

# A Framework for Designing a Healthcare Outcome Data Warehouse

*by Bambang Parmanto PhD,<sup>1,2</sup> Matthew Scotch,<sup>2</sup> and Sjarif Ahmad<sup>1</sup>*

## **Abstract**

Many healthcare processes involve a series of patient visits or a series of outcomes. The modeling of outcomes associated with these types of healthcare processes is different from and not as well understood as the modeling of standard industry environments. For this reason, the typical multidimensional data warehouse designs that are frequently seen in other industries are often not a good match for data obtained from healthcare processes. Dimensional modeling is a data warehouse design technique that uses a data structure similar to the easily understood entity-relationship (ER) model but is sophisticated in that it supports high-performance data access. In the context of rehabilitation services, we implemented a slight variation of the dimensional modeling technique to make a data warehouse more appropriate for healthcare. One of the key aspects of designing a healthcare data warehouse is finding the right grain (scope) for different levels of analysis. We propose three levels of grain that enable the analysis of healthcare outcomes from highly summarized reports on episodes of care to fine-grained studies of progress from one treatment visit to the next. These grains allow the database to support multiple levels of analysis, which is imperative for healthcare decision making.

Keywords: OLAP, healthcare, multidimensional database, data warehouse.

## Introduction

Many healthcare processes involve a series of patient visits or a series of outcomes. The modeling of outcomes associated with these types of healthcare processes is different from and not as well understood as the modeling of standard industry environments. For this reason, the typical multidimensional data warehouse designs that are frequently seen in other industries are often not a good match for data obtained from healthcare processes. Dimensional modeling is a data warehouse design technique that uses a data structure similar to the easily understood entity-relationship (ER) model but is sophisticated in that it supports high-performance data access.<sup>1</sup> Dimensional modeling refers to the process of designing the structure of a data warehouse through illustrations that show relationships between data tables. An example of dimensional modeling, also known as “star schema,” is illustrated below (Figure 1). The similarities between the traditional ER model and dimensional modeling are apparent: tables (or entities) have joins (or relationships) with other tables via primary keys. This method of data warehouse modeling has been used in standard industry for years for decision support in such areas as transportation, production, sales, and marketing. Data warehouse design for industries outside of healthcare is well understood and has been covered extensively.<sup>2-5</sup> Healthcare is well behind in the area of data warehouse management and decision support and needs to move forward in this direction.

The process of patient care in healthcare can be thought of as a value circle the center of this circle are data related to patient treatment.<sup>2,3</sup> The treatment is measured or generated by all the processes and organizations around the circle.<sup>2</sup> This is quite different from typical processes in other industries, which usually follow a linear chain model in which a product moves through a series of steps from raw material to finished goods or from customer order to delivery.<sup>2</sup> For this reason, healthcare does not match well with the standard methods used for multidimensional data warehouse design.

In much of healthcare, including outpatient rehabilitation services, the process involves a series of visits (as in many outpatient processes) or a series of outcome measurements (as in many inpatient processes). While this is not always the case in healthcare, multi-visit or multi-measurement processes constitute a significant portion of the industry. Healthcare processes that involve multiple health assessment measures across multiple patient visits make outcome analysis extremely difficult.

To address this problem, we implemented a slight variation of the dimensional modeling technique to make a data warehouse more appropriate for healthcare outcome research.<sup>2,3</sup> Our design process consists of the following steps:

- Define and choose the healthcare process to model.
- Choose the outcome grains of the healthcare process and define data marts and dimensions to capture the healthcare process.
- Define dimensions and their hierarchies.
- Define the facts (measures) that will populate each fact table, and design aggregation rules.
- Implement the design for a particular On-Line Analytical Processing (OLAP) system.

OLAP is decision-support software that allows the user to quickly analyze information that has been summarized into multidimensional views and hierarchies. A database that stores OLAP data has a multidimensional framework that can be represented as a data cube, instead of the typical tabular format seen in traditional databases. The cube contains dimensions, or types of information stored in the data warehouse. Figure 2 is an example of an OLAP cube for healthcare rehabilitation data. The dimensions are age, sex, clinic, ICD-9 code, year, and therapist. Each dimension has a hierarchy that defines the

dimension. For example, in this cube, the age dimension is defined as a series of age ranges and the specific ages that compose each age range. The ICD-9 dimension in this example is modeled after the International Classification of Diseases, Ninth Revision, with the lowest level representing the specific ICD-9 disease code and the highest level representing the broad disease category that the code falls under.

This paper will present a case study of a multidimensional database design for a data warehouse of healthcare rehabilitation outcomes at the Center for Rehabilitation Service at the University of Pittsburgh Medical Center. The primary purpose of the data warehouse is to support various outcome analyses of outpatient rehabilitation therapies. This paper presents a multidimensional database design that can be used as a blueprint for the development of a data warehouse for healthcare decision support.

