A Model for Quality Defects Detection in Record Sequences and its Application for Obstetric Data in Electronic Medical Records

TAL ABOUDY, ADIR EVEN, Ben-Gurion University of the Negev
YOAV BREZINOV, EDI VAISBUCH, Kaplan Medical Center

Abstract

Scenarios, in which real-world state transitions are documented by a sequence of data records, introduce unique data quality (DQ) challenges. Due to time and workload constraints, one might choose to update the values only for a subset of the required attributes, while replicating previous values of others. As a result, the record sequence might reflect the current state incorrectly and fail to capture critical transitions. Our model addresses such scenarios by evaluating record sequences and alerting on high likelihood of erroneous replication. The metrics consider attribute characteristics, the distances between consecutive values, and the likelihood of value transition. The potential contribution of the model is demonstrated with a preliminary evaluation of 200 real-world records collected in an Obstetrics unit in a large hospital. A model trained with 100 records reached accuracy level of 85% with detecting data acquisition flaws in the other 100. The manuscript introduces the model development, describes the preliminary evaluation with real-world data, and highlights directions for future research progress.

• **Information Systems → Database Management Systems** → Data Cleaning • **Medical Information Policy** → Medical Records

Key Words: Data Quality (DQ), Data Accuracy, Electronic Medical Record (EMR), Distance Metrics

1. INTRODUCTION

Many business scenarios require tracking transitions in the state of real-world entities, by maintaining a sequence of data records that reflect state audits at different time points – e.g., service representatives who verify customer details, mechanics who keep record of routine vehicle maintenance activities or, as in the evaluation context of this study, physicians who document repetitive patient visits using an electronic medical records (EMR) system. A common data quality (DQ) issue in such scenarios is failure to reflect correctly state transitions. Due to time and workload constraints, the person who reacquires the data may choose to update the values only for a subset of the required attributes, while replicating previous values of others. As a result, the record sequence might fail to capture critical transitions, and reflect the current state incorrectly.

An example for scenario as such is illustrated in Table 1, which reflects Obstetric data collection in an EMR system. Patients under risk are often required to visit the hospital a few times during their pregnancy period, for physician evaluation. The evaluation may require recording a large number of data items. To save physicians' time, the EMR system permits replication of records from previous visits, and the physicians are required to update mandatory attributes, so that they reflect transitions in the patient's state properly. However, due to time and workload constraints, unintended human errors might occur. In the sequence described in Table 1, for example, the first record reflects a patient's visit that was recorded correctly. However, the second record reflects some apparent data errors (marked red), which can be attributed to a replicated record (e.g., by a "Copy and Paste" operation) that was not updated carefully. The "Diagnosis" field still shows "Headache", while the "Visit Summary" no longer indicates headache complaints, and this diagnostic should have been removed. On the other hand, the "Visit Summary"

still suggests the patient is "Not taking any medications" however this sentence does not reflect the patient's state in the second visit since the "Medications" field indicates that a new medication is taken, and it should have been deleted.

Table 1: An Example of Incorrect Data Replication in a Sequence of Electronic Medical Records

Date	**Gestational Age**	**Medications**	**Diagnosis**	**Visit Summary**
28.3.14	32+2	-	Vertex; Headache	Pregnancy age 32+2. Not taking any medications. Complains on strong headaches in the last few days.
28.3.14	37+1	Zinnat, 500mg	Vertex; **Headache;** Reduced Fetal Movements	Pregnancy age 37+1. **Not taking any medications.** Complains on reduced fetal movements.

DQ defects, such as those described above might have negative consequences; hence, scenarios that might lead to such defects out to be addressed and investigated. With the rapidly-increasing accumulation of data resources and the growing reliance on those resources for supporting business operations and managerial decision-making - DQ is broadly recognized as a major issue in information systems (IS) management. Research has addressed DQ from many different perspectives – organizational, economical, procedural, technical, and others [Madnick at al. 2009]. DQ is typically perceived as a complex multidimensional concept [Panahy et al. 2013], where each DQ dimension – e.g., completeness, validity, and currency – reflects a different type of quality hazard [Even and Shankaranarayanan 2007]. The scenario described above can be associated with the dimension of accuracy, or correctness – the extent to which data values reflect correctly the state of the associated real-world entity [Batini et al. 2009]

