2007

Uncertainty in the information supply chain: Integrating multiple health care data sources

Monica Chiarini Tremblay
University of South Florida

Uncertainty In The Information Supply Chain:

Integrating Multiple Health Care Data Sources

by

Monica Chiarini Tremblay

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Information Systems and Decision Sciences
College of Business Administration
University of South Florida

Co-Major Professor: Donald J. Berndt, Ph.D.
Co-Major Professor: Alan R. Hevner, Ph.D.
Rosann Webb Collins, Ph.D
Joni L. Jones, Ph.D

Date of Approval:
July 13, 2007

Keywords: data quality, health care information systems, decision making, focus groups,
information supply chain, design science

Dedication

Without the support of my family, this work would never exist. Foremost, I am appreciative of the patience, understanding and sacrifices of my husband Tom, and my children Andrew and Laura. I am proud Andrew and Laura could say comprehensive exams and dissertation in the second grade. My mother, Orsola Rinaudo continuously encouraged me to remain in the Ph.D program and was always available to help during all the crucial moments, cancelling many plans at the last minute, and celebrating each step. I am forever grateful for her generosity. My father, Mario Chiarini, taught me the importance of hard work and humility and continuously reassured me that I made the right career choice. Papi, I wish that you could be here to read this; I hope my work does justice to your intellect and what you taught me. I dedicate this dissertation to my family who gave me the possibility of achieving my dream.

Acknowledgements

I am deeply indebted to my committee for their assistance. I am especially grateful to Dr. Berndt. I like to joke that I got my Masters in Dr. Berndt, since I took four courses from him during the masters program. His enthusiasm and unbelievable imagination and creativity are what attracted me to the Ph.D. program and are the source of the majority of my research interests. Dr. Berndt steadfastly and patiently endured my millions of questions and panic attacks and helped me complete the program in a tough year, one in which we both suffered through the loss of a parent. Dr. Hevner was one of the first professors I met at USF, and his work is a great source of inspiration. It was an honor to have him serve on my committee. I am very grateful to Dr. Collins for her unbelievable insight, keen eye for detail and excellent advice. She is a great mentor and I will really miss her. Dr. Jones was also an immense source of inspiration and advice. Many of the important decisions I made this year were made under her guidance. Thanks Joni for believing in me, sometimes more then I did.

I would also like to acknowledge the excellent faculty of the Information Systems Department, in particular Drs. Chari and Birkin; both served as department chairs during my studies and often provided valuable recommendations. The department has a superb group of teachers and researchers. I also extend a sincere thank you to the office staff: Judy Oates, Annette Weiser and Nadia Khouri were tremendously helpful in every type of situation, from bureaucratic to personal. I am grateful for the generous financial support of the Gaiennie Fund. Finally, I want to thank VISN 8 Patient Safety Center of Inquiry at James Haley VA and the Health Council of East Central Florida. Their insight was invaluable in this research.

Table of Contents

List of Tables

List of Figures

Uncertainty in the Information Supply Chain:
Integrating Multiple Health Care Data Sources

Monica Chiarini Tremblay

ABSTRACT

Similar to a product supply chain, an information supply chain is a dynamic environment where networks of information-sharing agents gather data from many sources and utilize the same data for different tasks. Unfortunately, raw data arriving from a variety of sources are often plagued by errors (Ballou et al. 1998), which can lead to poor decision making. Supporting decision making in this challenging environment demands a proactive approach to data quality management, since the decision maker has no control over these data sources (Shankaranarayan et al. 2003). This is true in health care, and in particular in health planning, where health care resource allocation is often based on summarized data from a myriad of sources such as hospital admissions, vital statistic records, and specific disease registries.

This work investigates issues of data quality in the information supply chain. It proposes three result-driven data quality metrics that inform and aid decision makers with incomplete and inconsistent data and help mitigate insensitivity to sample size, a well known decision bias. To design and evaluate the result-driven data quality metrics this thesis utilizes the design science paradigm (Simon 1996; Hevner, March et al. 2004). The metrics are implemented within a simple OLAP interface, utilizing data aggregated from several healthcare data sources, and presented to decision makers in four focus groups. This research is one of the first to propose and outline the use of focus groups as a technique to demonstrate utility and efficacy of design science artifacts.

Results from the focus groups demonstrate that the proposed metrics are useful, and that the metrics are efficient in altering a decision maker's data analytic strategies. Additionally, results indicate that comparative techniques, such as benchmarking or scenario based approaches, are promising approaches in data quality.

Finally, results from this research reveal that decision making literature needs to be considered in the design of BI tools. Participants of the focus groups confirmed that people are insensitive to sample size, but when attention was drawn to small sample sizes, this bias was mitigated.

Chapter One: Introduction

The Information Supply Chain is based on the studies of supply chain management (SCM), which have been widely used in management science (Sun and Yen 2005). A supply chain fulfills its customer's demand by a network of companies, mainly including suppliers, manufactures, and distributors. Similarly to a supply chain, an information supply chain (ISC) fulfills users' information requirements by a network of information-sharing agents (ISA) that gather, interpret, and satisfy the requirements with proper information(Sun and Yen 2005).

Regardless how rigorous the data cleansing processes by the ISA, there will still be data errors and peculiarities in the information supply chain data which are probably, but not necessarily, due to inaccuracy in the data. Information about these errors is not generally presented to decision makers, who will make choices and decisions based on the available data. In fact, most database queries are run without any data quality information. This is an especially troublesome issue in analytic databases (compared with transactional systems). Tracing and correcting these errors can be expensive, and at times impossible, but the threats to decision quality can be reduced by informing the information consumer about the data quality at decision time (Parssian 2006). Decision makers can be further aided by having some flexibility in the consideration of the effect of these data quality problems on different scenarios.

This thesis presents methodologies that communicate data quality information at decision time with simple and comprehensible metrics that can be calculated when the final information product is created. These metrics are evaluated with the use of focus groups comprised of several types of decision makers: healthcare analysts, database and data warehouse administrators, systems analysts, and graduate students familiar with data analytics.

Motivation

Practitioners have recognized the need for comprehensive knowledge management and decision support tools, and these tools have grown in sophistication in recent years. In industry, tools such as those produced by Micro Strategies, Business Objects, Hyperion, and Cognos ReportNet improve business performance by providing information within a single architecture. These tools are important for what Tom Davenport (2005, pg. 9) of the Babson Executive Information Center describes as the "emergence of a new form of competition based on the extensive use of analytics, data, and fact-based decision making".

To successfully compete on fact-based decision making, accurate data are needed. Yet, most companies assume that once data are collected from the information supply chain, cleaned and safely stored in a database, queries deliver the "correct" information (Trembly 2002). This is an incorrect assumption made by many, according to a recent Data Warehouse Institute study based on interviews with industry experts, leading-edge customers, and survey data from 647 respondents:

> "… a significant gap (exists) between perception and reality regarding the quality of data in many organizations, current data quality problems cost U.S. businesses more than $600 billion a year" (pg 48).

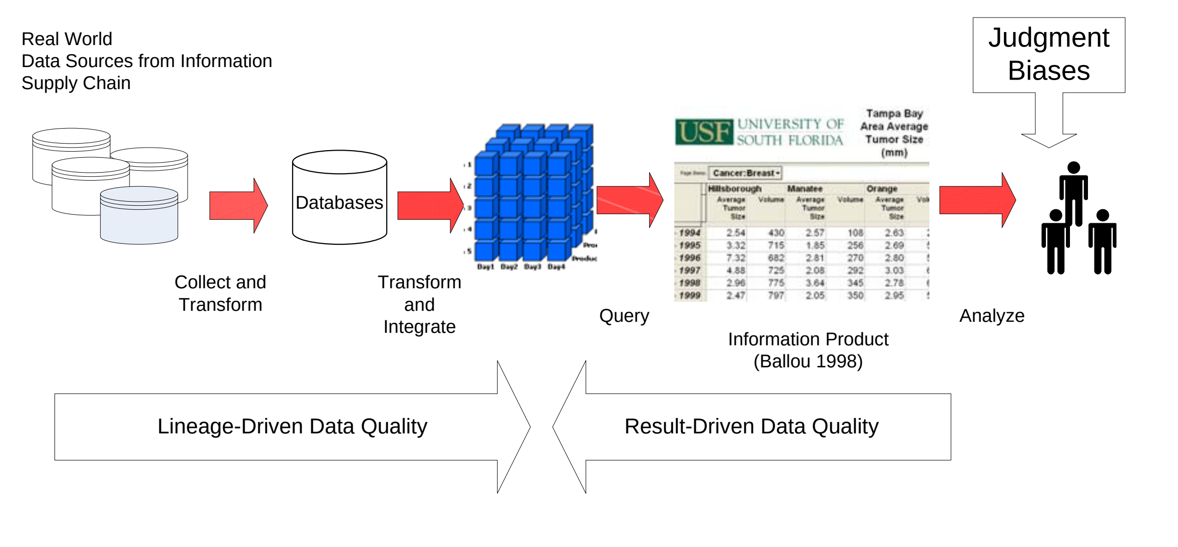
Context

Like other business organizations, the healthcare sector is increasingly becoming an information-driven service (Friede, Blum et al. 1995; Al-Shorbaji 2001; Derose, Schuster et al. 2002; Derose and Petitti 2003), particularly for public policy and health planning. The practice of evidence-based medicine, which is defined as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" (Sackett, Rosenberg et al. 1996), requires the emergence of technologies that support knowledge management. To improve public health's efficacy and profile, both practitioners and researchers need reliable and timely information to make information-driven or evidence based decisions (Friede, Blum et al. 1995).

The context of this research is that of public policy decision making and an extensive healthcare information supply chain in the state of Florida. Florida's health planning agencies develop evidence-based health plans at the district level that assess the heath status of communities and influence the policies and interventions to improve the delivery of care. For example, they can make decisions about the location of a new clinic for the uninsured and the type of services it should provide, based on the needs of the particular community.

Problem Statement

Information supply chains can be complex, multi-step processes that include the collection of raw data from many sources, intermediate transformations, compositions, and standardizations that ultimately supply the raw data for insightful analysis. The chain is anchored at one end in real-world data sources that define the history available for all subsequent analyses. The endpoints of the information supply chain are the various information products that support business processes and decision making activities. Data quality efforts can grow from either anchor point, often offering complementary capabilities.


Real World
Data Sources from Information
Supply Chain
Judgment
Biases
Databases
Collect and
Transform
Transform
and
Integrate
Query
Analyze
UNIVERSITY OF SOUTH FLORIDA
Tampa Bay Area Average Tumor Size (mm)
Cancer:Breast
Information Product
(Ballou 1998)
Lineage-Driven Data Quality
Result-Driven Data Quality

	Hillsborough		Manatee		Orange
	Average Tumor Size	Volume	Average Tumor Size	Volume	Average Tumor Size
1994	2.54	430	2.57	108	2.63
1995	3.32	715	1.85	256	2.69
1996	7.32	682	2.81	270	2.80
1997	4.88	725	2.08	292	3.03
1998	2.96	775	3.64	345	2.78
1999	2.47	797	2.05	350	2.95

Figure 1 – Research Landscape

As shown in Figure 1, data quality can be assessed as part of the original data collection process and propagated through transformations and compositions made by the ISA as part of *lineage-driven data quality* measurement. In contrast, *result-driven data quality* proceeds from the information product endpoint, with knowledge of the context, and works backward to provide measures that assist decisions makers in understanding uncertainties accounting for possible poor decision-making due to well-known judgment biases.

Lineage-Driven Data Quality

Data lineage refers to the body of metadata that is useful in understanding the origin and processing steps used to create data items for analysis, as well as for long-term storage to maintain a historical perspective (Cui and Widom 2000). While a data lineage can contain many metadata items of interest, the focus here is a on a subset of these items that are useful for measuring data quality. For instance, imagine that a healthcare information supply chain includes data on hospitalizations, including patient demographics and diagnostic codes. The data lineage might contain information about who collected the data, timestamps for various steps, and any algorithms used to modify the data. Among the lineage-driven data quality measures might be the number of missing values, descriptive statistics on patient demographics, or even the expected rate of diagnostic miscodes. These data quality measures are all independent of subsequent use. Of course, the general importance of data items and associated quality measures can be gleaned from their ultimate use. The distinction here is that no knowledge from the information product itself is necessary to calculate lineage-driven data quality metrics.

One the difficult challenges for lineage-driven data quality approaches is to propagate quality measures along a complex information supply chain. For example, combining two data items would also require combining their respective data quality information. What does it mean to take two (or more) error rates and combine them? In order to meet this challenge head-on, some approaches define data quality operators that precisely define the rules for data transformations (Ballou and Pazer 1985; Ballou, Wang et al. 1998; Cui and Widom 2000; Galhardas, Florescu et al. 2001; Shankaranarayan, Ziad et al. 2003; Shankaranarayan and Cai 2006). This is no small task since data transformations can include almost any possible algorithm! The hope is that the most common transformations can be handled, providing useful data quality information. Another approach is to have users provide quality metadata, in the form of weights or even intermediate quality measures that reflect the eventual use of the data. While this approach certainly involves the end user, the burden could become an obstacle to quality metric usage.

Result-Driven Data Quality

In contrast to lineage-driven data quality, result-driven data quality starts from the formulation of an information product and works backward to define data quality metrics. Some quality measures might limit the scope to easily available precursor data, while others might use the results of lineage-based transformations. However, with knowledge of the query, which indirectly captures aspects of a user's interests, very useful data quality metrics may be more easily obtained. For example, a report based on the hospitalization data described above might group patients by age categories and gender, counting hospital admissions and calculating length of stays. Each gender and age category defines an item in the final report that rests on a distinct set of values from the fine-grained data. The detailed data are unlikely to be uniformly distributed and the characteristics that relate to length of stay are also likely to differ for each combination. These independent subsets defined by the query (or report) provide the framework for calculating specific data quality measures. These measures could be as simple as assessing the sample size for each of the reporting combinations or as complex as understanding the effect of missing values on each of the aggregations. Whatever the quality measure, the calculations require knowledge of the grouping and filtering criteria, as well as the aggregation functions.

This thesis presents methodologies that communicate *result-driven data quality* (RDQ) information at decision time with simple and comprehensible metrics that can be calculated when the final IP is created. The decision maker is not involved in the calculation of the metric, but considers the metrics as they formulate a context-specific decision. Result-driven data quality is especially important in an environment where managers and decision makers utilize aggregated data (summary information) retrieved from several data sources in the information supply chain to make tactical decisions. This is true in health care, and in particular in health planning, where health care resource allocation is often based on summarized data from a myriad of sources such as hospital admissions, vital statistic records, and specific disease registries. These data are utilized to justify investments in services, reduce inequities in treatment, and rank health care problems to support policy formulation (Berndt et al. 2003).

Research Questions

In this research we investigate the communication of *result-driven data quality* (RDQ) information as calculated when a final IP is created. We investigate the design and evaluation of these metrics with a complex health care information supply chain that is queried in an OLAP environment, which is a common approach in many business intelligence tools. Three research questions are addressed: the first focuses on the design of the metrics, and the second and third on the evaluation of the metrics.

1) What is the design of result-driven data quality metrics that will aid decision-makers with the analysis of data from multiple data sources with varying levels of data quality in the health care information supply chain?

2) What is the utility of the data quality metrics?

3) What is the efficacy of the data quality metrics in altering a decision maker's data analytic strategies?

Research Approach

To design and evaluate the result-driven data quality metrics, this thesis utilizes the design science paradigm (Simon 1996; Hevner, March et al. 2004). The metrics are *built* with the intention to solve an identified organizational problem and are *evaluated* in an appropriate context to both provide feedback to the design process and a better understanding of the business process. In this research, the artifact will be a design and instantiation (Markus, Majchrzak et al. 2002; Hevner, March et al. 2004) of result-driven data quality metrics which will aid health planners in the process of the comparing data from multiple data sources.

The design of these metrics is informed by database theories on data quality, as well as behavioral decision-making theories. The metrics are evaluated in two phases. The first phase, exploratory focus groups, helps the researcher better understand the problem and will provide feedback for improvement of the design of the metrics. The second phase, confirmatory focus groups, evaluate the metrics' functionality, completeness, and usability and seeks to understand the impact of these metrics on the data analysis strategies of the decision-makers; thus, addressing the second and third research questions.

Foundation Theories

This study in primarily concerned with two streams of research: data quality and behavioral decision making. Literature in the area of data quality outlines the important conceptualization of proprietary data assets as off-the-shelf data products, or *information data products.* It also provides a framework to define data quality attributes and initial research in algorithms to calculate data quality metrics.

Data Quality

Information data products are manufactured much like any other product (Wang, Reddy et al. 1993; Wang, Storey et al. 1995; Strong, Lee et al. 1997; Ballou, Wang et al. 1998; Pipino, Lee et al. 2002; Shankaranarayan, Ziad et al. 2003; Parssian, Sarkar et al. 2004). Information producers generate and provide the "raw material" which is stored and maintained by information systems (or custodians) and accessed and utilized by information consumers for their tasks (Wang, Storey et al. 1995; Strong, Lee et al. 1997), creating a data product. As in manufacturing, the data products are in turn the raw material for a different data manufacturing process (Wang, Storey et al. 1995). Thus, just like the inputs and outputs of several manufacturing processes create a supply chain, the input and output of a series of data manufacturing processes create an information supply chain (Sun and Yen 2005).

Similar to the way a consumer purchasing an off-the-shelf product wishes to know information about the product (such as the ingredients, instructions for use, or date of expiration), data consumers should be informed about the quality of data products (Wang, Reddy et al. 1993).

The are many data quality attributes in the literature (at one point Wang identified more than 100 (Wang and Strong 1996)) the most common include: *usefulness, relevancy, timeliness, usage, interpretability, accessibility, believability, accuracy, completeness, credibility, consistency*. Wang (1997) classifies them into four categories:

1. Intrinsic Data Quality: Including the dimensions of accuracy, objectivity, believability, and reputation.
2. Accessibility Data Quality: Including the dimensions of accessibility and access security.
3. Contextual Data Quality: Including the dimensions of relevancy, value-added, timeliness, completeness, and amount of data. This category is considered in the context of the task at hand.
4. Representational Data Quality: Including the dimensions of interpretability, ease of understanding, representational consistency, and concise representation.

Judgment under Uncertainty

The general heuristics and biases that people use in making judgments are well researched. Though this study is mainly interested in strategies of data retrieval and representation that minimize these biases, it is important to understand the heuristics knowledge workers may use for decision making, as well as the possible biases that could result from the use of these heuristics. Heuristics are based on past experience and generally give good results, but they can also lead to severe and systematic errors (Tversky and Kahneman 1982). Tversky and Kahneman identify three heuristics that are used to access probabilities of an event that lead to biases in decision making: representativeness, availability, and anchoring and adjusting. This study investigates one factor of the bias of representativeness, insensitivity to sample size.

Research Description and Contributions

This work is one of the first to investigate issues of data quality in the information supply chain. The context for this study is health planning, where health care resource allocation is often based on summarized data from a myriad of sources (Berndt et al. 2003). Potential data quality measures and biases are identified by studying the existing literature and by conducting a field study in a Florida Health Planning Agency.

This work proposes three result-driven data quality problems: unallocated data, information volatility, and small sample size, and outlines metrics that aid decision makers in considering these problems in data analysis. These metrics are designed for use in an environment like health planning, where decision makers utilize aggregated data (summary information) retrieved from several data sources in the information supply chain to make tactical decisions. To design and evaluate the result-driven data quality metrics this thesis utilizes the design science paradigm (Simon 1996; Hevner, March et al. 2004). This research proposes the use of focus groups as a technique to evaluate design science research. It outlines a methodology for planning, selecting participants, conducting, analyzing and reporting the results of the focus groups to demonstrate utility and efficacy of the artifacts. The focus groups were comprised of several types of decision makers: healthcare analysts, database and data warehouse administrators, systems analysts, and graduate students familiar with data analytics.

The metrics were found to be useful and efficient in altering a decision maker's data analytic strategies. Supplying decision makers with information about the reliability of the data improved the quality of their decisions. Additionally, it was found that comparative techniques, such as benchmarking or scenario based approaches are promising approaches in data quality.

Results from this research indicate that decision making literature should be considered in the design of Business Intelligence (BI) tools. Participants of the focus groups confirmed that people are insensitive to sample size, but when attention was drawn to small sample sizes, this bias was mitigated.

Overview of Remaining Chapters

The remainder of this thesis is divided into ten chapters. Chapter two contains the Literature Review. The Research Design is covered in chapter three. Chapters four, five, and six describe the design of the data quality metrics. A description of the focus groups and the template coding are in chapter seven. Chapter eight describes the evaluation of the metrics though the use of focus groups. Finally, contributions, limitations of the research, as well as future research goals, are discussed in chapter nine.

Chapter Two: Literature Review and Theory

Introduction

The literature review begins with a description of the research paradigm utilized for this thesis: providing both a description of the design-science research and a summary of the framework and guidelines for conducting design-science research from Hevner et al. (2004). Next, the literature and theories that help describe the environment and the knowledge base which informs the design and evaluation of the metrics are reviewed: data quality, data quality in healthcare, healthcare planning, health planners as a form of knowledge work, uncertainty in knowledge work and decision making.

Research Paradigm: Design-Science Research

IS research is conducted in two complementary phases. Behavioral science research identifies a business need and develops and justifies theories that explain or predict phenomena related to this need. Design-science research builds and evaluates artifacts that address a particular business need. Behavioral science researchers search for the truth, while design-science researchers seek utility (Hevner, March et al. 2004). Hevner et al. (2004) stress that truth and utility are inseparable. The research of design-science researchers is informed by theories from behavioral science research, and the utility from design-science research provides information for behavioral science theories.

Hevner et al.'s (2004) information system research framework (Figure 2) illustrates how both research paradigms follow similar cycles. The knowledge base, consisting of prior IS research and results from reference disciplines, provides foundations and methodologies to be used in IS research. Foundations are the foundational theories, frameworks, instruments, constructs, models, methods, and instantiations used in the develop/build phase of a research study. Methodologies are the guidelines used in the justification or evaluation phase. Behavioral researchers develop theories using the foundations from the knowledge base and assess them using the methodologies in the knowledge base. Similarly, design-science researchers build artifacts based on the foundations in the knowledge base and evaluate them utilizing the methodologies in the knowledge base. Both research paradigms apply their findings to a business need in the appropriate environment and add them to the knowledge base.

The fundamental goal for design-science research is the provision and the demonstration of utility of an artifact. The design-science researcher designs an artifact that provides utility and provides evidence that this artifact solves a problem. There are two stages in design-science research: the development of the artifact and its evaluation (which cycles for refinement of the design).

Hevner et al. (2004) describe seven design-science research guidelines:

1. Design as an artifact: Design-science research requires the creation of an innovative artifact which can be in the form of a construct, model, method or instantiation. Rarely are artifacts complete information systems ready to be used in the business world. Rather they are innovations that help improve the information systems steps of analysis, design and implementation. For example, the entity-relationship diagram (Chen 1976) provides a set of constructs to describe data which has revolutionized data base design.

2. Problem Relevance: Design-science research should be relevant in a specified business domain.

3. Design Evaluation: Artifacts must be evaluated, in order to demonstrate they provide utility. Evaluation is crucial in design-science research. This requires that the artifact be evaluated within the technical infrastructure of the business environment. There are several methods to evaluate designs, which are available in the knowledge base: observation, analytics, experiments, testing or descriptive (Table 1 from Hevner et al. (2003) describes these methods). The evaluation should provide feedback to the design stage.

4. Research Contributions: Design-science research should contribute to the areas of design of artifacts, design foundations, and design methodologies. It must also contribute to the business environment by solving important, unsolved business problems.

5. Research Rigor: Rigor in design science research is attained by applying rigorous methodology, both in the design phase and in the evaluation phase.

6. Design as a Search Process: Design-science research progresses in an iterative manner. The design process follows a generate/test cycle (Simon 1996). For very large or difficult (wicked) design problems, as is often the case in information systems design problems, a search strategy is more appropriate. With each iteration, pieces of the puzzle are solved, and the scope of the problem grows.

7. Communication of Research: Design-science research should be communicated both to the technical and managerial audiences.

Figure 2 – Information Systems Research Framework

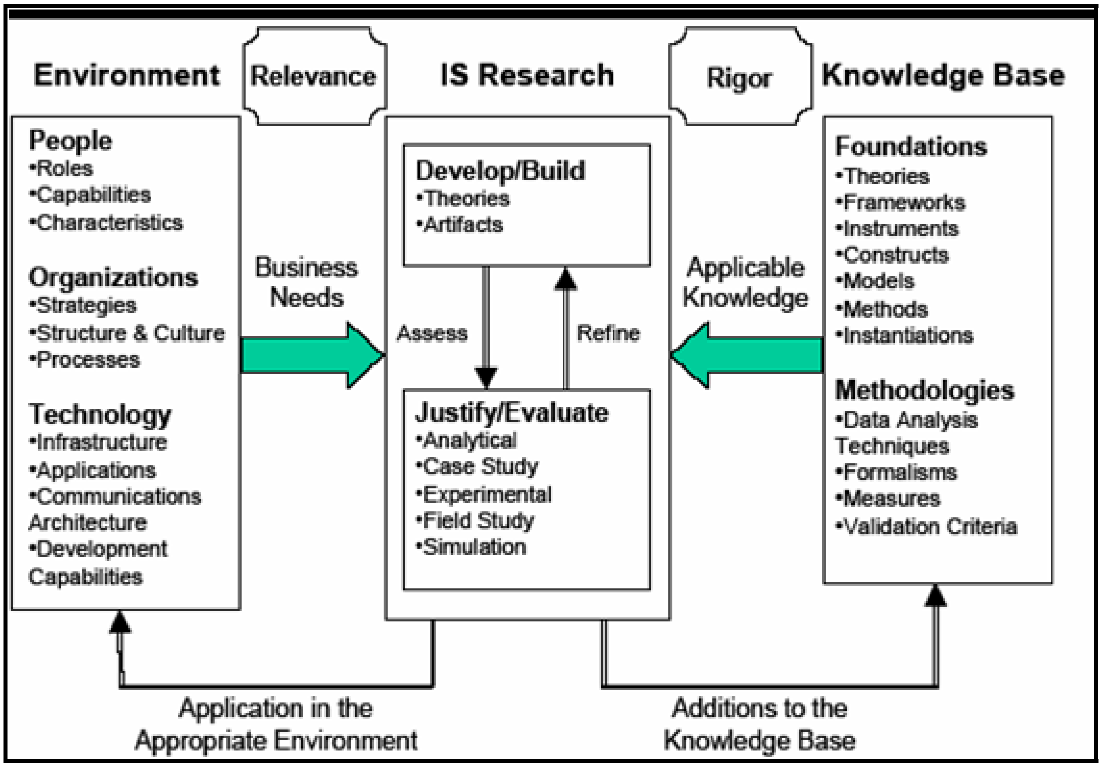
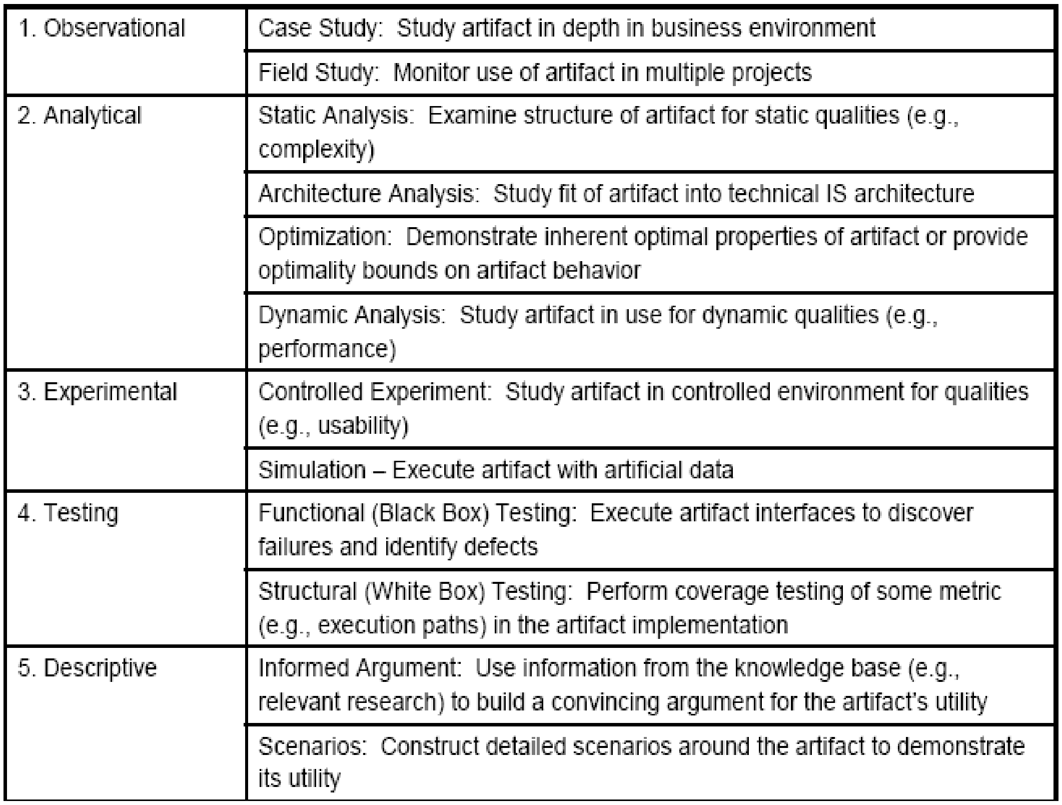


Table 1 – Design Evaluation Methods (from Hevner et al. 2004)



1. Observational	Case Study: Study artifact in depth in business environment
	Field Study: Monitor use of artifact in multiple projects
2. Analytical	Static Analysis: Examine structure of artifact for static qualities (e.g., complexity)
	Architecture Analysis: Study fit of artifact into technical IS architecture
	Optimization: Demonstrate inherent optimal properties of artifact or provide optimality bounds on artifact behavior
	Dynamic Analysis: Study artifact in use for dynamic qualities (e.g., performance)
3. Experimental	Controlled Experiment: Study artifact in controlled environment for qualities (e.g., usability)
	Simulation – Execute artifact with artificial data
4. Testing	Functional (Black Box) Testing: Execute artifact interfaces to discover failures and identify defects
	Structural (White Box) Testing: Perform coverage testing of some metric (e.g., execution paths) in the artifact implementation
5. Descriptive	Informed Argument: Use information from the knowledge base (e.g., relevant research) to build a convincing argument for the artifact's utility
	Scenarios: Construct detailed scenarios around the artifact to demonstrate its utility

Data Quality

The information supply chain creates a dynamic environment where a decision maker can retrieve data from many sources and utilize the same data for different tasks, as well as sharing the data and decision outcomes with others (Shankaranarayan, Ziad et al. 2003). Supporting decision making in this environment demands a proactive approach to data quality management, yet the decision maker has no control over these data sources (Shankaranarayan, Ziad et al. 2003). Raw data arriving from a variety of sources is often plagued by errors, which can lead to poor decision making. Yet, it is unclear what the effect of poor quality data are on the results of queries and reports used for decision making (Ballou, Wang et al. 1998; Parssian, Sarkar et al. 2004).