## Case Description

The Center for Rehabilitation Service at the University of Pittsburgh Medical Center is the leading provider of outpatient rehabilitation in the greater Pittsburgh metropolitan area (Pennsylvania, USA). The center offers physical, occupational, and speech therapies in over 40 clinics throughout southwestern Pennsylvania. Patients are referred to the center by physicians based on a diagnosis or particular disability. A therapist is assigned to evaluate the patient's physical condition during the first visit and manages the therapy throughout the course of treatment. A typical episode of care consists of three to 10 visits with an average of two visits per week. When the goal of the treatment is achieved, the patient's therapy through the center ends.

Data are recorded in the database in two ways: by the primary therapist caring for each patient, and by the patients themselves. The data include basic demographic information such as gender, age, and ethnicity; diagnosis; and type of treatment provided. To monitor the progress of the therapy, all patients are expected to complete the standard form of the Medical Outcomes Study 36-Item Short Form Health Survey, popularly known as SF-36, which measures health-related quality of life.<sup>7, 8</sup> An episode of care is defined as a course of outpatient treatment that the patient receives, beginning with an initial visit and ending with a final visit (noted by the caregiver). Ideally, patients complete SF-36 at three stages of their outpatient episode of care: at the beginning (intake), at an interim visit, and at the end (discharge). SF-36 queries the patient on eight different health outcomes: energy/fatigue, general health perception, mental health, bodily pain, physical functioning, role limitation due to emotional problems, role limitation due to physical problems, and social functioning. Acceptable reliability and validity of SF-36 have been reported for its use in aggregate analysis.<sup>9-11</sup> All patients complete SF-36 regardless of diagnosis or area of treatment.

In addition to the general health measure of SF-36, patients are also expected to fill out a disease-specific questionnaire for each visit. Some of the instruments used are the Activities of Daily Living Scale (ADLS), a knee-specific outcome survey), the Oswestry Index (a measure of disability imposed by pathologies and impairments that affect the cervical spine), the Foot and Ankle Disability Index (FADI), and the Disabilities of the Arm, Shoulder, and Hand (DASH) questionnaire.<sup>12-13</sup> In all, 12 disease-specific outcome measures are used in the rehabilitation database. Patients' conditions are measured using only one of these outcome measures for each diagnosis.

Although healthcare rehabilitation outcomes are often used for quality assurance and many quality improvement techniques can be adopted from other industries, the study of healthcare rehabilitation outcomes is quite different from quality assurance in other industries. While other industries can use relatively simple measurements, such as number of financial transactions, number of total cellular minutes, or quantity of inventory received, healthcare outcome measurement is complex: SF-36 contains eight different measurements, and different diseases have different outcome measures with completely different scales.

## Outcomes Associated with Healthcare Rehabilitation Services

The processes that we will model for outcome analysis are episodes of care of patients who underwent treatment in all the rehabilitation facilities in the network. An episode of care begins when a patient arrives in the clinic with a referral from a doctor. Each patient can have one or many diagnoses that require treatment. Diagnoses are given codes called ICD-9 codes. The International Classification of Diseases (ICD), maintained by the World Health Organization (WHO), is used to assign codes to diagnoses and procedures. Each ICD-9 code can be associated with more than one body region. Each visit within an episode of care includes the recording of some type of outcome measure (evaluation of the patient). For this study, a complete episode of care is when outcome measures from different visits are available for a patient within the specified time frame. For each patient, two types of outcome measures were recorded: general health measures, such as SF-36, that apply to all patients, and disease-specific measures that apply only to patients with relevant diagnoses. For our analysis purposes, if there is no definite caregiver-noted final visit, an episode of care ends when there has been a lapse of visits for at least 45 days.

In other industries, dimensional modeling usually involves designing multiple star schemas to support a business with many processes (value chains). In modeling healthcare services, including outpatient rehabilitation, there are only episodes of care (value circles). These individual episodes, however, might have multiple levels of analysis. This is especially true since “multilevel” approaches to understanding and improving healthcare outcomes have gained acceptance.<sup>14</sup> Multilevel analyses provide facilities to link different levels of data analysis, from general-health-level outcomes, to disease-specific outcomes, to fine-grained analysis of progress over time. Outcomes measured for many healthcare processes can be divided into three grains (or levels) of information:

