Success by Design: Effective Data Quality Measurements within a Hospital Data Warehouse

Richard E. Biehl, CSQE, CSSBB
Data Warehouse Architect
Data-Oriented Quality Solutions
Orlando, FL, USA rbiehl@doqs.com

2008 HIMSS Annual Conference, Session 142

This presentation describes a specific data quality strategy that was designed and implemented as part of the Mount Sinai Data Warehouse at the Mount Sinai Medical Center in New York. Requirements definition for the warehouse began in spring 2006, with design beginning in the winter 2006-2007. Development began in spring 2007, with Release 1.0 launch in December 2007. Release 1 started with major feeds from the ADT, CPOE, Labs, and Financial systems. By spring 2008, 12 additional systems are scheduled to be on-line as ETL feeds. There are over 100 systems targeted for inclusion over the next 3-5 years.

Rick Biehl can be reached at:

Data-Oriented Quality Solutions
2105 Whitfield Lane
Orlando, FL 32835 USA

407-296-6900

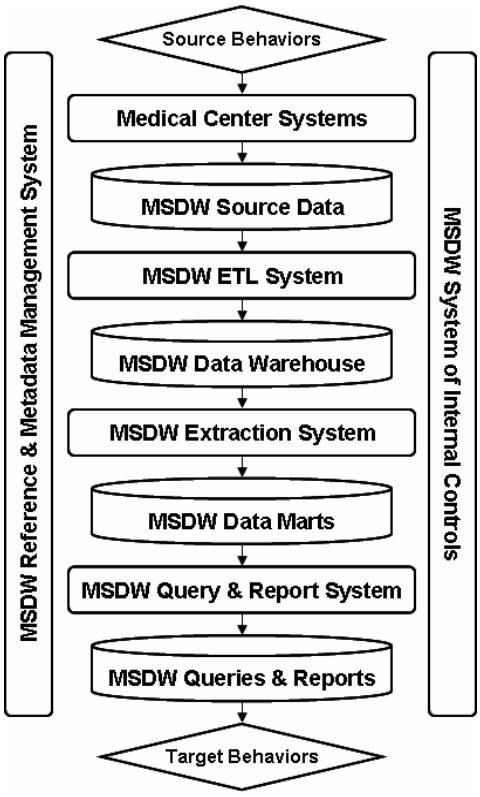
rbiehl@doqs.com

Learning Objectives

1. Recognize quality issues that affect the viability of healthcare data warehouses.
2. Define overlapping mechanisms that can minimize data warehouse quality risks.
3. Describe why bad data can't always be corrected at the source applications.
4. Articulate criteria for balancing data quality errors against data availability needs.
5. Adapt example generic quality rules to specific local settings and needs.

2008 HIMSS Annual Conference, Session 142

2



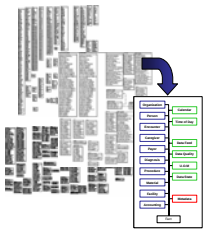
The intent of this presentation is to share the principles of design that allowed the Mount Sinai Data Warehouse (MSDW) to become an integral component of the medical center's strategy for monitoring and improving the quality of data and information across all of the IT systems in use in the provision of care.

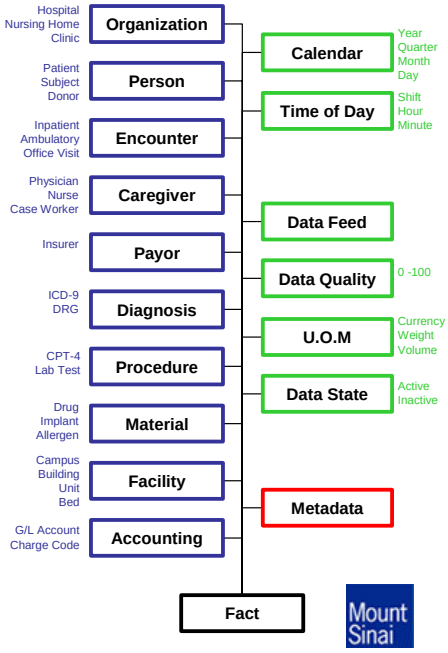
Data quality is part of the system of internal controls for the MSDW. (see left) Any care setting should be able to adapt the information presented here for inclusion in a local warehouse application.

For some smaller settings, the design described here might be considered an over-design; but our premise is that this design is generic enough to be implemented in almost any care-related organization, and that adapting an over-designed model would be far more effective than designing from scratch.

Mount Sinai Data Warehouse

- 17 dimensional star
- Oracle
- Business Objects
- ~1-2 million nightly facts
- ~2-7 years of history
- HIPAA-compliant





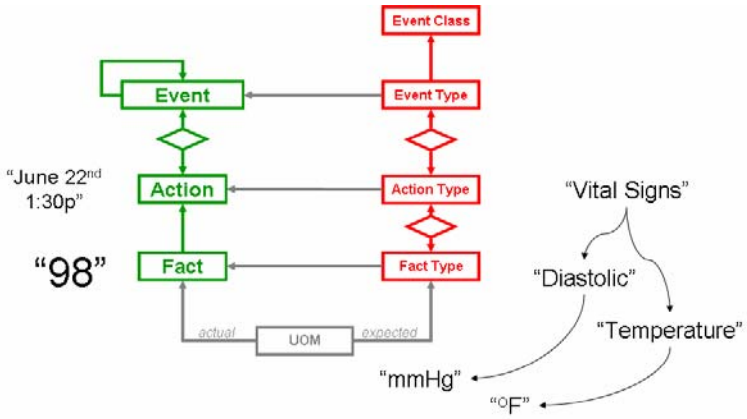
2008 HIMSS Annual Conference, Session 142