The issue of DQ is of major concern in the healthcare sector, in which the common practices typically involve comprehensive data collection [Weiskopf and Weng 2013]. Clinical data, such as documentation of patient visits and evaluations, are typically documented in EMR systems. Quality defects in clinical data, and particularly in EMR systems that are used for many purposes across the board, might endanger patients' safety, introduce major risks to their well-being, and decrease the quality of care [Bowman and Rhia 2013] – that's besides other major negative implications for healthcare management and clinical research. Clinical data in EMR systems is typically collected by variety of officials – e.g., physicians, nurses, hospital clerks, pharmacy technicians, laboratory workers - with different levels of commitment to DQ [Kahn et al. 2012]. Moreover, clinical environments tend to be stressed and the medical-staff members are required to allocate their time between direct care of patients, communication with patients, and administrative tasks [Ammenwerth and Spötl 2009]. As a result of those time and workload constraints, the recording and updating of clinical data often suffers from severe DQ issues – such as accuracies, inconsistencies, and missing values [Wechsler et al 2013].

A common solution in EMR systems, for aiding clinical data entry under time constraints, is the ability to replicate previous records. This solution is often implemented with a built-in "Copy-and-Paste" utility that replicated a previously-stored record that describes a patient's visit, to be used as a template for recording the next visit. An assumption that underlies this feature is that the majority of details collected during a visit will not change between two consecutive visits of the same patient (e.g., patient's height, demographic details, or permanent medical conditions); hence, the replication may offer some major time-saving potential. Research has already pointed out that such a solution, while facilitating faster data collection, has the potential to harm DQ, due to incorrect or insufficient update of the copied values [O'Donnell et al. 2009]. Ideally, older diagnostics should be deleted while new ones added; however, in reality, the information is not always updated properly due to lack of time or attention [Hirschtick 2006]. It is typical for medical staff to replicate previous notes visit after visit "as is", even in cases where the previous contents should have been altered. As a result, the notes keep lengthening and DQ defects accumulate [Siegler & Adelman 2009].

This research was conducted in collaboration with the Obstetrics & Gynecology department of a large Medical Center. Patients with pregnancy complications often visit the Labor & Delivery ward a few times during their pregnancy period for consulting and medical evaluation, and their visits are documented in the hospital's EMR system. Members of the department's medical staff have pointed out the potential hazards of record replications in the EMR system that supports their ongoing operation, and the example in Table 1 above is actually based on a real-world case in that department. Errors such as that one are common, and the department's management seeks for solutions that will help reducing, or at least alerting upon the detection of such DQ defects in the EMR system.

This study contributes to that end by developing a model that will help detect DQ defects in series of records that describe states of the same entity at different time points. The development was motivated by the problem raised by the members of the Obstetrics & Gynecology department and will be evaluated in that context, using real-world data samples provided by the department. However, the model is stated in a generalized form that would potentially fit other data management scenarios. The model alerts on potential DQ errors, considering a few factors: a) Level of structuration: is the attribute under evaluation structured (e.g., a single value, selected from a list), semi-structured (e.g., the selection of multiple values, in random order), or unstructured (e.g., free-text). b) Distance between attribute values: measurements that reflect how different are the value recorded in a consecutive data records, and c) Likelihood of value transition: assessment of whether attribute values are expected to change dynamically over time, or stay permanent.

The following section presents the model development, stating its formulation and underlying assumptions. This is followed by a preliminary model evaluation, conducted with real-world data received from the Obstetrics & Gynecology department. The evaluation used 200 record sequences – 100 used for training the model, and 100 for testing that reached an accuracy level of 85% with detecting data acquisition flaws. As the study is still under progress, the concluding section summarizes its contributions so far, and proposes directions for future development.

2. MODEL DEVELOPMENT

The model aims at the detection of DQ defects in a sequence of data records that describe the same entity, with the same set of alpha-numeric attributes, at different points of time. The model addresses scenarios with possible record replications – i.e., a "Copy and Paste" of a record to be used as a template for the consecutive record. The model assigns each record in the sequence with a DQ grade that reflects an estimation of its correctness. Following a typical DQ measurement schema [Even and Shankaranarayanan 2007], the DQ grade is assigned with a value between 0 (low quality) and 1 (high quality). The model evaluates each alpha-numeric attribute independently, and grades the DQ of the entire record as a weighted-average of record grades. As further discussed in the concluding section - future extensions may consider attributes that are not alpha-numeric (e.g., pictures, document scans), and possible interactions between attributes.