- *The episode-of-care outcome for general health measures.* This level of analysis looks at improvement on general health measures from the beginning of care to discharge. Absolute improvement scores are calculated using intake and discharge SF-36 scores. Individual or population patient scores are then compared to standardized normal scores for SF-36. Here, we are interested in patients in the aggregate, not group by body area affected. Since the level of analysis is episode of care, we are interested to know the progress from intake to discharge. At this level of analysis, we are not interested in the detailed progress from visit to visit.
- *The episode-of-care outcome for disease-specific measures.* At this level, we are interested only in the group of patients with relevant diagnoses. What factors influence disease-specific outcome measures? Similar to the general health measure level (above), we are interested to know the progress from intake to discharge. However, it has detailed measurements for specific outcome type. Since each outcome type has its own scale and its own interpretation, aggregation can only be done for individual outcome type. Aggregating outcome score across outcome type dimension is not valid.
- *Detailed outcome and treatment analysis for individual visits.* This level of analysis tracks detailed outcome scores and treatments from one visit the next. A *treatment* is a technique applied to a patient by a therapist in order to care for or deal with a medical diagnosis. For example, a “hot pack” may be considered a treatment. During an episode of care, a patient can have one or many treatments. The measurement used at this level is also specific outcome type.

If additional data (such as staffing, timing, etc.) are perceived to have an impact on outcomes, the dimensional modeling structure allows them to be easily included in any of the three grains.

Other healthcare-related data warehouses use different information grain pyramids. Berndt, Hevner, and Studnicki also created a data warehouse that allows for three levels of data analysis.<sup>15</sup> The lowest

level is transaction-oriented data, such as hospital discharge information, that correspond to individual events. The middle tier consists of aggregate data that can be viewed at different levels using data exploration functions unique to OLAP such as “drill up” (go to a higher level of data) and “drill down” (go to a lower or more detailed level of data). The topmost level of information in their data pyramid consists of reports on local community health indicators for high-level assessment purposes.

## Methods

In order to provide better analysis at different levels, we designed a multidimensional data warehouse that provides three levels (or grains) of data for healthcare outcome analysis (previously discussed and represented in Figure 3).

The episode-of-care grain relates to outcome analysis for general health measures. This grain captures measurements related to the entire length of care, from the patient’s first visit until discharge. One record in the star schema of this grain contains an entire episode of care. The measures in this star schema characterize the entire episode rather than individual visits or individual outcome measurements. In order to capture the dynamics of the episode of care, we divide each episode into three phases: intake, interim, and discharge.

The outcome summary grain relates to outcome analysis for disease-specific health measures. Its star schema is at the same level of detail as the episode-of-care grain. With this grain, as with the episode-of-care grain, we are interested in analyzing the progress from intake to discharge. However, it has detailed measurements for disease-specific outcomes. Since each diagnosis has its own scale and its own interpretation, aggregating outcome scores across diagnoses is not sensible. For example, the disabilities of the arm, shoulder, and hand (DASH) outcome is only recorded on patients with these types of conditions and thus is not applicable to all patients receiving treatment in the clinic. Thus, summarizing this outcome for all patients is not appropriate.

The transactional grain relates to outcome and treatment analysis for individual patient visits. It contains two star schemas: detailed outcome and detailed treatment. The most basic view of our healthcare system is at the individual transaction level, represented in healthcare as a patient visit. A patient visit generates several outcome measurements and several treatments. The purpose of the transactional grain is to allow detailed analyses that cannot be done with summarized data (for example, viewing a patient’s treatment on the initial visit to the clinic). Transactional schemas are also important in a healthcare data warehouse because traditional statistical analysis and data mining need granular, and unaggregated data that are only available in individual transactions.

## Findings

We implemented a rehabilitation outcome information pyramid with an OLAP system consisting of 78,000 episodes of rehabilitation care.<sup>16</sup> This large data warehouse supports analysis from the three different healthcare grains. For example, Figure 4 represents the middle level in the pyramid, summarizing outcomes for entire episodes of care. This particular example shows the average scores for the wrist and hand outcome at intake, interim, and discharge. This figure suggests improvement over the course of treatment, especially between interim and discharge.

While Figure 4 shows data related to a particular disease-specific outcome measure, Figures 5 and 6 represent the highest level in the healthcare information grain pyramid by displaying summaries of general health outcome measurements for all episodes of care. One important strength of this system is

that it allows researchers to conduct multilevel analysis by analyzing data at the general health level and the disease-specific level simultaneously. The system can be used to compare improvements in health-related quality of life across diagnostic categories and across different body areas. Figure 5 shows standardized SF-36 scores of patients at intake, at an interim visit, and at discharge. A baseline of SF-36 scores for general population (the general public) across different ages has been developed.<sup>6-7</sup> A standardized effect size of the outcome score can be calculated against this baseline score. The population is represented in the graph with a value of zero, and thus any score below zero represents an outcome worse than the population, while a greater score is better than the population. It shows in general how the patients in the clinic are progressing from disability to wellness. The figure illustrates that at the beginning, patients were substantially worse off than the general population in the physical and social components of the health-related quality of life measurement. By the time of discharge, patient conditions were almost identical to the population.