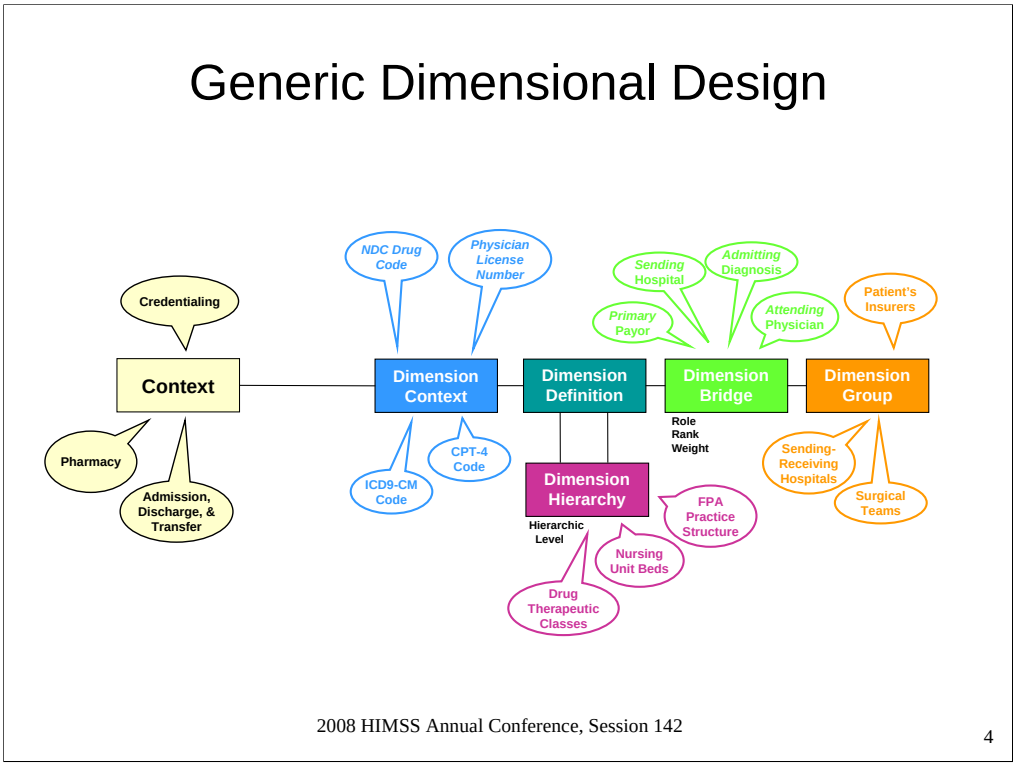
3

The data warehouse is built on a standard 17-dimensional star schema with a single Fact table containing a single Value column. All of the fact data loaded into the warehouse is loaded into this single column, and the foreign key pointers in the fact table establish the dimensional context for each fact.

Each of the dimensions falls into one of two categories: 1) Primary dimensions that provide clinical, financial, operational, and administrative context; and 2) Secondary dimensions that provide structural context, including date, time, and unit of measure.

Because all of the facts reside in the single Value column in the Fact table, the Metadata secondary dimension provides the business definition. Without the Metadata dimension, users of the warehouse would be unable to differentiate different data elements within the Fact table.

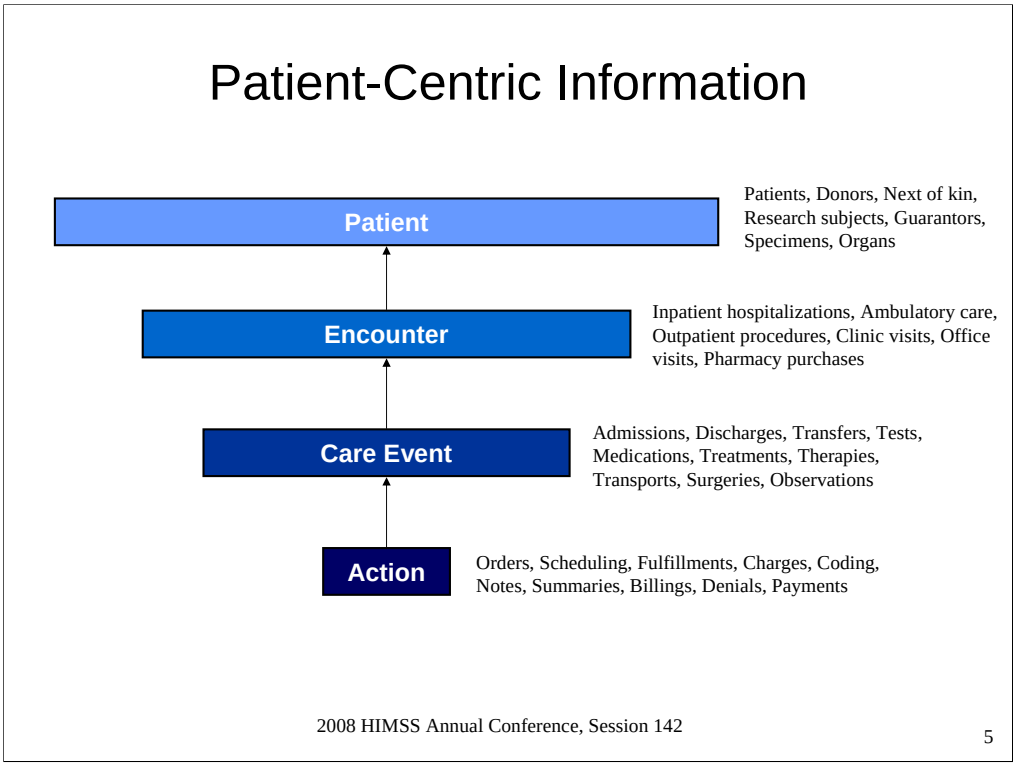




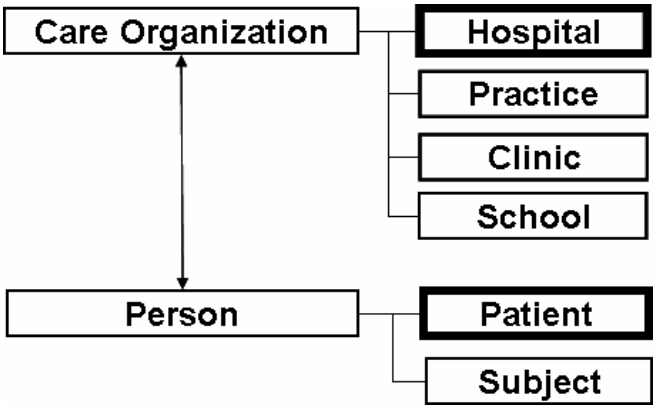
A central tenet of the design of the warehouse is that all dimensions should be structured in the same way, and exhibit the same behaviors. The overall understandability of the warehouse would be seriously impeded if differences occurred across the structures and functions of the dimensions. With so many distinct dimensions, the commonality of all design features and functional components allows users to understand and navigate the warehouse with a minimum of training and effort.

The common structure allows for definition of distinct items in the dimension (Dimension Definition), the clustering of multiple instances of dimensional items for reference by a single fact (Dimension Group) as well as the assignment of roles and ranks for items within those groups (Dimension Bridge), and mapping each instance in the dimension to the source system from which stored facts are received (Dimension Context).

These common capabilities are available for all of the warehouse dimensions, but are used to differing degrees within each dimension in the context of various forms of fact data stored in the warehouse.

