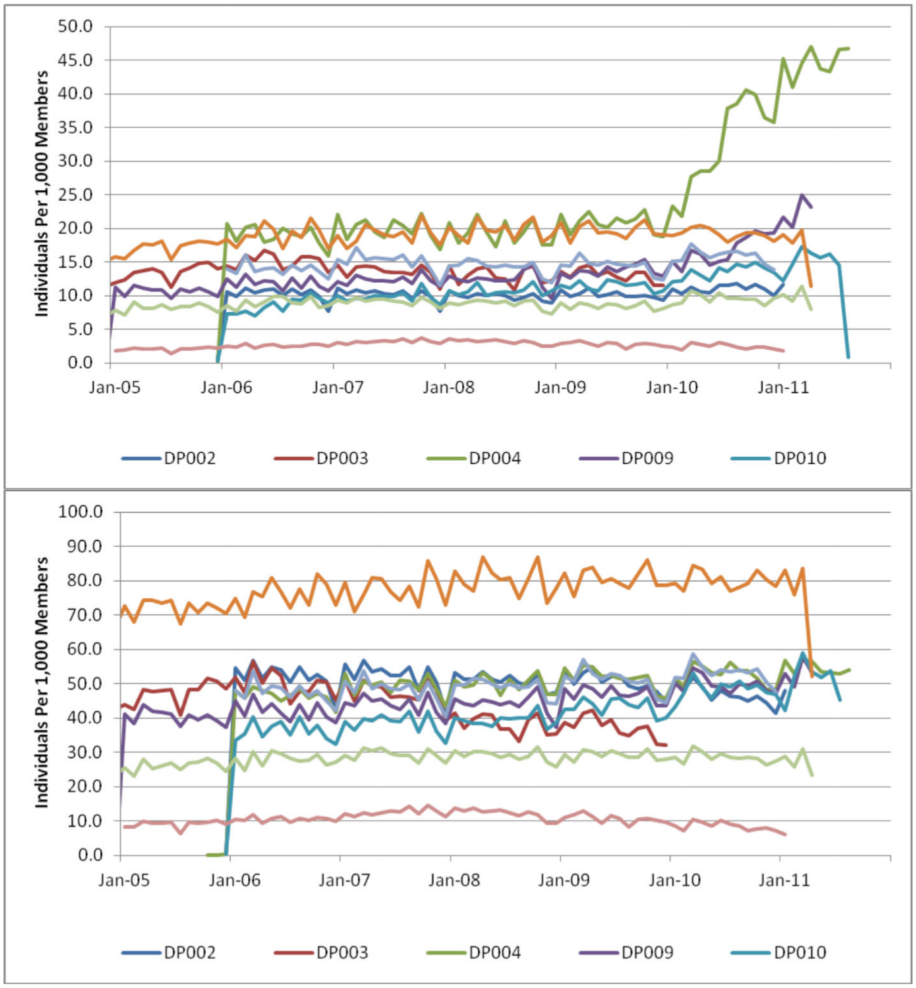

Figure 3.

Individuals with one more laboratory test results per 1,000 enrollees: Example of two different laboratory tests

Table 1

Patient matching across tables in the Mini-Sentinel common data model, by site

	DP2	DP4	DP5	DP8	DP9	DP10	DP12	DP13	DP14	DP15	DP17	Total
% of Patients Matched with Enrollment												
Demographic	96.3%	78.0%	39.6%	100%	89.4%	96.8%	92.9%	84.5%	49.6%	98.7%	49.7%	94.5%
Dispensing	97.9%	87.5%	99.6%	100%	94.6%	98.9%	99.7%	96.0%	93.9%	98.8%	92.6%	98.7%
Encounter	96.1%	76.6%	28.8%	100%	89.7%	98.2%	99.5%	90.6%	95.1%	98.3%	89.6%	96.9%
Diagnosis	97.0%	79.4%	28.9%	100%	91.9%	98.3%	99.5%	91.3%	95.1%	98.3%	90.0%	97.0%
Procedure	96.4%	77.7%	29.0%	100%	90.7%	98.2%	99.5%	92.0%	95.1%	98.3%	89.7%	97.1%

Note: Adapted from the Mini-Sentinel Data Quality and Characterization Procedures and Findings Report (2011).[42]