The design of the data warehouse enables researchers to create subsets of the total population and compare them statistically. Thus, patients' progression can be compared between groups based on age, diagnostic conditions, clinic, therapist, number of visits, race, gender, and so forth. Figure 6 illustrates the progression shown in Figure 5 for a particular ethnic group (African-Americans). Although the graph shows good discharge scores in general health, mental health, and vitality, it shows less impressive improvement in other physical and social measurements. The improvements in outcome that do occur on the remaining measurements still do not bring the patients in line with the general public, which is represented by zero on the graph.

The system also allows researchers to analyze the subtle and detailed progression of patient outcomes by tracking scores on disease-specific measurements. Figure 7 demonstrates the comparison of patient scores based on initial severity, and represents the lowest level of the healthcare information pyramid. The classification of initial severity is based on the method developed by Stineman and Granger.<sup>17</sup> The score used in this knee-specific outcome is the Activities of Daily Living Scale (ADLS). ADLS is a measure of functional limitation imposed during activities of daily living by pathologies and impairments affecting the knee. Higher scores indicate higher levels of physical function. Knee patients' initial conditions were divided into three categories based on their mean ADLS score when compared to one another. The categories are:

- "Most severe" patients, those receiving an ADLS score that ranks them in the lowest 25 percent when compared to the other knee patients
- "Moderate" patients, those receiving an ADLS score that ranks them at the 25th to 75th percentiles when compared to the other knee patients
- "Mild" patients, those receiving a score that ranks them in the highest 25 percent (or better than 75 percent) of knee patients

The figure shows that patients in the most severe group (red bars in the graph) exhibited significant improvement during the first few weeks of treatment, while patients in the upper quartile (yellow bars) exhibited no improvement. All patients' scores appear to level off with time.

Figure 8 shows a different pattern of progression, also based on initial severity; this type of analysis can be useful for making decisions related to the treatment of patients based upon interim condition scores<sup>17</sup>. Lower scores indicate higher levels of disability, and therefore positive absolute improvements indicate improvement (or reduction in disability). Thus, in this example, patients with less severe and mild initial conditions show no improvement or even negative effects of treatment; only patients with the most severe conditions show significant improvements over time. The numerical grouping of patients by severity enhances the decision-support system by going beyond categorical groupings that are predefined by dimension attributes and allowing researchers to see how patient groups defined by health severity respond to treatment. For outcome researchers, this type of analysis is essential for the decision-making

process.

## Conclusion

One of the key aspects of designing a multidimensional database for healthcare is finding the right grains for different levels of analysis. We have proposed three levels of grain that enable the analysis of healthcare outcomes from highly summarized reports on episodes of care to fine-grained studies of progress from one visit to the next. This hierarchy of grains allows the database to support multiple levels of analysis, which is imperative for quality healthcare outcome research. The powerful outcome analysis supported by this data warehouse design has the potential to reduce the length of patients' episodes of care, increase the quality of care, and lead to better health-related outcomes.

Bambang Parmanto, PhD is an assistant professor of health information management in the School of Health and Rehabilitation Sciences at the University of Pittsburgh in Pittsburgh, PA.

Matthew L. Scotch, MA, is a National Library of Medicine fellow in Bethesda, MD.

Sjarif Ahmad, MS, is a database manager at the Department of Health Information Management and Center for Clinical Pharmacology, University of Pittsburgh in Pittsburgh, in Pittsburgh, PA.

## Acknowledgement

We would like to thank Drs. Jay Irrgang and Anthony Delitto for their collaboration in this project.

