Published in final edited form as:
Int J Med Inform. 2016 June ; 90: 40–47. doi:10.1016/j.ijmedinf.2016.03.006.

Data Quality Assessment Framework to Assess Electronic Medical Record Data for use in Research

Andrew P. Reimer, PhD, RN[a,b,*], **Alex Milinovich**[b], and **Elizabeth A. Madigan, PhD, RN**[a]

Andrew P. Reimer: axr62@cwru.edu; Alex Milinovich: milinoa@ccf.org; Elizabeth A. Madigan: elizabeth.madigan@case.edu

[a]Frances Payne Bolton School of Nursing, Case Western Reserve University, 10900 Euclid Ave, Cleveland, OH 44106, United States

[b]Cleveland Clinic, 10900 Euclid Avenue, Cleveland, OH 44195, United States

1. Introduction

Medical transport plays an integral role in supporting health care delivery as patients often present to hospitals or clinics that do not provide the necessary services that acutely ill or injured patients require. Patient transfers primarily can be categorized as emergent or non-emergent. A growing body of evidence supports transfer of patients experiencing time sensitive emergencies such as trauma, stroke, or heart attack. However, there is sparse evidence to support the decisions of if, how, and when to transfer non-emergent patients who oftentimes experience poor outcomes [1–3].

Investigating patient transfers has presented multiple challenges due to the many facets involved in moving patients between hospitals and sometimes across health systems. A primary limitation is the lack or accessibility of data that are required to adequately assess the effect of the transfer on patient outcomes. Until recently, most research efforts investigating transferred patients have remained isolated to individual units or hospitals, producing limited insight and restricting our overall understanding of how transfer influences patient outcome.

The proliferation and use of electronic medical records (EMR) in the clinical setting now provides a rich source of clinical data that can be leveraged to support research on patient outcomes, comparative effectiveness, and health systems research. Particularly, reusing EMR data provides the distinct ability to study patients and interventions in actual clinical

[*]Corresponding author at: CFRN, Frances Payne Bolton School of Nursing, Case Western Reserve University, 2120 Cornell Rd, Cleveland, OH 44106, T: 216-368-7570, Fax: 216-368-5303.

Authors' Contributions

APR, AM, and EM were responsible for the study design and methods. APR and AM conducted data analyses. APR and EM drafted the manuscript. AM provided critical review of the data results reporting and manuscript.

Conflict of Interest

APR, AM, and EM have no conflicts of interest to report.

practice as they naturally occur [4], facilitating rapid translation of findings back into practice. Most research efforts now include EMR data abstraction to support individual studies, or more generally to support aggregation of large volumes of data in disease specific registries or clinical data repositories. Such is the case for the Transport Data Mart (TDM) [5] that we developed to support comprehensive outcomes research efforts that aggregates patient data across the entire episode of care for a patient that is transported from one hospital to another. However, amassing the large volume and variety of data that robust clinical EMRs provide is only the first stage. Once the appropriate data are identified and aggregated, the suitability of the data for research purposes [6–8] must be addressed.

Typically initial efforts for assessing the quality of EMR data abstracted for research purposes are focused on identifying and validating that the correct patient population was identified and abstracted. One approach, proposed by Faulconer and Lusignan's [9] eight step approach to assessing diagnostic data quality, provides an example of the steps necessary to accurately identify a patient for inclusion in a specific disease registry. However, for transported patients this potentially complicated problem of identifying patients that are transported does not exist. A patient either undergoes transfer or does not, creating a singular inclusion criterion for abstracting that particular episode of care into the TDM. Another approach is the Data Quality Probe method proposed by Brown and Warmington [10] that identifies cases in an EMR system that are not successfully matched (concordance), or contain errors between one item and another, that when applied longitudinally can improve data entry practice and overall quality.

Transported patients present a different problem related to matching the individual encounters across the multiple admissions and discharges that represent the entire episode of care. Capturing the entire episode of care includes linking EMR data from the referring hospital EMR, the transport EMR – if available, and the receiving hospital EMR. While the definition of data completeness can vary depending on the task at hand [11], an overarching assessment of data completeness, or in this case, inclusion and concordance across data sources, must be evaluated to assess the overall integrity of data inclusion and integration within the TDM. Therefore, the purpose of this paper is two-fold. First, we present a stepwise framework capable of guiding initial data quality assessment when matching multiple data sources regardless of context or application. Then, we demonstrate a use case of initial analysis of a longitudinal data repository of electronic health record data that illustrates the first four steps of the framework, and report results.