A large body of literature focuses on the definition of data quality (Wang, Storey et al. 1995; Redman 1996; Wand and Wang 1996; Wang and Strong 1996), how to categorize data quality into dimensions (Wang, Storey et al. 1995; Redman 1996; Wang and Strong 1996), and how to model information systems in order to track data quality (Morey 1982; Wang, Reddy et al. 1995; Ballou, Wang et al. 1998; Shankaranarayan, Ziad et al. 2003; Parssian, Sarkar et al. 2004). This review defines information data products, outlines available data quality frameworks, and describes data quality metrics in the literature.

The Information Supply Chain and Information Data Products

One of the methods suggested in the literature to determine data quality is to compare the provision or creation of data to the manufacturing of a product (Wang, Reddy et al. 1993; Wang, Storey et al. 1995; Strong, Lee et al. 1997; Ballou, Wang et al. 1998; Pipino, Lee et al. 2002; Shankaranarayan, Ziad et al. 2003; Parssian, Sarkar et al. 2004; Sun and Yen 2005). This is a valid analogy because it allows for the transfer of knowledge from the field of production quality.

In a manufacturing environment, a process consists of utilizing raw materials to create a product. There are both producers and consumers of a certain product. Similarly, information producers generate and provide the "raw material" which is stored and maintained by information systems (or custodians) and accessed and utilized by information consumers for their tasks (Wang, Storey et al. 1995; Strong, Lee et al. 1997), creating a data product.

Like in manufacturing, the data products are the raw material for a different data manufacturing process (Wang, Storey et al. 1995), data consumers are the data producers.

Thus, just like the inputs and outputs of several manufacturing processes create a supply chain, the input and out of a series of data manufacturing process create an information supply chain.   Table 2 provides an analogy between physical products and data products (Wang, Storey et al. 1995).

Table 2 – Analogy between physical products and data products

	Product Manufacturing	**Data Manufacturing**
Input	Raw Materials	Raw Data
Process	Materials Processing	Data Processing
Output	Physical Products	Data Products

An information data product (IP) is defined as a compilation of data items that is packaged in a way that it can be readily used. The IP term implies that this product has a certain value which is transferred to the customer. The creation of this IP can be thought of as *information manufacturing.* Information systems that produce these predefined IPs are referred to as *information manufacturing systems* (Ballou, Wang et al. 1998). Most recently this has been coined as the *Information Supply Chain* (Ballou, Wang et al. 1998; Sun and Yen 2005).

A data item can be as simple as a data string or a complex as a detailed report (Wang, Reddy et al. 1993). Like in the manufacturing environment different IPs can be standard products which can be manufactured in an assembly line (an invoice) or built to order (a specialized report). Ballou et al. (1998) use the term data quality for intermediate data products (those that experience additional processing) and reserve the terms *information product* and *information quality* for the final product that a customer receives.

It is assumed that the raw data needed to create an IP is available. This is analogous to made-to-stock in manufacturing. Made-to-stock items are either available in inventory, or can be assembled using raw materials available in inventory. Some IPs may share a production process and data inputs, with small variations that distinguish them. This analogy allows us to adapt proven methods for quality management such as total quality management (TQM) and International Organization for Standardization (ISO) from the manufacturing environment to that of data quality management (Ballou, Wang et al. 1998; Shankaranarayan, Ziad et al. 2003).

Ballou et al. (1998) outline some limitations to this analogy which arise from the nature of the raw material:

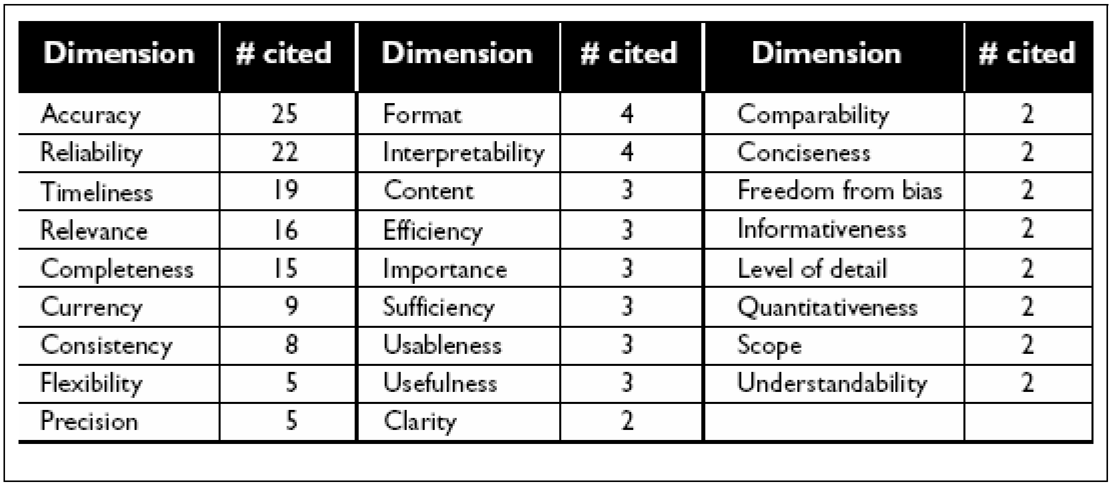
1. Raw input data, unlike raw materials, is not consumed. Stored data can be used indefinitely.
2. Producing multiple copies of an information product is inexpensive, which is not usually the case with a manufactured product.

Manufacturers request information about the quality of the raw materials they utilize in their manufacturing process, and correspondingly, data producers should be informed of the quality of the data products they utilize (Wang, Reddy et al. 1993). The challenge lies in deciding how to communicate this information, since different producers will have different data quality requirements, and different consequences for poor data quality.

Data Quality Frameworks

Several frameworks in the literature define a set of characteristics of data quality, referred to as data quality attributes (Wang, Reddy et al. 1993; Wang, Storey et al. 1995; Wang and Strong 1996). The most commonly used attributes to measure data quality include *interpretability, credibility* and *timeliness* (Wang, Reddy et al. 1993). In Wand and Wang (1996) a comprehensive review of the data quality literature shows that the most often cited data quality constructs listed in Table 3.

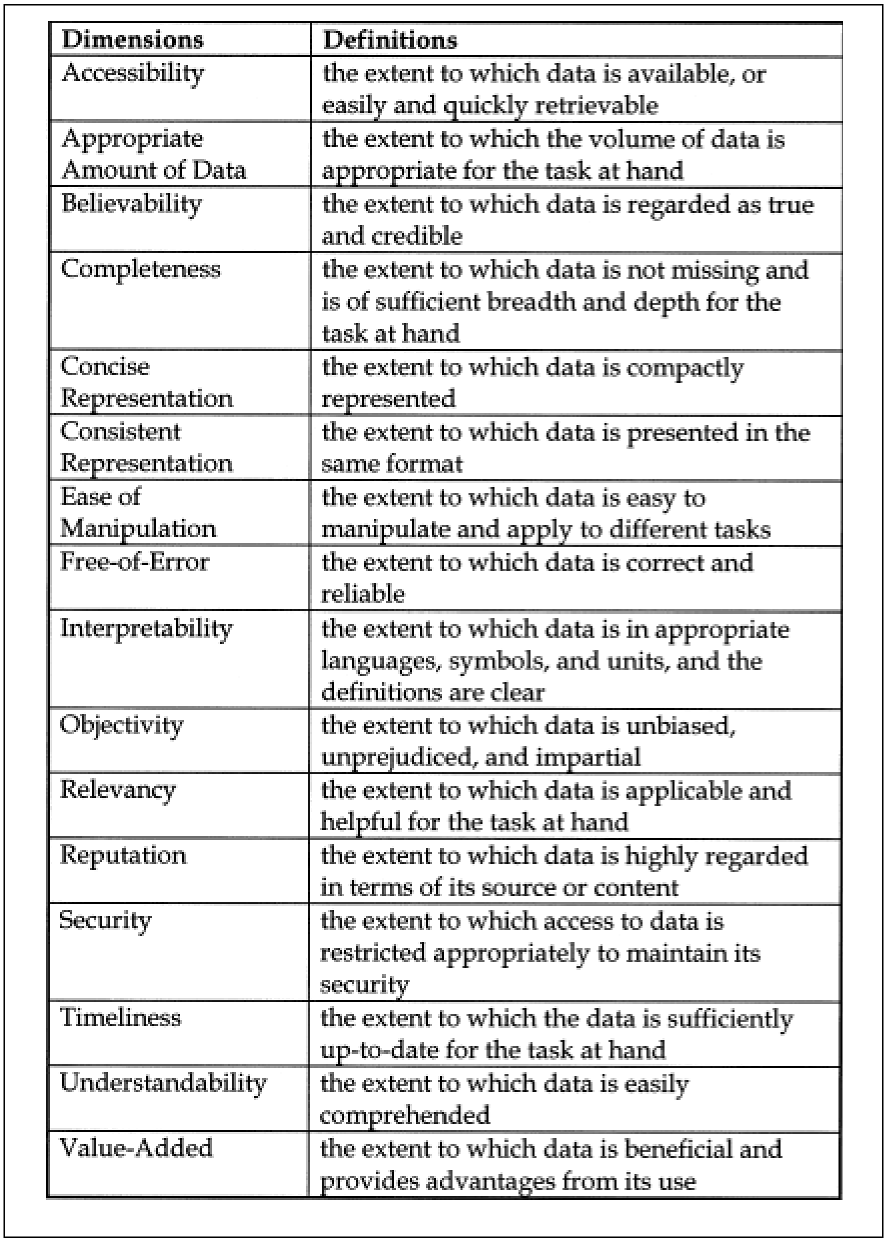
Table 3 – Frequency of Data Quality Dimensions (Wand and Wang 1996)



Dimension	# cited	Dimension	# cited	Dimension	# cited
Accuracy	25	Format	4	Comparability	2
Reliability	22	Interpretability	4	Conciseness	2
Timeliness	19	Content	3	Freedom from bias	2
Relevance	16	Efficiency	3	Informativeness	2
Completeness	15	Importance	3	Level of detail	2
Currency	9	Sufficiency	3	Quantitativeness	2
Consistency	8	Usableness	3	Scope	2
Flexibility	5	Usefulness	3	Understandability	2
Precision	5	Clarity	2		

Wang et al. (1994) surveyed IS professionals and researchers and gathered 179 data quality attributes. Their list of attributes was collapsed into fifteen data quality dimensions. These definitions are shown in Table 4 (Wang and Strong 1996; Kahn, Strong et al. 2002; Pipino, Lee et al. 2002). They continue this hierarchical approach and collapse these dimensions into four categories, resulting in a framework of data quality from the data consumers' perspectives (Figure 3)**.**

Table 4 – Definition of Data Quality Dimensions



Dimensions	Definitions
Accessibility	the extent to which data is available, or easily and quickly retrievable
Appropriate Amount of Data	the extent to which the volume of data is appropriate for the task at hand
Believability	the extent to which data is regarded as true and credible
Completeness	the extent to which data is not missing and is of sufficient breadth and depth for the task at hand
Concise Representation	the extent to which data is compactly represented
Consistent Representation	the extent to which data is presented in the same format
Ease of Manipulation	the extent to which data is easy to manipulate and apply to different tasks
Free-of-Error	the extent to which data is correct and reliable
Interpretability	the extent to which data is in appropriate languages, symbols, and units, and the definitions are clear
Objectivity	the extent to which data is unbiased, unprejudiced, and impartial
Relevancy	the extent to which data is applicable and helpful for the task at hand
Reputation	the extent to which data is highly regarded in terms of its source or content
Security	the extent to which access to data is restricted appropriately to maintain its security
Timeliness	the extent to which the data is sufficiently up-to-date for the task at hand
Understandability	the extent to which data is easily comprehended
Value-Added	the extent to which data is beneficial and provides advantages from its use

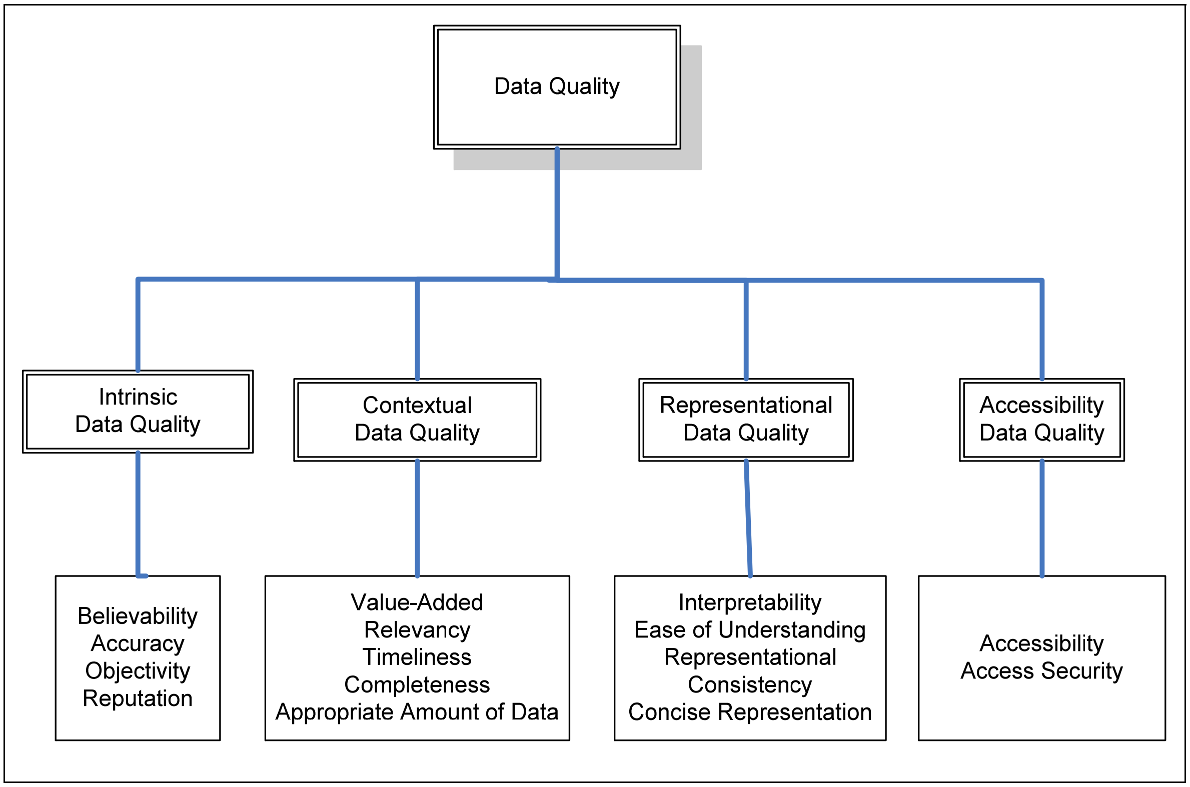


Figure 3 – Conceptual Framework of Data Quality (Wang and Strong 1996)

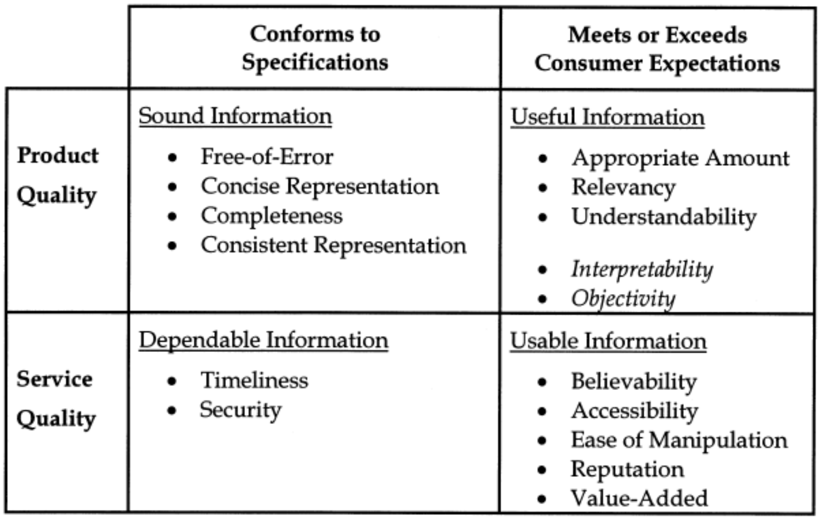
These dimensions capture what data consumers *ideally* need to know about their data, however much work needs to be done on how to technically operationalize these dimensions in a way that data quality could be objectively assessed. Furthermore, these hierarchies are not utilized in later literature, not even by the authors. Rather the 15 dimensions are often re-categorized. For example, Strong et al. (1997) utilize these same data quality categories and describes the following "potholes" in data collection that lead to data quality problems:

1. Multiple Data Sources
2. Subjective Judgment and Techniques in Data Production
3. Bypassing Input Rules and Too Strict Input Rules
4. Large Volumes of Data
5. Distributed Heterogeneous Systems
6. Complex Data Representations such as Text and Image

7. Coded Data From Different Functional Areas
8. Changing Data Needs from Information Consumers
9. Security-Accessibility Tradeoff
10. Limited Computing Resources

Pipino et al. (2002) describe how to assess these 15 metrics in practice, though at a very high level. They describe two types of assessments: objective and subjective. Objective assessments can be task-independent or task-dependent. Task-independent metrics can be applied to any data set, regardless of the tasks at hand, since they contain no contextual knowledge of the application. Task dependent metrics, are developed in specific application contexts. Kahn et al. (2002) take these 15 metrics and map them to a two-by-two conceptual model for describing IQ (see Table 5).

Table 5 – Mapping IQ dimensions into the PSP/IW Model



	Conforms to Specifications	**Meets or Exceeds Consumer Expectations**
Product Quality	Sound Information • Free-of-Error • Concise Representation • Completeness • Consistent Representation	Useful Information • Appropriate Amount • Relevancy • Understandability • *Interpretability* • *Objectivity*
Service Quality	Dependable Information • Timeliness • Security	Usable Information • Believability • Accessibility • Ease of Manipulation • Reputation • Value-Added

Redman (2001) takes a different approach. Twenty seven dimensions are mapped into three activities that correspond with the “define a view”, “obtain values” and “present results” activities of the life-cycle model. This grouping is associated with data usage alone, without considering other database issues such as storage and security. He defines these activities as: *conceptual view*, *data values* and *data representation*. A summary is shown in Table 6.

Table 6 – Summary of 27 Data Quality Dimensions (Redman 1996)

The Conceptual View			
Content	relevance	obtainability	clarity of definition
Scope	comprehensiveness	essentialness	
Level of Detail	attribute granularity	essentialness	
Composition	naturalness	identifiability	
	homogeneity	minimum unnecessary redundancy	
View Consistency	semantic consistency	structural consistency	
Reaction to Change	robustness	flexibility	
Values			
	accuracy	completeness	
		(entities and attributes)	
	consistency	currency/cycle time	
Representation			
Formats	appropriateness	format precision	efficient use of storage
	interpretability	format flexibility	
	portability	ability to represent null values	
Physical Instances	representation consistence		

Wand and Wang (1996) take an ontological perspective. They provide a design-oriented definition of data quality that reflects the intended use of the information. This analysis is based on the conflicts of two views of an application domain (also termed the real-world system). Representation deficiencies are defined in terms of the difference between the view of the real-world system as inferred from the information system and the view that is obtained by directly observing the real-world system. These dimensions are outlined in Table 7.

Table 7 – Intrinsic Data Quality Dimensions (Wand and Wang 1996)

DQ Dimension	Nature of Associated Deficiency	Source of Deficiency
Complete	Improper representation: missing IS states	Design failure
Unambiguous	Improper representation: multiple RW states mapped to the same IS state	Design failure
Meaningful	Meaningless IS state and garbling (map to a meaningless state)	Design failure and operation failure
Correct	Garbling (map to a wrong state)	Operation failure

Data Quality Metrics

This section summarizes previous research on the provision of data quality metrics in past literature is classified by the approach described in chapter one: metadata, lineage-driven or result-driven data quality. These are summarized in Table 8.

Metadata

The easiest approach to calculating data quality attributes and dimensions aforementioned in the data quality frameworks would be to simply receive this information in the form of metadata from the information producer. Data Quality metadata should describe the quality of the data. This is assuming that the metadata itself does not have data quality problems (such as missing or incorrect values). But even given this perfect condition, the literature only begins to define *which* metrics are objective measures of data quality. Practitioners have also analyzed this problem giving the following examples of possible metadata: accuracy, change management, definition changes, what actions are taken when data are "bad", missing, and duplicate. Data quality metadata are then tracked using data quality tools, repositories, and traditional documentation types. (Seiner 2000) suggests that data quality metadata should answer the following questions:

- How have the accepted values of the data changed over time?
- When did the accepted values change?
- How has the definition of the data changed over time?
- When did the definition of the data change?
- What constitutes "bad" data?
- What quality checks were performed against my data?
- What are the quality check procedures? Who wrote and executed them?
- Who analyzed the results?
- With what level of confidence can I trust my data?
- What is the accepted level of confidence before the data are considered "low quality" data?

Lineage Driven Data Quality Metrics

Several studies (Ballou, Wang et al. 1998; Jarke, Jeusfeld et al. 1999; Cui and Widom 2000; Shankaranarayan, Ziad et al. 2003) consider lineage information within a data warehouse infrastructure. Cui and Widom (2000) consider the data lineage problem within a multi-source data warehouse environment: for a given data item in a materialized warehouse view, they identify the set of source data items that produced the view item by presenting a lineage tracing algorithm for relational views with aggregation

(Cui and Widom 2000). In Jarke et al. (1999), metadata on data quality is derived from source data and is stored in a repository to be utilized to model data quality and to set data quality goals. Ballou et al. (1998) consider timeliness, accuracy and cost throughout the process of information manufacturing. Shankaranarayan et al.'s (2003) research, extends the manufacturing of data paradigm and suggest the following metadata to be captured at each step: (1) a unique identifier, (2) the composition of the data unit when it exits the stage, (3) the role and business unit responsible for each stage, (4) the individual(s) that may assume that role, (5) the processing requirements for that manufacturing step, (6) the business rules/constraints associated with it, (7) a description of the technology used and, (8) the physical location where the step is performed.

Wang et al. (1995) propose an attribute-based approach to data quality. The authors suggest augmenting data at the cell level with quality indicators. This would allow for a data consumer to judge the quality of the data without having to inspect the data manufacturing process. The authors suggest the following dimensions: accessibility, interpretability, usefulness and believability. Yet the research never directly outlines *how* these metrics would be captured and measured.

Result Driven Data Quality

Less research attention has been given to results driven data quality. Imielinski and Lipski (1984) consider how to represent incomplete data within the relational model, and find that it is heavily dependent on the processing of the information is to be performed, or in other words, what relational operators will be allowed. Parssian (2006) presents a methodology to estimate the effects of data accuracy and completeness on the relational aggregate functions: Count, Sum, Average, Max, and Min, using sampling strategies to estimate the maximum likelihood values.

Table 8 – Data Quality Metrics Type

Authors	**Metric**	**Approach**	**Type**
(Jarke, Jeusfeld et al. 1999)	Several	No Mathematical techniques suggested/ Paper suggests framework	Lineage
(Wang, Reddy et al. 1995)	Several	Cell Level Tagging	Lineage
(Imielinski and Lipski 1984)	Nulls (Incomplete Data)	Extend Codd's Relational Model	Result
(Ballou, Wang et al. 1998)	Timeliness, Accuracy, Cost	Information Manufacturing	Lineage
(Shankaranarayan, Ziad et al. 2003)	Accuracy, Completeness	IP MAPS	Lineage
(Parssian, Sarkar et al. 2004)	Accuracy, Completeness	Estimate mis-membership, accuracy, incompleteness	Lineage/ Result
(Parssian 2006)	Accuracy, Completeness	Sample to estimate the maximum likelihood values	Result
(Cui and Widom 2000)	Data Source	Data Lineage	Lineage

Data Quality and Healthcare

Information Systems are becoming an integral part of public health decision making. Information acquisition can now be transacted rapidly (Maibach and Holtgrave 1995; U.S. 1995; Chapman and Elstein 2000) and from several sources. Health policy decision makers need reliable, timely information with which to make information-driven decisions, and improved tools to analyze and present new knowledge (Friede, Blum et al. 1995). Public health agencies recognize the need to formally and quantitatively assess and improve the quality of their programs, information, and policies (Derose, Schuster et al. 2002). Yet, traditional software and hardware developed for laboratory science or business often lack features required for public health. For example, standard statistical packages do not facilitate standardization, fit models to certain disease patterns, or calculate sample sizes for case-control studies(Friede, Blum et al. 1995). Friede et al. [1995, pg 240] point out:

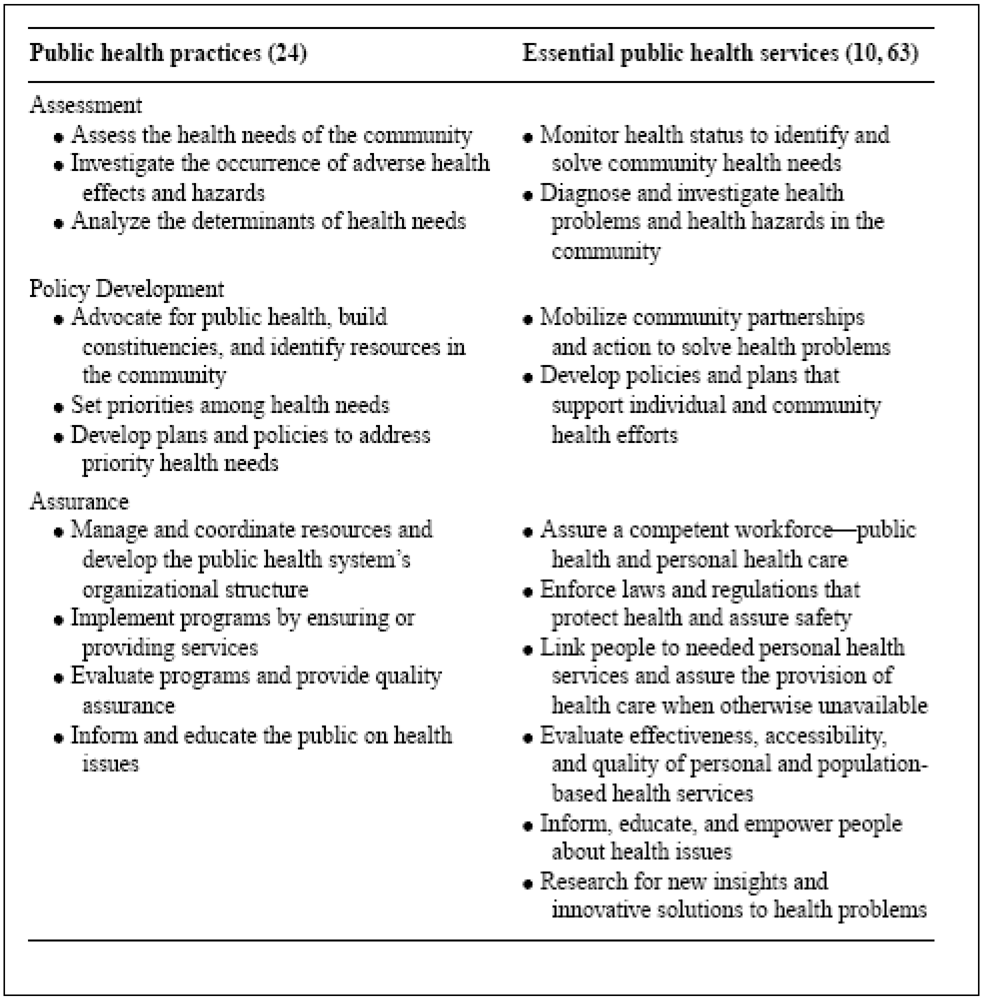
> "The combination of the burgeoning interest in health, combined with health care reform and the advent of the Information Age, represent a challenge and an opportunity for public health. If public health's efficacy and profile are to grow, practitioners and researchers will need reliable, timely information with which to make information-driven decisions, better ways to communicate, and improved tools to analyze and present new knowledge."