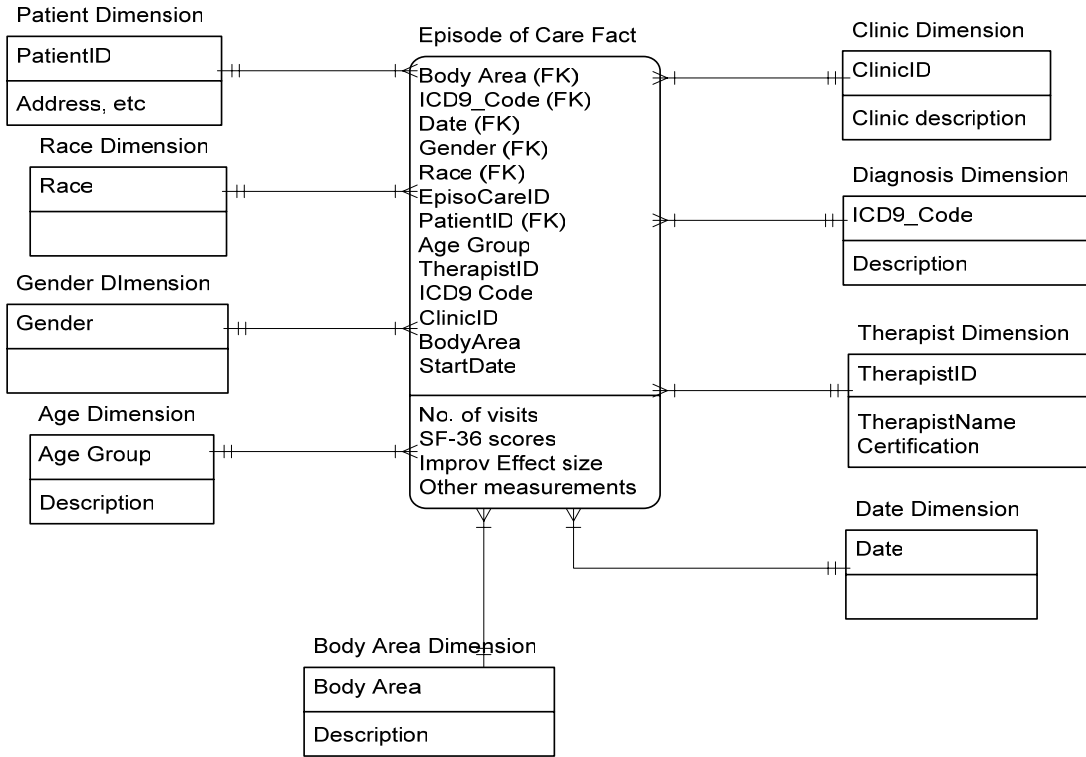
## Notes

1. Kimball, R. "A Dimensional Modeling Manifesto." *DBMS* 10, no. 9 (1997): 58–60, 69–70.
2. Kimball, R. *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses*. New York: Wiley, 1998.
3. Kimball, R. and M. Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. 2nd ed. New York: Wiley, 2002, pp. xxiv, 436.
4. Barquin, R. and H. Edelstein (Editors). *Planning and Designing the Data Warehouse*. Upper Saddle River, NJ: Prentice Hall, 1997, p. 311.
5. Sperley, E. *The Enterprise Data Warehouse: Planning, Building, and Implementation*. Vol. 1. Upper Saddle River, NJ: Hewlett-Packard, 1999, p. 333.
6. Scotch, M. and B. Parmanto. "SOVAT: Spatial OLAP Visualization and Analysis Tool." In *Proceedings of HICSS-38*. Waikoloa, HI: IEEE, 2005.
7. Ware, J. E., Jr. "SF-36 Health Survey Update." *Spine* 25, no. 24 (2000): 3130–39.
8. Ware, J. E., Jr., and C. D. Sherbourne. "The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual Framework and Item Selection." *Medical Care* 30, no. 6 (1992): 473–83.
9. Brazier, J. E. et al. "Validating the SF-36 Health Survey Questionnaire: New Outcome Measure for Primary Care." *British Medical Journal* 305, no. 6846 (1992): 160–64.
10. McHorney, C. A. et al. "The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of Data Quality, Scaling Assumptions, and Reliability across Diverse Patient Groups." *Medical Care* 30, no. 1 (1994): 40–66.
11. Ruta, D. et al. "The SF 36 Health Survey Questionnaire: A Valid Measure of Health Status." *British Medical Journal* 307, no. 6901 (1993): 448–49.

12. Fairbank, J. C. et al. "The Oswestry Low Back Pain Disability Questionnaire." *Physiotherapy* 66, no. 8 (1980): 271–73.
13. Fairbank, J. C. and P. B. Pynsent. "The Oswestry Disability Index." *Spine* 25, no. 22 (2000): 2940–52; discussion 2952.
14. National Institutes of Health, *Toward Higher Levels of Analysis: Progress and Promise in Research on Social and Cultural Dimensions of Health* (Publication No. 01-5020). Washington, DC: US Government Printing Office, 2001.
15. Berndt, D. J., A. R. Hevner, and J. Studnicki. "The Catch Data Warehouse: Support for Community Health Care Decision-Making." *Decision Support Systems* 35, no. 3 (2003): 367–84.
16. Parmanto, B. and M. Scotch. "Mining Information from Mountains of Electronic Health Records: Unique Challenges and Solutions." *75th AHIMA National Convention and Exhibit Proceedings*, Minneapolis, MN, October 2003.
17. Stineman, M. G. and C. V. Granger. "Outcome, Efficiency, and Time-Trend Pattern Analyses for Stroke Rehabilitation." *American Journal of Physical Medicine and Rehabilitation* 77, no. 3 (1998): 193–201.