2. Materials and methods

2.1 Framework

The conceptual framework guiding this study is based on the five dimensions presented by Weiskopf and Weng[6] (completeness, correctness, concordance, plausibility and currency), with a specific focus on guiding deep evaluation of completeness and concordance of variables used in record linkage, and then assessment of the variables required for analysis across multiple data sources. The five dimensions as defined by Weiskopf and Wang are: 1) completeness– "is a truth about a patient present in the EHR?" 2) correctness – is an element that is present in the EHR true?" 3) concordance –"is there agreement between elements in the EHR, or between the EHR and another data source?" 4) plausibility – "does an element in the EHR make sense in light of other knowledge about what that element is measuring?" and 5) currency – "is an element in the EHR a relevant representation of the patient state at a given point in time?"

Concordance becomes particularly important during data quality assessment in a longitudinal data repository. Due to the need to assess several elements of data presence and agreement across multiple records or data sources simultaneously in one distinct step, it is useful to think of concordance as a construct - "*longitudinal concordance*." The construct longitudinal concordance is defined as assessing data element presence, agreement, and source agreement of specified variables across multiple data sources. The first concept, data element presence is defined as the minimum required data elements to facilitate matching data across sources. For example, matching across medical record sources (as illustrated in the case example provided in detail later) would entail identifying the necessary variables required that might include: medical record number, patient name, admission date/time, etc.; while matching across disparate data sources such as Twitter accounts (twitter handle, date/time), to weather data (date/time, location), entails different data to assess for initial feasibility of matching across domains or sources. Then the second concept, data element agreement, is defined as "two or more elements within the target data domains or sources are compared to see if they report the same or compatible information" [6]. Lastly, data source agreement is an extension of data element agreement and is defined as "data from the data domains or sources are compared with data from another domain or source to determine if they are in agreement"[6]. By definition data element presence, data element agreement, and data source agreement are considered as individual data quality assessment methods. For the purpose of longitudinal data quality assessment, they are combined into one methodological step, and can be applied to any sets of variables within and across any two or more data sources. Therefore, the construct *longitudinal concordance* subsumes the dimensions of completeness and correctness within it. The importance of assessing longitudinal concordance can be illustrated by the patient demographics that are present in each episode of care record. The primary questions to be addressed when assessing longitudinal concordance include: 1) are each of the data elements of interest present in each record, and 2) do those same data elements agree across record sources and domains? Operationalizing the currently proposed conceptual framework requires incorporating four definitions of data completeness [12] (documentation, breadth, density, and, prediction) that offer specific assessment measures to develop a standard stepwise approach that can be replicated in other projects. Merging the data quality dimensions with the data quality assessment definitions yields a six step process – due to the addition of a preliminary assessment step for including external data sources (i.e. patient log), and the breakout of data element presence as a discreet and significant assessment step - to conducting data quality assessment for a longitudinal data repository as displayed in Figure 2 with each step more fully described in the Analytical Approach section 2.4.

2.2 Guiding Aim

The guiding aim for this investigation is to identify the total number of patient episodes of care that include data across the entire episode of care for all patients transferred by helicopter (sending hospitalization, transport, and receiving hospitalization). We choose this subgroup because it represents the most restrictive patient population with the most amount of records across data sources to be matched. Conducting this data quality and matching assessment entails three distinct phases of assessment and groups of variables. The first assessment phase consists of the steps and variables of interest necessary to match the records across domains or sources (refer to Figure 2). Then, once records are successfully matched across sources, the second assessment consists of identifying and assessing if the analytical variables of interest are present and sufficiently represented in the available data to answer the research questions. Lastly, though not demonstrated in this paper, phase three consists of analyzing the data to predict absence of data (missingness — e.g. missed medical appointment) or the anticipated generation of new data in the future (e.g. hospital readmission). Each patient episode of care must be successfully matched across care encounters and include specific clinical variables. For the purpose of this investigation specific categories of variables of interest include: procedures, laboratory results, medications, and vital signs.