The use of healthcare information technology (HIT), and in particular the study of decision support and knowledge management, is a fertile field of study for IS Researchers. Healthcare enterprises can be regarded as 'data rich' (Abidi 2003) as they generate massive amounts of data, such as electronic medical records, clinical trial data, hospital records and administrative reports.

Healthcare information systems can be viewed as a continuum, beginning with individual patient level data and their interaction with health services, moving to aggregated data (Berndt, Hevner et al. 2003), to knowledge-based data, and then to community data used for policy development (Al-Shorbaji 2001). However, these data are rarely transformed into a strategic decision-support resource. Like other business organizations, the healthcare sector is increasingly becoming an information-driven service (Friede, Blum et al. 1995; Al-Shorbaji 2001; Derose, Schuster et al. 2002; Derose and Petitti 2003), particularly for public policy and health planning. The practice of evidence-based medicine requires the emergence of technologies that support knowledge management.

To improve public health's efficacy and profile, both practitioners and researchers need reliable and timely information to make information-driven or evidence based decisions (Friede, Blum et al. 1995). Thus, there is endless opportunity to transform raw empirical data into the kind of knowledge that can impact strategic decision-making, planning and management of the healthcare enterprise. Table 9 summarizes the activities in public health and services (Derose, Schuster et al. 2002), all of which are data-intensive. There are many different tasks in the assessment, development and assurance of public health policy. This study focuses on this rich health planning domain and focuses on a set of specific decision making activities related community needs assessment.

Table 9 – Public Health Practices and Public Health Services



Public health practices (24)	Essential public health services (10, 63)
Assessment	
• Assess the health needs of the community • Investigate the occurrence of adverse health effects and hazards • Analyze the determinants of health needs	• Monitor health status to identify and solve community health needs • Diagnose and investigate health problems and health hazards in the community
Policy Development	
• Advocate for public health, build constituencies, and identify resources in the community • Set priorities among health needs • Develop plans and policies to address priority health needs	• Mobilize community partnerships and action to solve health problems • Develop policies and plans that support individual and community health efforts
Assurance	
• Manage and coordinate resources and develop the public health system's organizational structure • Implement programs by ensuring or providing services • Evaluate programs and provide quality assurance • Inform and educate the public on health issues	• Assure a competent workforce—public health and personal health care • Enforce laws and regulations that protect health and assure safety • Link people to needed personal health services and assure the provision of health care when otherwise unavailable • Evaluate effectiveness, accessibility, and quality of personal and population-based health services • Inform, educate, and empower people about health issues • Research for new insights and innovative solutions to health problems

Health Policy Makers

Health policy has two distinct areas: clinical health policy and social health policy. Matchar and Samsa (2000) define clinical health policy as policies that focus on "the *clinical enterprise* (e.g. should women between the ages of 40 and 49 have a mammogram)…, as well as the structures that support those decisions (e.g. when electronic medical records should be used…)" (pg 146). Social health policy focuses on decisions "that relate to the *context of the clinical enterprise*, including law, reimbursement, access to care, and so forth". In health planning, communities are inspected and individuals are targeted for interventions based on known or predicted risk.

Success indicators are calculated from population statistics in the form of quality of care and organizational performance measures (Derose and Petitti 2003).

Matchar and Samsa (2000) define a health policy maker as “anyone who either makes health related decisions directly or influences the health related decisions of others” (pg. 147). Health policy makers, like other decision makers, have to balance multiple considerations: understanding the causes and consequences of death, disease, and disability. Attempting to put that understanding to work for our collective well-being is a difficult task (Oliver 2006). Though the optimal solution for patient outcome is desirable, other factors such as budgetary constraints and politics are also relevant (Matchar and Samsa 2000; Oliver 2006). This is aggravated by the fact that their decisions need to be defendable under scrutiny (Matchar and Samsa 2000). Decision biases can complicate this decision process for public policy decision making. Public health communication specialists recognize cognitive factors involved in decision-making about health behaviors, and are becoming more sophisticated in addressing them (Maibach and Holtgrave 1995).

In the state of Florida local health councils have been established as a network of non-profit agencies that conduct regional health planning and implementation activities. The Boards of Directors of these councils are composed of health care providers, purchasers and nongovernmental consumers. Florida's eleven councils (ranging in size from one county to 16 counties) develop district health plans containing data, analysis and recommendations that relate to health care status and needs in the community. The recommendations are designed to improve access to health care, reduce disparities in health status, assist state and local governments in the development of sound and rational health care policies and advocate on behalf of the underserved. Local health councils study the impact of various initiatives on the health care system, provide assistance to the public and private sectors, and create and disseminate materials designed to increase their communities' understanding of health care issues.

Health Planning as Knowledge Work

It is useful to view health policy decision making as knowledge work in order to understand the challenges to decision making in this context. When considering a task that is classified as knowledge work, the principal activities performed revolve around "the acquisition, creation, packaging, or application of knowledge"(Drucker 1993; Davenport, Jarvenpaa et al. 1996; Drucker 1999). Acquisition entails the activities required to understand knowledge requirements, searching for the requirements, and then preparing the knowledge for transfer to a requester or user. Creation consists of the research activities and creative tasks that are performed to generate new knowledge. Packaging involves the preparation and assembly of knowledge for consumption by a requester or user. Finally, application consists of the activities that involve the use of existing knowledge in a situation. While the general activities are described in literature on knowledge work, the specific tasks performed by knowledge workers tend to vary greatly from task to task, such that there is little routine or repetition and a greater emphasis on creativity (Davenport, Jarvenpaa et al. 1996; Tremblay, Fuller et al. 2006).

Knowledge workers are defined as employees who apply their own knowledge, acquired through experience and education, to develop new knowledge or apply existing knowledge (Drucker 1999). While most information economy jobs in the 21st century have some knowledge work components, the focus of this research area is on high-level knowledge workers. High-level knowledge workers are highly skilled professionals, who are seen as having three main tasks: (1) the specific tasks in a job that produce valued results for the organization, (2) building their individual knowledge and expertise, through their work and learning efforts, and (3) self-management of their work (Davis 2002). This means that they have both short term (accomplish the task requirements) and long term (maintaining and increasing their own knowledge base) goals and desire autonomy (self management).

The health planners at the agencies that supply the context to this study fit the characteristics of high-level knowledge workers. They are quasi-statisticians and are experts in their domain. Their tasks are not predefined tasks, and often require them to use previous experience as well some intuition to find data that is not easily attained. Though they learn from each task, no two tasks are the same. Most have a master's degree in health services administration. Their work requires that they have a good level of familiarity with word processors and spreadsheets, as well as some basic usage of queries to databases.

Often these health planners apply their own experience and their tasks are not predefined and they have autonomy in how they perform their work(Tremblay, Fuller et al. 2006).

Uncertainty in Knowledge Work

Knowledge workers often make choices under uncertainty, often with inconsistent and incomplete information. Studies (Tversky and Kahneman 1982) have shown that humans will use heuristics ("rules of thumb") to solve such problems, possibly introducing biases and resulting in sub-optimal decisions.

The general heuristics and biases that people use in making judgments are well researched. Though we are mainly interested in strategies of data retrieval and representation that minimize these biases in the health planning context, it is important to understand which heuristics knowledge workers could possible use for decision making as well as the possible biases that could result from the use of these heuristics. Though these heuristics are based on past experience and generally give good results, they can also lead to severe and systematic errors (Tversky and Kahneman 1982). Tversky and Kahneman identify three heuristics that are used to access probabilities of an event that lead to biases in decision making: representativeness, availability and adjustment and anchoring.

Representativeness

Representativeness describes a heuristic used by decision makers in which the probability of an event is judged by how closely it resembles examples that they have available from past experience or memories. Thus, if an event appears similar to a past experience or event it is judged to belong to that event. In some cases this may result in an accurate classification of an event, however several problems exist with this strategy. The decision maker often overlooks factors that should be considered, for example sample size or sample distribution. Tversky and Kahneman identify six factors that lead to incorrect classifications due to representativeness: insensitivity to prior probabilities of outcomes, misconceptions of chance, insensitivity to predictability, the illusion of validity, misconceptions of regression, and insensitivity to sample size.

Insensitivity to Prior Probabilities of Outcomes

Prior probability of outcomes usually will not have any effect on representativeness, but has a strong effect on probability. In experiments conducted by Tversky and Kahneman, subjects were given stereotypical descriptions of certain individuals, allegedly sampled from 100 professionals (either engineers or lawyers). Regardless of information given to the subjects on the amount of engineers and lawyers that made up the sample (one group was told there were 70 engineers and 30 lawyers and another group was told the reverse), subjects gave almost identical probability judgments, paying little attention to the prior probabilities. Interestingly enough, when subjects were not given the individual descriptions they applied the information on prior probability correctly. So, in the case where they were told that the group consisted of 70 engineers and 30 lawyers, they correctly identified the probability that the individual was an engineer to be 70% and a lawyer 30%. Kahneman and Tversky concluded that when no specific evidence was given, proper prior probabilities were used, and when worthless evidence was given prior probabilities were ignored (Tversky and Kahneman 1982).

Misconceptions of Chance

When making decisions, people expect that a sequence of events generated by a random process will represent the essential characteristics of that process (even when the sequence is short). So, for example, if a coin is fair, subjects expect HHH to be followed by a T (also referred to as the gambler's fallacy). Tversky and Kahneman discuss how misconceptions of chance are often present in the evaluation of results of even the most experienced research psychologists, in what to they refer to as the "law of small numbers". Investigators expected that a hypothesis about a population is represented by a statistically significant result, regardless of sample size. This results in too much weight given to results, with overinterpretation of findings from small samples.

Insensitivity to Predicatability

Insensitivity to predictability is similar to insensitivity to sample size in that people do not account for the probability of events. However, insensitivity to predictability refers to ignoring the differential probabilities of the future events. Some events are much more likely to occur than others, but people often view all predictions to be equally likely, or they underestimate the relative differences in predictability.

The illusion of validity

When people feel that an outcome is representative to an input they are confident of their result, regardless of the quality of the input. This confidence is based entirely on the level of fit between the predicted outcome and the information they receive. This overconfidence is most observed when the input information is highly redundant (for example predicting that a student that has all B's his first semester will have a B grade point average at graduation), when, statistically speaking redundancy among input variable *decreases* the accuracy of a prediction.(Tversky and Kahneman 1982) .

Misconceptions of Regression

When examining and comparing samples, one may notice that extreme outliers tend to regress toward the mean (for example, performance on consecutive examinations, height of fathers and sons). However, people intuitively expect succeeding trials to be representative of the previous trial. So often, when they encounter this phenomenon they tend to invent spurious causal explanations(Tversky and Kahneman 1982). For example, predicting that any extremely depressed client will not feel as depressed in the next session is much more likely than the client will become more depressed (Tracey and Rounds 1999).

Insensitivity to Sample Size

When people evaluate the probability of obtaining a certain result from a sample drawn from a certain population they apply a representativeness heuristic. People assess the likelihood of a sample result by asking themselves how similar that sample result is to the properties of the population from which the sample was drawn, regardless of the size of the sample. Tversky and Kahneman outline the following example:

A certain town is served by two hospitals.

- In larger hospital, 45 babies born per day.
- In smaller hospital, 15 babies born per day.
- 50% of babies are boys, but the exact percentage varies from day to day. For a period of 1 year, each hospital recorded the days on which more than 60 percent of the babies born were boys.
  - Which hospital do you think recorded more such days?

Most subjects judged the probability to be about the same for both hospitals, without taking into account that the larger hospital (because it has a much larger sample size) is less likely to stray from 50%.

In this study, the representativeness factor of insensitivity to sample size is selected. An initial field study found this to be a frequent problem (Tremblay, Fuller et al. 2006). In health planning, rates are utilized to present information, but the denominator may not be the same. For example, breast cancer rates may appear to exceed lung cancer rates, when in reality the breast cancer rate is reported per 100,000 women, and the lung cancer rate may be per 100,000 people.

Availability

In certain cases, people may judge the frequency or probability of an event by the ease with which occurrences can be brought to mind, often ignoring other facts that may be relevant. Tversky and Kahneman outline several biases besides frequency and probability that affect availability: biases due to retrievability of instances, biases due to the efficacy of a search set, biases of imaginability and illusory correlation.

Biases Due To Retrievability of Instances

Availability of certain instances may bias a person's judgment on the frequency or probability of a certain event. For instance, a person may see a house burning on their way home from work. This will have more impact on a person's subjective probability of this accident then reading a story in the paper about a house burning (Tversky and Kahneman 1982).   This bias of exposure is one especially relevant to health care. Clinicians use their past and current clients as comparisons so the quality of any decision rests upon the completeness of this sample and our ability to access it completely. Because of the ease of retreivability, few clients serve as an inappropriate basis of clinical comparison for decision-making.(Tracey and Rounds 1999)

Biases Due To the Efficacy of a Search Set

People tend to judge as more probable those events that are easier to search for. For example, a person may think it will be easier to sample at random from a dictionary more words that begin with the letter ‘r’ then words whose third letter is ‘r’ simply because it is easier to search for words by their first letter.

Illusory Correlation

Illusory correlation refers to when two events as having a strong associative bond between them we are likely to judge them as frequently co-occurring. Changinminds.org has a good example:

> “I meet people from around the world. One of the ways I assess people is how generous they are. I meet a person who is very generous. I like them and ask where they are from, which turns out to be Iceland. I later meet another generous person who also turns out to be from Iceland. I assume that most people from Iceland are, by my standards, generous. In fact, I've spoken to many people from Iceland before who were not that generous, but I did not pay attention to their origins.”

Biases of Imaginability

Bias of imaginability refers to the tendency to retrieve information that is plausible without regard to its probability. For example, a certain expedition may be judged as risky because of the description given, even though the probability of the imagined disaster is low (Tversky and Kahneman 1982). Thus, people use imaginability as a flawed indicator of probability of occurrence (Tracey and Rounds 1999). In health planning, this could lead to incorrectly inflating the probability of event due to their imaginability, and the adoption of a very conservative approach toward prevention even in the face of highly unlikely events (Tracey and Rounds 1999). The common occurrence of this factor in health planning makes it a good candidate for this study as an example of a bias of representativeness.

Adjustment and Anchoring

In many situations people are biased in their decision by a value that is specified in the formulation of the problem or by an incomplete calculation carried out in the person's head. This phenomenon is called anchoring (Tversky and Kahneman 1982). Individuals tend not to sway to far from initial information or impressions (their anchor), even when presented with very different information (Tracey and Rounds 1999).

Prospect Theory (Kahneman and Tversky 1979) can help explain this phenomenon. People tend to value a certain gain versus one that is less certain, even when the expected value of each is the same. Tversky and Kahneman told people to assume there was disease affecting 600 people and they had two choices (example from changingminds.org):

- Program A, where 200 of the 600 people will be saved.
- Program B, where there is 33% chance that all 600 people will be saved, and 66% chance that nobody will be saved.

The majority of people selected A, showing a preference for certainty. They then offered them another choice:

- Program C, where 400 people will die.
- Program D, where there is a 33% chance that nobody will die, and 66% chance that all 600 people will die.

Most people now selected D, seeking to avoid the loss of 400 people. This manipulation illustrates how the framing influences the decision made.

Summary

There is a renewed interest in evidence-based business with a focus on competition rooted in analytic capabilities (Davenport, Cohen et al. 2005). The emphasis on business analytics rests on a foundation of sophisticated database technologies and dramatic growth in online data. This is particularly true in public health, where Information Systems are becoming an integral part of evidence-based decision making. Health policy decision makers need reliable, more detailed, and higher quality timely information.

One way to organize data quality efforts is to use the information supply chain to provide a perspective. Information supply chains can be complex, multi-step processes that include the collection of raw data from many sources, intermediate transformations, compositions, and standardizations that ultimately supply the raw data for insightful analysis. Research in data quality provides very limited guidance and rarely has consensus on which data quality metrics should be provided as data quality metadata, how to quantitatively calculate these metrics (Fisher 2002) and where to store them. Several frameworks exist, though often the dimensions suggested conflict with those of other frameworks and often are not mutually exclusive.

Part of a data quality effort is providing information to data consumers so that well known decision biases are not aggravated. Behavioral decision making literature, in particular judgment under uncertainty literature, outlines the heuristics utilized by decision makers. The intersection of the data quality, behavioral decision making and public health literature forms the basis for this research.

Chapter Three: Research Design

Introduction

The literature on data quality (Wang, Reddy et al. 1993) and on information manufacturing systems (Ballou and Pazer 1985; Wang, Reddy et al. 1995; Ballou, Wang et al. 1998; Ballou and Pazer 2003; Shankaranarayan, Ziad et al. 2003; Parssian, Sarkar et al. 2004; Shankaranarayan and Cai 2006; Shankaranarayanan and Cai 2006) has considered the quality of data derived from the information supply chain. These supply chains may rely on human or automated agents to gather and transform data for analytic use directly on the desktop, or indirectly through a more integrated data warehouse infrastructure. Regardless of the path through the information supply chain, the end-user is presented (or helps create) an information product, and examines how data quality characteristics from source data affect the quality of the final information product.

As described in chapter one, many of the metrics proposed in the literature are lineage driven. As information products are created and reused for the creation of new information products, data quality information is tracked and calculated at each step. Maintaining a chain of quality metrics through multiple data transformations and compositions is a challenging task. The context of use or importance of specific data items can add an additional dimension to data quality calculations. This is often handled by having the end user or data quality administrator express their judgment through the assignment of weights or other factors that influence the quality metric calculations. At an abstract level this seems appropriate, but at a practical level, tagging cell-level data is both time and cost prohibitive.

This thesis presents methodologies that communicate *result-driven data quality* (RDQ) information at decision time with simple and comprehensible metrics that can be calculated when the final IP is created. The decision maker is not involved in the calculation of the metric, but considers the metrics as they formulate a context-specific decision. Result-driven data quality is especially important in an environment where managers and decision makers utilize aggregated data (summary information) retrieved from several data sources in the information supply chain to make tactical decisions. This is true in health care, and in particular in health planning, where health care resource allocation is often based on summarized data from a myriad of sources such as hospital admissions, vital statistic records, and specific disease registries. These data are utilized to justify investments in services, reduce inequities in treatment, and rank health care problems to support policy formulation (Berndt et al. 2003).

To design and evaluate the result-driven data quality metrics this thesis utilizes the design science paradigm (Simon 1996; Hevner, March et al. 2004). Figure 4 helps illustrate the research sequence, which is based on Hevner, March et al. (2004) conceptual framework for information systems research. To identify potential data quality measures and biases, a field study is conducted in a Florida Health Planning Agency. Results from the field study and a review of the literature help with the selection of the data quality issues and biases on which to focus this research. The data quality metrics are designed and implemented with simple Online Analytical Processing interfaces in order to present these metrics to decision makers. Evaluation methods are explored, and the focus group technique is selected as the evaluation technique. The metrics were evaluated using two types of focus groups: exploratory and confirmatory.

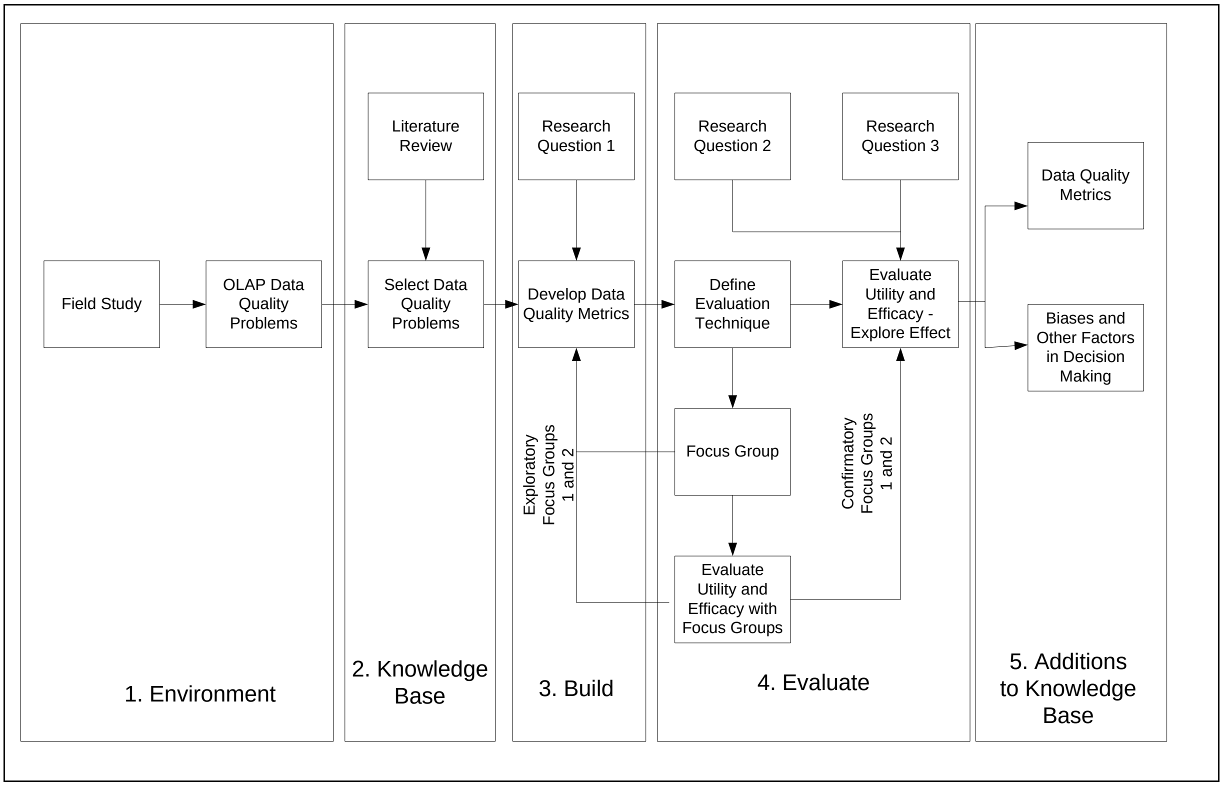

Literature Review
Research Question 1
Research Question 2
Research Question 3
Data Quality Metrics
Field Study
OLAP Data Quality Problems
Select Data Quality Problems
Develop Data Quality Metrics
Define Evaluation Technique
Evaluate Utility and Efficacy - Explore Effect
Biases and Other Factors in Decision Making
Exploratory Focus Groups 1 and 2
Focus Group
Confirmatory Focus Groups 1 and 2
Evaluate Utility and Efficacy with Focus Groups
1. Environment
2. Knowledge Base
3. Build
4. Evaluate
5. Additions to Knowledge Base

Figure 4 – Research Agenda

This chapter is organized as follows: an outline of the research model is given, followed by a description of the field study and the theoretical development of the metrics (Chapters 4, 5, 6 cover each of these metrics and bias mitigation techniques individually), and concludes with an explanation of the evaluation method.

Research Model

The research model for this study is based on Hevner et al.'s (2004) framework for information systems research (Figure 5). Utilizing the design research cycle, this research *builds* an artifact with the intention to solve an identified organizational problem and *evaluates* the artifact in an appropriate context to both provide feedback to the design process and a better understanding of the process (Hevner, March et al. 2004). In this research, the artifact consists of the result-driven data quality *metrics* (RDQM) that are *instantiated* (Markus, Majchrzak et al. 2002; Hevner, March et al. 2004) and *evaluated* with the use of focus groups consisting of healthcare and data warehousing experts. The RDQMs improve decision making for health planners utilizing an OLAP (Online Analytical Processing) environment by providing information about the quality of aggregated data (summary information). The design of these RDQMs is informed by database theories on data quality, finance literature on time series data and behavioral decision making theories.

This research employs multiple methods of inquiry: one field study, two exploratory focus groups and two confirmatory focus groups. The field study (Tremblay, Fuller et al. 2006) helps the researcher better understand the technical problem and the context of healthcare planning, in particular how these knowledge workers utilized currently available business intelligence tools, to identify issues in data quality in the health panning context, and finally helps to design the focus group tasks which are utilized to prototype and evaluate the RDQMs. The exploratory focus groups provide feedback for improvement of the design of the artifact (Markus, Majchrzak et al. 2002; Hevner, March et al. 2004) and help refine the coding scheme for the confirmatory focus groups. Finally, two confirmatory focus groups evaluate the utility of the RDQMs.

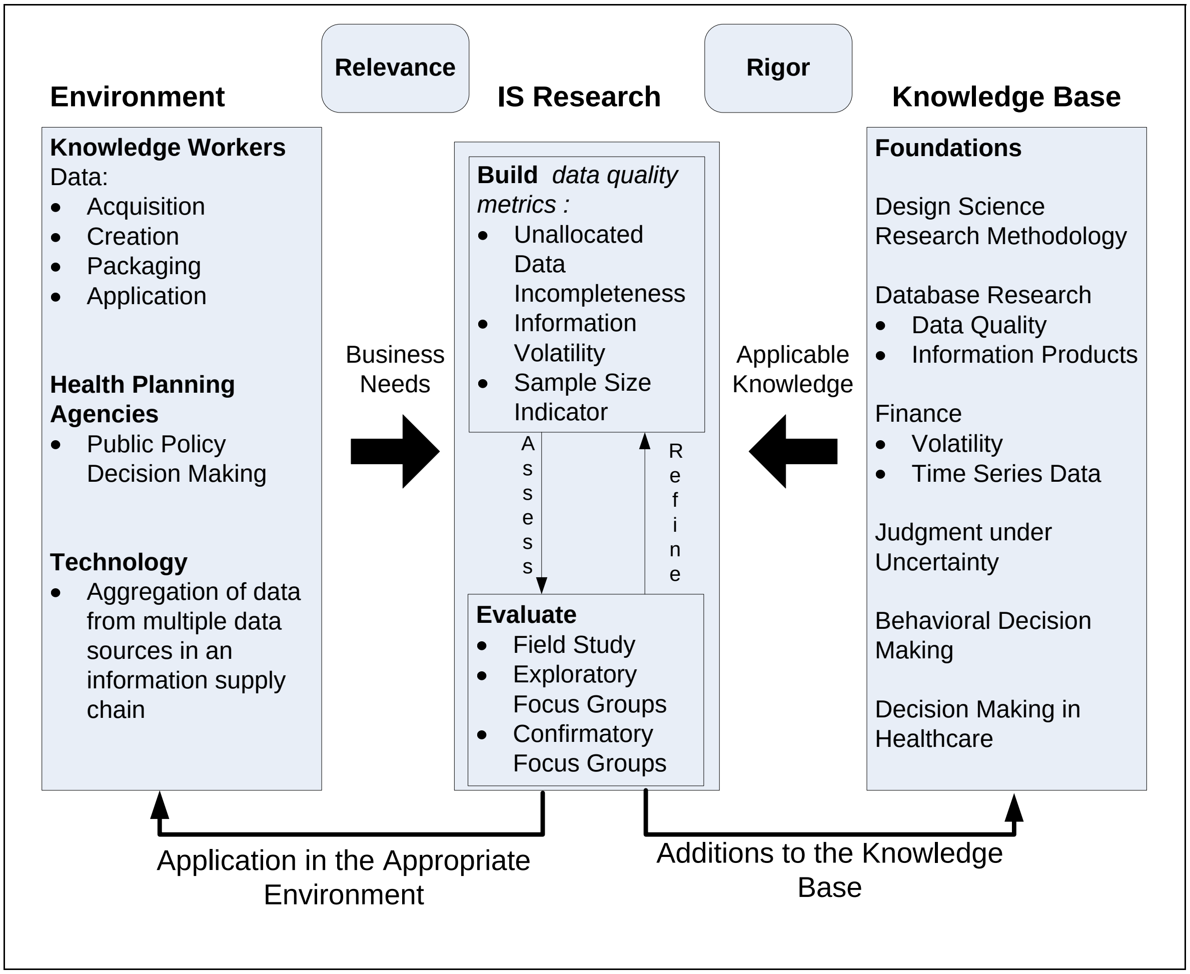

Relevance
Rigor
Environment
IS Research
Knowledge Base
Knowledge Workers
Data:
• Acquisition
• Creation
• Packaging
• Application
Health Planning Agencies
• Public Policy Decision Making
Technology
• Aggregation of data from multiple data sources in an information supply chain
Business Needs
Build data quality metrics :
• Unallocated Data Incompleteness
• Information Volatility
• Sample Size Indicator
Assess
Refine
Evaluate
• Field Study
• Exploratory Focus Groups
• Confirmatory Focus Groups
Applicable Knowledge
Foundations
Design Science Research Methodology
Database Research
• Data Quality
• Information Products
Finance
• Volatility
• Time Series Data
Judgment under Uncertainty
Behavioral Decision Making
Decision Making in Healthcare
Application in the Appropriate Environment
Additions to the Knowledge Base

Figure 5 - Research Model (adapted from Hevner et al. 2004)

Field Study – Understanding the Environment

On-Line Analytical Processing (OLAP) is an example of a new breed of business intelligence tools that give decision makers the flexibility to customize the selection, aggregation, and presentation of data. In an OLAP environment, analytic information is typically represented as data cubes. Business analysts can then slice through the data cube in many ways, creating unique information products with each cut. Appendix A provides an introduction to Data Warehousing and Online Analytic Processing.