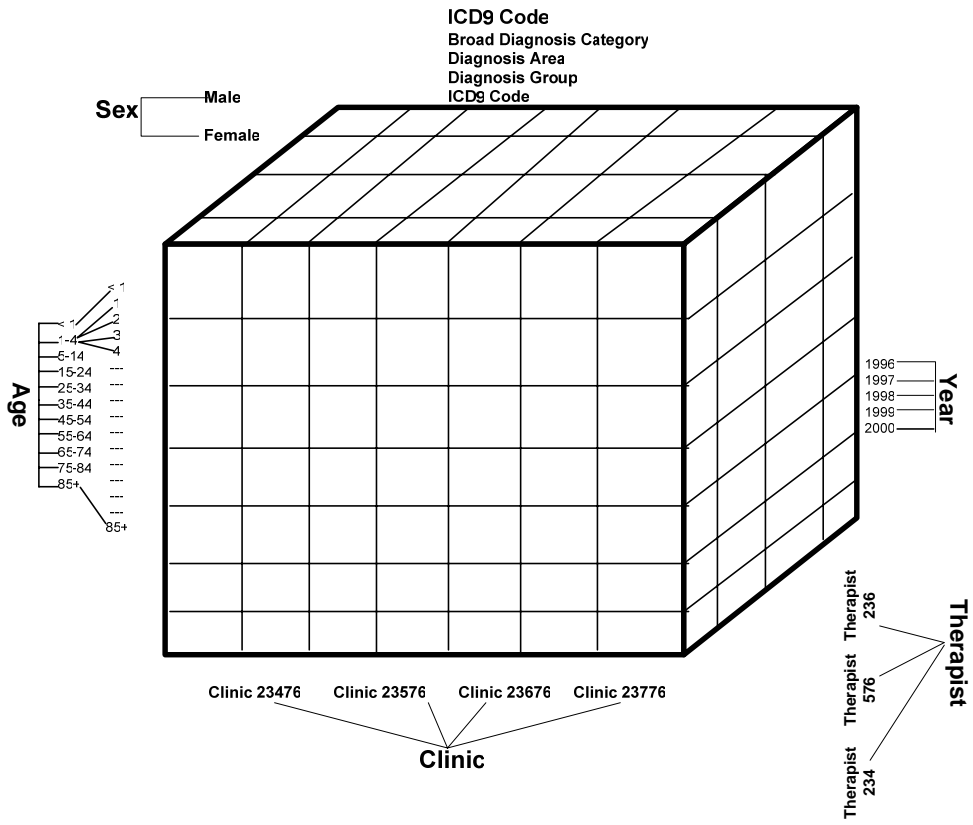
**Figure 1**

Star schema (dimensional modeling) of a data tables for a data warehouse



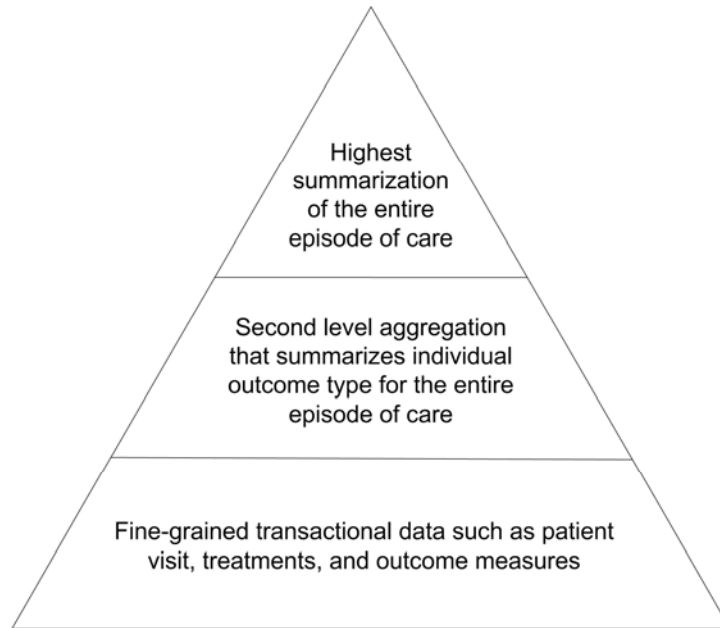
**Figure 2**

A multidimensional OLAP cube for healthcare rehabilitation data. The cube contains the different dimensions (or types of data in the data warehouse) and their hierarchies. Adapted from SOVAT.<sup>6</sup>