2.3 Data

The TDM is a comprehensive data mart that incorporates records for all patient transfer requests for the Cleveland Clinic Health System. Cleveland Clinic is a not-for-profit health system in Northeast Ohio that consists of a main campus in Cleveland and eight community hospitals throughout the region. The health system handles approximately 32,000 patient referrals for transport, and moves approximately 5,300 of those patients annually by ambulance, helicopter, and jet. The TDM consists of patient EMRs from three primary sources: referring hospital, transport EMR, and the receiving hospital (Figure 1). Each source contains multiple data domains primarily abstracted from the patient's EMR (interventions, lab results, vital signs, medications), financial, and risk mortality prediction data. Inclusion of patient encounters is generated by the transport request log that is maintained by the transport team. Every request for possible transfer is logged by capturing basic patient demographics, patient location, and reason for transfer. Each data abstraction is initiated from the transport request log and linked to a matching patient encounter from the three primary sources. However, not every log entry is successfully matched to an inpatient hospitalization (e.g. the transfer never took place). Additionally, not every encounter can be matched to both a sending and receiving hospital if the patient is either transferred-in from a non-health system hospital, or transferred out to a non-health system hospital. Encounters that are not machine linked via the matching algorithm are entered into a queue for manual review by system administrators for matching or permanent disposition as an unmatched record via a data mart graphical user interface. For purposes of demonstrating this project, we assessed all data from 2014.

2.4 Analytical Approach

Phase 1 – Recorded Matching Assessment

Step 1

Preliminary analysis: The first assessment to conduct is whether the appropriate records are included into the TDM. The initiating source for record matching and curation is the transport log. Each entry into the log is potentially mapped to a hospital encounter at the receiving hospital, and potentially to the sending hospital if the transfer originated within the health system. However, creation of a transport log entry does not necessarily mean that a patient was actually transferred into the health system, primarily due to the patient never being transferred, or the patient was transferred to a non-health system hospital. Step 1 may not be applicable to all data quality assessments of longitudinal data repositories (e.g. studies using data from only one EMR will not need to complete this step as it does not apply).

Step 2

Documentation – Longitudinal Concordance: Simply assessing if records are present for each transport log entry is insufficient for assessing documentation completeness as a particular set of variables are used to perform this initial assessment. Of particular importance to assessing the quality of data matching in a longitudinal registry is the added dimension of longitudinal data agreement. Longitudinal data source agreement as used in this study is defined as the agreement of those basic demographic elements that must be contained in each record to assess likelihood of a positive match between records. The number of required data elements to assess an appropriate match will vary depending on the data sources and overall data management practices related to your own institution. For example, the Cleveland Clinic health system maintains an enterprise medical record number for each patient that is unique, and capable of matching any number of varying medical record numbers that may exist for a given patient at different hospitals within the health system. Although a unique patient identifier, the enterprise medical record number does not provide sufficient record linking capabilities in itself. The TDM requires a temporal order of a minimum of one patient encounter, either at the sending or receiving hospital within the health system. The admission and discharge date and times must be incorporated to match the records as the patient progresses through the episode of care. Therefore, as a second level assessment of documentation, longitudinal registries must include assessment of completeness to include the verification of demographic data agreement across the necessary data sources. Conceptually, including demographic variables that are commonly associated with every record (i.e. last name, first name, date of birth, medical record number) is not measuring breadth - as previously described by Weiskopf et al. [12] - of available records for that particular patient as those data elements are generally included in every record by default and do not represent volume of available data.

To measure longitudinal data agreement in the TDM we identified the common demographic data elements that are used to match each patient hospital encounter record across the TDM to include: (1) last name, (2) first name, (3) date of birth, (4) gender, (5) medical record number linked to patient identification number, then in combination (6) transport date and (7) transport time, matched to (8) receiving hospital and (9) receiving hospital admission time, and/or (10) sending hospital and (11) sending hospital discharge time (Table 1). Each of the data elements are compared by the matching algorithm to assess appropriateness of matching candidate records individually and in combination. Step 2 is likely to apply to most data adequacy evaluations as record matching across EMR types is critical to most studies of this type.

Step 3

Breadth: The next level of assessment for records that are correctly identified is to assess the breadth of available records linked to that particular patient encounter. Types of records that can be included in this assessment include: labs results, medications, procedures or interventions, vital signs, financial or administrative records, and others. The first level of assessment is identifying whether there are records that are successfully matched to that specific encounter, understanding that there are three encounters per matched record, and the depth and type of data that those records include. The second level of assessment includes accounting for the availability of each of the desired record domains for each hospital day (i.e. laboratory results, medications, vital signs, procedures). To assess breadth in the TDM, we assessed the presence of the following records for every patient hospital day for each encounter: vital signs, laboratory results, procedures, and medications. Breadth requirements will vary by study; some studies may have more periodic (i.e. less frequent) record domains of some types. For example, in a home health care record, vital signs would be expected on each visit whereas laboratory results would be less frequent.

