WHITE PAPER

The Role of Data Profiling In Health Analytics

An Encore Point of View | *Randy L. Thomas* | *October 2013*

AN ENCORE POINT OF VIEW

Using data captured in the course of providing healthcare has never been more important. The shift from fee-for-service to fee-for value demands that healthcare organizations understand the quality and cost of the care they provide to the communities they serve. To succeed in the new world of at-risk contracts and responsibility for not only clinical outcomes but the overall health status of defined populations, data-driven decisions in support of performance improvement has never been more critical. Healthcare organizations must harvest the data from across its portfolio of implemented applications and put it to use – re-purpose it – in pursuit of healthier people at optimal cost. This requires that organizations integrate and aggregate data from across the continuum, even including data from affiliated organizations. And the shift to electronic clinical quality measures (eCQM) included in at-risk contracts and required by regulation requires discrete, consistent and reliable data.

AN ENCORE POINT OF VIEW, cont.

Health care organizations have been building data warehouses of various sorts (clinical, financial, enterprise) for a number of years. These initiatives typically start with high hopes for the reporting, measurement and analytics the collected data and associated technology will support. The reality, sadly, often does not meet expectations. Usually, as the first reports and dashboards start rolling out from the warehouse, the organization will hear cries of "this can't be right" and "I don't believe the data." Why? What goes wrong? Is it the data model or reporting tools or ability to access the data?

The answer to the last question is most likely "no." No data model is perfect, every reporting tool has certain limitations and data accessibility must be balanced with HIPAA, other privacy and compliance requirements; but these issues aren't the primary reasons that expectations are unmet. Often the data populating the warehouse just isn't as expected. Data dictionaries, organizational standards, and pick lists for data entry fields may describe the intent of a particular data field but don't guarantee that the data captured in the source system actually reflects that intent. Front-end users are incredibly inventive at finding ways to circumvent edit checks and data field requirements to expedite completing a transaction (e.g., registering a patient) as the pressure to speed throughput increases. It is also unwise to assume that data complying with an HL7 transaction standard is 100% as expected. Yet eCQMs and the analytics supporting performance improvement require consistent, reliable data. To ensure the data landing in a data warehouse –or even in an analytic application– is as expected, an organization needs to profile and evaluate that data prior to first moving it.

Key Points for Data Reusability

- ***Direction*** *– have an analytics roadmap*
- ***Governance*** *– enterprise-wide accountability for data assets*
- ***Integrity*** *– profile data to assess reliability*
- ***Improvement*** *– workflow remediation to improve data reliability*

WHAT IS DATA PROFILING

Data profiling is the activity that examines the format and values of data and helps an organization understand exactly the state it's in. Data profiling describes data in a way in which the data's strengths and weaknesses become apparent.[1] It is the statistical analysis and assessment of the data values in source systems (e.g., EHR, ADT) for consistency, uniqueness and logic. Profiling evaluates the actual content, structure and quality of the data by exploring relationships that exist between value collections both within and across data sets (e.g., valid phone numbers or ICD-9 procedure codes).

Data profiling will convey:

- Characteristics of the data compared to what is expected in a field
- Facts about the data (e.g., occurrence of null values)
- Fields that need further investigation

Data profiling does have limitations, however. It will not convey:

- Accuracy of the data (e.g., procedure miscoded)
- Rules that apply to the data
- Efficiency of data capture process

There are two aspects of data profiling – quantitative and qualitative. The first can be done with any number of data profiling tools – or even spreadsheets – but the second requires subject matter expertise specific to the type of data (e.g., a clinician to evaluate validity of clinical data).

TWO ASPECTS OF DATA PROFILING

There are two aspects of data profiling – quantitative and qualitative. The first can be done with any number of data profiling tools – or even spreadsheets – but the second requires subject matter expertise specific to the type of data (e.g., a clinician to evaluate validity of clinical data).