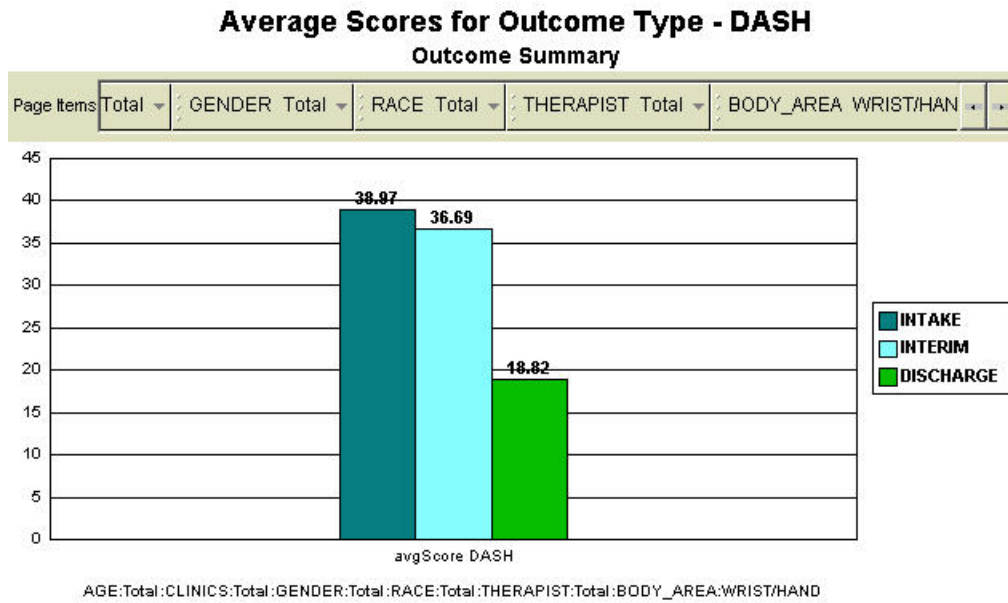
**Figure 3**

Information grain pyramid



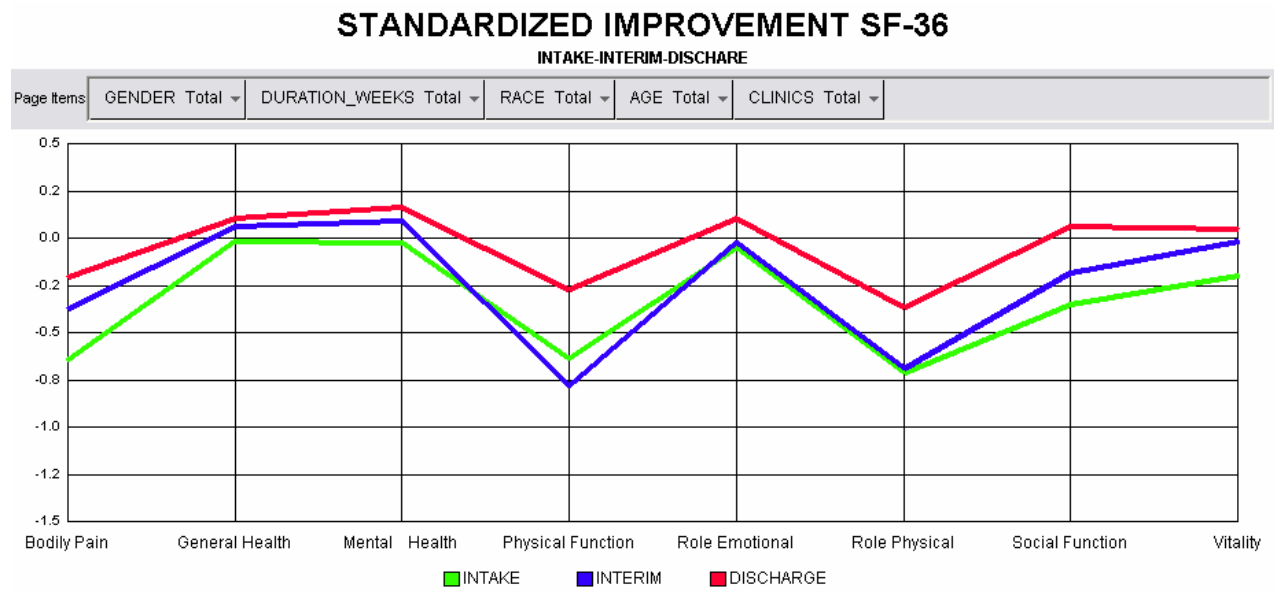
**Figure 4**

Middle level of information grain: improvement of wrist and hand outcomes from intake to discharge, as measured by the disabilities of the arm, shoulder, and hand (DASH) questionnaire. The score ranges from 0 to 100. Higher scores indicate higher levels of disability.