Phase 2 – Analytical Variable Assessment

Step 4

Data element presence: Once it has been established that sufficient records exist and the breadth of those records covers the range of data necessary to answer the research question at hand, an assessment of data element presence should be conducted. Assessing data element presence, while done in step 2 for the specific matching variables related to assessing correct matching of records across encounters, must also be completed as a standalone step to enable assessment of specific variables of interest. Assessing data element presence can consist of either generally assessing for several data elements of interest or every data element required to answer the research questions such as general descriptive statistics or pre-specified prediction models. Several data elements of interest might be specifically targeted to assess for their presence. To establish metadata for the TDM, we assessed data element presence by measuring the presence of a diagnosis, set of vital signs (systolic/diastolic blood pressure, heart rate, breath rate, pulse oximetry – resting rate), and lab panels – comprehensive metabolic panel, complete blood count for each hospital day. As noted in Step 3, this step will vary depending on the nature of the study. The important element is that this is specified apriori.

Step 5

Density: Although not directly associated with this study due to the immediate nature of care that transported patients' progress through, assessment of density can be applicable. For instance, performing longitudinal analyses that assess patient trajectory over time may be useful, or in the case of assessing predictability – an additional dimension of data completeness not covered here – assessment of density is necessary. Therefore density can be measured by assessing if the desired number of records or data points are available over a set period of time [12]. To perform a longitudinal assessment of patient trajectory in the TDM we would first assess the number of records that had prior records available leading up the incident hospitalization for the transferred patient. Then, we would assess the number and types of records available for each patient. Other ways of assessing density can be accomplished as described by Weiskopf and Hripcsak's article on assessing data completeness in which they implemented an adjustment calculation more completely described by Sperrin et al. [13]. For other studies, there may be apriori density measures established. For example, some post-acute care sites like home health care require the completion of a standardized data collection (OASIS) at least every 60 days. If the OASIS density measure was not met, it would raise issues relative to the data quality.

Phase 3 - Analysis

Step 6

Prediction: Prediction is also not assessed in this study but may be applicable in other applications. Prediction is genreally considered as the ability to resue EMR data to "predict something or find an association" [12]. The abiity to use EMR data to predict future events has received increased attention as hospital and health care systems are leveraging their data to identify high risk patients to receive increased care or targeted interventions (i.e. patients most likely to be readmitted). Current applications include using EMR data to run risk calculators or nomograms to aid clinical decision making for patients and providers and is accomplished through applying regression models to predict the desired event or outcome [12]. Specific examples include computational frameworks that can formulate and query EMR data to identify adherance to long-term medications [14], or develop a prognostic model capable of predicting trauma patient outcomes after damage control surgery based on EMR data available at the time the model is run [15]. Applying prediction to TDM data could yield clinical decision support via models that recommend the appropriate transfer method (helicopter or ambulance), or if the patient should be transferred at all.

3. Results

Step 1 – Preliminary analysis—There were a total of 9,557 log entries for 2014. Of these, 8,139 log entries were successfully matched to corresponding records in the health system EMR with only 537 cases requiring manual matching due to errors in patient demographics entered in the transport log that are later corrected in the permanent hospital EMR. Of those successfully matched cases, 2,832 were successfully mapped to both the sending and receiving hospital encounters within the Cleveland Clinic health system for the specific episode of care that included transport (both ground and air), with 196 records requiring manual review and mapping (resulting in a 93% automatic matching rate). The last group consisted of those records successfully mapped across the entire episode of care to include sending and receiving hospital encounters as well as the transport encounter (helicopter trips only), of which there were 590 complete cases. For purposes of demonstrating each of the following assessment steps longitudinally, the default working dataset will consist of the 2,832 patients that were successfully matched to both sending and receiving hospitalizations.

Step 2 – Documentation - Longitudinal Concordance—Assessing longitudinal concordance was completed in two steps (figure 3). The first step was assessing those records that were automatically matched via exact matching criteria on three progressive levels of record inclusion across the episode of care with the current working dataset of 2,832 cases remaining from step 1. The first group was matched between the transport log with any corresponding patient encounter in the EMR: last name, first name, date of birth, gender, and medical record resulting in 1,557 of 2,832 records automatically matched (55% automatic matching rate). To assess matching from the transport log to a corresponding post-transport hospital encounter, group two consisted of matching the variables included in group one, with the additional variables: combination of transport date and time plus receiving hospital name, and transport date and time plus receiving hospital admission time to assess the temporal order of transfer. The combination of matching variables in group two resulted in 1,113 automatic matches (39% automatic matching rate). Group three consisted of matching records across the episode of care to include sending hospital records. In addition to the variables included in group two, group three added: combination of transport date and time plus sending hospital name, and transport date and time plus receiving hospital name. Group three resulted in 1,104 of 2,832 cases being automatically matched (39% automatic matching rate).