To understand the impact of this type of tool, we studied an implementation of an OLAP interface on the Comprehensive Assessment for Tracking Community Health

(CATCH) data warehouse used by knowledge workers at a regional health planning agency in the State of Florida (Berndt, Hevner et al. 2003). The field study provided an exceptional opportunity to study knowledge workers in a real life context. It offered a rich understanding of the health planner's tasks and their use of Business Intelligence technology. The results of this field study are forthcoming in a special issue on Decision Support in Medicine of Decision Support Systems (Tremblay, Fuller et al. 2006).

Several ideas emerged from observing the health planners interact with this business intelligence technology. We observed that their individual and task characteristics evolved, as did the outcomes. In addition, as their level of expertise with the OLAP tool increased, their job roles began to change. Rather than remain data collectors, they began performing more as consultants. Before the implementation of a data warehouse and decision support system, their job consisted of finding data and providing it to their customers. Using the OLAP interface, they began providing their users with highly detailed data, along with interpretations and descriptions. They also used individual judgment to advise their clients which data they really needed. The health planners were no longer only acquiring and packaging knowledge, but creating and applying it as well.

Observations from this field study are the motivation behind the selection of the metrics and presentation methods, which were identified as problematic issues in the use of BI tools in the health planning context. In particular, we noticed that decision makers did not have information about the level of completeness, data consistency and amount of data utilized in the reported summarized data, which led to incorrect decision making. Table 10 contains some example quotes from the field study.

Table 10 – Example Quotes – Data Quality Problems

Data Consistency: “Not that I mistrust it, but sometimes there are goofy things with the data. For example, for some reason for the year 2000 the rates were just double what they were for the previous 2 years and the following 2 years. What caused this, are you sure it’s right, because as soon as I go and present this data they are going to ask what happened in 2000? Why is this like that? We were not able to find out why – it was correct – maybe for that year they were counted differently, maybe due to some piece of legislation.”
Amount of Data Used: Another case they provided the data to help locate a screening center for cervical cancer. In this case a lower granularity proved not to be as valuable: “one person wanted to do some kind of special oncology care for women. Basically breast cancer, or cervical cancer. She asked for late stage cervical cancer by zip code. Unfortunately, we don’t have that. Because the cervical cancer rates are really pretty low, that once you break them down to zip codes the numbers are useless. They are too low to be significant, so if you have ‘1’ in a zip code you really can’t use it. I was able to give it to her [the data] by county, and by zip code for other cancers, but not for late stage cervical cancer.”

Theoretical Development from the Knowledge Base and the Field Study

Table 11 illustrates data quality pitfalls observed in the healthcare field study for the three data quality dimensions These pitfalls are described by the following data quality dimensions in the literature: *completeness, representational consistency* (Wang, Reddy et al. 1995; Wang and Strong 1996; Jarke, Jeusfeld et al. 1999), and whether the IP contains the *appropriate amount of data* needed for the decision (Kahn, Strong et al. 2002; Pipino, Lee et al. 2002). Figure 6 shows these dimensions in Wang et al.’s data quality framework and where in the framework these quality dimensions fall.

Table 11 – Potential Data Quality Pitfalls

Data Quality Dimension	Example Pitfall
Completeness	Hospital discharges occur continuously. But not all hospitals choose to send their data at the same rate. Hospitals are continuously collecting data; but they may differ in their batching and transmission strategies. Some hospitals may send incomplete data, filling in information with later transmissions. Some of the data may be purposely set to null because of privacy and security issues (such as sensitive information on the location of AIDS cases). When decisions are made with incomplete data, knowledge workers should know the extent of incomplete data.
Representational Consistency	Different data sources report data with different definitions for their calculations. Furthermore, a change in IT staff could result in definitional changes within a single data source. Somewhat unpredictable trends may emerge, when in fact they are due to the volatility of the data. For example health planners noticed a trend in heart disease that looked like a sine wave. This trend was due to changes in the data definitions. Another cause of inconsistency is seasonal changes (as is the case in Florida with migrant workers and "snowbirds") or scarcely populated groupings.
Appropriate Amount of Data	Disease rates and averages are often compared across regional areas, or by time periods. Attention should be paid to the volume. For example, large counties should not be compared to smaller counties where the volumes are low. Furthermore, past literature shows that even if volumes are reported most people tend to ignore this information (Kahneman and Tversky 1979).

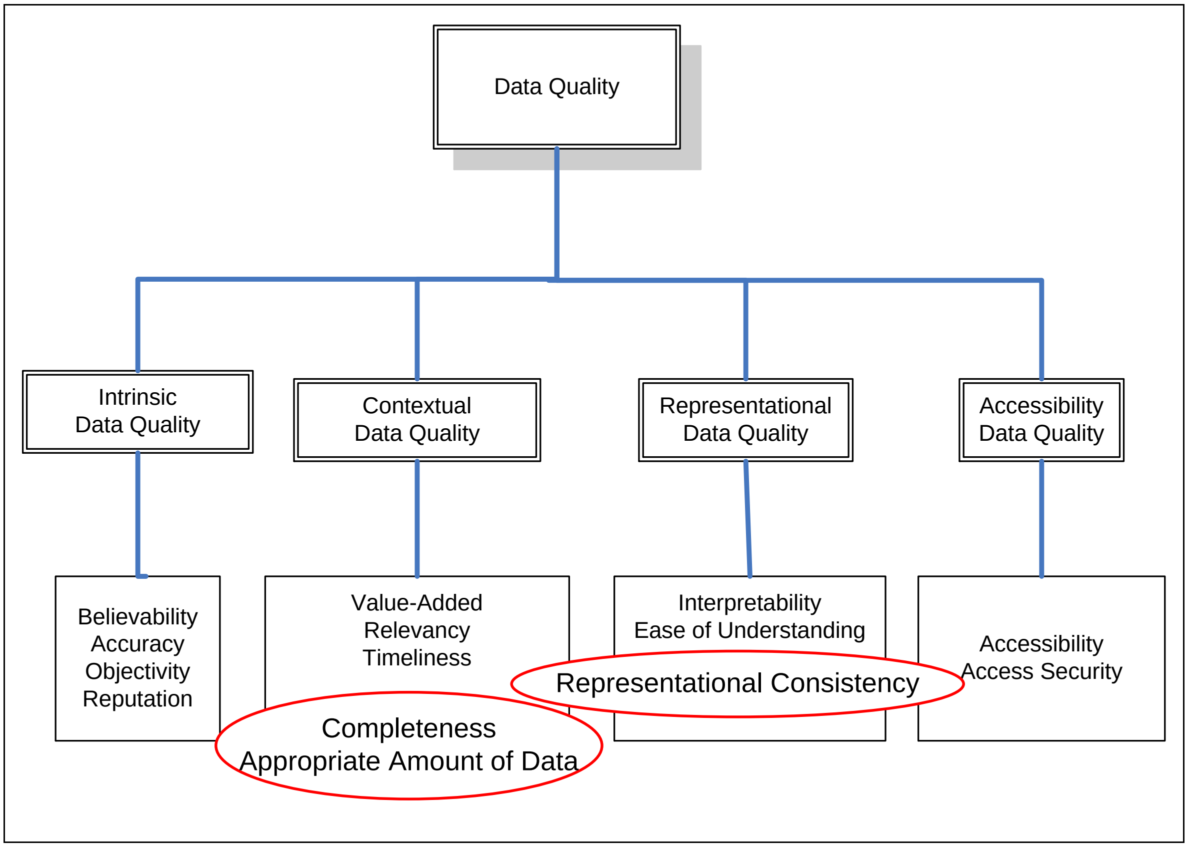


Figure 6 - Selected Dimensions from Conceptual Framework of Data Quality

Building the Result-driven Data Quality Metrics

Regardless of data cleansing processes in an information supply chain, there will still be data errors and peculiarities in the data which are probably, but not necessarily, due to inaccuracy in the data. In fact, in healthcare, there are many possible reasons why data from the information supply chain can be incorrect when aggregated (some examples are shown in Table 11). Information about these errors is not generally presented to decision makers, who will make choices and decisions based on the available data. In fact, most database queries are run without any data quality information. This is an especially troublesome issue in analytic databases (as compared with transactional systems), because tracing and correcting these errors can be expensive, and at times impossible. However, threats to decision quality can be reduced by informing the information consumer about the data quality at decision time (Parssian 2006). Decision makers can be further aided by having some flexibility in the consideration of the effect of these data quality problems on different scenarios.

We consider how to present information on the three data quality dimensions for any unique information product in an OLAP environment. This thesis proposes three data quality measures and associated metrics (DQMs) which are summarized in Table 12.

Table 12 – Data Quality Metrics

Data Quality Dimension	**DQM**	**Problem**
Completeness	Unallocated data	Null values in any of the grouping variables
Representational Consistency	Information Volatility	Inconsistency in data values
Appropriate Amount of Data	Sample Size Indicator	Insensitivity to sample size by decision makers when considering/comparing groupings

Unallocated data

Past research has considered the some of the effects of inaccurate or missing data on information products (Imielinski and Lipski 1984; Ballou and Pazer 1985; Ballou, Wang et al. 1998; Ballou and Pazer 2003; Parssian, Sarkar et al. 2004; Parssian 2006). Parssian (2006, 2004) defines two types of nulls: existential nulls and non-existential nulls. Existential nulls are values that arrive as incomplete from the supplier; the non-existential nulls are data that do not exist in the real world (for example number of live births for a male). It is possible that the attributes are null either for an identifier[1] attribute or a non-identifier attribute (Parssian, Sarkar et al. 2004). This could also compromise the accuracy of aggregated fields. Several studies suggest methodologies to estimate the correct value (Ballou, Wang et al. 1998; Shankaranarayan, Ziad et al. 2003; Parssian, Sarkar et al. 2004; Burdick, Deshpande et al. 2006; Timko, Dyreson et al. 2006).

The unallocated data (UD) metric considers the effects of null values in any of the grouping or filtering variables, providing an operational definition for aspects of incompleteness. When information products that contain aggregated data are created, the common strategy is to map null values to a single "unknown" category so the nulls will group together. The amount and size of unallocated data groupings will be different depending on how the knowledge worker cuts or slices through the data (and navigates through a dimensional hierarchy). This can be fairly complex, especially as the number of group by variables used in the information product becomes large. The UD metric and presentation methods are intended to highlight the impact of incompleteness on data cubes.

Information Volatility

A second important data quality concept is data consistency and the related concept of volatility. There are several definitions of data volatility in the literature.

[1] In a relational model, if referential integrity is enforced there should be no nulls in the identifiers, but in this case nulls are coded as a field that maps to "unknown" in the corresponding look up table or dimension.

Most quality frameworks consider volatility as a part of timeliness of the data (Wang, Reddy et al. 1995; Wang, Storey et al. 1995; Ballou, Wang et al. 1998). From this information product perspective, volatility is analogous to shelf life. Shelf life is less important when products do not spoil; while critical when they need to be sold within a certain window. Similarly, raw data or information products have a length of time during which they are valid. Highly volatile data have a short shelf life, while others are infinite (Ballou et al. 1998b).

This study differs from the existing literature in the definition of data volatility. Here it is defined as a measure of consistency in data values, rather than relating to timeliness or shelf life. This thesis proposes a measure of reliability called *information volatility*. Information volatility is defined as the rate of change in the values of stored data. It follows that data that exhibits unpredictable changes are considered highly volatile. Business intelligence tools rarely offer any form of reliability measures. When considering aggregated data, or when observing trends decision makers rely on point estimates, such as an average, when, in fact, these aggregated values may be biased by noisy data. Supplying decision makers with information about the reliability of the data should improve the quality of their decisions. A descriptive analysis of the data can often provide an understanding of any unusual patterns.

Sample Size Indicator

Insensitivity to sample size is a form of the well-known representativeness bias (Kahneman and Tversky 1979). When people evaluate the probability of obtaining a certain result from a sample drawn from a certain population they apply a representativeness heuristic. People assess the likelihood of a sample result by asking themselves how similar that sample result is to the properties of the population from which the sample was drawn, regardless of the size of the sample. Choices of presentation of the data are essential to effectively mitigate well known judgment biases. This metric differs from the other two, in that it is not a calculation, but rather investigates drawing the attention of a decision maker to aggregated data based on small sample sizes. This research investigates how in a BI tool, the data presentation can be utilized to mitigate the bias of insensitivity to sample size by drawing attention to sample sizes.

Evaluation of the Data Quality Metrics

Several methods exist to evaluate designs: observation, analytics, experiments, testing or descriptive, and more recently action research (Baskerville and Myers 2004; Hevner, March et al. 2004; Jakob, Lars et al. 2004; Rikard, Ola et al. 2004; Cole, Purao et al. 2005). This study employs focus groups, in particular two exploratory focus groups and two confirmatory focus groups.

Focus Groups are well known both in management disciplines and healthcare research (particularly by clinicians) (Morgan 1988; Krueger and Casey 2000; Anonymous 2006; Stewart, Shamdasani et al. 2007). Several software engineers have also suggested their use as an evaluation and knowledge elicitation technique (Massey and Wallace 1991; Nielsen 1997; Kontio, Lehtola et al. 2004; Anonymous 2006).

A focus group is a moderated discussion among 6-12 people who discuss a topic under the direction of a moderator, whose role is to promote interaction and keep the discussion on the topic of interest (Stewart, Shamdasani et al. 2007). A typical focus group lasts about two hours and covers a range of topics that are decided on beforehand.

The focus group technique is utilized in social research to study ideas in a group context (Morgan 1988). The term *focus* in the title refers to the fact that interview is limited to a small number of issues (Stewart, Shamdasani et al. 2007). It has been found effective both as a self-contained means of collecting data (as a primary research tool) or as a supplement to other methods of research (as a secondary research tool) (Krueger et al. 2000; Morgan 1988).

The focus group technique is particularly useful as an exploratory method, when little is known about the phenomenon, but also can be used as a confirmatory method to test hypotheses (Stewart, Shamdasani et al. 2007). Focus groups can be valuable to gain shared understandings, but allow for individual differences in opinion to be voiced. There are several reasons focus groups seemed as an appropriate evaluation technique for this study (based on Stewart et al. (2007), pg.42):

1. Flexibility. Focus groups allow for an open format, and are flexible to handle a wide range of topics. Our study investigated three different metrics within the same context. Other evaluation methods (such as a designed experiment), would have been difficult to design, unless each of the metrics were considered separately.

2. Direct Interaction with Respondents. This allowed for the researcher to clarify any questions about the metrics as well as probing the respondents on certain issues.

3. Large Amounts of Rich Data. This data allowed a deeper understanding, not only on the respondents' reaction and use of the metrics, but other issues that accompany the use of data quality information.

4. Building on Other Respondent's Comments. The group setting allowed for the emergence of ideas or opinions that would not have been uncovered in individual interviews. Additionally, causes of disagreement pointed to possible problematic areas.

Information systems researchers have called for a broader variety of available empirical methods to improve relevance of research (Galliers 1991; Benbasat and Weber 1996), yet few have embraced the focus group approach. IS research has mostly utilized focus groups in conjunction with other empirical methods (Mantei and Teorey 1989; Manning 1996; Smith, Milberg et al. 1996; Debreceny, Putterill et al. 2003; Baker and Collier 2005; Jarvenpaa and Lang 2005; Xia and Lee 2005; Torkzadeh, Chang et al. 2006). Very few utilize focus groups to evaluate a design science artifact (Mantei and Teorey 1989), though this could be due to the novelty of the method in the IS field. Table 13 outlines some examples of how focus groups have been utilized in the major IS journals literature.

Similarly, the software engineering community has suggested a need for a wider availability of empirical methods to improve validity and generalizability of their designs, and several have utilized focus groups (Basili 1996; Kontio, Lehtola et al. 2004). In the IT industry, focus groups are widely used in information systems usability studies, as a replacement for usability testing or contextual interviews and produce different kinds of information. For example, usability.gov is a U.S. Government Web site managed by the U.S. Department of Health & Human Services that outlines the use of Focus groups in the design of web pages (see http://www.usability.gov/methods/focusgroup.html).

Table 13 – Focus Group use in Information Systems

Authors	**Journal / Year**	**Primary Research Tool**	**Design Science**	**Focus Group Use**
Mantei, M. M. and T. J. Teorey	MIS Quarterly/1989	No	Yes	Generate ideas about problems of database retrievals (to be utilized in interface design)
Xia, W. and G. Lee	Journal of Management Information Systems/2005	No	No	Conceptual Development of information systems development project (ISDP) complexity
Jarvenpaa, S. L. and K. R. Lang	Information Systems Management / 2005	Yes	No	Understand mobile technology use. Focus groups were utilized to capture shared reactions, issues, experiences and opinions.
Debreceny, R., M. Putterill, et al.	Decision Support Systems / 2002	No	No	Identify electronic commerce issues (managers in firms contemplating electronic commerce activity)
Torkzadeh, G., J. C.-J. Chang, et al.	Decision Support Systems / 2006	No	No	Generate scenarios and issues to barriers to CRM success
Smith, H. J., S. J. Milberg, et al.	MIS Quarterly/1996	No	No	Development of instrument that identifies and measures the primary dimensions of individuals' concerns about organizational information privacy practices

Focus Group Methodology

The focus groups consisted of the six steps outlined below (Kontio, Lehtola et al. 2004):

1. Defining the research problem
2. Selecting participants
3. Planning the event
4. Conducting the Focus Groups
5. Data analysis

6. Reporting

Research Problem

The goal of the focus groups was to address research questions two and three: the evaluation of the utility and efficacy of the developed metrics and to understand if these metrics were salient to a decision maker, in particular, if they would help alter, enhance or sway a decision by changing the way decision makers analyzed the data (data analytic strategies). Two types of focus groups were used: exploratory and confirmatory. The exploratory focus groups had two roles: 1) the provision of feedback to be utilized for design changes to both the metrics and to the focus group script, 2) the identification of the constructs to be utilized in the coding scheme. The confirmatory focus groups were used to understand the particular implications to the research question (Stewart, Shamdasani et al. 2007)

Planning the Focus Group Sessions

A total of four focus groups were planned. The planning process included creating a carefully planned script; in which all three of the metrics were presented to the participants (Script is included as Appendix B). This research utilized the "rolling interview guide" (Stewart, Shamdasani et al. 2007) for the first two focus groups, which are referred to as "exploratory focus groups". With a rolling interview guide a script was created for the first group, based on the outcome of the first exploratory focus group the guide is revised for use in the second exploratory focus group. Based on the outcome of the second exploratory focus group, both the script and the metrics are revised. One of the advantages of this approach is that it allows information to unfold over time as we discovered more about how people would understand and use the metrics (Stewart, Shamdasani et al. 2007). There were no revisions made after the confirmatory focus groups, since continuous change makes comparisons across focus groups difficult (Stewart, Shamdasani et al. 2007).

"Vignettes" or story lines were used to create fictitious decision scenarios based on current healthcare situations (in recent news reports) and data from a sample healthcare ISC. This ISC includes data from Florida's statewide cancer registry, which has been collecting incidence data since 1981, county data from the US Census Bureau, demographic data from commercial sources, and an internally generated time dimension. The strategy was to present the data with and without the metric information in order to detect differences in the collective decision making process. A PowerPoint presentation was also used to help describe the vignettes and the metrics. The moderator outlined the decision context. An example scenario was "Imagine that you are in a position where you help define public policy. For example, you are making decisions about where in the state you may open a cancer center, or whether a certain ethnicity or race is underserved…"

Participants

Participants were recruited with a phone call in which the study was described and their participation was requested. Some of the participants were enrolled in a Data Warehousing or Data Mining course. Other participants were part of the local VA hospital. Thus, the selection of participant was not completely randomized, but rather a convenience sample. Copy of the telephone call script is in Appendix C. The participants had to have the following requirements: previous experience with decision support software, a college degree (many had advanced degrees), some training in statistics and healthcare experience was preferred. Most of the participants were people that completed the data warehousing and/or the data mining course in the ISDS department in the USF College of Business, or acquaintances whose job requires a high use of data analytics (spreadsheets, business intelligence tools, statistics packages).

Conducting the Focus Groups

The focus groups were held in conference rooms. The participants were seated in a U-shape arrangement to encourage collaboration (Krueger and Casey 2000) and allow space for the moderator to demonstrate the tool and PowerPoint presentation. The moderator presented the scenarios but tried to include as much flexibility as possible, in order to approximate individual use. For example, participants were encouraged to ask the moderator to drill down or roll up, to observe data for different counties as part of their decision making process. The sessions were recorded and professionally transcribed. As recompense the participants received lunch or dinner.

Data Analysis

The interpretation of the focus group discussions is an important step. The content of the focus groups was analyzed, carefully selecting techniques that emphasize the reliability and replicability of the observations and results (Stewart, Shamdasani et al. 2007). To accomplish this, the focus groups were coded using template analysis. This technique was selected because of its flexibility. Unlike a grounded theory approach (Desanctis and R. 1987), template analysis normally starts with at least a few predefined codes which help guide analysis. The first step in template analysis is creating an initial template by exploring the focus group transcripts, academic literature, the researchers own personal experience, anecdotal and informal evidence, and other exploratory research (King 1998). In template analysis, the initial template is applied in order to analyze the text, but is revised during ongoing analysis (King 1998).

The best approach to create the initial template is to begin with a few pre-defined codes, which usually revolve around the topic guide (King 1998) – which in our case was the data quality metrics. The contents of the discussions are also examined for the meaning and its particular implications for the research questions, such as changes in data analytic strategies and evidence or counter-evidence of the metric's usefulness. Individual constructs were investigated, looking for common themes and variations within the constructs that would provide rich description of the participants' reactions to design features and attitudes to decision making with data quality. In addition, several other coding categories were created during coding to explore the entire range of participants' reactions. Cohen's Kappa was used to assess inter-rater reliability. Cohen's Kappa is thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance (Cohen 1960)

Reporting

King (1998) outlines three common approaches to present the researcher's interpretation of the data. The approach taken was to create an account structured around the main themes identified (usefulness, efficacy), drawing illustrative examples from each transcript as required. As recommended by King (King 1998) direct codes from participants are included: short quotes to aid in the specific points of interpretation and longer passages of quotation to give a flavor of the original discussions. A final stage of the analysis was to look at key relationships between the constructs. One important set of relationships investigated the change in decision making strategies once the participants received data quality information.

Summary

This chapter outlines the methodology utilized to design three simple and comprehensible result-driven data quality metrics. A field study and the literature are used in the identification of these metrics. Once these metrics were designed they are evaluated through the use of focus groups. The focus groups are transcribed and analyzed though the use of content analysis.

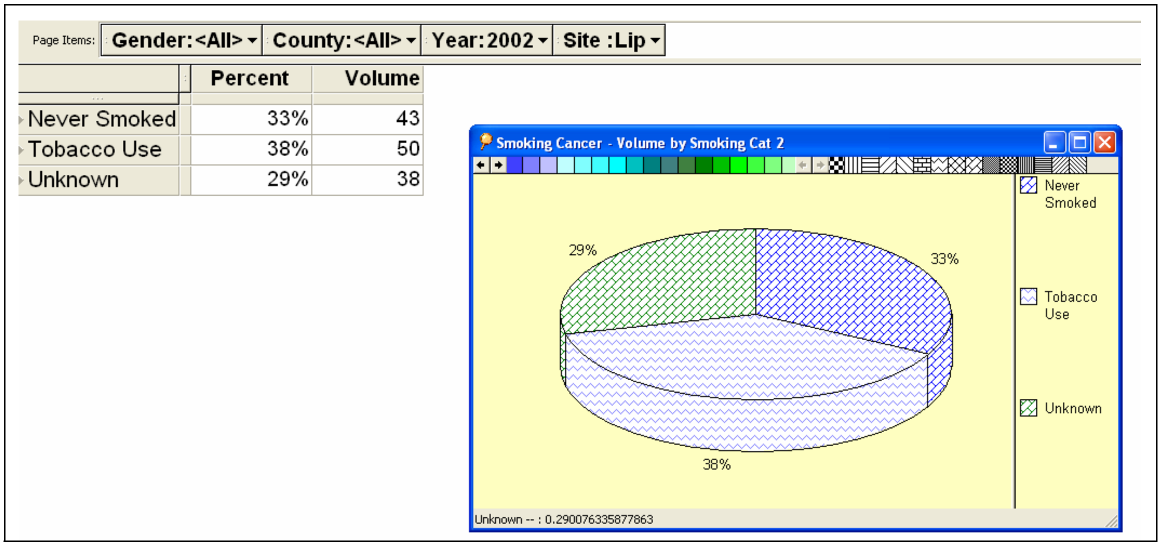
Chapter Four: Unallocated Data

Introduction

One of the problems encountered when combining or aggregating data from multiple sources in the information supply chain is missing codes and codes that do not match other sources of data, which results in data that is not assigned to any of the possible cells in a data cube. The unallocated data (UD) metric considers the effects of null values in any of the grouping or filtering variables, providing an operational definition for aspects of incompleteness. When information products that contain aggregated data are created, the common strategy is to map null values to a single "unknown" category so the nulls will group together. The amount and size of unallocated data groupings will be different depending on how the knowledge worker cuts or slices through the data (and navigates through a dimensional hierarchy). This can be fairly complex, especially as the number of group by variables used in the information product becomes large. The UD metric and presentation methods are intended to highlight the impact of incompleteness on data cubes. We first utilize an example to explain our approach for handling unallocated data and then discuss the detailed calculations.

Figure 7 illustrates a simple example of UD. In this case, the cube was formed with data from several sources in the ISC, including Florida's statewide cancer registry, which has been collecting incidence data since 1981. The University of Miami Miller School of Medicine has been maintaining the FCDS (Florida Cancer Data System at http://fcds.med.miami.edu) since that time. In addition, county data from the US Census Bureau, demographic data from commercial sources, and an internally generated time dimension were used to construct data cubes.

We consider UD along only one attribute to simplify the first example. In this case, we consider the smoking status of patients diagnosed with cancer. In this particular view of the cube the decision maker has selected a single year of 2002 and cancer of the lip. In Figure 7, a large share (29%) of the data on smoking is unknown. This could threaten any conclusions one might draw linking smoking to health issues of lip cancer.