QUANTITATIVE DATA ASSESSMENT

Quantitative assessment identifies the following:

- **Type** *(e.g., numeric or text)*
- **Format** *(e.g., nn.n or mm/dd/yyyy)*
- **Frequency** *(e.g., count of null values, count of all 9s in a field)*
- **Reference Table Matches** *(e.g., discharge disposition codes)*

This quantitative evaluation results in a set of statistics about the data. Example:

Field Name	NULL	Missing	Actual	Completeness	Cardinality	Uniqueness	Distinctness
Customer ID	0	0	3,338,190	100.00%	3,338,190	100.00%	100.00%
Account Number	0	0	3,338,190	100.00%	3,254,735	97.50%	97.50%
Customer Name 1	50,072	16,690	3,271,428	98.00%	2,997,864	89.81%	91.64%
Customer Name 2	2,450,670	53,077	834,443	25.00%	798,531	23.92%	95.70%
Tax ID	886,703	41,444	2,410,043	72.20%	2,120,837	63.53%	88.00%
Gender Code	1,204,060	50,264	2,083,866	62.43%	8	0.00%	0.00%
Birth Date	627,019	0	2,711,171	81.22%	25,275	0.76%	0.93%
Postal Address Line 1	196,536	5,193	3,136,461	93.96%	2,886,753	86.48%	92.04%
Postal Address Line 2	2,349,569	42,966	945,655	28.33%	875,578	26.23%	92.59%
City Name	171,517	15,171	3,151,502	94.41%	29,876	0.89%	0.95%
State Abbreviation	723,865	0	2,614,325	78.32%	72	0.00%	0.00%
Zip Code	925,591	0	2,412,599	72.27%	48,731	1.46%	2.02%
Country Code	0	0	3,338,190	100.00%	5	0.00%	0.00%
Telephone Number	515,781	0	2,822,409	84.55%	2,624,840	78.63%	93.00%
E-mail Address	1,204,608	0	2,133,582	63.91%	2,037,570	61.04%	95.50%

Figure 1. Quantitative Evaluation Results

The data source has been processed by a data profiling tool, which has provided the above counts and percentages that summarize the following field content characteristics:

- **NULL** – count of the number of records with a NULL value
- **Missing** – count of the number of records with a missing value but not null (e.g., space instead of a number or letter or punctuation mark)
- **Actual** – count of the number of records with an actual value (i.e., non-NULL and non-missing)
- **Completeness** – percentage calculated as Actual divided by the total number of records
- **Cardinality** – count of the number of distinct actual values (i.e., value occurs only once, not repeated; e.g., unique patient ID)
- **Uniqueness** – percentage calculated as Cardinality divided by the total number of records
- **Distinctness** – percentage calculated as Cardinality divided by Actual

The variety of both source systems and the data captured in these systems is quite broad in healthcare.

QUALITATIVE DATA ASSESSMENT

Qualitative aspects of data profiling involve a subject matter expert manually examining the data values for rational values. For example, a field in an EHR labeled "temperature" may contain data in the correct type and format (nn.n – nnn.n) but the range of values may exceed what is logical from a clinical perspective – such as a temperature exceeding 500. Or a field labeled "vaccine dose" might comply with reference table values and expected percentage of null values, but the dose amount may be incorrect (e.g., 0.5 ml vs. 5 ml). Qualitative analysis can also identify data fields that capture similar information in different ways (duplicative fields). For example, a field in an EHR labeled "Tobacco Type" may have a pick list for the type of tobacco utilized (e.g., cigarettes, oral, cigar, pipe). The same EHR may also contain individual fields that indicate a Yes/No response for field names "Cigar use," "Oral Tobacco Use," "Pipe use." A subject-matter expert would identify these duplicate fields during a qualitative assessment.

Other industries may not have the same need for the qualitative aspects of data profiling as healthcare. The variety of both source systems and the data captured in these systems is quite broad in healthcare. Other factors have contributed to the variability of actual data values compared to expected data values, particularly when it comes to clinical systems or any system when time is scarce and the pressures to encourage adoption or complete transactions is great.