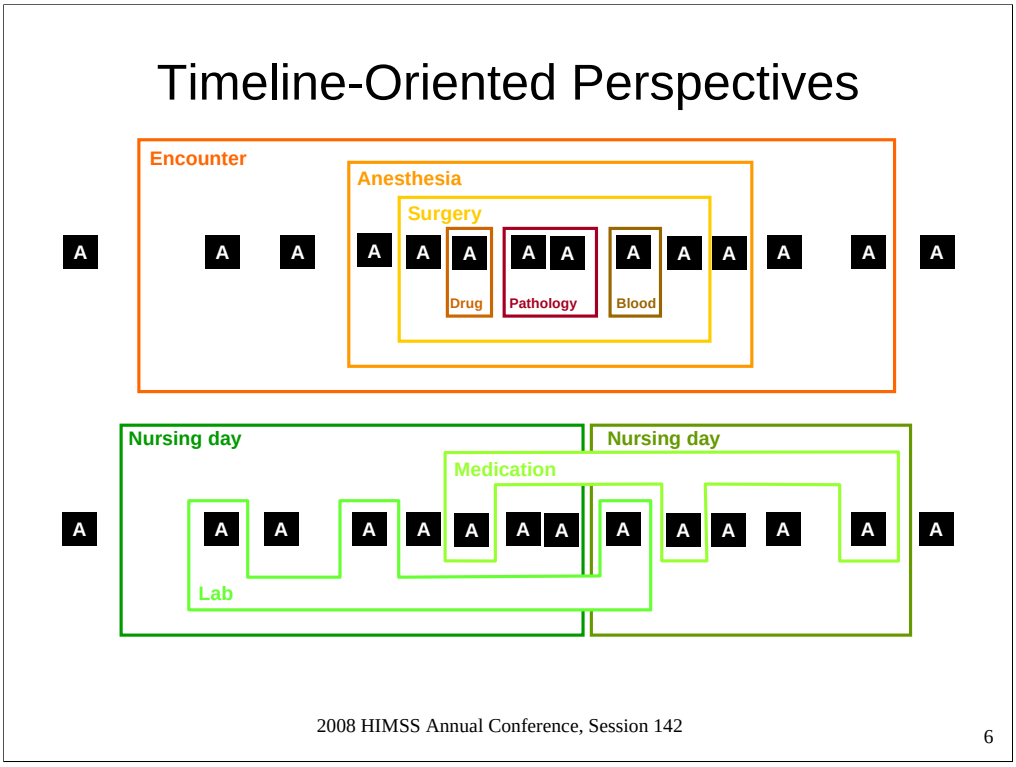


The warehouse provided a generic structure for storing information within and across the 17 dimensions. No dimension is preferred or highlighted in the capabilities offered by warehouse interfaces and tools.



However, the data in the warehouse is highly patient-centric today. For example, a user can ask about time intervals for patients moving in and out of beds, but can not yet ask about time intervals involved in cleaning of beds. This latter data is non patient centered, and so is not yet available in the warehouse.

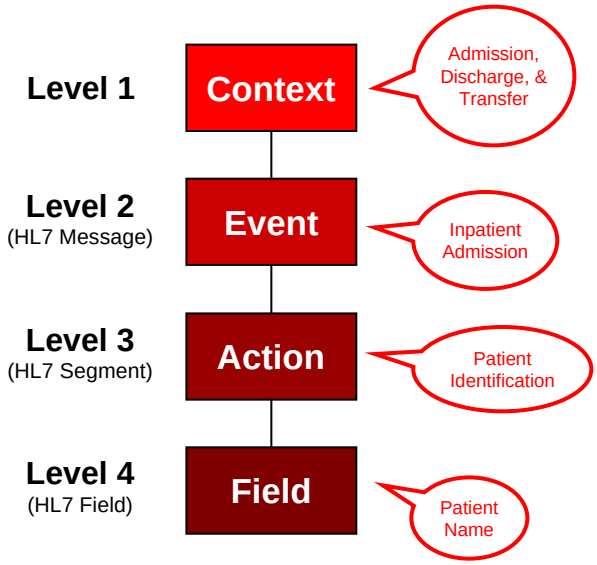
While the warehouse can be asked virtually any data-oriented questions, the warehouse can only answer questions that relate to patients today. In the future, we'll expand beyond patient data; but that focus will remain a centerpiece of the warehouse for the next several years.

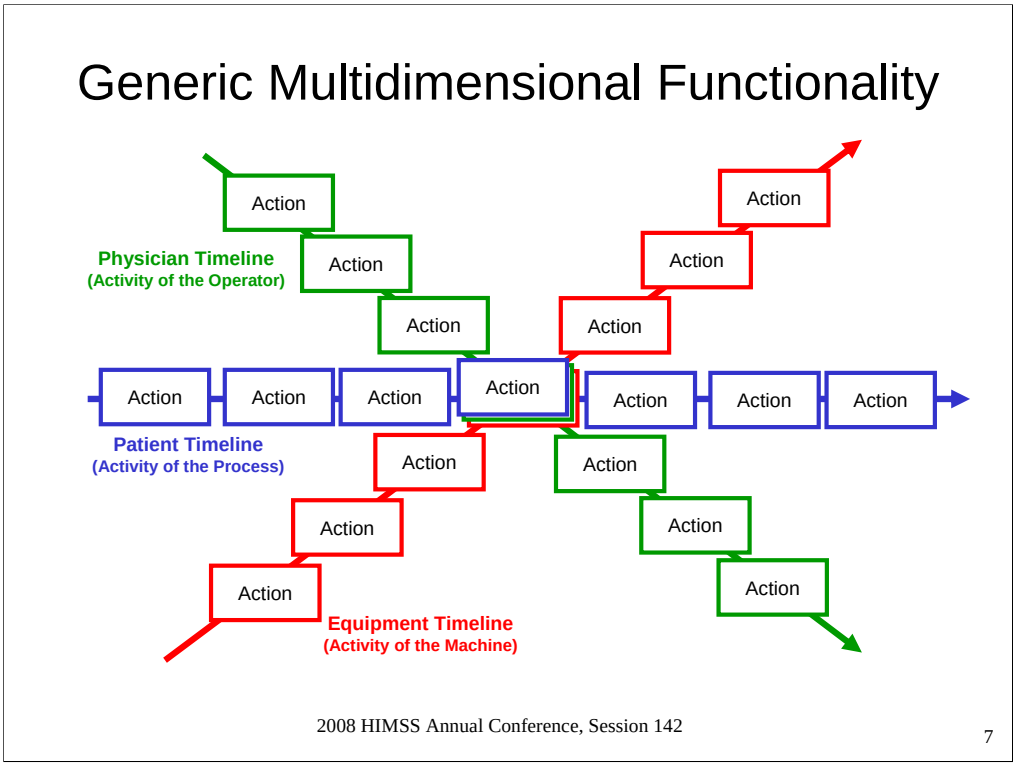


The warehouse design is centered around storing data for individual care actions. These care actions combine into events at multiple and varied levels of detail, including some very complex nested structures. Simple events nest into aggregate structures along a timeline.