Page Items: Gender:<All> County:<All> Year:2002 Site :Lip

	Percent	Volume
Never Smoked	33%	43
Tobacco Use	38%	50
Unknown	29%	38

Smoking Cancer - Volume by Smoking Cat 2

Unknown -- : 0.290076335877863

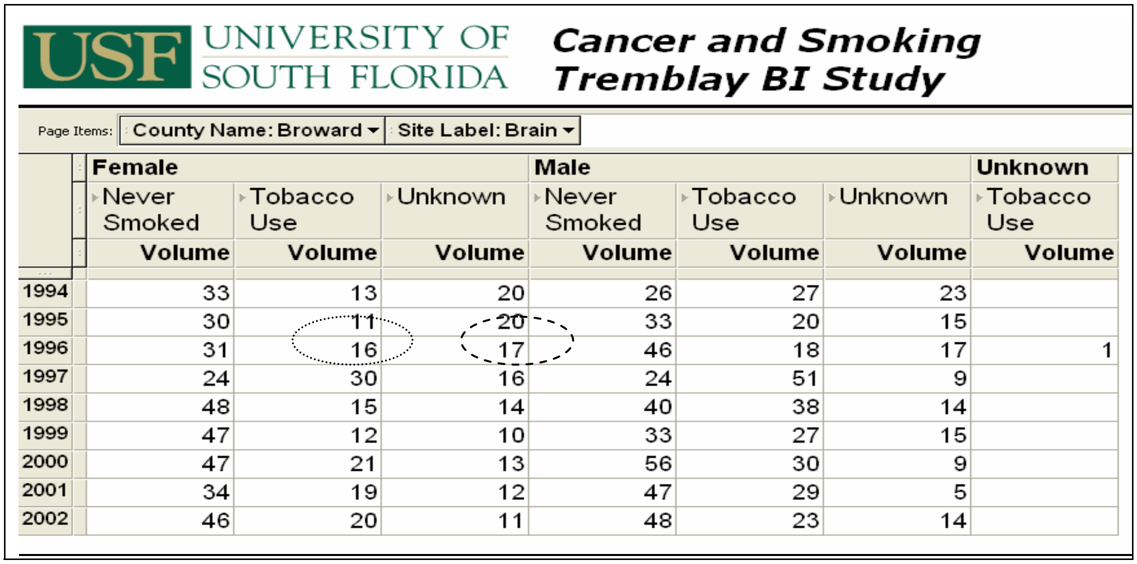
Figure 7 – Unallocated Smoking Data

Another example IP is formed using the query in Figure 8 . The resulting IP is shown in Figure 9. A decision-maker wishes to compare yearly cancer volumes for smokers and non-smokers broken down by gender. For this particular data cube, the decision maker has the ability to filter by county, by the type of cancer, and additionally to drill along the time dimension (for example drilling to monthly data for a particular year). There could be unknowns in the data for the aggregations fields: year (or at the lower granularities along this dimension), tobacco use, gender, county, or cancer site, either in one field or for all the possible combinations.

```
SELECT      tobacco_category,
            cat_year,
            county_name,
            gender_name,
            site_label,
            sum(cat_count)
FROM        counties,
            fcds_tobacco,
            fcds_cancers,
            fcds_genders,
            fcds_sites
WHERE       tobacco_code =fcds_tobacco AND
            county_id = cat_county  AND
            gender_code = gender  AND
            code = fcds_site_grp
GROUP BY    tobacco_category,
            cat_year,
            county_name,
            gender_name,
            site_label;
```

Figure 8 – Query for Smoking Cube

Figure 9 illustrates how the amount of unallocated data can be shown by labeling nulls as unknowns and including counts for these fields (in this example there were no unknowns for year and the unknown for the filtering fields are not shown). For example, for 1996, in Broward County, there were 16 women that smoked that were diagnosed with brain cancer. Yet for 17 women it was unknown whether they smoked or not. Furthermore, there is unallocated data for gender, and possibly for county, year and cancer site.



Cancer and Smoking
Tremblay BI Study

Page Items: County Name: Broward | Site Label: Brain

	Female			Male			Unknown
	Never Smoked	Tobacco Use	Unknown	Never Smoked	Tobacco Use	Unknown	Tobacco Use
	Volume	Volume	Volume	Volume	Volume	Volume	Volume
1994	33	13	20	26	27	23	
1995	30	11	20	33	20	15	
1996	31	16	17	46	18	17	1
1997	24	30	16	24	51	9	
1998	48	15	14	40	38	14	
1999	47	12	10	33	27	15	
2000	47	21	13	56	30	9	
2001	34	19	12	47	29	5	
2002	46	20	11	48	23	14	

Figure 9 – More Complex Unallocated Smoking Data Example

Two questions arise: what is the best way to present the information on unallocated data to a decision-maker and how does that information affect a decision-maker's decision. The approach illustrated in Figure 7- Figure 9, is simply to display the amount of unknown data. This can be done by replacing the nulls with unknowns, thus that data can be displayed with the use of a query[2]. However, this can be cognitively taxing for a decision maker, since the he needs to consider many unallocated fields in the formulation of his decision.

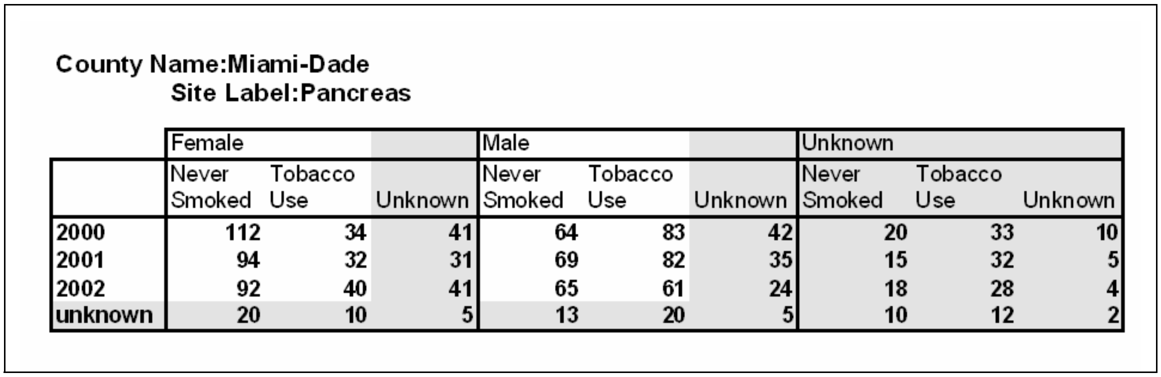
The approach we suggest is to proportionally distribute the unallocated data using the dimensions form that the cube. Though this is a reasonable assumption, one potential downside is that if there is some sort of systematic bias in the data it will be magnified by this distribution heuristic. For example, if a certain ethnic group or gender is culturally

[2] In the case of counts, nulls values in the non-identifying attribute are not an issue, since they are also are replaced with nulls.

embarrassed by a certain disease, and tends not to seek care, proportionally distributing this data according to all the dimensions will result in not accurately attributing some of the unallocated data to this group. Another approach is to allocate the data according to a subset of the data (not considering all the dimensions). For example, in our sample cube, we may ignore the disease dimension (type of cancer) and may choose to distribute the data according to the county and time dimension (proportion of males and females in a certain year for a certain county). Certainly, these are valid concerns, and decisions should be made after careful investigations on a case by case basis. When implementing proportional allocation as a design feature, these may be options that the decision maker sets. However, some reasonable default policy for distributing unallocated data should available without burdening the user. In addition to unknown grouping attributes, data may be null within the measure (non-identifier) field. Some past literature has considered imprecise data within the grouped measure (Burdick, Deshpande et al. 2006), but the focus here is on the group-by attributes which determine the pattern of allocation resulting from a data cube query.

Proportional Data Allocation

In this study, we propose a method for proportionally allocating data when faced with unknown values in all grouping dimensions, though the algorithm can be used for any subset of these dimensions as well. Also, we have to consider that there are many possible aggregations, including counts, averages, rates, min or max. In this study two are considered: counts and averages, since these are among the most common aggregation methods and illustrate the issues when simply counting or calculating within a measure. Other aggregation functions are left to future research.



County Name:Miami-Dade
Site Label:Pancreas

	Female			Male			Unknown		
	Never Smoked	Tobacco Use	Unknown	Never Smoked	Tobacco Use	Unknown	Never Smoked	Tobacco Use	Unknown
2000	112	34	41	64	83	42	20	33	10
2001	94	32	31	69	82	35	15	32	5
2002	92	40	41	65	61	24	18	28	4
unknown	20	10	5	13	20	5	10	12	2

Figure 10 – Unallocated Data Values for Pancreatic Cancer

We apply the proportional allocation approach and illustrate the results with an example, including the use of "unallocated data cubes" or UDCs. UDCs show which cells are affected by the unallocated data, and help visualize how the probabilities can be used to allocate this data. Figure 10 shows a sample IP with the unallocated data highlighted in grey (unknown county or unknown site of cancer are not shown–we omit them in order to simplify our example).

Unallocated Data Cubes

Figure 11 illustrates one version of the UDC for the data from Figure 10. Though some data points fall in a cell, the overlapping circles illustrate data that could be allocated to any of the cells it touches. Since the cube is not transparent, it is impossible for all allocations of unknown data to be shown. Thus, we "spin" the cube, or remove some of the layers of the cube to reveal other unallocated data. Several views are needed to fully illustrate where unallocated data could be assigned.

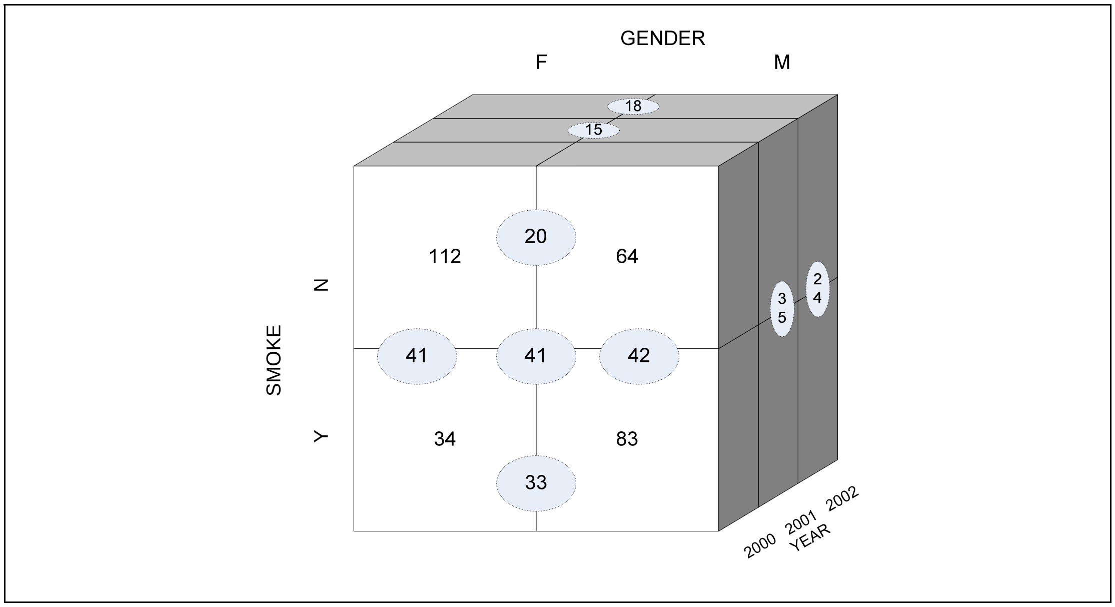


Figure 11 – Unallocated Data Cube for IP in Figure 10

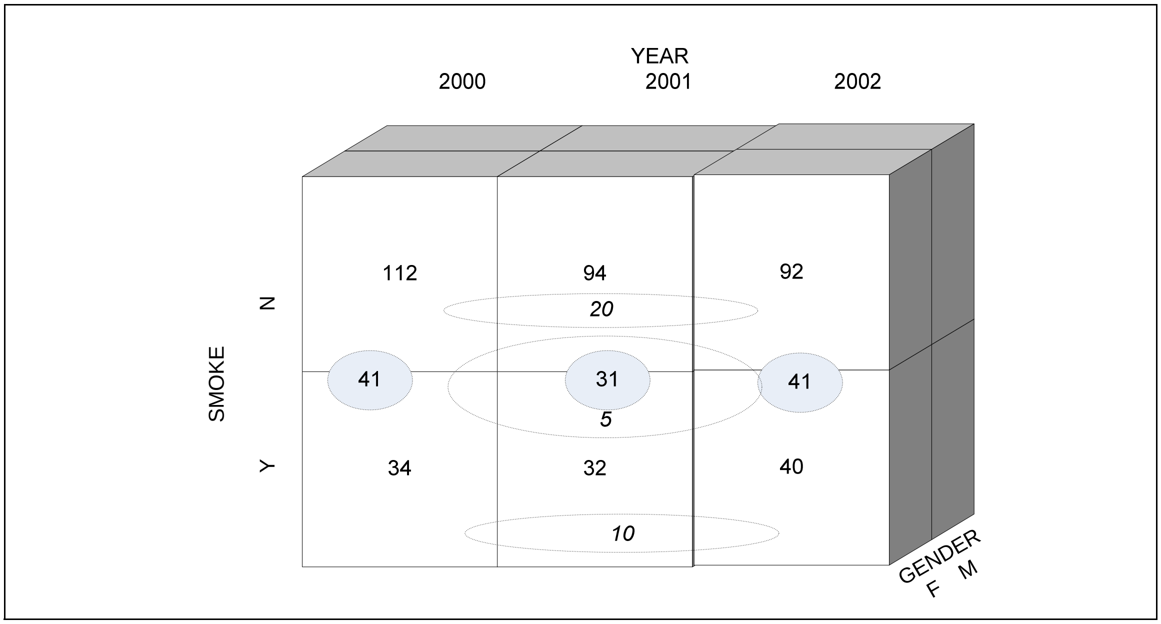


Figure 12 – Unallocated Data for Gender=F

Figure 12 shows the cube from a smoke/year attribute perspective. Notice that this view shows the female data only, the following figure illustrates how one could look beneath the female layer to see male data (Figure 13).

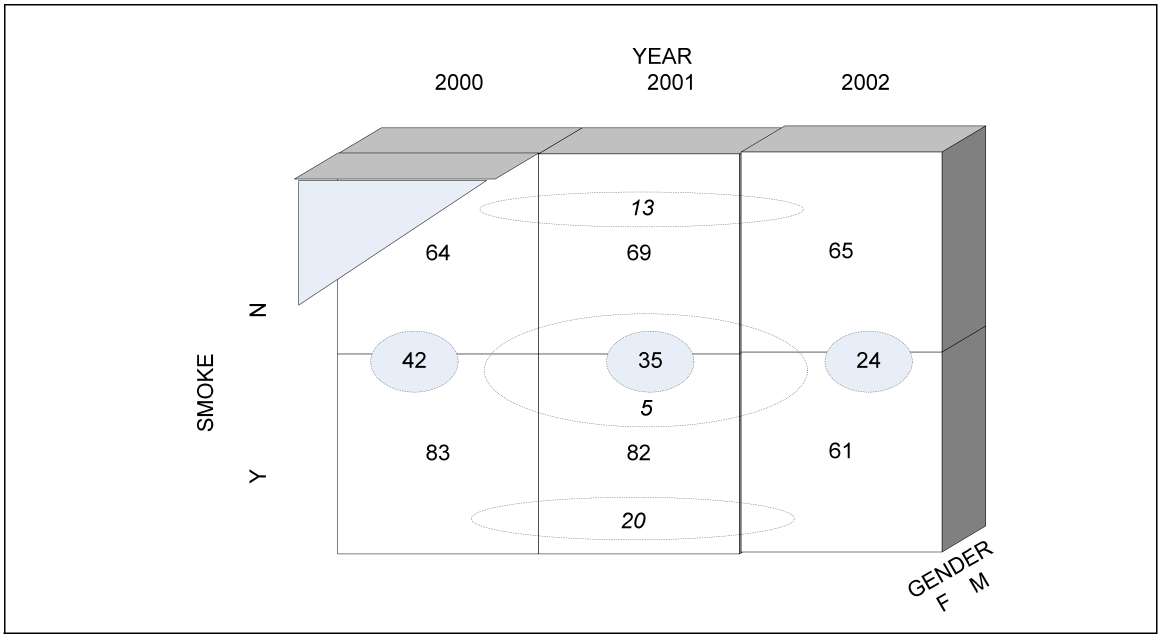


Figure 13 – Unallocated Data for Gender=M

We now consider the calculations needed for the proportional allocation of these data. We define the terms needed and illustrate the calculations for one cell with at table and an example.

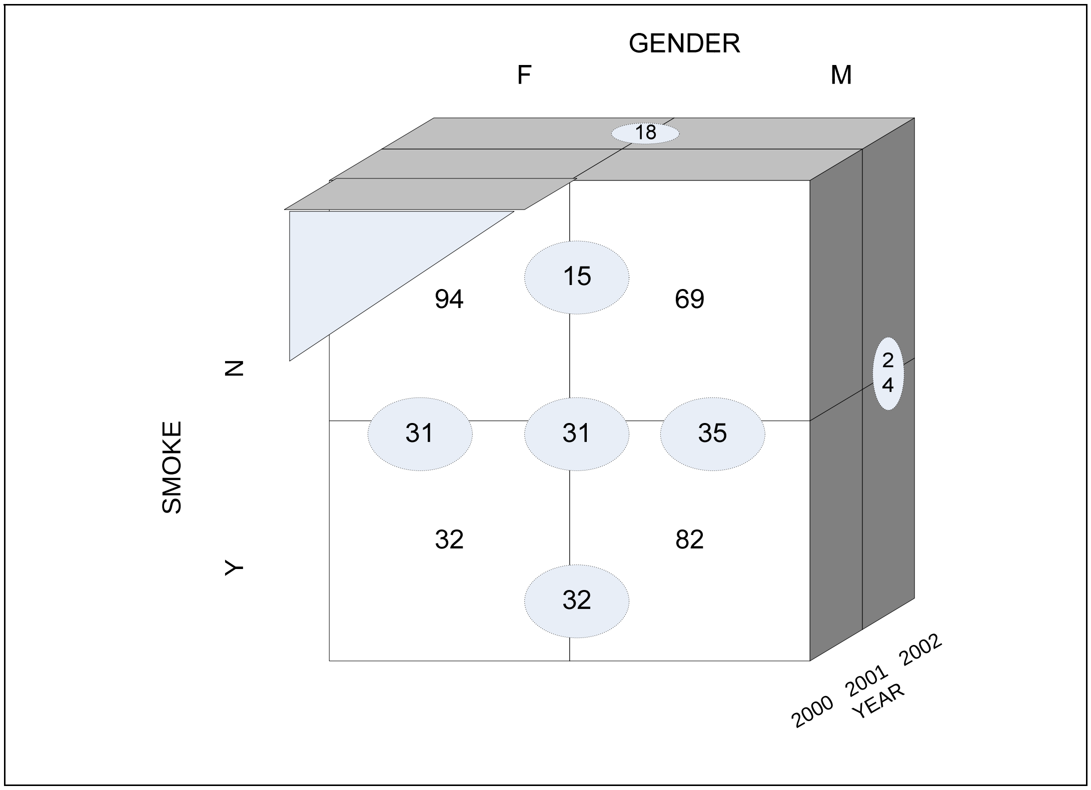


Figure 14- Unallocated Data for Year=2001

Definitions

Table 14 summarizes the notation used in this section. The cube shown for this example is a 2-by-2-by-3 cube. Each lattice represents the dimensional attribute used for aggregation. We refer to a measure by its coordinates on the cube $M_{Y,G,S}$ where Y=year, G=Gender and S= Smoking status, which are the possible dimensional attributes used for grouping (D_1, D_2, D_3). We refer to the unallocated data using coordinates as well, when the lattice information is known, we use a subscript containing the value of the data, when it is unknown we use X. Thus $U_{2000,F,X}$ is unallocated data for females in the year 2000 where the smoking status is unknown. We refer to the data we will allocate to the cell of interest as *a* where only the subscript for the missing aggregation data is shown. a_{yi} refers to data that is being allocated from the unallocated data with missing information for the dimensional attribute year.

Table 14 – Variable Definitions

VARIABLE	DEFINITION
D d_i	**Dimensional Attributes For A Fact Table** *D* – In the set of all possible dimensions used for aggregation, which is made up d_i In our example, d_1,d_2,d_3 can have the following values: • Y - Year: 2000,2001,2002 • G - Gender: F,M • S - Smoking Status: Y,N • X - Coordinate for unknown dimensional attribute value
M_{d_i}	**Measures** - Aggregated Fact Table Data In An OLAP Cube – The subscripts are actual values that correspond to d_i. In our three dimensional cube $M_{2000,F,N}$ refers to the cell containing the aggregated data for year 2000, Females that did not smoke – in Figure 10 this value is 112.
U_{d_i}	**Unallocated Data**. In our example: $U_{2000,X,N}$ refers to unallocated data where the year is equal to 2000, the gender grouping variable is unknown and the smoking status is = N
A_{d_i}	**Allocated Data**. For a certain Measure – this is the amount of unallocated data assigned to it. $A_{2000,F,S}$ is the unallocated data allocated to $M_{2000,F,S}$.
a_{d_i}	**Amount** of data U_{d_i} assigned to M_{d_i}
p_{d_i}	**Proportion** of data U_{d_i} assigned to M_{d_i}
S_{d_i}	**Sum** of all cells with no missing data along a particular combination of dimensions

Sample Calculation - Count

In order to allocate the data based on probability one needs to calculate, at the cell level, the probability that the unallocated data that touches that cell belongs in that cell. For example, if we wish to re-allocate data in the non-smoking, female, year 2000 cube ($M_{2000,F,N}$) we must consider seven unallocated cells and the probabilities that the unallocated data belongs in the cell of interest.

To calculate *how many* unallocated fields could impact a cell, all combinations of the remaining attributes must be considered. A combination is an un-ordered collection of unique elements and is calculated by

$$C_n^k = \frac{n!}{k!(n-k)!} \quad \textbf{(1)}$$

where *n* is the number possible attributes, and k is the number of missing attributes.. Setting j to the number of attributes, we calculate the amount of unallocated data fields that impact one cell as

$$\textbf{Number of allocated fields to consider} = \sum_{i=1}^{j} C_j^i = C_3^1 + C_3^2 + C_3^3 = 7 \quad \textbf{(2)}$$

Table 15 – Summary of Unallocated Data that is considered for *M*

Missing Grouping Attributes	**Combinations**	**Example**	**Proportion Assigned**
1	$C_1^3 = 3$	Missing Year ($U_{X,F,N}$)	$a_{X,F,N}$
		Missing Gender ($U_{2000,X,N}$)	$a_{2000,X,N}$
		Missing Smoking Status ($U_{2000,F,X}$)	$a_{2000,F,X}$
2	$C_2^3 = 3$	Missing Year and Gender ($U_{X,X,N}$)	$a_{X,X,N}$
		Missing Year and Smoking Status ($U_{X,F,X}$)	$a_{X,F,X}$
		Missing Gender and Smoking Status($U_{2000,X,X}$)	$a_{2000,X,X}$
3	$C_3^3 = 1$	Missing Year, Gender and Smoking Status($U_{X,X,X}$)	$a_{X,X,X}$
Total	$\sum_{i=1}^{j} C_j^i = C_3^1 + C_3^2 + C_3^3 = 7$		$A_{2000,F,N}$

Table 15 illustrates that there are seven collections of unallocated data to consider. We consider when the unallocated data are due to one aggregation field missing, then two, and finally three: $C_3^1 + C_3^2 + C_3^3$=3+3+1=7.

We use the Measure of one cell as an example: $M_{F,2000,N.}$ We account for all the unallocated groupings that would impact the aggregated field for a female, year 2000, non smoking data (with a value of 112). Equation 3 outlines how the proportion of unallocated data to allocate to our cell is calculated; Equation 4 calculates the amount.

$$P_{d_i} = \frac{\mathrm{M}_{\mathrm{d_i}}}{\sum\limits_{\forall S_d} M_{S_d}} \tag{3}$$

$$a_{d_i} = P_{d_i} * U_{d_i} \tag{4}$$

Three views of the data cube are used to visualize how the unallocated data fields impact our cell of interest. Figure 15 illustrates the first three unallocated fields considered.

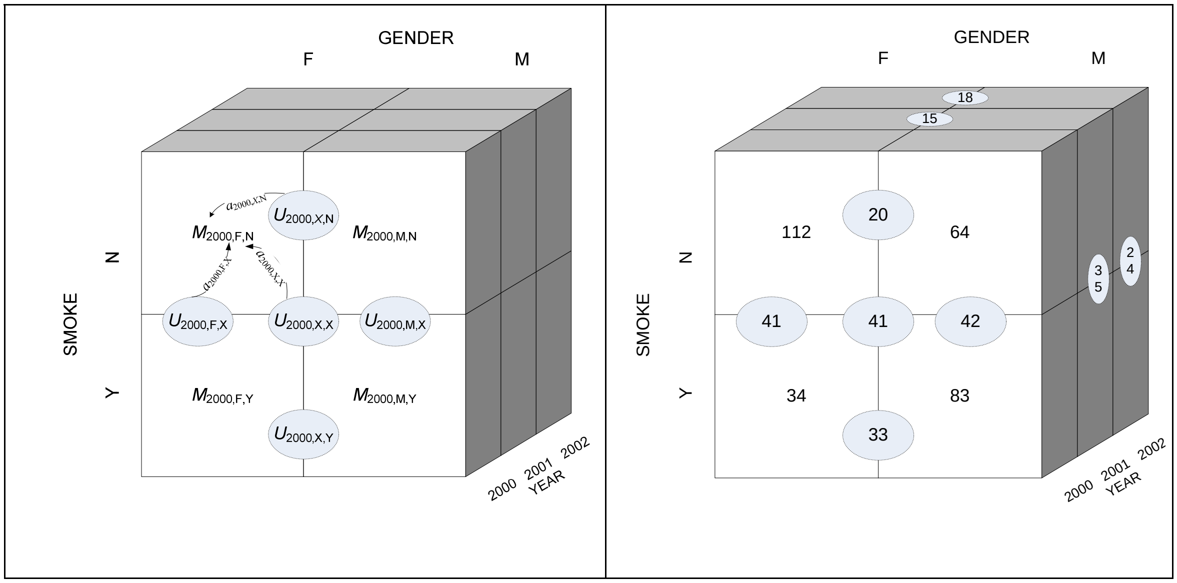


Figure 15 – UDC Year 2000

1. For $a_{2000,F,X}$ we calculate

$$P_{2000,F,X=}\frac{\mathrm{M}_{2000,\mathrm{F,N}}}{\sum_{\forall S} M_{2000,F}}=\left(\frac{112}{(112+34)}\right)=\left(\frac{112}{146}\right)=.77$$

$$a_{2000,F,X}=P_{2000,F,X}*U_{2000,F,X}=.77*41=31.5$$

2. For $a_{2000,X,N}$ we calculate

$$P_{2000,X,N=}\frac{\mathrm{M}_{2000,\mathrm{F,N}}}{\sum_{\forall G} M_{2000,N}}=\left(\frac{112}{(112+64)}\right)=\left(\frac{112}{176}\right)=.64$$

$$a_{2000,X,N}=P_{2000,X,N}*U_{2000,F,X}=.77*20=12.73$$

3. For $a_{2000,X,X}$ we calculate

$$P_{2000,X,X=}\frac{\mathrm{M}_{2000,\mathrm{F,N}}}{\sum_{\forall G,S} M_{2000}}=\left(\frac{112}{(112+64+34+83)}\right)=\left(\frac{112}{293}\right)=.38$$

$$a_{2000,X,x}=P_{2000,X,X}*U_{2000,X,X}=.38*10=3.80$$

For the next two groups of unallocated data we utilize a different view of the cube shown in Figure 16. One of the amounts has already been calculated, but two remain and are outlined in step 4 and 5.

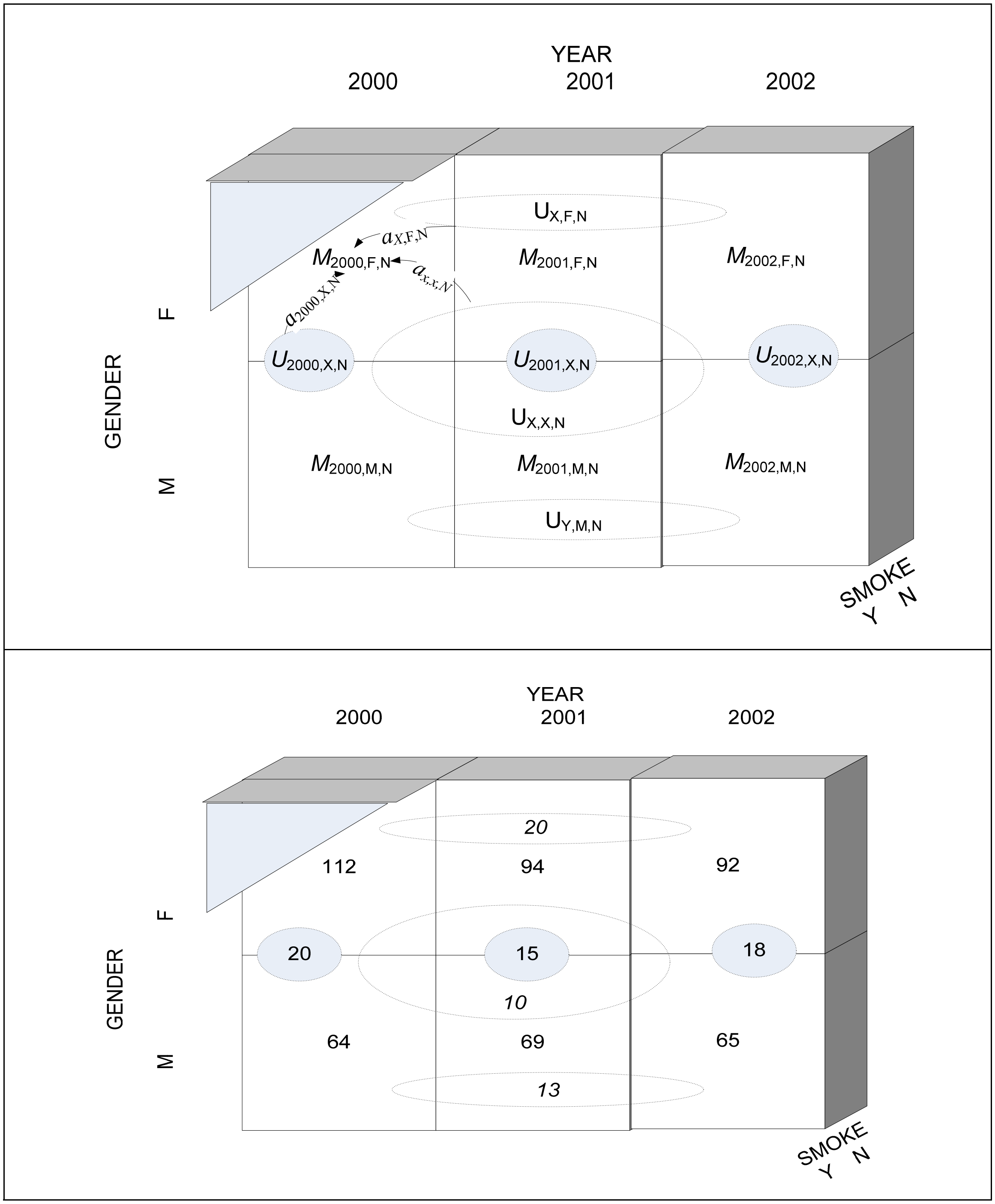


Figure 16 –UDC for Non-Smokers

4. For $a_{X,F,N}$ we calculate

$$P_{X,F,N=}\frac{\mathrm{M}_{2000,\mathrm{F,N}}}{\sum_{\forall Y} M_{F,N}} = \left(\frac{112}{(112+94+92)}\right) = \left(\frac{112}{298}\right) = .38$$

$$a_{X,F,N} = P_{X,F,N} * U_{X,F,N} = .38 * 20 = 7.52$$

5. For $a_{X,X,N}$ we calculate

$$P_{X,X,N=} \frac{\mathrm{M}_{2000,\mathrm{F,N}}}{\sum\limits_{\forall Y,G} M_N} = \left(\frac{112}{(112+94+92+64+69+65)} \right) = \left(\frac{112}{496} \right) = .23$$

$$a_{X,X,N} = P_{X,X,N} * U_{X,X,N} = .23 * 10 = 2.3$$

We consider another view for unallocated fields shown in Figure 17. Again, two of the amounts have already been calculated, but one remains and is outlined in step 6.