The DQ grade per attribute considers its characteristics, the distance between consecutive attribute values, and the likelihood of state transition over time. Some more detailed explanations on these factors are given in the next sections.

2.1 Attribute Characteristics

The model considers the following characteristics of alpha-numeric data attributes:

Level of Structure: Attributes can be classified, at a high level, according to their level of structure (examples in Table 2).

1. **Structured:** Attribute with a single value that conforms to a pre-specified value domain.
2. **Semi-Structured:** Attribute that contains multiple-values, taken from a pre-defined list, with no limitations on order or number of items selected.
3. **Unstructured:** A free-text attribute that does not necessarily conform to a rigid value domain.

Table 2: Levels of Attribute Structure

Height (cm.)	**Diagnosis**	**Visit Summary**
165	Pregnancy; Vertex; Reduced Fetal Movement	31 years old. Felt reduced fetal movements in the last week.
Structured	Semi-Structured	Unstructured

Level of Volatility: Attributes can be classified, at a high level, according to the likelihood of their values to transition over time, within a given sequence of data records (examples in Table 3).

1. **Stationary:** Attribute with values that are likely to remain unchanged.
2. **Dynamic:** Attribute with values that tend to transition over time.

Table 3: Stationary versus Dynamic Attributes

Date	**Drug Allergies**	**Gestational Age**
28.1	Penicillin	24 + 6
11.2	Penicillin	26 + 1
25.3	Penicillin	35 + 2
	Stationary	*Dynamic*

2.2 Distance

The distance between the values of the same attribute in two consecutive records is a quantitative measurement of the variation, expressing the degree of difference between the values. The model defines distance as a non-negative number, normalized to a [0, 1] range. The model accommodates different forms of distance definitions, depending on the attribute's level of structure. The following definitions relate to the distance between the values of a certain attribute (indexed [k]), in two consecutive records (indexed [j] and [j-1], where j > 1), in a certain sequence of records (indexed [i]) that reflects samples of the same entity at different time points.

1. **Structured Attributes:** for an attribute [k] with a value domain defined as a bounded range $[X^k_{min}, X^k_{max}]$, the distance between $X^k_{i,j}$ and $X^k_{i,j-1}$ is defined as:

$$d^k_{i,j} = \left| \frac{X^k_{i,j} - X^k_{i,j-1}}{X^k_{\max} - X^k_{\min}} \right| \quad (1)$$

   If the value domain of attributes [k] is a discrete set of N possible values $\{V^k_n\}_{n=1..N}$, the distance can be defined as d=0 if $X_1 = X_0$, or d=1 otherwise.

$$d^k_{i,j} = \begin{cases} 0, & if \quad X^k_{i,j} = X^k_{i,j-1} \\ 1, & otherwise \end{cases} \quad (2)$$

2. **Semi-Structured Attributes:** in this case, the value domain of attributes [k] is also defined as a discrete set of N possible values $\{V^k_n\}_{n=1..N}$. Both $X^k_{i,j}$ and $X^k_{i,j-1}$ contain a subset of those values, and the distance is defined as:

$$d^k_{i,j} = \left| \frac{I\left(X^k_{i,j}, X^k_{i,j-1}\right) + R\left(X^k_{i,j}, X^k_{i,j-1}\right)}{I\left(X^k_{i,j}, X^k_{i,j-1}\right) + R\left(X^k_{i,j}, X^k_{i,j-1}\right) + U\left(X^k_{i,j}, X^k_{i,j-1}\right)} \right| \quad (3)$$

   Where

   I - The number of "inserted" values, which exist in $X^k_{i,j}$ but not in $X^k_{i,j-1}$

   R - The number of "removed" values, which exist in $X^k_{i,j-1}$ but not in $X^k_{i,j}$

   U - The number of "unchanged" values, which exist both in $X^k_{i,j-1}$ and $X^k_{i,j}$

   Notably, if the number of items in both $X^k_{i,j}$ and $X^k_{i,j-1}$ is 1, this formulation folds back to the distance definition in Equation 2, for a structured attribute with a discrete value domain (i.e., d = 0 if values are identical, 1 if not).