The upper example on the slide illustrates a cascade of Encounter, Anesthesia, Surgery, Drug, Pathology, and Blood Events within each other. The result of this cascade is that certain actions, such as the action in the Blood Event, can be viewed as being part of multiple events depending upon the level of focus used in conducting analysis.

The warehouse is also designed to recognize inter-event relationships that don't represent simple cascades as described above. As illustrated on the lower example on the slide, events might overlap in more complex ways, particularly when events are defined by temporal boundaries rather than inter-action chains.





The warehouse stores facts about events and actions that take place throughout the medical center. Because each event or action takes place at a point in time, extracting multiple events or actions results in a time series of data for analysis or reporting. The density of data on the timeline will depend upon the breadth of the sources of data feeding the warehouse and the narrowness of the queries that are used for reporting the data.

While the data in the Mount Sinai Data Warehouse is patient-centric, not all timelines need to be. Timelines can be generated along any of the non-time dimensions defined for the data warehouse. Because the warehouse is defined along multiple dimensions, each fact in the warehouse could be viewed as occurring along a timeline for each dimension.

Each fact represents the convergence of the 17 dimensions that provide it context and structure, and it can be queried from the viewpoint of any of those dimensions.

Supplier-Customer Model (SIPOC)

SUPPLIERS	INPUTS	REQ'TS	PROCESS	OUTPUTS	REQ'TS	CUSTOMERS
Standards Organizations	Metadata	Stable	Mount Sinai Data Warehouse	Data marts	Accessible	MSDW Users
		Complete			Controlled	
	Reference data	Stable			Compliant	
		Aligned			Appendable	
Mount Sinai Leadership	Data policies	Comprehensive		Data queries & reports	Controlled	
	MSDW resources	Available			Meaningful	
		Adequate			Clear	
Data Administration	Performance & control settings	Stable			Compliant	
		Optimal			Traceable	
Application Systems (includes appended data marts)	Metadata	Stable		Data issues & hypotheses	Relevant	Data Owners
		Complete			Timely	
	Reference data	Stable			Actionable	
		Aligned			Traceable	
	Factual data	Compliant				
Data Owners	Reconciled data issues	Correct		Performance & control data	Auditable	Data Administration
		Authorized			Actionable	
					Compliant	

2008 HIMSS Annual Conference, Session 142

8

This slide shows the Six Sigma SIPOC for the Mount Sinai Data Warehouse. It shows all of the suppliers and inputs to the warehouse, as well as all of the outputs and customers of the process. Requirements are defined for each.

While the paramount purpose of the warehouse is to provide incoming factual data to users through queries and reports (green lines in slide), the primary focus of this presentation is on a feedback loop involving outputs of data issues and hypotheses that are then converted by data owners into reconciled data issues. The warehouse includes an expert workstation specifically design to aid data owners in identifying, tracking, and resolving data quality issues.

Data Quality Measurement

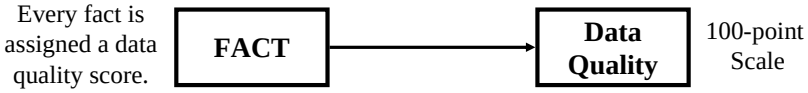
- Fact-level data quality indicators
- Generalized notification mechanism
- Hierarchy exception analysis
- Activity measurement error
- Time-series exception analysis
- Statistical process control

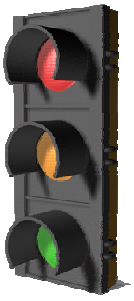
The warehouse functions and features include these six categories of tools and capabilities. Each interacts to provide an integrated and comprehensive layer of data quality monitoring and control. The sixth, SPC, is still in its infancy in our current release, but will continue to expand and broaden in the near future.

These defenses form a Swiss cheese pattern of controls through which data exhibiting quality problems can not pass. None of the controls is perfect, but collectively they form a robust quality shield that protects warehouse users from quality problems in the data and source applications.

Fact-Level Data Quality Indicators

Every fact is assigned a data quality score.





Red (0-30) – Significant problems known. Data must be explicitly requested, otherwise it is automatically omitted.

Yellow (31-70) – Problems suspected to be significant enough that data should not be used for certain analysis or reporting. Included unless requested to be excluded.

Green (71-100) – No significant problems.

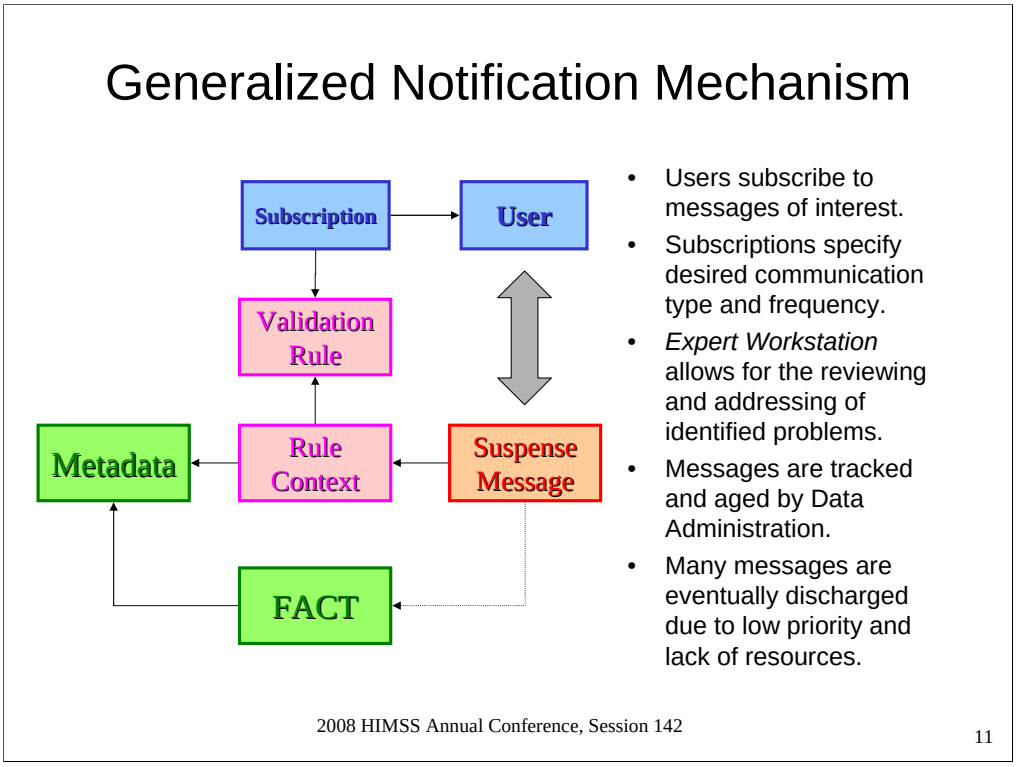
2008 HIMSS Annual Conference, Session 142

10

The first line of defense against users encountering data quality problems is the Data Quality indicator associated with every fact in the warehouse. Using a scale from 0 to 100, every fact is rated for data quality as it is extracted from its source and transformed for loading into the warehouse. Most data quality problems are minor, and would be readily recognized by healthcare IT professionals from almost any institution.