6. For $a_{X,F,X}$ we calculate

$$P_{X,F,X=} \frac{\mathrm{M}_{2000,\mathrm{F,N}}}{\sum\limits_{\forall Y,S} M_F} = \left(\frac{112}{(112+34+94+32+92+40)} \right) = \left(\frac{112}{404} \right) = .28$$

$$a_{X,F,X} = P_{X,F,X} * U_{X,F,X} = .28 * 5 = 1.4$$

We consider a final view for unallocated fields shown in Figure 18.

7. For $a_{X,X,X}$ we calculate

$$P_{X,X,X=} \frac{\mathrm{M}_{2000,\mathrm{F,N}}}{\sum\limits_{\forall Y,G,S} M} = \left(\frac{112}{(112+34+94+32+92+40+64+83+69+82+65+61)} \right) = \left(\frac{112}{828} \right) = .14$$

$$a_{X,X,X} = P_{X,X.X} * U_{X,X,X} = .14 * 2 = .27$$

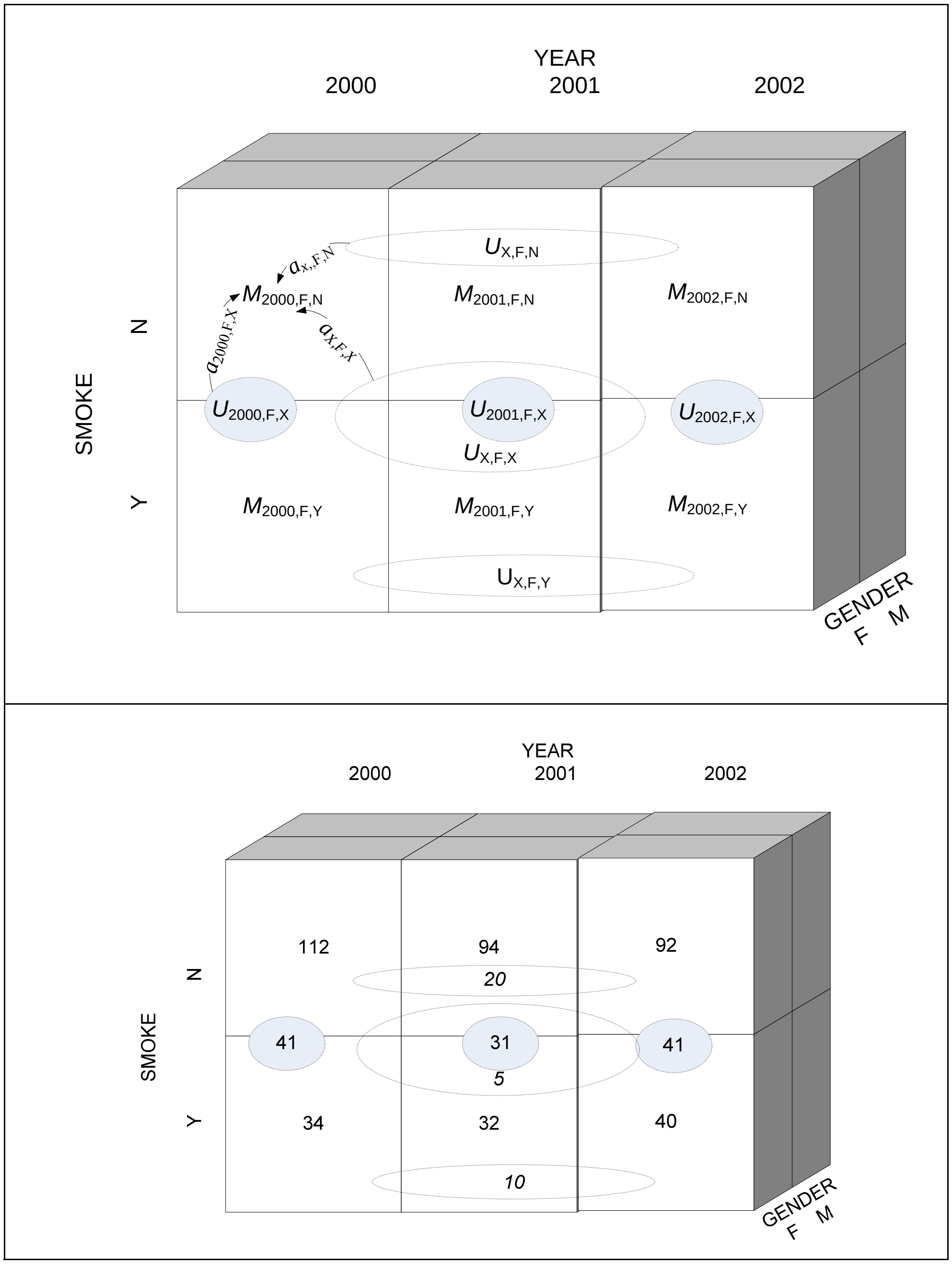


Figure 17 – Unallocated Data for Gender=F

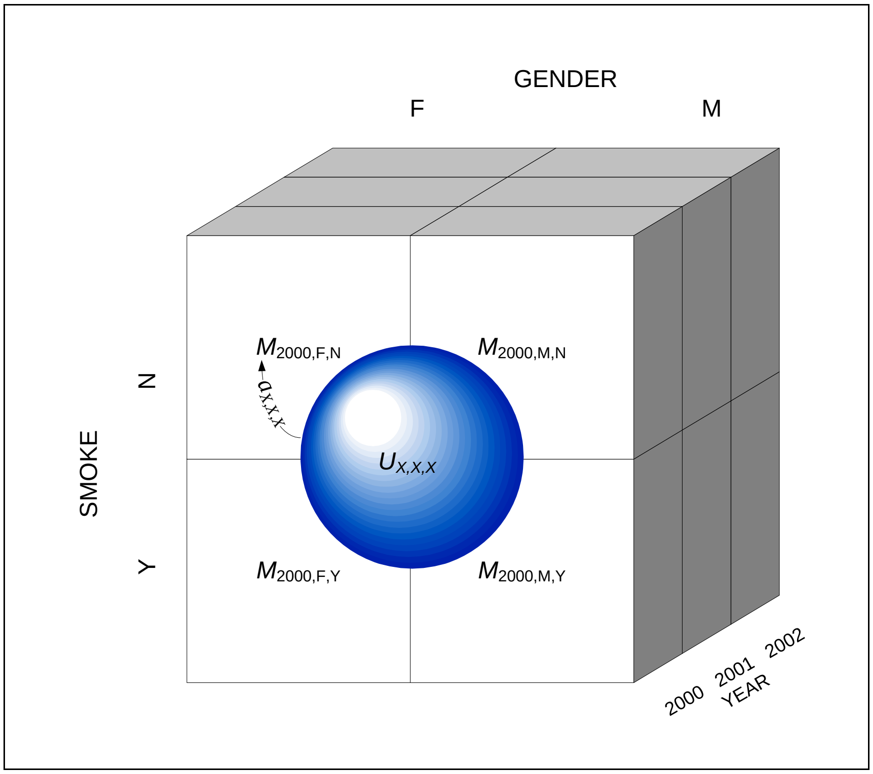


Figure 18 – Unallocated Data for missing data for all grouping variables

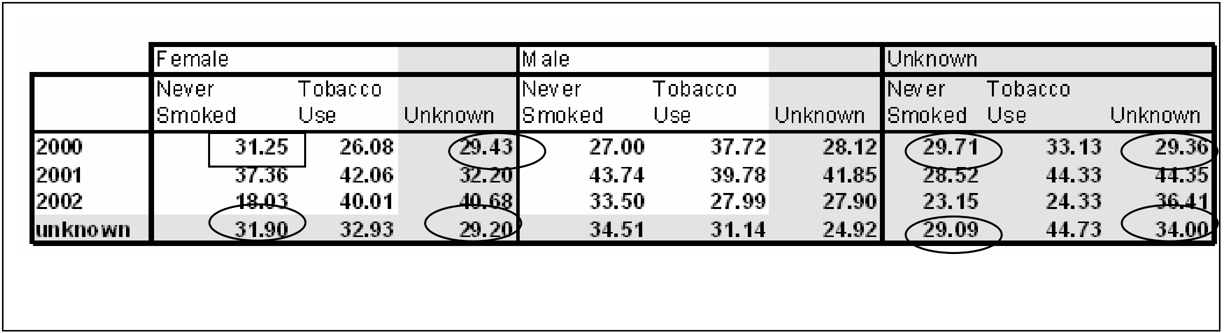
Finally we sum the values from steps 1-7 for the total points that we add to $M_{2000,F,N}$.

Table 16 – Summary of Example Calculations

	Considered Unallocated Data	**Proportion Assigned**	**Amount Assigned**	**Value**
1	$\mathrm{U}_{2000,\mathrm{F},\mathrm{X}}$ – Unknown Smoking Status	$P_{2000,F,X=}\dfrac{\mathrm{M}_{2000,\mathrm{F},\mathrm{N}}}{\sum\limits_{\forall S} M_{2000,F}}$	$a_{2000,F,X} = P_{2000,F,X} * U_{2000,F,X}$	*31.5*
2	$\mathrm{U}_{2000,\mathrm{X},\mathrm{N}}$ Unknown Gender	$P_{2000,X,N=}\dfrac{\mathrm{M}_{2000,\mathrm{F},\mathrm{N}}}{\sum\limits_{\forall G} M_{2000,N}}$	$a_{2000,X,N} = P_{2000,X,N} * U_{2000,F,X}$	*12.73*
3	$\mathrm{U}_{2000,\mathrm{X},\mathrm{X}}$ Unknown Gender and Smoking Status	$P_{2000,X,X=}\dfrac{\mathrm{M}_{2000,\mathrm{F},\mathrm{N}}}{\sum\limits_{\forall G,S} M_{2000}}$	$a_{2000,X,X} = P_{2000,X,X} * U_{2000,X,X}$	*3.8*
4	$\mathrm{U}_{\mathrm{X},\mathrm{F},\mathrm{N}}$ Unknown Year	$P_{X,F,N=}\dfrac{\mathrm{M}_{2000,\mathrm{F},\mathrm{N}}}{\sum\limits_{\forall Y} M_{F,N}}$	$a_{X,F,N} = P_{X,F,N} * U_{X,F,N}$	*7.52*
5	$\mathrm{U}_{\mathrm{X},\mathrm{X},\mathrm{N}}$ Unknown Year and Gender	$P_{X,X,N=}\dfrac{\mathrm{M}_{2000,\mathrm{F},\mathrm{N}}}{\sum\limits_{\forall Y,G} M_{N}}$	$a_{X,X,N} = P_{X,X,N} * U_{X,X,N}$	*2.3*
6	$\mathrm{U}_{\mathrm{X},\mathrm{F},\mathrm{X}}$ Unknown Year and Smoking Status	$P_{X,F,X=}\dfrac{\mathrm{M}_{2000,\mathrm{F},\mathrm{N}}}{\sum\limits_{\forall Y,S} M_{F}}$	$a_{X,F,X} = P_{X,F,X} * U_{X,F,X}$	*1.4*
7	$\mathrm{U}_{\mathrm{X},\mathrm{X},\mathrm{X}}$ All unknown	$P_{X,X,X=}\dfrac{\mathrm{M}_{2000,\,\mathrm{F},\,\mathrm{N}}}{\sum\limits_{\forall Y,G,S} M}$	$a_{X,X,X} = P_{X,X.X} * U_{X,X,X}$	*.27*
8	Total Unallocated Data for $M_{2000,\mathrm{F},\mathrm{N}}$	$A_{2000,F,N}$		*59.52*

Sample Calculation – Average

To consider proportional allocation for an average we use a cube built from the Cancer Data previously described (Figure 19). We consider the average age based on the same filtering and grouping variables. We use the same example, re-allocating data in the non-smoking, female, year 2000 cube ($M_{2000,\mathrm{F,N}}$). We have already calculated how much of the data to allocate from each unknown cell (Table 16). To consider the amount of records to be allocated, a, round up to the nearest whole number and select the a highest and a lowest values in that unallocated field U_{d_i} and include those values for the calculation of the average in $M_{2000,\mathrm{F,N}}$.



	Female			Male			Unknown		
	Never Smoked	Tobacco Use	Unknown	Never Smoked	Tobacco Use	Unknown	Never Smoked	Tobacco Use	Unknown
2000	31.25	26.08	29.43	27.00	37.72	28.12	29.71	33.13	29.36
2001	37.36	42.06	32.20	43.74	39.78	41.85	28.52	44.33	44.35
2002	18.03	40.01	40.68	33.50	27.99	27.90	23.15	24.33	36.41
unknown	31.90	32.93	29.20	34.51	31.14	24.92	29.09	44.73	34.00

Figure 19 – Sample OLAP Screen for UD Average Calculation

In Figure 19 , $M_{2000,\mathrm{F,N}}$ is highlighted with a square. The circled values are the series that need to be considered in proportionally allocating values. For example, $\mathrm{U}_{\mathrm{X,F,N}}$, Unknown Year has a total of 20 values. The rounded value of $a_{X,F,N}$ is 8, so we take the eight highest and eight lowest values from that series and include them in the calculation of the average for $M_{2000,\mathrm{F,N.}}$ In the case where a is higher then the available values we take all the values in that series.

Scenario Based Allocation

In uncertain situations, “What if...?” questions can be helpful in considering several alternatives. In fact, scenario-based decision making is widespread in business and organizations (Harries 2003), and is a natural extension in the case of unallocated data. In this research we propose presenting decision-makers three scenarios:

1. Ignoring unallocated data
2. Proportional allocation
3. Worst/Best case scenario

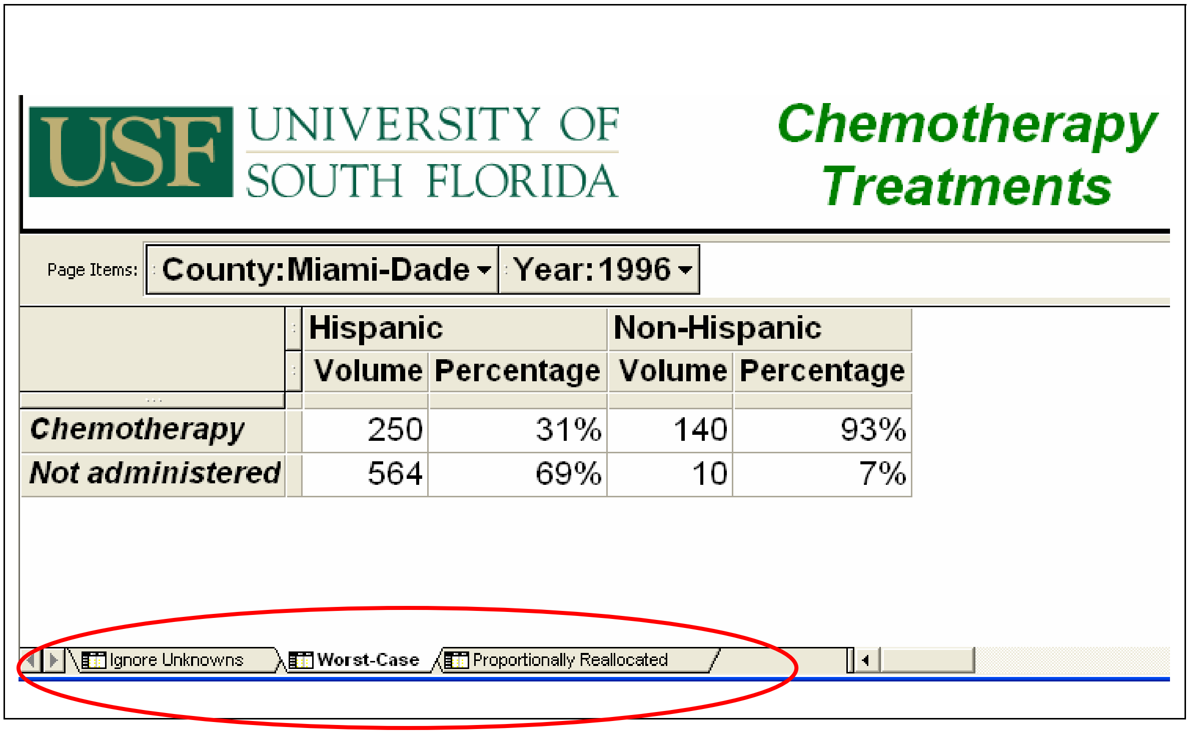
The first scenario simply ignores any missing data, the second is the approach described in the previous section. The third approach is exploratory, in that we leave a full implementation of an algorithm to calculate best/worst-case scenarios based on decision-maker’s input for future research, but we present a prototype to the confirmatory focus groups to begin to understand the usefulness of such an approach.

Example

Figure 20 shows an OLAP screen reporting volumes and percentages of cancer patients receiving chemotherapy when recommended by their physician, grouped by whether they are ethnically Hispanic or not[3]. The scope is to investigate whether there is a disparity in treatment for Hispanic patients. Across the bottom the decision maker can investigate the three scenarios. For this very simple example, the worst-case scenario

[3] This is a fictitious cancer.

assigns all data where ethnicity is unknown to Hispanic, but ignores unknowns in the year, county, and whether chemotherapy was administered.



Chemotherapy Treatments

Page Items: County:Miami-Dade Year:1996

	Hispanic		Non-Hispanic	
	Volume	Percentage	Volume	Percentage
Chemotherapy	250	31%	140	93%
Not administered	564	69%	10	7%

Figure 20 – Example of Scenario-Based Approach

Summary

When combining or aggregating data from multiple sources in the information supply chain missing codes and codes that do not match other sources of data result in data that is not assigned to any of the possible cells in a data. This chapter describes the UD metric which considers the effects of null values in any of the grouping or filtering variables for counts and for averages. It also proposes a case-based approach for presenting unallocated data to a decision-maker, which gives flexibility for the decision maker to consider different “what if” scenarios.

Chapter Five: Information Volatility

Introduction

Business Intelligence tools rarely offer any form of reliability measures. When considering aggregated data, or when observing trends decision makers rely on point estimates, such as an average, which may be biased by noisy data. Supplying decision makers with information about the reliability of the data should improve the quality of their decisions, as descriptive analysis of the data can often provide an understanding of any unusual patterns. Yet, reliability of data is difficult to quantify, in that it is highly subjective and dependent of the context of the decision being made. This chapter proposes a measure of reliability called *information volatility* and introduces the notion of benchmarking the reliability of data.

Definition

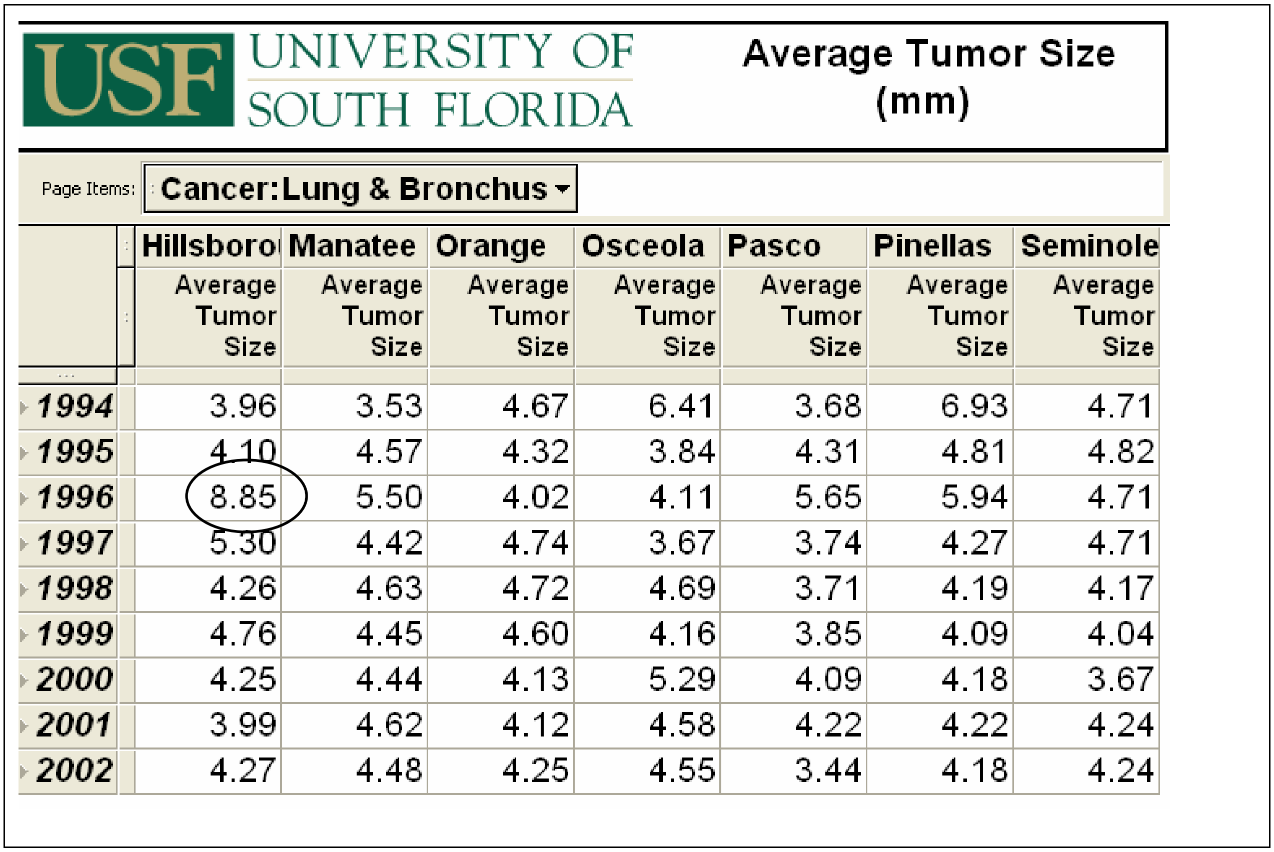
Information Volatility is defined as the rate of change in the values of stored data. Assessment of reliability is outlined by presenting the decision maker with a metric and benchmarking. Two forms of *information volatility* are identified: inter-cell information volatility and intra-cell information volatility (Table 17).

Table 17 – Types of Information Volatility in Dimensional Modeling

Information Volatility	Definition
Intra-cell	In aggregated data, for example an average, the information volatility within the series of numbers that form that calculation
Inter-cell	When comparing values across groupings, the information volatility across those values

Intra-Cell Volatility

In an OLAP tool, aggregated data are calculated from a series of numbers, and represents a summarized value for a particular set of grouping variables. The values of these aggregated fields can be deceiving. Take, for example, summarized data being shown as an average. This average may be compromised of a series of numbers arriving from various sources, with various levels of accuracy. In cases where the data are not tightly distributed around the mean, central tendency may not be descriptive. The values that make up this average could have several outliers, or fluctuate significantly, thus an average would not be an accurate representation of the data. The OLAP screen shown in Figure 21 shows the average tumor size for lung cancer (based on real data), by county and year for lung cancer. In the year 1996 for Hillsborough, it is highly unlikely that this average of 8.85 is an accurate representation of central tendency. It is probable that there are some outliers or some issues with data quality that compromise this average.



Average Tumor Size (mm)

Page Items: Cancer:Lung & Bronchus

	Hillsboro	Manatee	Orange	Osceola	Pasco	Pinellas	Seminole
	Average Tumor Size	Average Tumor Size	Average Tumor Size	Average Tumor Size	Average Tumor Size	Average Tumor Size	Average Tumor Size
1994	3.96	3.53	4.67	6.41	3.68	6.93	4.71
1995	4.10	4.57	4.32	3.84	4.31	4.81	4.82
1996	8.85	5.50	4.02	4.11	5.65	5.94	4.71
1997	5.30	4.42	4.74	3.67	3.74	4.27	4.71
1998	4.26	4.63	4.72	4.69	3.71	4.19	4.17
1999	4.76	4.45	4.60	4.16	3.85	4.09	4.04
2000	4.25	4.44	4.13	5.29	4.09	4.18	3.67
2001	3.99	4.62	4.12	4.58	4.22	4.22	4.24
2002	4.27	4.48	4.25	4.55	3.44	4.18	4.24

Figure 21 – Example of Intra-Cell Volatility

Inter-Cell Volatility

Summarized values are frequently utilized to observe trends across a dimension, with the most obvious being the time dimension. Decision makers may also make comparisons along other dimensions, for example across geographical regions. The decision maker should be warned about interpreting or drawing any conclusion about trends that are sporadic or unstable.

Causes of Information Volatility

A previous study (Berndt, Hevner et al. 2003; Tremblay, Fuller et al. 2006) observed three causes for the presence of unusual or unpredictable trends in aggregated data from the information supply chain. The first was inconsistent data definitions.

Different data sources report data with diverse definitions for their calculations. Furthermore, even within a single data source, a change in IT staff can result in definitional changes. For example, health planners noticed a trend in heart disease that looked like a sine wave. This trend was due to changes in the data definitions. Another cause is seasonal changes, as is the case in Florida with migrant workers or "snowbirds". A final cause is scarcely populated groupings, where even a small change may seem very significant. All three of these scenarios indicate the presence of some sort of instability in the data.

Information Volatility Metric

Stability of data from a certain source in the information supply chain can be examined by considering the rate of change and impact of change in the values it provides over a grouping variable or by its dispersion about a central tendency. It follows that data that exhibit unpredictable changes are considered highly volatile. Assuming a normal distribution, a confidence interval can give a decision maker a feel for the dispersion of the data. A large confidence interval is indicative of data that are not tightly distributed along the mean, thus volatile in its values.

For normally distributed data, the Coefficient of Variation (CV) is useful to compare the standard deviations of different variables that are in different units of measure. This statistic measures the ratio of the standard deviation of a variable relative to its mean. We define this as the unit of measure of information volatility when dealing with data that are normally distributed.

$$CV = [\frac{(100) \bullet \sigma}{\mu}] \qquad \textbf{(5)}$$

In healthcare data, data are frequently not normally distributed; in fact, the data are often time-series data. In order to judge the volatility of this sort of data we transform the data in order to achieve a more “well-behaved” distribution. This problem is well researched in the field of finance. In financial analysis, volatility is a standard measure of financial vulnerability, and is used to assess the risk/return tradeoffs in option pricing (Hotopp 1997).

An example of this is the Black and Scholes (1973) model, which is utilized for option pricing, and considers six inputs: current stock price, strike price, time to expiry, risk free interest rate, dividends and volatility (Kotze 2007). The first three inputs are known, and the last three are estimated. Black and Scholes outline the importance of the volatility parameter in their model. Thus, much research in finance has focused on estimating volatility.

Volatility is defined as a measure of uncertainty or risk based on the size of changes in a security’s value (McClave, Benson et al. 2005) . A fund's volatility indicates the tendency of the returns to rise or fall in a short period of time. Thus, a volatile security is considered high risk because its performance may change quickly in either direction at any moment (Croome 2003).