The second level consisted of assessing each variable individually for a successful match between the transport log and corresponding patient encounter in the EMR (Figure 3). Automatic matching consists of the TDM running an algorithm to match each variable based on relaxed temporal constraints when applicable. Last name, first, gender, medical record number, sending hospital name and receiving hospital name were based on exact matching. Date of birth is captured in years in the transport log and was therefore matched on the date of birth transformed into years from the exact date of birth contained in the EMR records. Transport date and transport time was matched on ± 2 days of the specified date/time contained in the transport log. As presented in figure 3, the results vary across the individual variables with the sending hospital yielding the highest automatic match rate of 2,765 of 2,832 records matching (98% automatic match rate), and receiving hospital admission time yielding the lowest automatic match rate of 2,249 of 2,832 records automatically matched (79% automatic match rate).

Step 3 – Breadth—Medications (93%) were recorded most often for each encounter, and the least frequently recorded were laboratory results ranging from 59–65% (Table 2).

Step 4 – Data Element Presence—Vital signs averaged the highest number of mean recordings per encounter ranging from 106–143, with the widest range of 0–9896 (Table 1). Laboratory results exhibited the smallest number of repeated measures averaging only 7–8 panels per admission. Procedures and medications also varied widely between encounters.

4. Discussion

Completing data quality assessment via the proposed stepwise framework yielded several insights. We choose to use inclusion criteria that would identify the smallest subgroup of potential subjects to be included while demonstrating the application of the data quality framework. This is evident in the final study sample of 590 subjects that demonstrates how quickly the number of eligible research subjects can reduce based on how broadly or narrowly the inclusion criteria are defined. These findings also suggest that there is a certain absence of records, that for whatever reason, can be considered a false record (i.e. an encounter note generated but not actually filled out and still remains in the system) and thus not matched across data sources. Therefore, completing step one quickly provides a way to identify potentially eligible study subjects out of the total population available, and if there are adequate subjects to move forward. If, for instance we were conducting a study on all patients referred for transport and their associated clinical outcomes, then the study population would be much larger.

Applying exact matching criteria in the first level of assessment in Step 2 resulted in poor automatic matching efficiency ranging from 55% down to 39% as the number of matching variables increased. These results demonstrate the variability inherent in data entry practices across record sources and that a lot of work is necessary to develop appropriately relaxed temporal constraints to improve automatic matching efficiency. After initiation of efficient matching algorithms, matching efficiency remained relatively consistent. Additionally, there appears to be relatively limited drop-off of additional records when the criteria for matching increases after the first group of matching variables is initiated. This also seems to be reflected in results from Step 3 in that 92% of the records contained a complete set of vital signs, suggesting that there are a high proportion of records that contain nearly complete data.

A primary problem when analyzing EMR data is the high dimensionality of the data — the number of covariates is far greater (often in multiples) than the sample size [16]. The measures of central tendency reported in Step 4 indicate that a smaller proportion of records contain the most amount of data, exemplifying the necessity to assess data element presence as a discreet step during initial data quality assessment. For example, the average hospital encounter contains around 100 vital sign measurements, 7 laboratory testing measurements, 21 procedures, and 55 medication administrations. Yet when assessing the range for each of these categories, especially for vital signs (0 – 9896) and procedures (1–981), we see that the averages are skewed dramatically to the right. This skewness can have implications for planning data inclusion and statistical analyses of how to handle those cases with voluminous data points.

Another issue is how to handle those encounters that are not included due to not meeting the final matching criteria. Although a record may not be included in the final sample, it may contain enough data to enable running comparative analyses to assess for potential differences between the final sample and those records not included. Conducting comparative analyses may or may not identify potentially significant differences in the two samples. Decisions can be made as to which variables must absolutely be matched or if there is room for adjusting the temporal matching criteria around each matching variable (e.g. discharge date/time from the sending hospital and the transport EMR may not temporally match — there can be overlaps in time, or major gaps between documented discharge and pickup times ranging anywhere from several minutes to hours). Alternatively, different matching approaches such as probabilistic matching could be implemented that has shown high accuracy when compared to the deterministic approach reported here, especially when human review is not possible [17] or when missing data entries are reducing reliability [18]. Depending on the study questions and study purpose, it might be advantageous to include records that do not contain all variables of interest as this more accurately reflects real-life clinical practice where not all information is known or available at the time decisions are being made. These are important considerations as we move forward and develop methods of handling real-world generated data to develop decision support systems capable of supporting a true learning health system [19].