The Data Quality Indicator was specifically calibrated in recognition of the idea that the quality of the data across our source systems is obviously good enough to support normal day-to-day operations. While problems are numerous, few problems rise to a severity where use of our data is compromised. Recognizing this, the Data Quality Indicator was specifically calibrated to rarely fall below our Green, or acceptable, range.

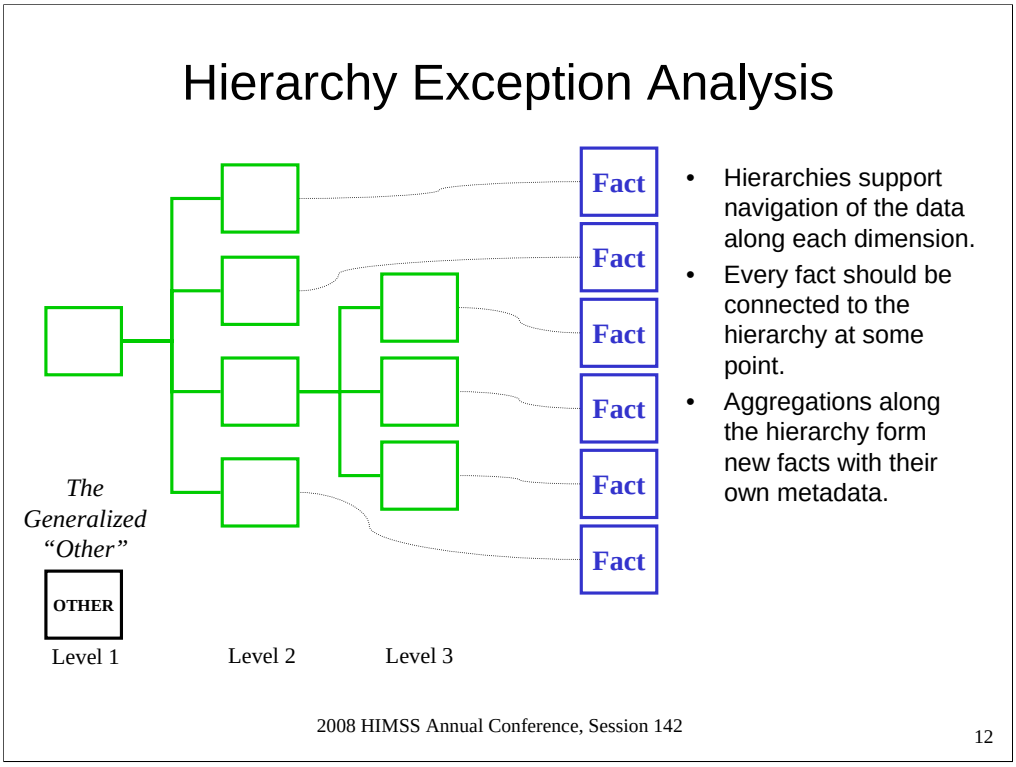
Typical data quality issues that result in a score of less than 100 include having to remove special characters from free-text entered clinical data (2 points), clinical activity by physicians not through the credentialing system yet (5 points), or orders and administration of drugs not yet in the material master system feeds (10 points). This data remains perfectly usable, and some of the errors automatically correct themselves as other data arrives in the warehouse.



The second line of defense is to let the appropriate users know when data quality issues have been identified. To accomplish this, individual users subscribe to the rule-based messages that they are interested in, and receive notifications when messages are created against individual rules.

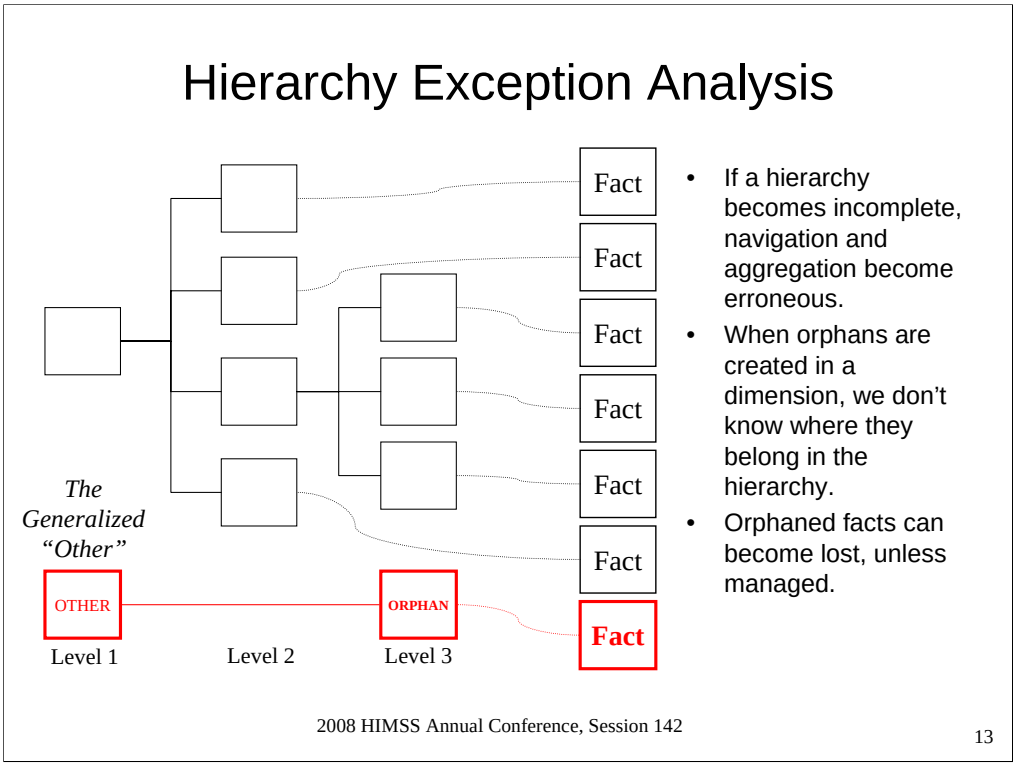
Each rule that might result in a data quality score decrement in the fact table also generates a suspense message whenever that decrement occurs. Users can browse the suspense registry to review individual or grouped messages, and can traverse the warehouse through these message to see the actual data that has each quality problem. Authorized users can even correct the data.

Data Administration staff monitor the aging of every message to assure that data owners are acting on suspense messages to which they are subscribed. It is also from this base of data quality messages that data quality reports are prepared for our Steering Committee and management.