3. **Unstructured Attributes:** the Edit Distance defines the distance between two strings ($X^k_{i,j-1}$, $X^k_{i,j}$) as the minimum number of insertions (I), removals (R) or substitutions (S) of characters needed to transform string $X^k_{i,j-1}$ into string $X^k_{i,j}$ [Backurs and Arturs 2015]. In this study, the distance is defined at the word-count level, rather than at the character-by-character level. The distance is normalized to a [0, 1] by diving the total number of insertions, removals, and substitutions by W, the number of words in the union of both values.

$$d_{i,j}^{k} = \left| \frac{I\left(X_{i,j}^{k}, X_{i,j-1}^{k}\right) + R\left(X_{i,j}^{k}, X_{i,j-1}^{k}\right) + S\left(X_{i,j}^{k}, X_{i,j-1}^{k}\right)}{W\left(X_{i,j}^{k}, X_{i,j-1}^{k}\right)} \right| \quad (3)$$

The distance effect on the DQ grade is influenced by the attribute's level of volatility. With static attributes the value is not expected to change; hence, the DQ grade generally decreases with greater distance. On the other hand, with dynamic attributes the value is expected to change; hence, the DQ grade is likely to increase with greater distance. Obviously, the sensitivity to distance may change between attributes; hence the distance effect is formulated as:

$$Q_{i,j}^{k} \propto 0.5\left(1 - Y^{k}\right) + Y^{k} * \left(d_{i,j}^{k}\right)^{\alpha_k} \quad (4)$$

Where

$Q^{k}_{i,j}$ - Quality grade for attribute [k] of record [j], in sequence [i]

Y^{k} - Indicates whether attribute [k] is static ($Y^{k} = -1$) or dynamic ($Y^{k} = 1$)

$d^{k}_{i,j}$ - Distance between two consecutive values, $X^{k}_{i,j}$ and $X^{k}_{i,j-1}$

α^{k} - Distance sensitivity parameter, $0 \leq \alpha^{k} \leq 1$

This formulation has the form of $1-d^{\alpha}$ (Figure 1a) for static attributes and d^{α} (Figure 1b) for dynamic attributes, where the sensitivity grows with a greater α.

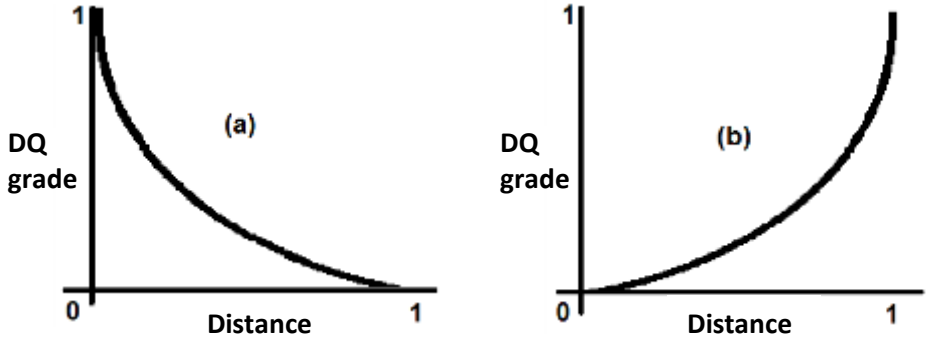

Figure 1. The Distance Effect on Quality for Static (a) versus Dynamic (b) Attributes

2.3 Time Sensitivity

Dynamic attributes can be also characterized by their time sensitivity – the likelihood to change within a certain period of time. Attributes that are highly time sensitive tend to change frequently, whereas attribute that are less time sensitive tend to change only after much longer time periods. For example, in the cases of Obstetrics data – attributes such as "Gestational Age" and "Referral Cause" are likely to change between visits; hence, both should be assigned with high time sensitivity. On the other hand, attributes such as "Drug Allergies" are not likely to change between visits; hence, should be assigned with relatively low time sensitivity. The effect of time is formulated as follows:

$$Q_{i,j}^{k} \propto \left(d_{i,j}^{k}\right)^{\Delta t_{i,j}^{\beta^k}} \tag{5}$$

Where

$Q^k_{i,j}$ - Quality grade for attribute [k] of record [j], in sequence [i]