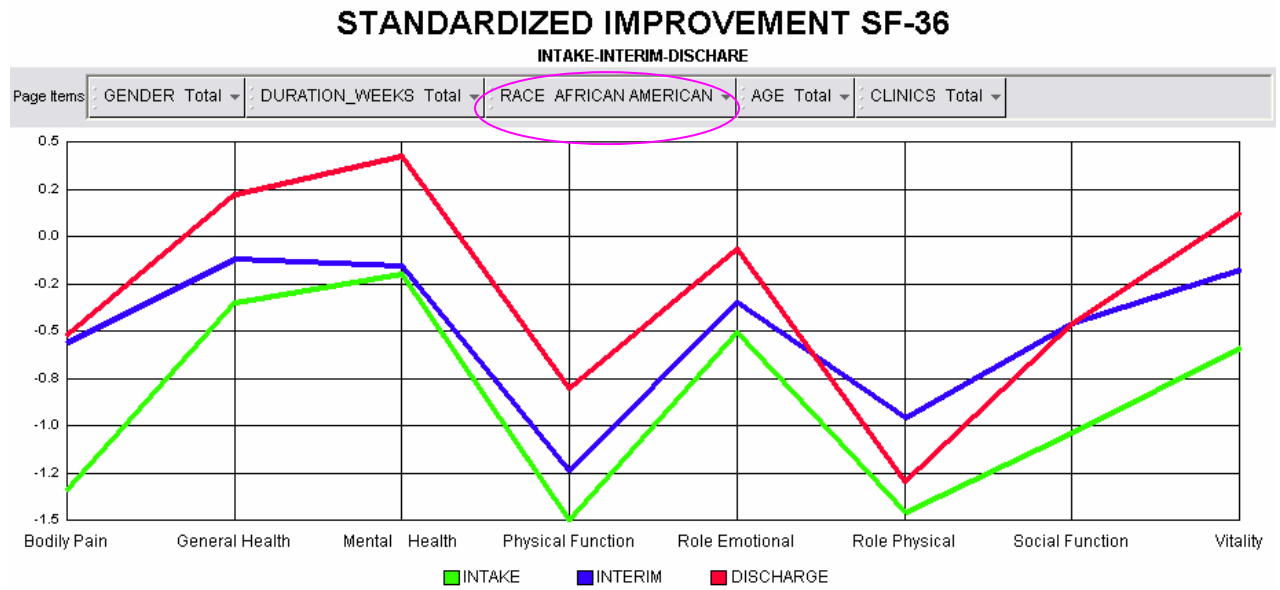
**Figure 5**

Comparison of health-related quality of life outcomes at intake, an interim visit, and discharge

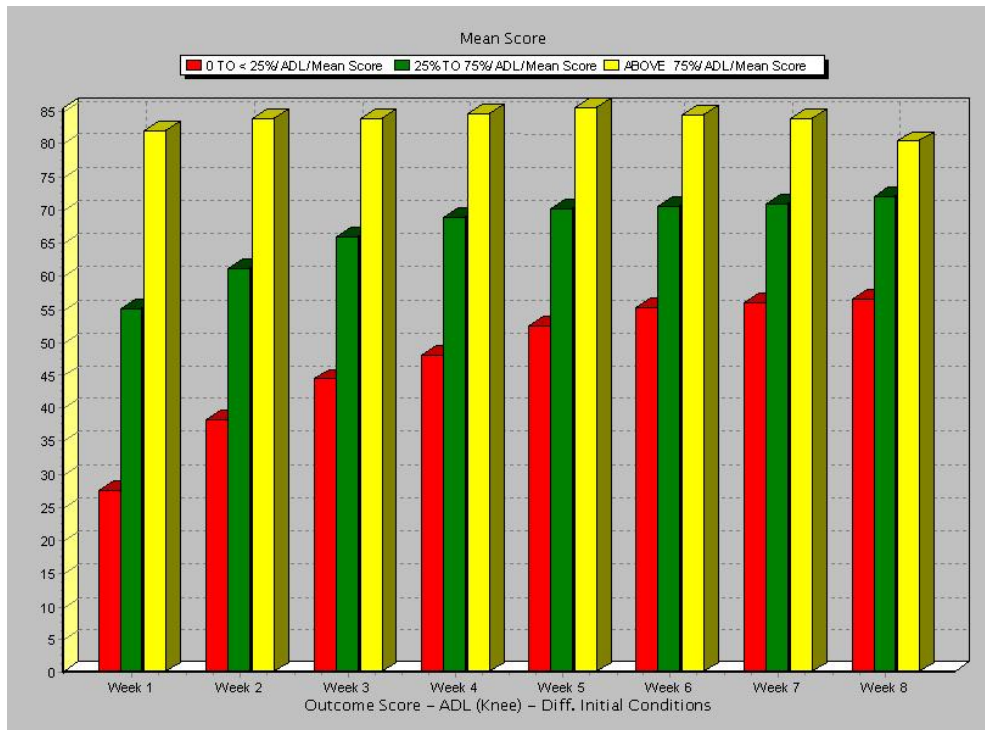


**Figure 6**

Comparison of health-related quality of life outcomes at intake, an interim visit, and discharge for African-Americans



**Figure 7**  
Disease-specific measurement of patient progress



**Figure 8**

Discovering subtle patterns and trends for disease-specific outcomes (low back pain, Oswestry Index)