A third line of defense is to ensure that every piece of data in the warehouse remains visible along every dimension, even when data quality problems are encountered. To do this, we've institutionalized an "Other" category within every dimensional hierarchy.

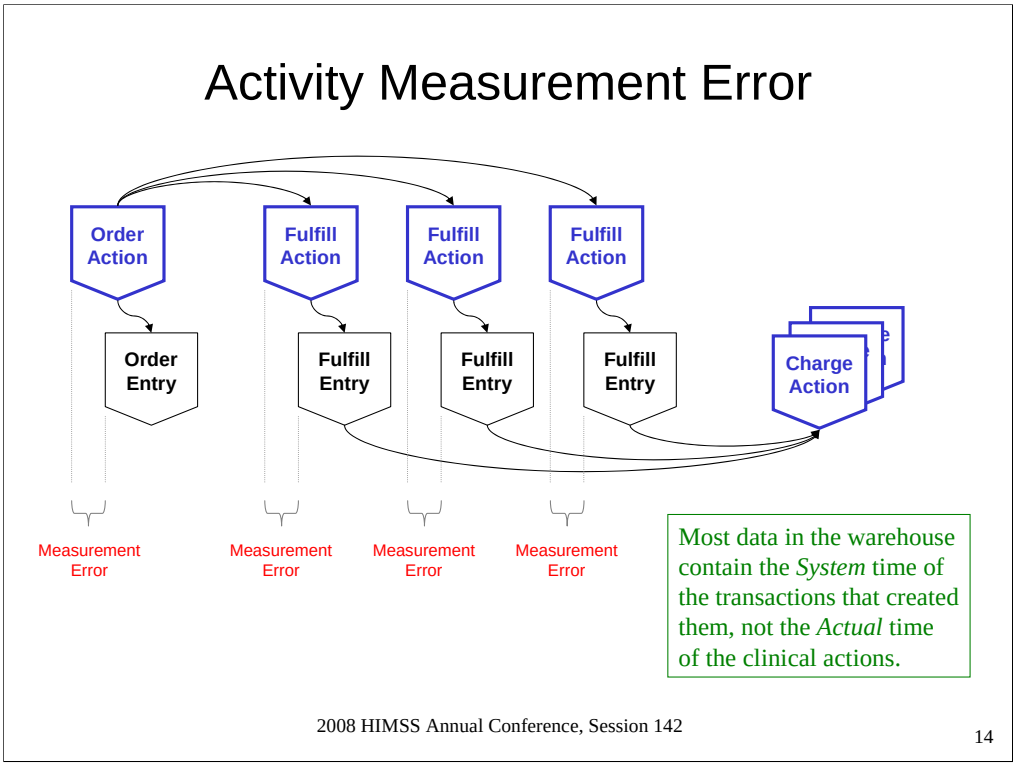
Typically, there are no facts associated with the branch of each hierarchy, and users would not even realize it is there. Now imagine that a new fact arrives with a data quality problem that prevents proper alignment of the fact to the hierarchy. Examples might include an unrecognized physician in the caregiver dimension, or an unrecognized drug in the material dimension. The warehouse will automatically create orphan entries in the necessary dimensions in order to store the incoming facts, and will decrement the quality scores of the arriving facts, thus generating suspense messages so that the data owners for the related data will be notified. But in the meantime, what happens to dimensional reporting? ….



…. These incoming facts are temporarily associated with the *Other* branch of the dimensional hierarchy. This is necessary because as orphan entries are created, the information needed to properly align the orphans with the hierarchy might not be present in the incoming transaction.

For orphan physicians, the physician's department or clinical specialty might not be known. For orphan drugs, the therapeutic class might not be known. Without this important reference data, new orphan entries can't be attached to the dimensional hierarchy at the correct points.

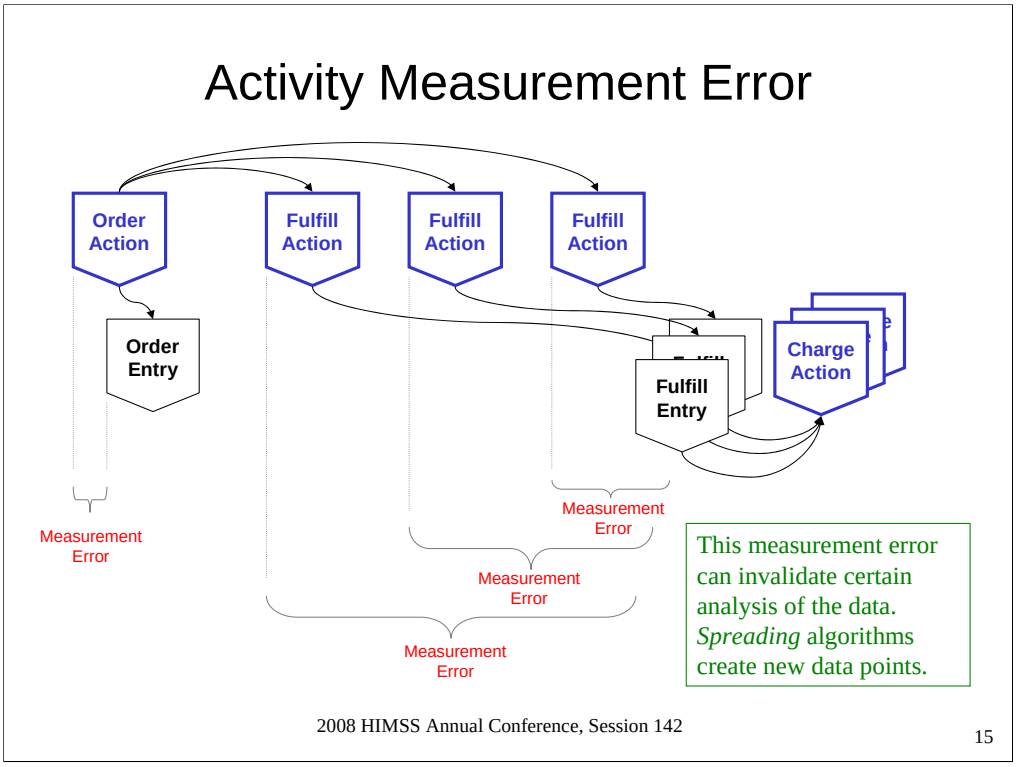
By connecting to the *Other* branch in the hierarchy, the data is not lost and users who query along traditional hierarchic lines can be warned if significant data begins to accumulate in the *Other* branch. As data owners adopt the orphan entries in response to suspense messages, all of this data eventually becomes connected to the correct branches of the hierarchy.