As in our case of a series of healthcare numbers, the most logical choice to describe central tendency of any series of stock prices would be its mean and standard deviation. However, frequently the average price of a stock will be different for each sub-period of history. In order to meaningfully measure volatility the mean around which the variability is measured has to be stable (Hotopp 1997). For this reason, a continuously compounded return is utilized. A continuously compounded return can be scaled over a longer time frame. For stock price volatility, therefore, it is preferable to compute the continuously compounded return (also referred to as the log relative return) by using formula 6, with the assumption is that the *returns* will be normally distributed. In, formula 6, r_t is return and p_t is the price at time t and p_{t-1} is the price one period earlier:

$$r_t = \ln(p_t / p_{t-1}) \qquad \textbf{(6)}$$

Volatility is calculated by using the text book definition of standard deviation, where n is the number of periods, where n is the number of data points in the historical sample, $\bar{r}$ is the mean return of the sample (calculated as log relatives as outlined above):

$$\sqrt{\frac{1}{n-1} \bullet \sum_{t=1}^{n} (r_t - \bar{r})^2} \qquad \textbf{(7)}$$

The major assumption is that financial asset prices are random variables that are lognormally distributed. The lognormal distribution is widely used in situations where values are positively skewed, for example in financial analysis for security valuation or in real estate for property valuation (Mun 2006). The lognormal distribution allows that prices could rise infinitely (though this would be a rare case), but cannot fall below zero. There is some disagreement on the assumption of log normality of stock price movements; however, the empirical data have supported the lognormal distribution, and it is generally accepted as a reasonable approximation (McMillan 1996).

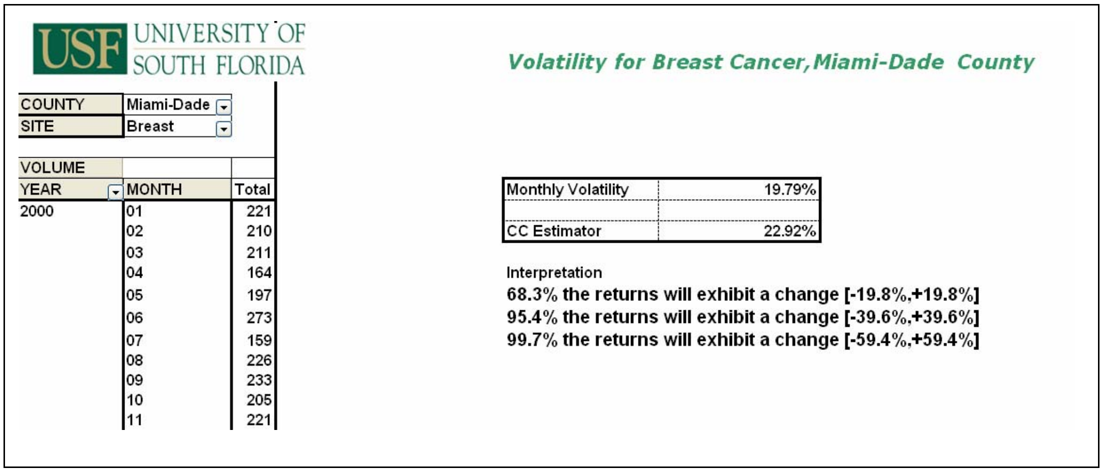
Intuitively, this makes sense. Stock prices are usually positively skewed rather than normally (symmetrically) distributed. Stock prices exhibit this trend because they cannot fall below the lower limit of zero but might increase to any price without limit (thus show a skewness). Other data have shown the same patterns, including property values and IQs The three conditions that underlie the lognormal distribution are (Mun 2006):

1. The uncertain variable can increase without limits but cannot fall below zero.
2. The uncertain variable is positively skewed, with most of the values near the lower limit.
3. The natural logarithm of the uncertain variable yields a normal distribution.

Interpretation of Information Volatility Metric

The volatility measure is interpreted as a percentage. For example, a volatility of 10%, has a the mean of 0 (a return of zero means no change in the values of the data), and due to the properties of a normal distribution, we say:

- With a probability of 68.3% (1 standard deviation from the mean) the returns will exhibit a change within [-10%,+10%]
- With a probability of 95.4% (2 standard deviations) the returns will exhibit a change within [-20%,+20%]
- With a probability of 99.7% (3 standard deviations) the returns will exhibit a change within [- 30%,+30%]



Volatility for Breast Cancer,Miami-Dade County

COUNTY	Miami-Dade
SITE	Breast

VOLUME		
YEAR	MONTH	Total
2000	01	221
	02	210
	03	211
	04	164
	05	197
	06	273
	07	159
	08	226
	09	233
	10	205
	11	221

Monthly Volatility	19.79%
CC Estimator	22.92%

Interpretation
68.3% the returns will exhibit a change [-19.8%,+19.8%]
95.4% the returns will exhibit a change [-39.6%,+39.6%]
99.7% the returns will exhibit a change [-59.4%,+59.4%]

Figure 22 – Volatility Example

All three interpretations (for one, two, or three standard deviations) can be provided to a decision maker, but for the ease of understanding, the first (a standard deviation) is sufficient to communicate the volatility of the data. Figure 22 illustrates the calculated volatility. In this particular example, the decision maker is examining the trend in monthly volumes of breast cancer diagnosis by county. The volatility measure of 19.79% explains the level of volatility in the data. The three interpretations are given also outlined for the decision maker, though it is probably sufficient for a decision maker to consider the one standard deviation interpretation.

Decision on Distribution

As we extend the volatility calculation to data quality as an approximation of stability in the data, we need to consider whether the lognormal distribution is an accurate assumption for all the possible series of numbers that a decision maker can encounter in the use of an OLAP tool. We adopt the following rule of thumb (Mun 2006): if the coefficient of variability (CV):

- Is greater than 30 percent, we assume a lognormal distribution.

- Is less then 30 percent, we use the normal distribution.

In an OLAP cube, this decision is made when the cube is formed. If the data are found to be normal, the Coefficient of Variation is used as a measure of information volatility, otherwise the standard deviation of the log returns is utilized. Once this is established, it does not change. For example a user may drill down to a lower granularity within the cube, the measure (whether normal or log return) for Information Volatility will be that predetermined at the cube formation, and will be calculated at the lower granularity. It would be possible when cutting or slicing in a cube whose data were determined to be normal, to end up with subset of data have a high CV, which indicates the data are highly volatile.

In the next sections consider this decision for each type of Information Volatility. First the distributions are illustrated, followed by the calculation and interpretion of the metric. Rather then simulating the data, we utilize real data from various healthcare data providers.

Example, Intra-Cell Volatility Non-Normal

For Intra-Cell we consider the information volatility within that calculation of aggregated data. This particular example is from a simple health care cube created with data extracted from the Florida Cancer Registry (see Chapter 3 for a detailed explanation). We wish to find the average tumor size for a certain cancer, for each county. Tumor Size can be used as a predictor of survival (we can argue that counties with smaller average tumor sizes are more successful at identifying cancers at an early stage and starting treatment). As an illustration we consider the occurrences of stomach cancers in Hillsborough County in 1997, by utilizing the query shown in Figure 23:

```
SELECT      avg(eod_tum_size)
FROM        fcds_cancers
WHERE       fcds_site_grp='012' (code for stomach cancer)
AND         cat_year='1997
AND         cat_county='12097' (code for Hillsborough County)
```

Figure 23 Average Tumor Size at Male Patients in Hillsborough

The average tumor size for this query is approximately 26 mm, but we wish to have a measure of how indicative or reliable this number is. The first step is to decide which distribution to utilize. The CV ($\mu = 40.5, \sigma = 73$) for this series of numbers is 180%, which is greater then 30%, thus the lognormal transformation is used. ARENA software was utilized to fit a lognormal distribution to these data (prior to transformation). As seen in Figure 24 the assumption that the data are lognormal is appropriate.

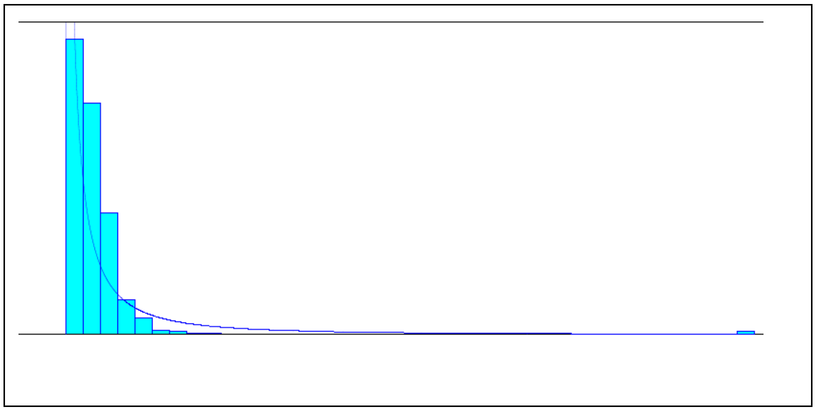


Figure 24 – Distribution of Tumor Size Data

When the data are transformed by taking the natural log of all the values (Figure 25), the distribution looks normal, but we take the approach of calculating the logrelative returns (Figure 26). When the returns are plotted, we see a much tighter normal distribution. We then calculate the standard deviation of the returns, which provides the information volatility for an average stomach cancer tumor size (1.18 – 118% volatility).

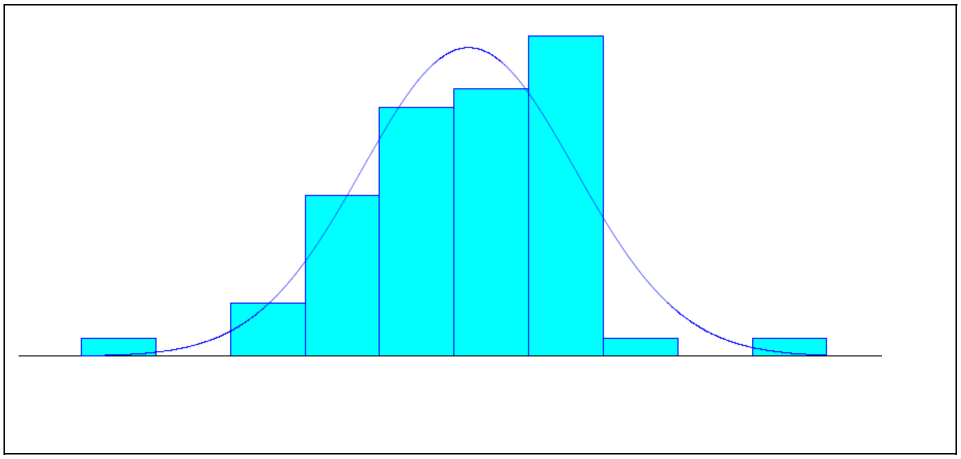


Figure 25 - Stomach Cancer Data Transformed by natural log

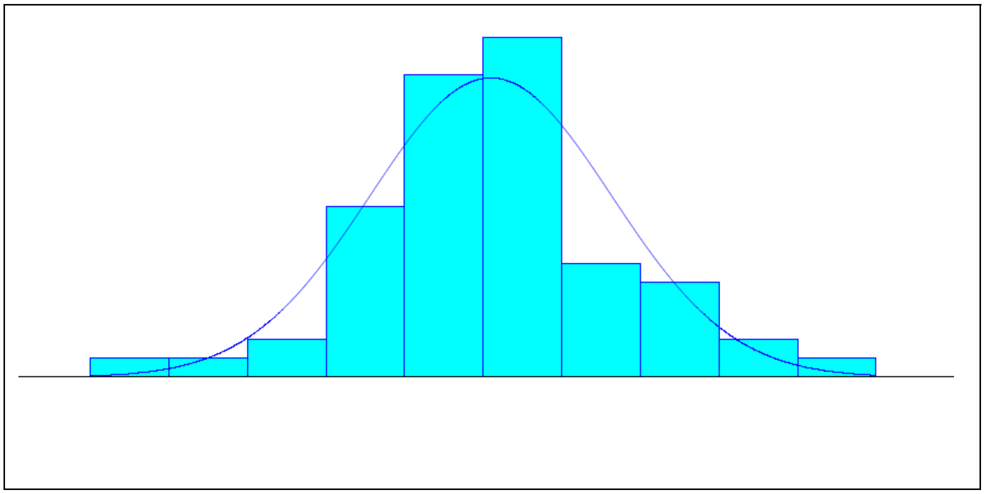


Figure 26 – Return Values for Stomach Cancer Data

Example, Intra-Cell Volatility Normal

Some of the data encountered in health care is well described with a normal distribution. A good example of this is birth weight. The following example is data from Florida's Vita Statistics records stored in a data warehouse. We query (Figure 27) to obtain the average weight of boys born in Hillsborough County in the year 2000.

```
SELECT      weight_grams
FROM        vs_births
WHERE       cat_county=12097 (Hillsborough County) AND
            cat_gender=1 (boys) AND
            cat_year=2000;
```

Figure 27 - Query for Hillsborough County Birth Weight

The data has an average of 3316 (grams) and a standard deviation of 570. The CV = (570/3316)*100 = 17%, which is below 20%, so we can assume the data are normally distributed. A histogram of the data shows that a normal distribution (Figure 28) is appropriate. Thus we utilize the coefficient of variation as a metric for volatility (17%).

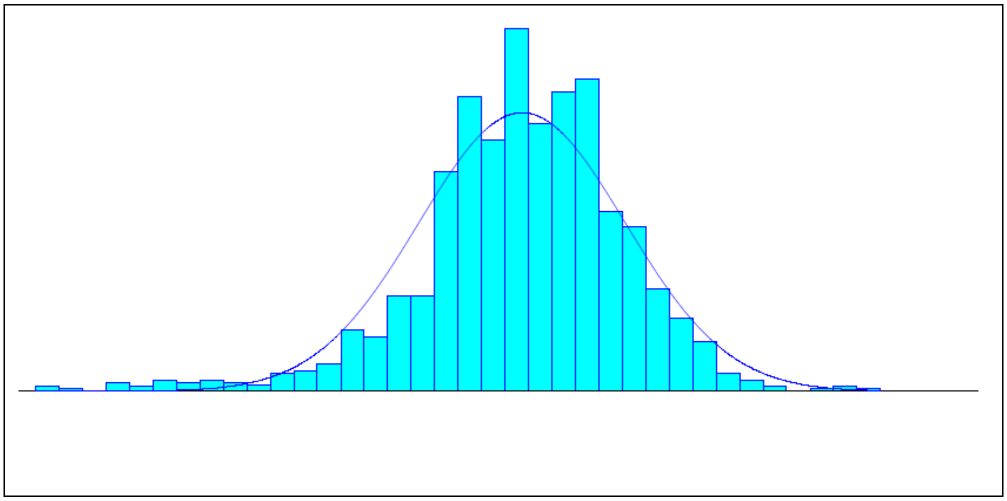


Figure 28 - Histogram of birth weight of males in Hillsborough County.

Example, Intra-Cell Volatility

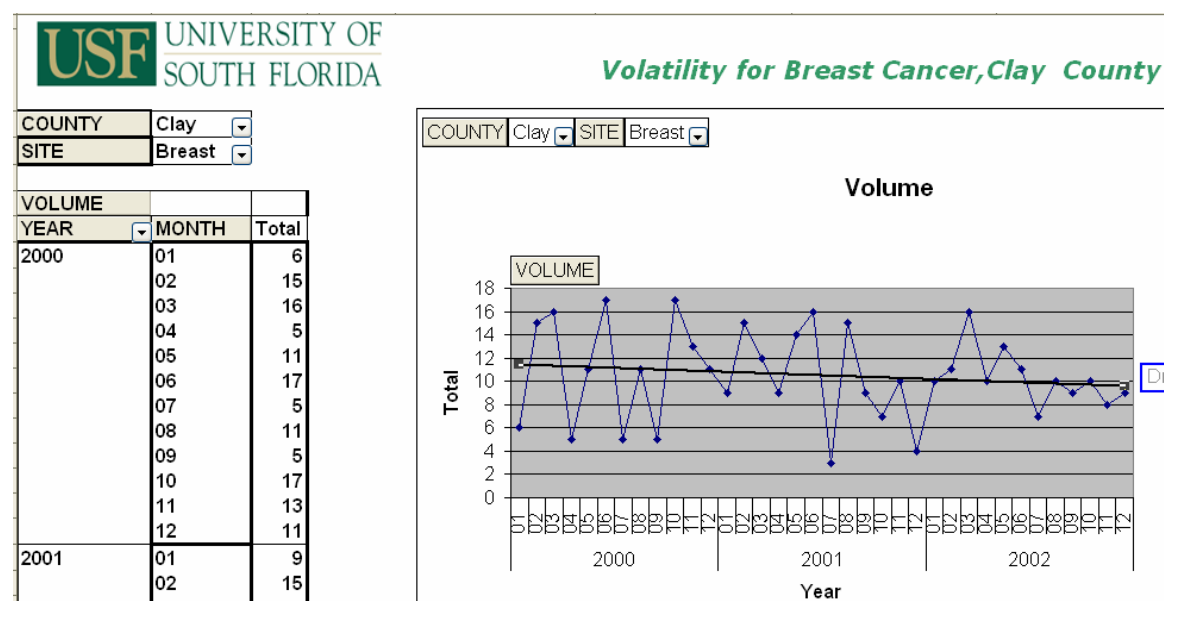
For intra-cell volatility the Information Volatility Metric is utilized to help a decision maker judge the stability of an observed trend. A decision maker may be utilizing trends to get a feel for the future behavior of data. Information Volatility can help a decision maker get a feel for the variability in the trend, and the trend's reliability for future prediction. In this case we illustrate data where we assume a lognormal distribution. For the majority of the cases trends are observed across time (these types of trends are very similar to stock data trends), thus the stability of the trend is an important consideration.

As an example we examine breast cancer volumes by county. The news has reported that the number of cases has been declining, and we wish to examine this by observing monthly volumes of breast cancer diagnosis for each county. We build a cube with the query shown in Figure 29.

```
SELECT     fcds_cancers.cat_count,
           fcds_cancers.cat_county,
           fcds_cancers. month,
           fcds_cancers.year,
           fcds_sites.site_label,
           counties.county_name
FROM       counties, fcds_sites, fcds_cancers
WHERE      counties.county_id = fcds_cancers.cat_county AND
           fcds_cancers.fcds_site_grp = fcds_sites.code
```

Figure 29 - Counts of cancer occurrences by month, county

In Figure 30 we examine the volumes for Breast Cancer in Clay County to understand if the downward trend is true for this particular county. We build the OLAP cube in EXCEL (linked to an ORACLE database), and include a chart with a linear trend line. There indeed seems to be downward trend, but from the chart the data seem volatile. In this example information on volatility may give the decision maker a feel for the "jumpiness" in this trend. Figure 31 illustrates how this metric can be presented to a decision maker. This particular trend has about 70% volatility, which indicates that the there is quite a bit of variation in this trend.

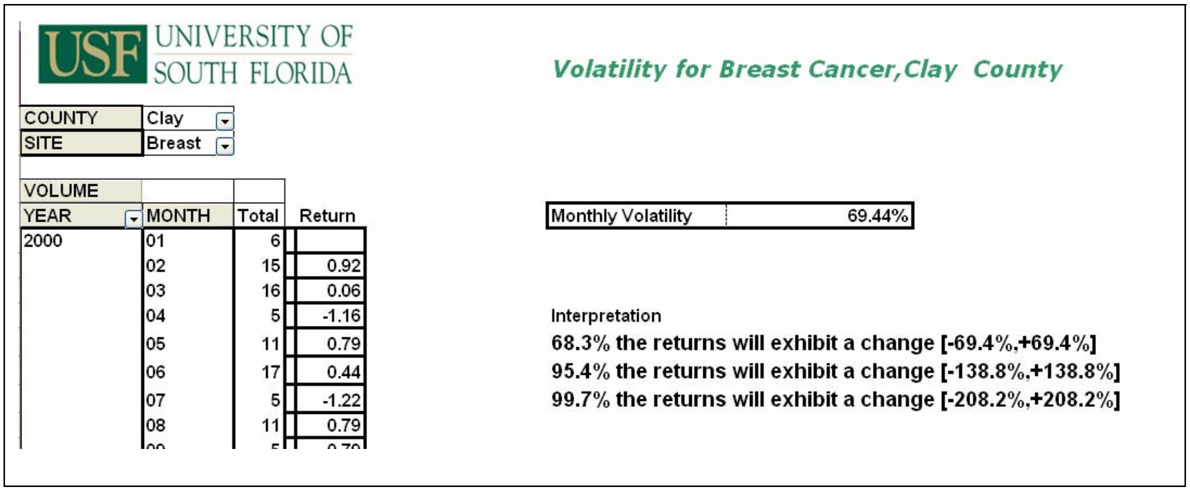

UNIVERSITY OF SOUTH FLORIDA

Volatility for Breast Cancer,Clay County

COUNTY	Clay
SITE	Breast

VOLUME		
YEAR	MONTH	Total
2000	01	6
	02	15
	03	16
	04	5
	05	11
	06	17
	07	5
	08	11
	09	5
	10	17
	11	13
	12	11
2001	01	9
	02	15

Figure 30 - Volatility in Breast Cancer Monthly Volumes



Volatility for Breast Cancer,Clay County

COUNTY	Clay
SITE	Breast

VOLUME			
YEAR	MONTH	Total	Return
2000	01	6	
	02	15	0.92
	03	16	0.06
	04	5	-1.16
	05	11	0.79
	06	17	0.44
	07	5	-1.22
	08	11	0.79

Monthly Volatility	69.44%

Interpretation

68.3% the returns will exhibit a change [-69.4%,+69.4%]

95.4% the returns will exhibit a change [-138.8%,+138.8%]

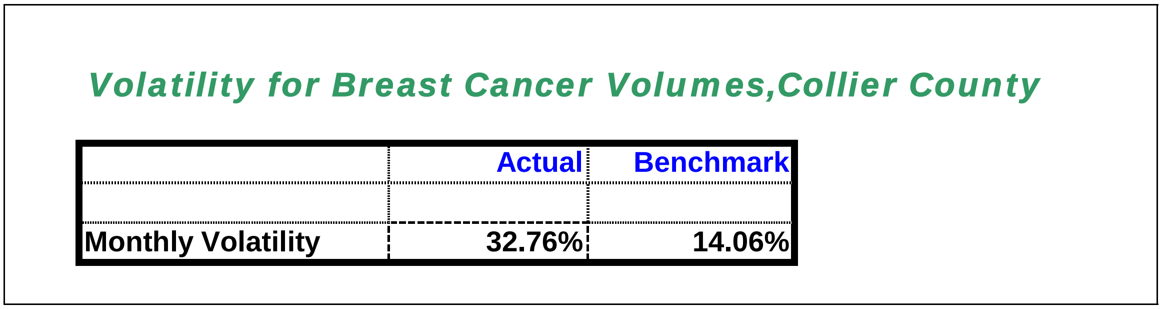
99.7% the returns will exhibit a change [-208.2%,+208.2%]

Figure 31 – Volatility Metric

Benchmarking

To extend our numerical information volatility metric, this research considers a local volatility model as a benchmark approach. This approach is also common in stock indices (Heath and Platen 2006). While future studies include the set of standard benchmarks for different types of healthcare data, the initial approach is to roll up to the largest granularity. For example, if considering a trend in monthly volumes of breast cancer occurrences for a certain county in Florida, we would calculate the volatility in the monthly volumes for the entire state of Florida as a benchmark. As a prototype three approaches were taken:

1. Numerical Benchmark. Reporting numerical values for volatility by also calculating the value for the benchmark. Figure 32 is an example.



Volatility for Breast Cancer Volumes,Collier County

	Actual	Benchmark
Monthly Volatility	32.76%	14.06%

Figure 32 – Numerical Volatility Benchmark

2. Graphical Presentation of Benchmark. By graphing the return both for the trend of interest and its benchmark (on the same scale). Figure 33 shows an example for the same data.

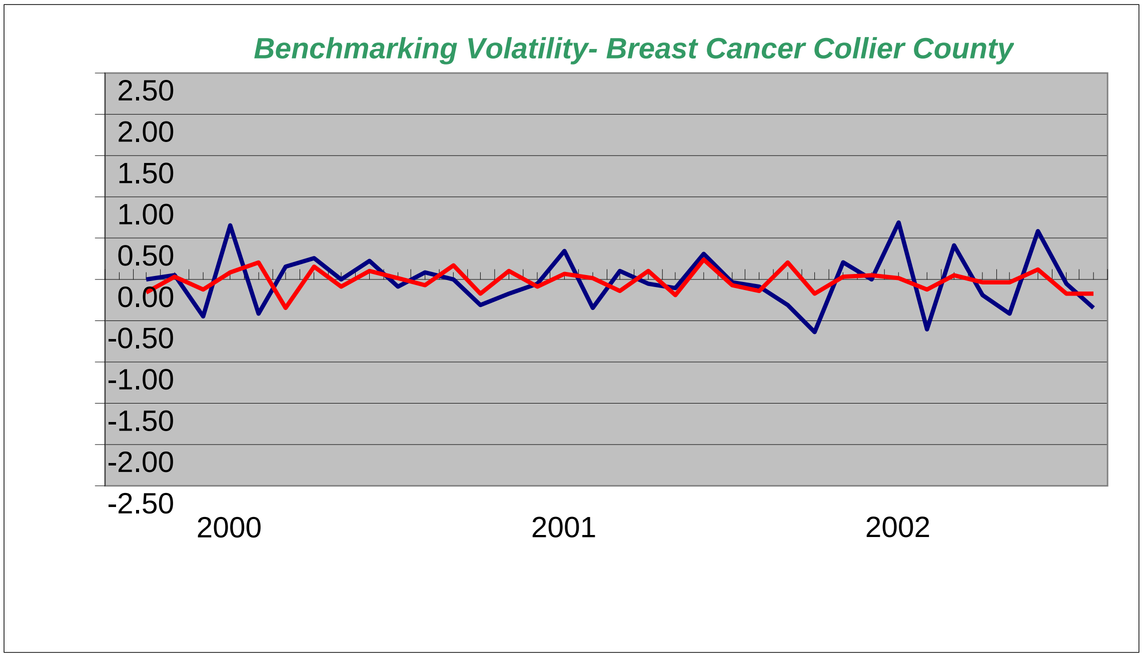
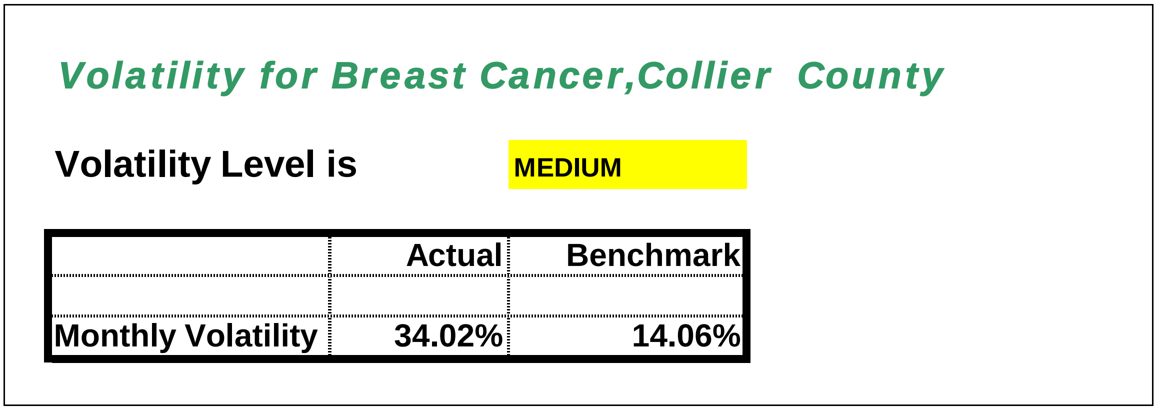


Figure 33 – Graphical Benchmarking of Volatility

3. Categorical Benchmarking. Assigning a category to level of volatility in comparison to the benchmark of Low, Medium, or High. For our example, shown in we

arbitrarily set 50% or higher as HIGH, 30% -50% as MEDIUM, and lower then 30% as LOW. Ideally these sensitivities would be set by the decision-maker.



Volatility for Breast Cancer,Collier County

Volatility Level is MEDIUM

	Actual	Benchmark
Monthly Volatility	34.02%	14.06%

Figure 34 – Categorical Volatility Benchmarking

Summary

A measure of reliability called *information volatility* is proposed as an addition to Business Intelligence tools when considering aggregated data, or when observing trends. Two types of information volatility are defined: intra-cell and inter-cell. For each, two types of distributions are considered: normal and lognormal, which is often the case for time series data. The calculations are created borrowing from the finance literature, since there are similarities in the types of data. In order to understand the information volatility metric the notion of benchmarking is introduced, with three propositions: numerical benchmarking, graphical benchmarking and categorical benchmarking.

Chapter Six: Sample Size Indicator

Introduction

Studies (Tversky et al. 1982) have shown that humans will use heuristics ("rules of thumb") when making choices under uncertainty. Heuristics are based on past experience and generally give good results, but they can also lead to severe and systematic errors (Tversky et al. 1982). When managers and decision makers utilize aggregated data (summary information) retrieved from several data sources in the information supply chain, it is important to understand the heuristics knowledge workers may use for decision making and the possible biases that could result from the use of these heuristics. An appropriate environment to study these heuristics and biases is health planning, since aggregated information supply chain data are frequently utilized to support public policy formulation (Berndt et al. 2003).