This study has several limitations. First, EMR data for this study is from only one health system, including only on EMR program (Epic), and one transport EMR (GoldenHour). Second, we only assessed one year of data (2014), including a larger sample of data may have yielded different matching results. However, during 2014 the hospital based transport log was transitioned from a SharePoint based transport log to an in-house developed SQL based log, providing a useful test case for the matching algorithms (that includes 40 individual abstraction and matching steps) across data sources.

Lastly, although we provide evidence that no one variable contributes to poor matching across patient encounters of care (Figure 3), in combination, even minimal missingness of a combination of matching variables can lead to reduced matching efficiency. One of the limitations in assessing data completeness is that there is no guidance as to what the metrics or benchmarks for data concordance, completeness and correctness should be. Currently, implementing approaches such as the Structured Data Quality Report method [20], or the Data Quality Probe [10] method over time have been shown to improve provider practice for clinical data entry and thus data completeness, additional assessment efforts are necessary. Future work should focus on identifying what constitutes acceptable missingness, and/or establish benchmarks of acceptable inclusion levels of patient records (e.g. what are acceptable confidence intervals for predictors within a decision support system that is developed and then runs on EMR data, and when can it be introduced into practice).

5. Conclusion

The proposed six-step data quality assessment framework is useful in establishing the metadata for a longitudinal data repository that can be replicated by other investigations. Additionally, the framework provides a pragmatic stepwise approach to identify a potential study population available from a given data repository for a research project to answer specific research questions. However, there are practical issues that need to be addressed including the data quality assessments identified here. As health care systems move from episodic payment to payment structures that follow patients across sites of care, the need for high quality EMR data that crosses sites of care will become increasingly important. In addition, as interoperability issues are resolved, the advantage of having high quality data for clinical decision support will offer new opportunities for research, but more importantly for clinical practice. Currently, the most prescient issue to address is the need to establish data quality metrics for benchmarking acceptable levels of EMR data inclusiveness through testing and application.

Acknowledgments

Funding Sources/Disclosures: Dr. Reimer was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant KL2TR000440;

and through the National Institutes of Health Loan Repayment Program, National Center for Advancing Translational Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Hill AD, Vingilis E, Martin CM, Hartford K, Speechley KN. Interhospital transfer of critically ill patients: demographic and outcomes comparison with nontransferred intensive care unit patients. J Crit Care. 2007; 22:290–295. [PubMed: 18086399]
2. Reimer AP, Schiltz N, Koroukian SM, Madigan E. National Incidence of Medical Transfer: Patient Characteristics and Regional Variation. Journal of Health and Human Services Administration. 2015 In Press.
3. Rosenberg AL, Hofer TP, Strachan C, Watts CM, Hayward RA. Accepting Critically Ill Transfer Patients: Adverse Effect on a Referral Center's Outcome and Benchmark Measures. Annals of Internal Medicine. 2003; 138:882–890. [PubMed: 12779298]
4. Sox HC, Goodman SN. The methods of comparative effectiveness research. Annu Rev Public Health. 2012; 33:425–445. [PubMed: 22224891]
5. Reimer AP, Madigan E. Developing a fully integrated medical transport record to support comparative effectiveness research for patients undergoing medical transport. EGEMS (Wash DC). 2013; 1:1024. [PubMed: 25848576]
6. 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; 20:144–151. [PubMed: 22733976]
7. Hersh WR. Caveats for the use of operational electronic health record data in comparative effectiveness research. Medical care. 2013; 51:S30–S37. [PubMed: 23774517]
8. Hersh WR, Cimino J, Payne PR, Embi P, Logan J, Weiner M, Bernstam EV, Lehmann H, Hripcsak G, Hartzog T, Saltz J. Recommendations for the use of operational electronic health record data in comparative effectiveness research. EGEMS (Wash DC). 2013; 1:1018. [PubMed: 25848563]
9. Faulconer. An eight-step method for assessing. Informatics in Primary care. 2004; 12:243–254.
10. 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]
11. Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems. 1996; 12:5–34.
12. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. Journal of biomedical informatics. 2013; 46:830–836. [PubMed: 23820016]
13. Sperrin M, Thew S, Weatherall J, Dixon W, Buchan I. Quantifying the longitudinal value of healthcare record collections for pharmacoepidemiology. AMIA. 2011; 2011:1318–1325.
14. Mabotuwana T, Warren J, Kennelly J. A computational framework to identify patients with poor adherence to blood pressure lowering medication. Int J Med Inform. 2009; 78:745–756. [PubMed: 19631581]
15. Demsar J, Zupan B, Aoki N, Wall MJ, Granchi TH, Beck JR. Feature mining and predictive model construction from severe trauma patient's data. International Journal of Medical Informatics. 2001; 63:41–50. [PubMed: 11518664]
16. Zhang H, Legro RS, Zhang J, Zhang L, Chen X, Huang H, Casson P, Schlaff W, Diamond M, Krawetz S, Coutifaris C, Brzyski RG, Christman G, Santoro N, Eisenberg E. Decision trees for identifying predictors of treatment effectiveness in clinical trials and its application to ovulation in a study of women with polycystic ovary syndrome. Human Reproduction. 2010; 25:2612–2621. [PubMed: 20716558]
17. Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a Probabilistic Record Linkage Technique without Human Review. AMIA Annual Symposium Proceedings. 2003; 2003:259–263. [PubMed: 14728174]