The fourth line of defense is the evaluation of the sequencing of similar events in the warehouse for measurement errors associated with data entry in the source application systems. Facts arriving at the warehouse for loading often contain the date and time that information was committed to the source system rather than the date and time at which the represented real-world action was taken.

In most situations the gap in time is negligible and a natural level of variation is expected and acceptable. However, larger time gaps can affect decision-making if the scale of measurement error becomes a significant relative to the time scales being analyzed by users.

The most extreme cases involve activity batching by system users…

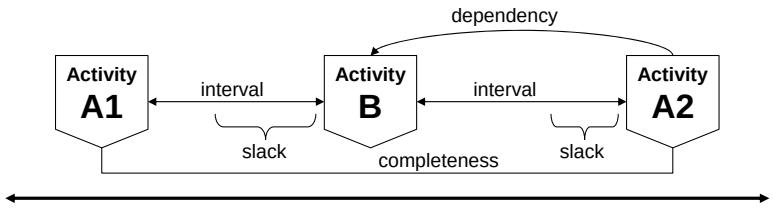


… In extreme cases, the time delays associated with committing transactions to source systems can overwhelm the actual data timings being recorded. Examples of such a situations could be a nurse on a floor unit dispensing medications to multiple patients on the unit before returning to the CPOE application to record the administrations, or a phlebotomist drawing blood from multiple patients before returning to the lab and accessioning the samples. The time delays between the actual activities and the recording of those activities become significant.

In the warehouse, these situations can be identified by the clustering of activity over timeframes that don't allow adequate time for the actual events to have occurred. When a single nurse enters seven or eight medication administrations in only one or two minutes, the warehouse recognizes that batching has occurred. By reviewing previous CPOE entries from the same nurse, a more accurate timeline can be estimated for when those events might actually have occurred. By spreading the data within the available window, the impact of the batching of transactions is reduced. Overspreading can reduce confidence more than the original batching might have. The spreading algorithm must be continually tuned in order to provide a natural spread that offers a high degree of confidence in the estimates.

Time-Series Exception Analysis

- Completeness Analysis
  – Are data for all the action steps there?
- Dependency Analysis
  – Are required previous events there?
- Interval Analysis
  – How far apart in time were actions accomplished, and did they occur in the expected order?
- Critical Path Analysis
  – Was there slack time that could have allowed things to happen sooner?



The fifth line of defense is the monitoring of activity sequences, dependencies, and timings. Suspense messages can be raised whenever the timing or patterns among recorded activities don't align with requirements or expectations.

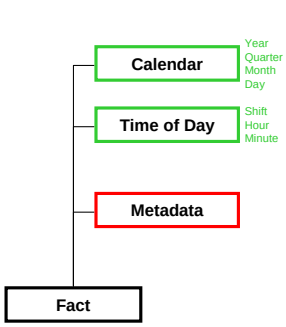
Common exceptions identified through these checks include encounter reservations for which patients don't arrive, clinical orders or results prior to admission, delayed transfers, orders not administered, administrations not charged, missing results, delayed or rescheduled procedures, delayed discharges, missing discharge plans, and late posted charges.

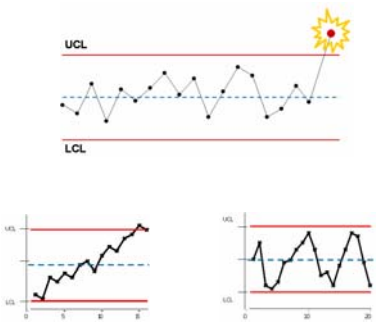
Particular attention is focused on event patterns that indicate that expenses are being incurred without clinical benefit. An example in this category is the conducting of lab tests, the results of which don't arrive until after discharge. Another, though harder to spot, is the ordering of unnecessary procedures that actually force a delay in discharge.

Many of the anomalies identified in this category can represent hypotheses that can ultimately be explained or rationalized, and are not actual errors. But the focus of the data quality program is to identify data quality problems. We prefer false signals to missed opportunities.

Statistical Process Control

- Control Analysis
  - Is the represented process under statistical control?
  - Are there multiple unexplained populations?
  - Are there outliers in the data?
  - Are there noteworthy trends occurring over time?