$d^k_{i,j}$ - Distance between two consecutive values, $X^k_{i,j}$ and $X^k_{i,j-1}$

$\Delta t_{i,j}$ - Time gap between records [j] and [j-1] in sequence [i], normalized to a [0, 1] range

β^k - Time-gap sensitivity parameter

The time gap $\Delta t_{i,j}$ between records [j] and [j-1] in sequence [i] is determined by the time stamp that indicates when the records were acquired. The model assumes that the time gap is normalized to a [0, 1] by dividing it with the maximum possible time gap - a constant determined by the characteristics of the specific evaluation context. For example, in the context of visits during pregnancy period - the time gap is measured in days, normalized by division by 280, the standard pregnancy duration in days. The general assumption is that as the greater is the time gap between records, the greater is the likelihood of transition. Attributes that are more likely to change within a given time gap, are said to have greater time sensitivity. The sensitivity of attribute [k] is represented by β^k, a non-negative parameter, where $0 < \beta^k < 1$ indicates high sensitivity, and $\beta^k > 1$ indicates weak sensitivity.

2.4 Model Formulation

Based on the definitions above, the DQ grade is defined as:

$$Q_{i,j}^{k} = 0.5 * \left(1 - Y^{k}\right) + Y^{k} * \left(d_{i,j}^{k}\right)^{\alpha^k * \Delta t_{i,j}^{\beta^k}} \tag{6}$$

Where

$Q^k_{i,j}$ - Quality grade for attribute [k] of record [j], in sequence [i]

Y^k - Indicates whether attribute [k] is static ($Y^k = -1$) or dynamic ($Y^k = 1$)

$d^k_{i,j}$ - Distance between two consecutive values, $X^k_{i,j}$ and $X^k_{i,j-1}$

α^k - Distance sensitivity parameter, $0 \leq \alpha^k \leq 1$

$\Delta t_{i,j}$ - Time gap between records [j] and [j-1] in sequence [i], normalized to a [0, 1] range

β^k - Time-gap sensitivity parameter

At the record level, the DQ grade $Q_{i,j}$ for record [j] in sequence [i] is defined as a weighted average of attribute grades:

$$Q_{i,j} = \sum_{k=1}^{K} C^k * Q_{i,j}^k \tag{7}$$

Each attribute [k] is assigned with a weight C^k that reflects its relative importance. For maintaining the record quality grade within a 0-1 range, total sum of weights is 1 ($\sum_{k=1..K} C^k = 1$).

3. PRELIMINARY EVALUATION

The model will be evaluated with a real-world dataset that reflects patients visits over a 4-year period at the Obstetrics & Gynecology department. The dataset was retrieved by the department's IT personnel, while maintaining strict anonymity and eliminating any details that could potentially identify the patients. Further, the evaluation was monitored by the hospital's 'Helsinki Committee', and received its approval. The dataset contained approximately 100,000 records grouped into sequences, where each sequence reflects repetitive visits of the same patient at different time points during her pregnancy period. Each record contains a large number of attributes that reflect the patient demographic details, clinical measurements, and physicians' evaluation during the visit.

For the purpose of preliminary evaluation, a total of 277 records from 77 sequences were chosen (3.6 records per pregnancy sequence on average), each record containing 18 different attributes. Three out of the 18 attributes were used as identifiers (Patient ID, Visit Date and Visit Reference Number); hence, were not included in the quality evaluation. The first record of each pregnancy sequence was not graded, but used as reference point for defining the DQ grade for the next record. The remaining 200 records were classified manually by a physician from the department into defected (a record with certain DQ issues) or correct (a record with no DQ issues). The records were divided randomly into training versus a test set (100 records in each). The training set was used to assess the model's parameter (α and β per attribute in Equation 6), and setting the alert threshold. The relative weights of each attribute ($\{C^k\}$ in Equation 7) were set after consulting with the department's physicians, based on their assessment of possible consequences for potential errors in each attribute. The alerting threshold value was set to 0.85, in order to balance between maximum defected records identified as defected, and minimize false alarms.