Often documentation and consistent training for the intent of a field is minimal, confusing or does not exist. This leaves the end user the role of interpreting the intent of the field – resulting in inconsistency across a health system. Inconsistent change control practices may also lead to creation of duplicate or ill-defined fields.

An example of how variation can occur in a seemingly simple and straightforward field is a telephone number.

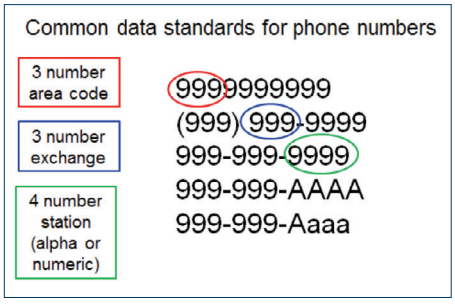

Figure 2. Telephone Number Data Standards

In these examples, "9" represents any digit (i.e., number), "A" represents any upper case alpha (i.e., letter) character, and "a" represents any lower case alpha character. All may be valid formats and data types in different applications but to bring this information into a common repository, this variation needs to be understood and transformed into a common format to enable consistent re-use of the data.[2]

Linked with appropriate enterprise-wide data governance and ongoing data management processes, data profiling will set any organization on the fast track to driving actionable information from the data it collects.

IMPROVING DATA RELIABILITY

In the case of EHRs, system adoption frequently trumps standardization of data capture. Often, as long as the providers can find the data they need to enable clinical decision-making and plan interventions, standardizing where and how data is captured isn't paramount.

> *But, when data is moved to an enterprise data warehouse to support analytics and measurement, inconsistencies are uncovered that render the information suspect. Data profiling can identify these issues prior to moving the data.*

Then decisions can be made about whether the data can be "fixed" as it is moved (e.g., multiple different code sets that mean the same thing mapped to a standard code set) or workflow can be modified at the front end to ensure consistent, accurate data capture moving forward.

Decisions can also be made about how firmly to "lock down" a data field – such as making it mandatory and prescribing a set of values (ensuring discrete data).

> *But this rigidity of data capture needs to be used judiciously; users will find their way around what they perceive are barriers to completing their tasks (e.g., registering a patient, documenting a clinical finding) as quickly as possible.*

Often, the path to convince end users to change workflow and accept perceived limitations on valid data values is through their frustration at not being able to obtain the analysis and measurement using the data they collect. Only when they can tangibly see the reward will the "pain" be worth it!

VALUABLE ENTERPRISE ASSET

Healthcare organizations are now coming to the realization that data is a valuable enterprise asset. Time and money invested in implementing EHRs and other systems at the front line of patient care are great sources of data to support analytics and measurement for performance improvement and other initiatives. To ensure the data is accurate, reliable and consistent, data profiling can establish a realistic picture of the current state of the data and help guide efforts to improve its re-usability.

Linked with appropriate enterprise-wide data governance and ongoing data management processes, data profiling will set any organization on the fast track to driving actionable information from the data it collects.

REFERENCES

1. Elliott King, Data Quality 101: The Ultimate Guide for Data Stewards, A MELISSA DATA eBook; Downloaded from http://www.melissadata.com/whitepaper/data-quality-stewards-ebook.asp
2. Data Profiling: The Foundation for Data Management; DataFlux; October, 2003.

ABOUT ENCORE

Encore, A Quintiles Company, is one of the most successful consulting firms in the health information technology (HIT) industry. Founded in 2009 and led by Encore CEO Dana Sellers and President Tom Niehaus, the company provides consulting services and solutions that assist its expanding client base with a wide range of HIT strategy, advisory, implementation, process-redesign, and optimization initiatives. Encore focuses on capturing the right data at the right time, establishing analytical capabilities that meet the evolving information and reporting needs of healthcare providers to document and improve clinical and operational performance. For more information about Encore, please visit *www.encorehealthresources.com*.