18. Ong TC, Mannino MV, Schilling LM, Kahn MG. Improving record linkage performance in the presence of missing linkage data. Journal of biomedical informatics. 2014; 52:43–54. [PubMed: 24524889]

19. I.I.o. Medicine. Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary. Washington, DC: 2011.

20. Taggart J, Liaw ST, Yu H. Structured data quality reports to improve EHR data quality. Int J Med Inform. 2015; 84:1094–1098. [PubMed: 26480872]

Highlights

- The proliferation of electronic medical records (EMR) provides a rich source of clinical data.
- Currently, no standardized approach has been proposed to assess data completeness.
- The six-step data quality framework can guide standardized metadata assessment.
- Applying the same assessment framework can lead to standardized reporting in research.

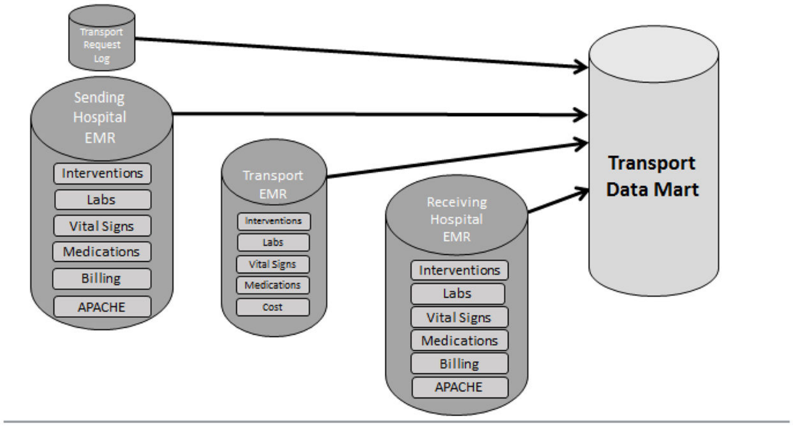
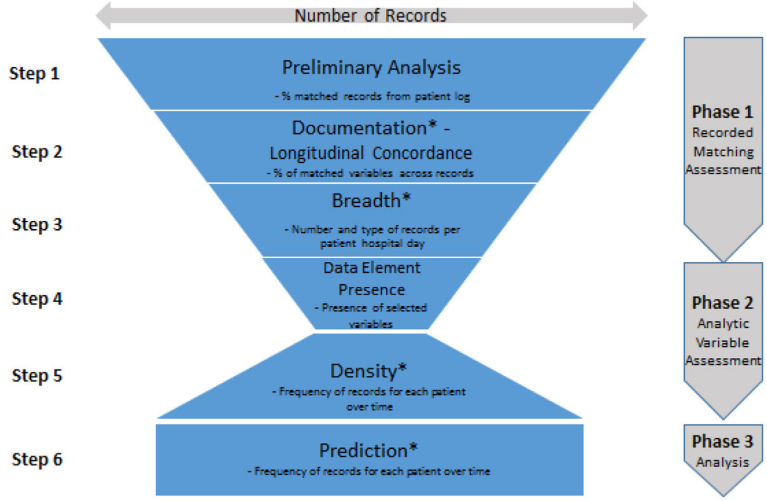


Figure 1.