Table 2. Model Classification versus Physician Classification

		Physician's Classification		
		No DQ Issues	**To be Corrected**	**Total**
Model Classification	**No DQ Issues**	TN: 74	FN: 7	**81**
	To be Corrected	FP: 8	TP: 11	**19**
	Total	**82**	**18**	**100**

The results of the preliminary evaluation are summarized in Table 4. Out of 100 records in the test set, 85 were classified correctly (classification accuracy of 85%). Out of the 18 record that were classified by the physician as suffering from DQ issues 11 could be detected by the model (classification sensitivity of 61%). Most FP cases were caused by a value of a dynamic attribute that was copied from a previous record (i.e., unchanged) and still reflects correctly the patient's state that remained unchanged. Most FN cases were caused by an error in an attribute's value that was caused by a different reason other than a data replication. Both FP and FN are expected to be improve by considering also the dependencies between attributes and will be further discussed in the conclusions section.

Obviously, while the preliminary results were encouraging, they are insufficient for a robust model validation. As the study progresses - the intention is to use the dataset received from the hospital for a comprehensive evaluation of the model and for developing it further.

4. CONCLUSIONS

With the rapid accumulation of data resources and the growing dependency of organizations on the use of those resources for running operations and supporting decision making – the need for tools and techniques that will help monitoring and improving DQ are on the rise. This study contributes to that end by developing a model that addresses the evaluation of a sequence of data records that reflect samples of the same certain entity at different time points. A preliminary evaluation of the model used a dataset received from the Obstetrics & Gynecology department of a large hospital. A subset of that dataset was used for training the model (i.e., to estimate the values of its parameters). Other records were then used to test the model, toward assessing its classification performance.

The study is still in progress, and will be further extended, where the following are considered as possible directions:

- Extending both the training and the test datasets, so that they reflect a much larger sample of real-world records.
- Extending and refining the model, so that it addresses possible dependencies between attributes (e.g., the values of a certain set of attribute will help detecting DQ defects in others).
- Considering other forms of distance metrics that will apply not only syntactic comparison between attributes values, but also on a semantic comparison, based on relationship between attributes.

Approaching those direction will hopefully help turning the model into a useful tool, and assessing its potential contribution.

REFERENCES

Ammenwerth, E., and Spötl, H.P. 2009. The time needed for clinical documentation versus direct patient care. Methods of Information in Medicine 48 (1): 84-91.

Backurs, A., and Indyk, P. 2015. Edit distance cannot be computed in strongly sub-quadratic time (unless SETH is false). In the Proceedings of the 47th Annual ACM on Symposium on Theory of Computing.

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. 2009. Methodologies for data quality assessment and improvement. ACM Computing Surveys (CSUR) 41 (3): 16.

Bowman, S.M.J, and Rhia, C.C.S. 2013. Impact of electronic health record systems on information integrity: Quality and safety implications. Perspectives in Health Information Management: 1c.

Even, A., and Shankaranarayanan, G.. 2007. Utility-driven assessment of data quality. ACM SIGMIS Database 38 (2): 75-93.

Hirschtick, R.E. 2006. Copy and Paste. JAMA 295 (20): 2335-2336.

Kahn, M.G., Raebel, M.A., Glanz, J.M., Riedlinger, K. and Steiner, J.F. 2012. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Medical Care 50 (Suppl. Jul): S21-9.

Madnick, S.E., Wang R., Lee, Y., and Zhu, H. 2009. Overview and framework for data and information quality research. Journal of Data and Information Quality 1 (1), Article no. 2.

O'Donnell, H.C., Kaushal, R., Barrón, Y., Callahan, M.A., Adelman, R.D., and Siegler, E.L.. 2009. Physicians' attitudes towards copy and pasting in electronic note writing. Journal of General Internal Medicine 24 (1): 63-8.

Panahy, P., Shariat, H., Sidi, F., Affendey, L.S., Jabar, M.A., Ibrahim, H., and Mustapha, A. 2013. A framework to construct data quality dimensions relationships. Indian Journal of Science and Technology 6 (5): 4422-31.

Siegler, E.L., and Adelman, R. 2009. Copy and paste: A remediable hazard of electronic health records. The American Journal of Medicine 122 (6): 495-6.

Wechsler, A., Even, A., and Weiss-Meilik, A. 2013. A model for setting optimal data-acquisition policy and its application with clinical data. In the proceedings of the 2013 International Conference on Information Systems (ICIS), Milan, Italy.

Weiskopf, N.G., and Weng, C. 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Information Association 2013 (20): 144-151