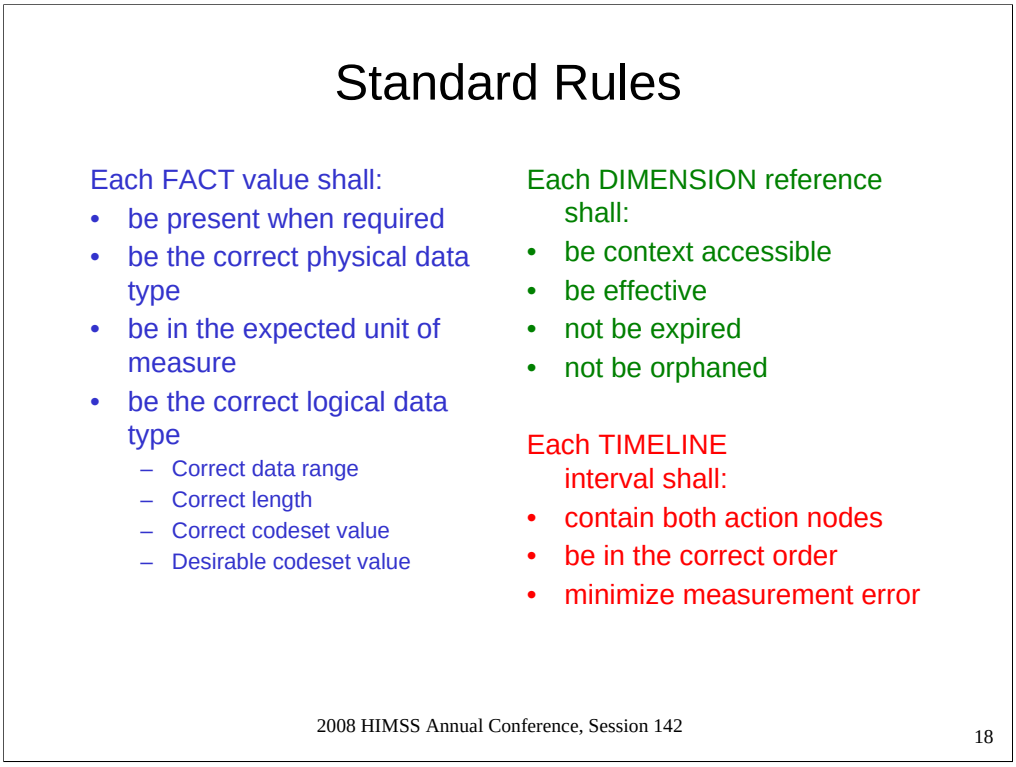


2008 HIMSS Annual Conference, Session 142

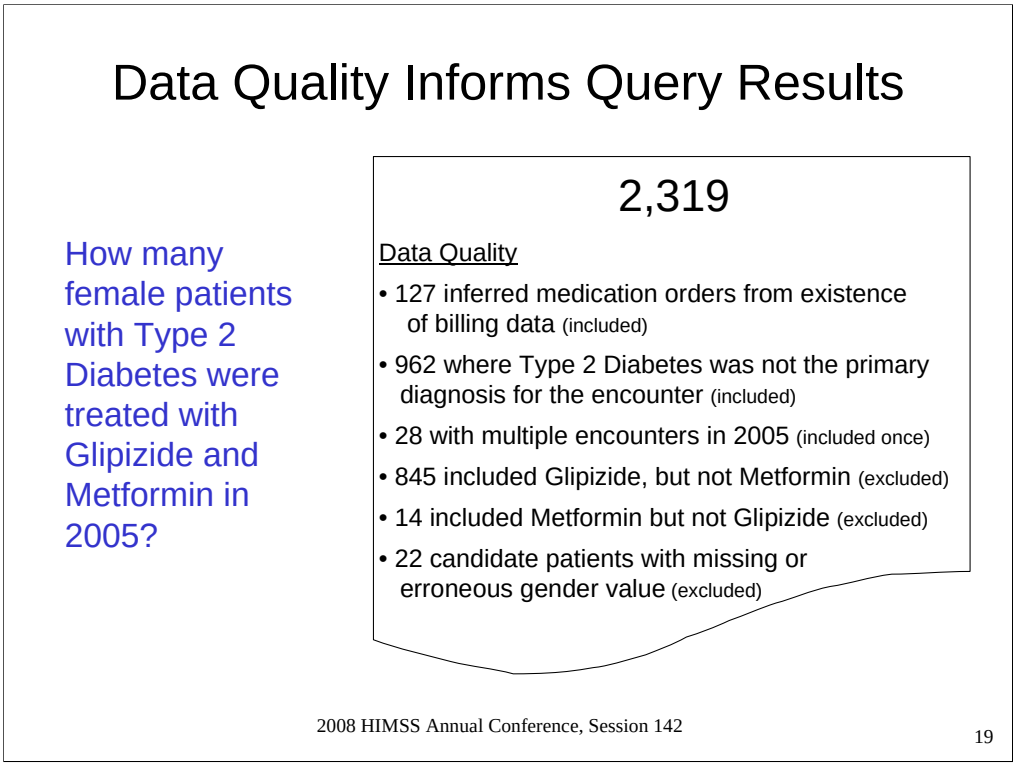
17

The sixth line of defense is Statistical Process Control (SPC). Because every fact occurs along each of the 17 dimensional timelines, all data in the warehouse participates in multiple time series. Quantitative data can be measured and monitored using SPC along each dimension to identify special causes in terms of trends and outliers. These special causes raise exception conditions that result in suspense messages being generated and sent to subscribing users.

Caveat – Full SPC capability is rudimentary today, and won't be fully available in production until summer 2008.



Because of the generic nature of the design of the warehouse, features available for the monitoring and control of data in one of the 17 dimensions is also available for the monitoring and control of data in the other 16 dimensions. As a result, all of the rules monitored and controlled in the warehouse can be described in only three standard layers: dimensional rules that control the structure and content of each dimension, fact rules that control what is placed in the fact table, and timeline rules that enforce the integrity of cross-action and multi-action events. Every data quality rule is a specific example of a rule in one of these three layers.



This slide shows a mock-up display for an example query in the warehouse. Every query in the warehouse has an opportunity to show data quality information that might be relevant to assessing the accuracy and usefulness of the query results obtained. It is up to the warehouse user to decide if the information provided is material to the usefulness of the query.

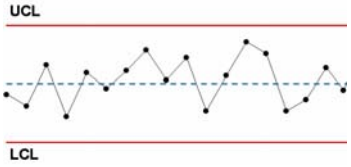
For example, the above query didn't specify whether or not Diabetes was to be included as a primary diagnosis or not. The 2,319 cases reported included 962 where Diabetes was not the primary. Since this number is large relative to the query result, a researcher might want to restate the query if primary diagnoses had been intended.

Also, the query had asked about patients, not encounters. The data quality report indicates that 28 patients that fit the criteria had multiple encounters. Since that number is very small relative to the query result, changing the query to ask for encounters rather than patients probably would not materially effect the researcher.

The 845 patients with only Glipizide might warrant further investigation before concluding this query is completely valid for its intended use.

Data Quality Informs Query Results

What kind of variability exists between the initial incision times recorded in the Anesthesia and Surgical Manager systems for surgeries within the hospital over the past month?



Data Quality

- 2,365 anesthesia-surgery events found (included)
- Anesthesia record not found for 12 surgical procedures (excluded)
- Incision times missing from 6 surgical procedure events (excluded)

2008 HIMSS Annual Conference, Session 142

20

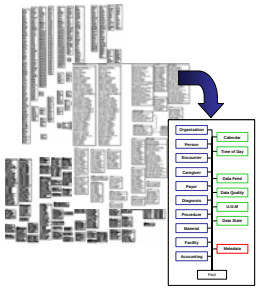
This is another example mock-up of a sample query with related data quality details. Since the data quality exceptions are minor relative to the overall data, a researcher might safely conclude that the data is in statistical control; and that variation between incision times recorded in our two systems is not uncontrolled. The differences in incision times appears systemic to the process.

Caveat – As of winter 2007-2008, we're still implementing the data quality report features. The capabilities demonstrated on this and the previous slide are currently working on a case-by-case basis, but a generalized data quality sub-report capability will not be available in production until summer 2008.

Our Goal: Self-Aware Proactive Data

"Declare the past, diagnose the present, foretell the future. As to diseases, make a habit of two things — to help, or at least to do no harm."

Hippocrates, *Epidemics*, Bk. I, Sect. XI



"To wrest from nature the secrets which have perplexed philosophers in all ages, to track to their sources the causes of disease, to correlate the vast stores of knowledge, that they may be quickly available for the prevention and cure of disease – these are our ambitions."

Sir William Osler
(1849-1919)

2008 HIMSS Annual Conference, Session 142

21

Taking our lead from the writings of Hippocrates and Osler, our focus in the Mount Sinai Data Warehouse is to first do no harm. While a great many data elements in the warehouse have an explicit or implicit margin of error associated with them, our mission includes making sure that those actual or latent defects don't impact our users, or their decision-making processes.

Users of the warehouse are free to use all of the data that the warehouse makes available, but that use is always tempered by the data quality measures and tools built into the data bases and query tools provided. Those tools continue to get better, and the algorithms used to detect data quality problems continue to improve. We're working to reduce the risk associated with decision-making using data that has been removed from the operational context of the source systems in which it was created.