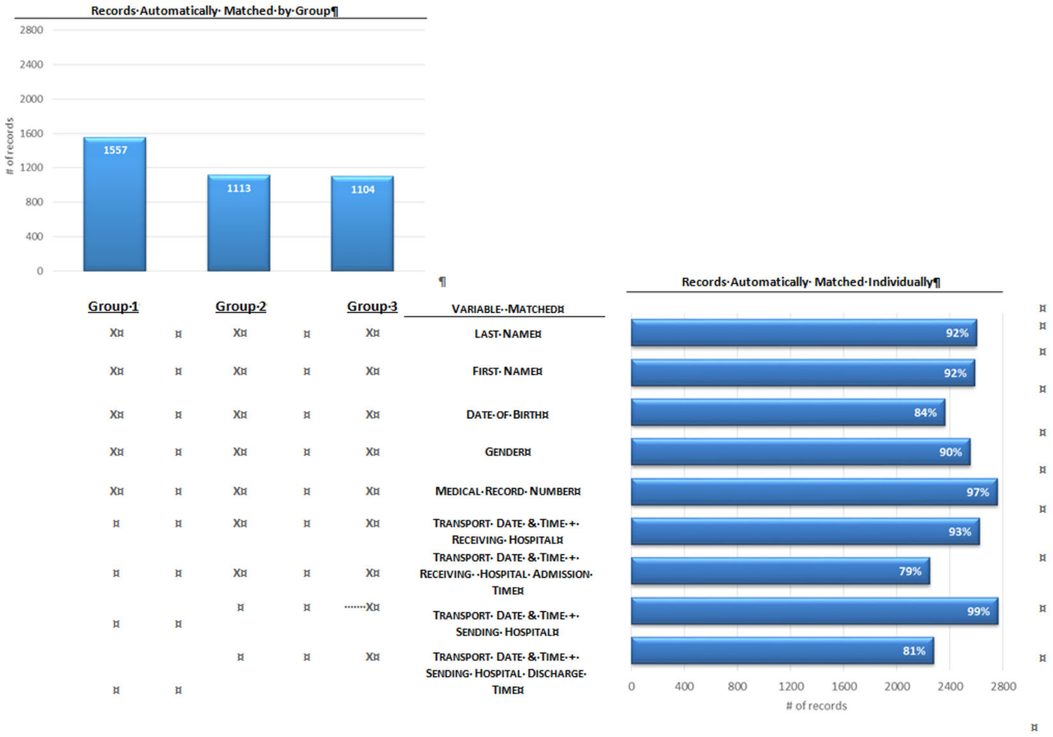
Transport Patient Data Mart: Data domains and source databases


Number of Records
Step 1
Preliminary Analysis
- % matched records from patient log
Step 2
Documentation* - Longitudinal Concordance
- % of matched variables across records
Step 3
Breadth*
- Number and type of records per patient hospital day
Step 4
Data Element Presence
- Presence of selected variables
Step 5
Density*
- Frequency of records for each patient over time
Step 6
Prediction*
- Frequency of records for each patient over time
Phase 1
Recorded Matching Assessment
Phase 2
Analytic Variable Assessment
Phase 3
Analysis

*Definition of data completeness as defined by Weiskopf et al. (2013).

Figure 2.

Steps of Operational Framework



Group 1	Group 2	Group 3	Variable Matched
X	X	X	Last Name
X	X	X	First Name
X	X	X	Date of Birth
X	X	X	Gender
X	X	X	Medical Record Number
	X	X	Transport Date & Time + Receiving Hospital
	X	X	Transport Date & Time + Receiving Hospital Admission Time
		X	Transport Date & Time + Sending Hospital
		X	Transport Date & Time + Sending Hospital Discharge Time

Figure 3.

Data Source Agreement

Table 1

Data Element Matching

	Data Source			
Data Element	**Transport Log**	**Sending Hospital**	**Transport EMR**	**Receiving Hospital**
Last name	X	X	X	X
First name	X	X	X	X
Date of birth	X	X	X	X
Gender	X	X	X	X
Medical record number	X	X	X	X
*Discharge date/time		X	X	
*Transport date/time	X		X	
*Admission date/time			X	X

*individual data elements of date and time are combined, collapsing 6 individual elements into 3 for purposes of displaying matching between sources

Table 2

Breadth and Data Element Presence

	Number of recordings per Patient per hospitalization			
	% with data	**mean**	**median**	**range**
Vital signs				
Pulse	92	143	55	0 – 9896
Blood Pressure	92	106	50	0 – 5022
Respirations	92	129	52	0 – 6459
Pulse Oximetry	92	140	50	0 – 9658
Laboratory Results				
Complete Metabolic Panel	65	8	5	0–225
Complete Blood Count	59	7	4	0–245
Procedures [1]	78	21	9	1–981
Medication [2]	93	55	37	1–910

[1] Measure of total procedures performed per hospitalization (e.g. intravenous catheter insertion, intubation, MRI scan, etc.)

[2] Measure of total medication administrations per hospitalization — can include multiple administrations of the same medication