An example of such a heuristic is insensitivity to sample size. People assess the likelihood of a sample result by asking themselves how similar that sample result is to the properties of the population from which the sample was drawn, regardless of the size of the sample. Studies have already shown that people are insensitive to sample size (Bar-Hillel 1982; Tversky and Kahneman 1982; Klein, Goodhue et al. 1997), so this research does not test this theory. Rather, it explores mechanisms to mitigate this particular bias in tools that are used to examine aggregated data, in particular OLAP tools. Initially, a simple design is suggested and prototyped to the focus groups. The feedback from the focus groups is intended for use in future research, both to improve the method used as well as to extend the findings and methodology to the study of other well known biases.

An Example

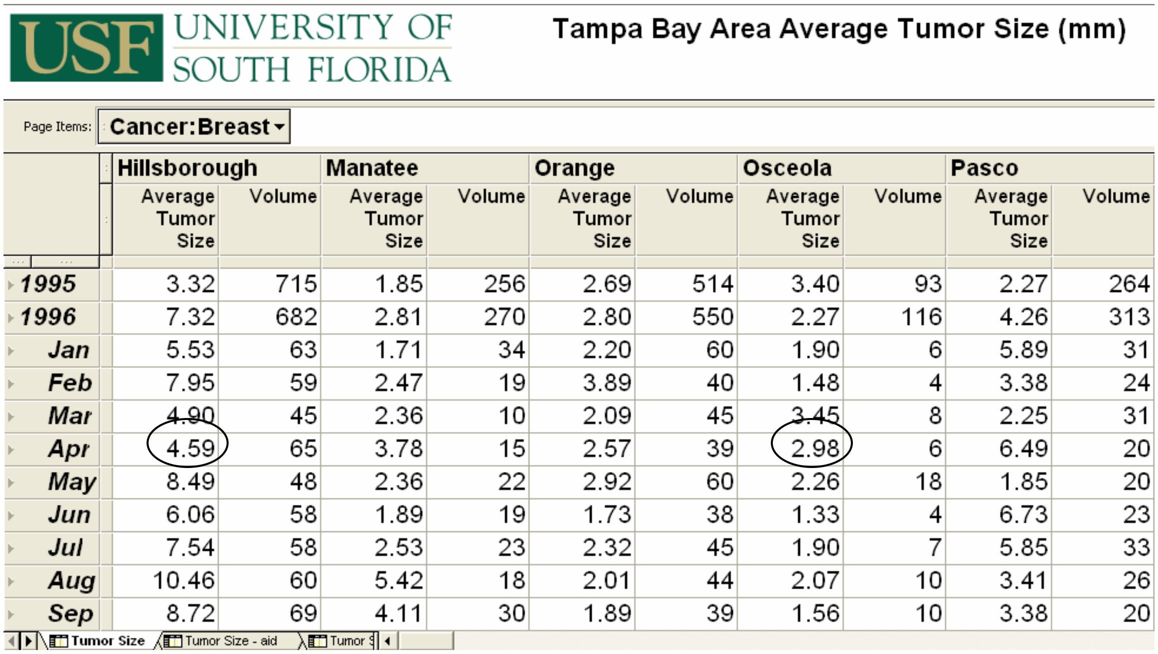
As an illustration we consider the average tumor size in the Tampa Bay Region using the query shown in Figure 23. We suggest that counties with smaller average tumor sizes are being more successful at identifying cancers at an early stage and starting treatment, thus tumor size can be used as a predictor of survival. This is an illustrative example, since in a realistic situation several other data would be considered.

```
SELECT     AVG(eod_tum_size)/10, COUNT(eod_tum_size)
FROM       fcds_cancers
WHERE      health_district = 'HRS5' OR
           health_district = 'HRS6' OR
           health_district = 'HRS7'
```

Figure 35 Average Tumor Size for Three Regions

The resulting cube is shown in Figure 36. As the decision-maker navigates the cube, he may drill down to the month level and compare this average among several counties. Rolling up or down along a dimension is a useful capability in an OLAP environment. In this case, however, for smaller counties the sample sizes tend to decrease as a user rolls down along the time dimension. Thus, the reported averages tend to be more likely to be influenced by outliers, and are less reliable depending on the amount of data utilized to calculate the particular average. For example, when comparing the average tumor size for breast cancer in April of 1996 between Hillsborough county and Osceola county, it may seem that Osceola County is doing a significantly better job at early detention. However, closely observing the volumes, Osceola County had six occurrences compared to the 65 in Hillsborough County. The difference in sample size is an important distinction when comparing these two averages, since the smaller sample size is more likely to be influenced by any outliers. Based on past studies, many people would tend to ignore this (Einhorn and Hogarth 1981; Bar-Hillel 1982; Tversky and Kahneman 1982; Klar 1990). Additionally, depending on how the cube used for decision making is formed, the volume may not even be reported.

In fact, many modern business intelligence tools use dashboards to give a summarized version of the data to managers or high level decision makers. In these cases, it is unlikely that volume would be shown or that any representation of sample size would be given.



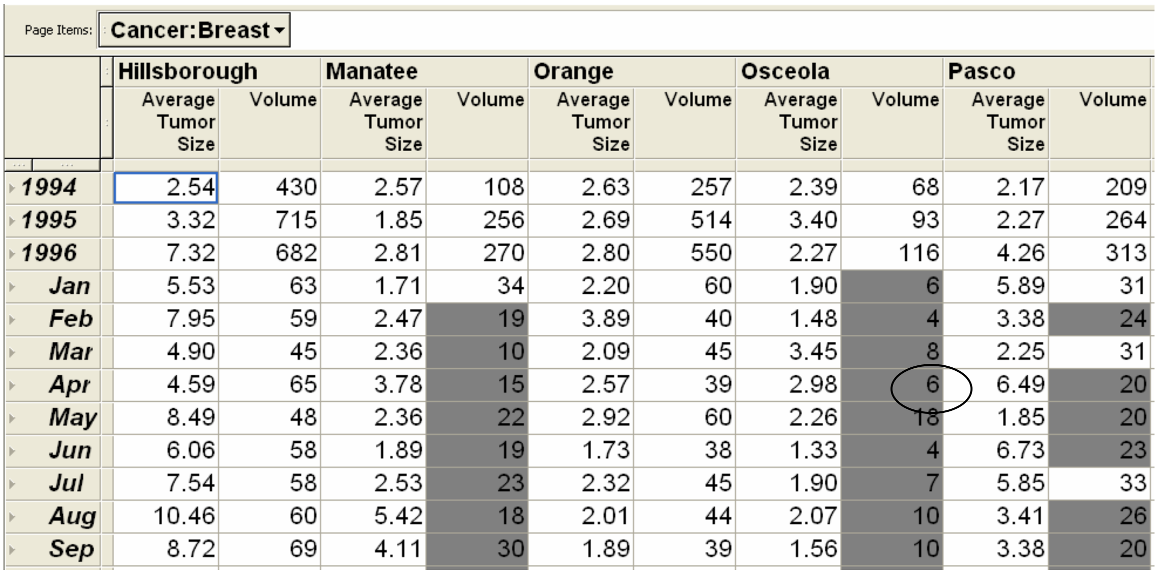
Tampa Bay Area Average Tumor Size (mm)

Page Items: Cancer:Breast

	Hillsborough		Manatee		Orange		Osceola		Pasco	
	Average Tumor Size	Volume	Average Tumor Size	Volume	Average Tumor Size	Volume	Average Tumor Size	Volume	Average Tumor Size	Volume
1995	3.32	715	1.85	256	2.69	514	3.40	93	2.27	264
1996	7.32	682	2.81	270	2.80	550	2.27	116	4.26	313
Jan	5.53	63	1.71	34	2.20	60	1.90	6	5.89	31
Feb	7.95	59	2.47	19	3.89	40	1.48	4	3.38	24
Mar	4.90	45	2.36	10	2.09	45	3.45	8	2.25	31
Apr	4.59	65	3.78	15	2.57	39	2.98	6	6.49	20
May	8.49	48	2.36	22	2.92	60	2.26	18	1.85	20
Jun	6.06	58	1.89	19	1.73	38	1.33	4	6.73	23
Jul	7.54	58	2.53	23	2.32	45	1.90	7	5.85	33
Aug	10.46	60	5.42	18	2.01	44	2.07	10	3.41	26
Sep	8.72	69	4.11	30	1.89	39	1.56	10	3.38	20

Figure 36 – Example OLAP Sheet, Insensitivity to Sample Size

A simplistic approach is to warn the decision maker by highlighting those average values that they are investigating is based on a small sample size. Business Intelligence tools have grown in sophistication, and so has research in Human-Computer Interaction. Several approaches can be taken, such as the use of small flags, or even changing the size of the font, but we leave the details of the best way to draw attention to these numbers to future research and concentrate on the simple task of just drawing attention to these problematic averages.



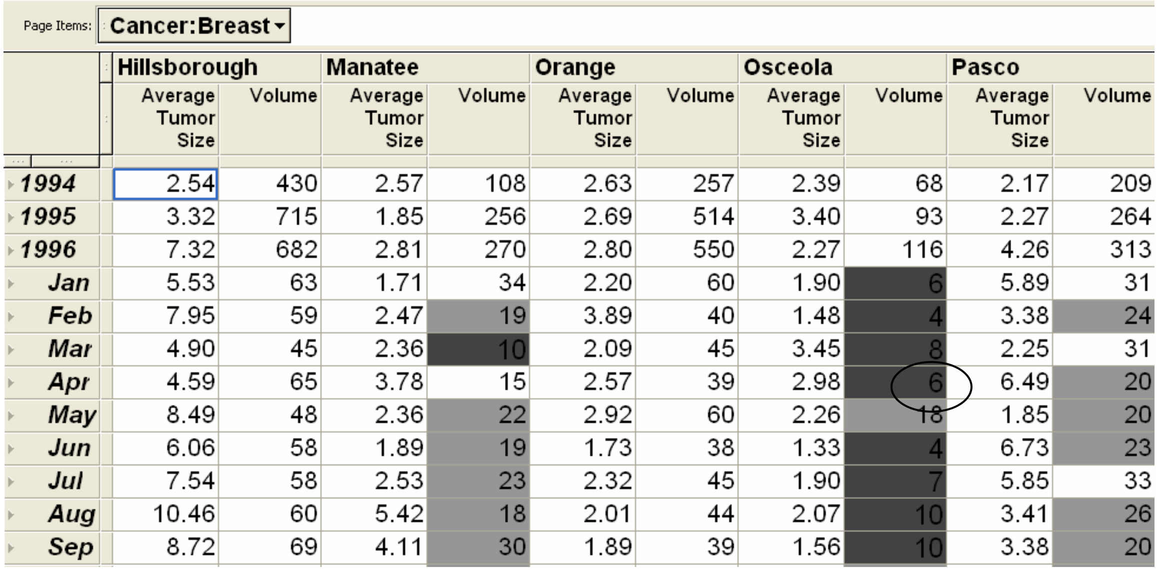
Page Items: Cancer:Breast

	Hillsborough		Manatee		Orange		Osceola		Pasco	
	Average Tumor Size	Volume	Average Tumor Size	Volume	Average Tumor Size	Volume	Average Tumor Size	Volume	Average Tumor Size	Volume
1994	2.54	430	2.57	108	2.63	257	2.39	68	2.17	209
1995	3.32	715	1.85	256	2.69	514	3.40	93	2.27	264
1996	7.32	682	2.81	270	2.80	550	2.27	116	4.26	313
Jan	5.53	63	1.71	34	2.20	60	1.90	6	5.89	31
Feb	7.95	59	2.47	19	3.89	40	1.48	4	3.38	24
Mar	4.90	45	2.36	10	2.09	45	3.45	8	2.25	31
Apr	4.59	65	3.78	15	2.57	39	2.98	6	6.49	20
May	8.49	48	2.36	22	2.92	60	2.26	18	1.85	20
Jun	6.06	58	1.89	19	1.73	38	1.33	4	6.73	23
Jul	7.54	58	2.53	23	2.32	45	1.90	7	5.85	33
Aug	10.46	60	5.42	18	2.01	44	2.07	10	3.41	26
Sep	8.72	69	4.11	30	1.89	39	1.56	10	3.38	20

Figure 37 – Example OLAP Sheet, Small Sample Sizes Highlighted

Figure 37 shows an example of this approach based on the OLAP sheet in screen from Figure 36. In this case, the volumes that were below 30 (number based on the central limit theorem) were highlighted the OLAP with the volume highlighted, though ideally this sensitivity could be set by the decision maker. Note that for the comparison previously described (average tumor size for breast cancer in April of 1996 between Hillsborough county and Osceola county), the value for volume for Osceola County is highlighted.

Another approach is to allow the decision-maker to control the sensitivity or to utilize some sort of gradient, utilizing some sort of marking which draws more attention to more severe cases.   Figure 38 shows an example of this, where very small sample sizes (less than or equal to 10) have darker highlighting in red then those between 11 and 30.



Page Items: Cancer:Breast

	Hillsborough		Manatee		Orange		Osceola		Pasco	
	Average Tumor Size	Volume	Average Tumor Size	Volume	Average Tumor Size	Volume	Average Tumor Size	Volume	Average Tumor Size	Volume
1994	2.54	430	2.57	108	2.63	257	2.39	68	2.17	209
1995	3.32	715	1.85	256	2.69	514	3.40	93	2.27	264
1996	7.32	682	2.81	270	2.80	550	2.27	116	4.26	313
Jan	5.53	63	1.71	34	2.20	60	1.90	6	5.89	31
Feb	7.95	59	2.47	19	3.89	40	1.48	4	3.38	24
Mar	4.90	45	2.36	10	2.09	45	3.45	8	2.25	31
Apr	4.59	65	3.78	15	2.57	39	2.98	6	6.49	20
May	8.49	48	2.36	22	2.92	60	2.26	18	1.85	20
Jun	6.06	58	1.89	19	1.73	38	1.33	4	6.73	23
Jul	7.54	58	2.53	23	2.32	45	1.90	7	5.85	33
Aug	10.46	60	5.42	18	2.01	44	2.07	10	3.41	26
Sep	8.72	69	4.11	30	1.89	39	1.56	10	3.38	20

Figure 38 - Example OLAP Sheet – with Sensitivity Analysis

Summary

This chapter outlines a simple method to mitigate a well known judgment bias: insensitivity to sample size. Health planning is an appropriate environment to study this since aggregated data from the information supply chain data are frequently utilized to support public policy formulation (Berndt et al. 2003), and sample size is not always reported or considered. Volumes or aggregated values that are comprised of 30 or less values (number based on the central limit theorem) are highlighted, though ideally this sensitivity will be set by the decision maker. Another possible approach is to allow the decision-maker to control the sensitivity or to utilize some sort of gradient, and drawing more attention to severe cases.   In this simple example we utilize highlighting, but if implemented in a BI tool, other methods can be investigated, such as font size or flags. These HCI issues are left for future research.

Chapter Seven: Focus Group Description and Coding

Introduction

This chapter describes the procedure, participants, coding and script changes for the four focus groups that were run to evaluate the metrics outlined in chapters four, five and six. There were two types of focus groups: exploratory and confirmatory which are described in chapter 3. The focus groups were recorded and transcribed and coded by two independent raters using template analysis. Using a “rolling interview” (Stewart, Shamdasani et al. 2007) approach, incremental changes were made after each of the exploratory focus groups’ script based on feedback from an observer and the focus group participants. Upon completion of both of the exploratory focus groups the metrics were improved, based on feedback from the participants.

Exploratory Focus Group One (EFG1)

The vignettes utilized in the first focus group are summarized in Table 18. The first column in Table 18 describes what feature of the metric was utilized in the example. There were a total of five cases presented. Three of the vignettes utilized the UD metrics, one the volatility metric and one the sample size indicator.

Table 18 – Summary of EFG1 Vignettes

Metric / Feature	Vignette	Decision	Problem
UD – Pie Chart	Studies have shown that smoking is responsible for most cancers of the larynx, oral cavity and pharynx, esophagus, and bladder. In addition, it is a cause of kidney, pancreatic, cervical, and stomach cancers, as well as acute myeloid leukemia.	Is there correlation between smoking and certain types of cancer	Large amounts of missing data in one grouping attribute (whether the patient smoked or not)
UD – Proportional Allocation	When Hispanics are diagnosed with a certain cancer (fictitious example), they're less likely to receive chemotherapy than non Hispanics.	Is there disparity in care?	Large amounts of missing data in several grouping attributes
UD – Proportional Allocation	Rates for Liver Cancer seem to be increasing for Hispanic and decreasing for all other ethnicities	Is there disparity in care?	Large amounts of data did not have ethnicity
Information Volatility – Numeric IV Metric	Counties neighboring Miami-Dade are better at early detection/prevention of Breast Cancer based on volumes of cases	Examine trend – is this a true claim?	Neighboring counties' data exhibit large jumps in values
Sample Size Indicator - Highlighting	Tumor size has been shown to be a good predictor of survival for certain cancers, including: breast, lung and endocrine. Compare average tumor size in Hillsborough to that of neighboring counties	How does Hillsborough compare to other counties?	Neighboring counties may have averages based on very small sample sizes

Participants

There were a total of three participants in Exploratory Focus Group 1, whose demographics characteristics are summarized in Table 19. This group did not have health care experience, but all had just completed a data warehousing class, and all had jobs were they conducted data analysis. The examples were simple enough that health care experience was not necessary. This group was small due to some participant absences. Since this was the initial exploratory focus group, the decision was made to proceed and collect the data. The cases were presented on an overhead screen, with the moderator navigating the OLAP interfaces. The moderator explained the vignettes, the OLAP screen overhead and the decision problem. After the explanation, the participants, for the most part guided the moderator, asking for greater/lower granularity, or different filtering and grouping variables as they formulated their decision. EFG1 was recorded (sound only) and professionally transcribed.

Table 19 – EFG1 Participants

Gender	Age	Last Degree	Current Position	Course in Statistics?	Years of work Experience	Years of Health-care Experience	Self Reported Comfort with Data Analysis (7 point scale)
F	29	MS MIS	Quality Assurance Analyst	Y	5	0	4
M	34	MS MIS	Programmer Analyst	Y	8	0	5
M	33	MS MIS	Systems Analyst	Y	5	0	5

Creating Initial Coding Template

The initial template was created by two of the researchers after an initial read of the transcript by both coders, and taking into consideration the scope of the focus groups, which was the evaluation of the proposed metrics. The initial template is shown Table 20.

Table 20 – Initial Coding Scheme

Category	Construct	Definition
Unallocated Data	*Unallocated Data - Data Analysis Tactic Before*	Strategies to deal with unallocated data prior to receiving metric.
	Unallocated Data - Interpretation Before	Interpretation prior to receiving metric.
	Unallocated Data - Data Analysis Tactic After	Strategies to deal with unallocated data after to receiving metric.
	Unallocated Data - Interpretation After	Interpretation after receiving metric.
	Design Feature Unallocated Data	Mention of the UD feature, design improvement suggestion.
Volatility	*Volatility - Data Analysis Tactic Before*	Strategies to deal with volatility prior to receiving metric.
	Volatility - Interpretation Before	Interpretation before receiving metric.
	Volatility - Data Analysis Tactic After	Strategies to deal with volatility after receiving metric.
	Volatility - Interpretation After	Interpretation after receiving metric.
	Design Feature Volatility	Mention of the Information Volatility feature, design improvement suggestion.
Insensitivity To Sample Size	*Insensitivity To Sample Size - Data Analysis Tactic Before*	Strategies to deal with sample size prior to receiving metric.
	Insensitivity To Sample Size - Interpretation Before	Interpretation prior to receiving metric.
	Insensitivity To Sample Size - After	Strategies to deal with sample size- after receiving metric - Data Analysis Tactic.
	Insensitivity To Sample Size - Interpretation After	Interpretation after receiving metric.
	Design Feature Sample Size	Mention of the Sample Size Indicator feature, design improvement suggestion.
Other [4]	*Overall Data Quality*	Perceptions of data quality, Other issues of data quality.
	Speculation	Speculation on DQ problems.
	Other Factors In Decision Making	Including stakeholder issues.

[4] These constructs are utilized for post-hoc analysis

Coding and Inter-rater Reliability

Once the template was completed and agreed upon by the researchers, the transcripts for EFG1, were coded by identifying sections that were relevant and annotating the appropriate codes from the initial template. Cohen's Kappa was used to measure inter-rater reliability. Cohen's Kappa for EFG1 was 78% indicating a satisfactory level of inter-rater reliability.

EFG1 was then reconciled between coders. The two independent coders discussed the areas of disagreement, stopping when agreement was reached on all higher ordered codes and most lower order codes (King 1998). The transcripts were then recoded based on the reconciliation between the two coders.

Changes Made Prior to Conducting Exploratory Focus Group 2

One of the researchers involved in this study participated in EFG1 as an observer. He carefully judged peoples understanding of the scenarios, their reaction to the metrics and the flow of the conversation and took notes. From these notes, changes were made to focus group script, and summarized below:

1. Clarification of the goal of research, and description of who normally would utilize these types of tools and for what sorts of tasks. This was done with the creation and inclusion of a PowerPoint presentation:
    a. Description of the use of OLAP interfaces for decision support, and how frequently, in the use of these tools, the assumption is made that the data are correct.
    b. Outlining of research goal - understanding how confident people are with data at the cell level.

c. Description of the three data quality issues under consideration: unallocated data, information volatility and insensitivity to sample size.

2. Provision of more direction in each vignette scenario, and more probing of participants for final conclusion or judgment for each vignette.

3. For the volatility metric examples, addition of a graph to the OLAP screen for a pictorial feel of the variability in the data (along with the numeric value of volatility).

Exploratory Focus Group 2 (EFG2)

The same vignettes utilized in the first focus group were utilized. The changes described in the previous section were implemented (a copy of the PowerPoint presentation is in Appendix D).

Participants

There were a total of four participants, whose demographics characteristics are summarized in Table 21. This group consisted of data warehouse developers and database administrators involved in the implementation and support of a healthcare data warehouse. This group's duration was twice that of the first group, primarily because the participants were knowledgeable in the technical aspects of data aggregation and had many questions and comments. Secondly, there was a more detailed explanation of the study and of the vignettes. EFG2 was recorded (sound only) and professionally transcribed.

Table 21 – EFG2 Participants

Gender	Age	Last Degree	Current Position	Course in Statistics?	Years of work Experience	Years of Health-care Experience	Self Reported Comfort with Data Analysis (7 point scale)
M	36	MS MIS	Senior Network Administrator	Y	12	0	6
M	28	MS MIS	SQL Server DBA/ETL Developer	Y	6	6	6
F	33	MS MIS	SQL Server DBA	Y	9	1	6
M	33	MS MIS	Director of Application Development/ Assistant CIO	Y	9	9	5

Coding and Inter-rater Reliability

Utilizing the template created for EFG1, the transcripts for EFG2, were coded by identifying sections that were relevant and annotating the appropriate codes from the initial template. Cohen's Kappa for EFG2 was a disappointing 43%. Two reasons were identified for the low inter-rater reliability:

1. The second coder was not familiar enough with the vignettes and did not realize when the discussion was shifting to a new case.
2. There was not enough difference between lower level codes.

Two decisions were made to both remedy the current situation, and avoid it when coding the confirmatory focus groups. The coding template for this and all subsequent focus group was restricted to higher level codes for the metrics. The final coding scheme is shown in Table 22. When Cohen's Kappa was calculated using the new coding scheme it was 63%, which is still slightly lower then the recommended 70%.

Table 22 – Final Coding Scheme

Higher Level Construct	**Construct**	**Definition**
Unallocated Data	*Unallocated Data Before*	• Strategies to deal with unallocated data prior to receiving metric. • Interpretation prior to receiving metric.
	Unallocated Data After	• Strategies to deal with unallocated data after to receiving metric. • Interpretation after receiving metric.
	Design Feature Unallocated Data	Mention of the UD feature, design improvement suggestion.
Volatility	*Volatility Before*	• Strategies to deal with volatility prior to receiving metric. • Interpretation before receiving metric.
	Volatility After	• Strategies to deal with volatility after receiving metric. • Interpretation after receiving metric.
	Design Feature Volatility	Mention of the Information Volatility feature, design improvement suggestion.
Insensitivity To Sample Size	*Insensitivity To Sample Size Before*	• Strategies to deal with sample size prior to receiving metric. • Interpretation prior to receiving metric.
	Insensitivity To Sample Size After	• Strategies to deal with sample size- after receiving metric - Data Analysis Tactic. • Interpretation after receiving metric.
	Design Feature Sample Size	Mention of the Sample Size Indicator feature, design improvement suggestion.
Other[5]	*Overall Data Quality*	Perceptions of data quality, Other issues of data quality.
	Speculation	Speculation on DQ problems.
	Other Factors In Decision Making	Including stakeholder issues.

[5] These constructs were utilized for post-hoc analysis

Additionally, the transcripts given to the coders included screen shots and explanations of what was being discussed, as well as highlighting the passages that were to be coded (for example, ignoring side chatter and conversations).

EFG2 was then reconciled between coders. The two independent coders discussed the areas of disagreement, stopping when agreement was reached on all higher ordered codes and most lower order codes (King 1998). The transcripts were then recoded based on the reconciliation between the two coders.

Changes Made Prior to Conducting Confirmatory Focus Groups

After conducting this final exploratory focus group, two types of changes were made based on participant's comments: 1) changes to the focus group methodology and 2) changes to the metrics and their presentation.

Changes to the Focus Group Methodology

1. Included a statement which clarified that decisions were to be made with available information – though in a realistic situation such a decision would not be made without considering other sources of data. The participants of EFG1 were expert data warehouse developers, and many had trouble making a blanket decisions based on limited data. For example one participant stated:

   > *"Well, one of the things that I – but see, I'm a numbers person. So I would say, well, how many variables do you want to throw in to your analysis? And then once you have a decent number of variables, then you could say – instead of guessing"*.

   Another participant suggested that he would like more in-depth analysis:

> *"I mean, if I was going to make this – if I was going to make this leap of faith, and say hey, cancer is caused by smoking, or whatever – this particular cancer is caused by smoking, I'd want to know how they are indicated, and then take them out one by one, and see if that changes anything"*

This point was clarified in the confirmatory focus groups.

2. Removed third unallocated data example. This example was complicated and difficult to explain. Too much time was spent explaining and answering questions about this example. Since, there were already two UD examples, this vignette was dropped.

3. Showed unallocated data cube in PowerPoint presentation to explain unallocated data. Proportional allocation was difficult to explain without a visual aid. Most of the participants in the exploratory groups wanted a better explanation on how the unallocated data was distributed. They questioned reallocating the data based on the volumes on the cube, rather then basing in on general population (for example in the disparity in chemotherapy treatments example, reallocating according to the sample population):

> *"Would that proportional allocation be different if you did it where you took the total of people receiving or not receiving chemo, and ignored the unknowns for now? And then, both for Hispanics and non-Hispanics, and figured out what the percentage was on the total of the sample, since it's from the same area. And see if that percentage holds."*

4. Explained Volatility by showing a chart of volatility in the stock market, and the effects of certain historical events on stock market returns.

Changes to Metrics

1. *UD Metric.* Rather than just presenting proportional allocation, a case based scenario was created which showed: the cube without allocation, the proportional allocation and the worst-cased scenario (based on the context of the decision).

2. *Information Volatility Metric.* Benchmarking was added, which included a graphical presentation, a numerical presentation and a categorical presentation (medium, high, low) of benchmarking data.

3. *Sample Size Indicator.* Gradients were added, trying to emulate sensitivity analysis set by the analyst based on conversations in the focus groups where participants indicated they would want to know the severity on the sample size issues.

Confirmatory Focus Group 1 (CFG1)

The vignettes and metrics were altered as a result of the exploratory focus and are summarized in Table 23. There were a total of four cases presented. Two of the vignettes utilized the UD metrics, one the volatility metric and one the sample size indicator. The focus group took place in the VA Hospital VISN 8 Patient Safety Center of Inquiry and lasted over two hours.

Table 23 – Summary of Confirmatory Focus Group Vignettes

Metric / Feature	Vignette	Decision	Problem
UD – Pie Chart	Studies have shown that smoking is responsible for most cancers of the larynx, oral cavity and pharynx, esophagus, and bladder. In addition, it is a cause of kidney, pancreatic, cervical, and stomach cancers, as well as acute myeloid leukemia.	Is there correlation between smoking and certain types of cancer?	Large amounts of missing data in one grouping attribute (whether the patient smoked or not).
UD – Proportional Allocation, Worse Case Scenario	When Hispanics are diagnosed with a certain cancer (fictitious example), they're less likely to receive chemotherapy than non Hispanics.	Is there disparity in care?	Large amounts of missing data in several grouping attributes.
Information Volatility – Numeric IV Metric and Benchmarking	Counties neighboring Miami-Dade are better at early detection/prevention of Breast Cancer based on volumes of cases.	Examine trend – is this a true claim?	Neighboring counties' data exhibit large jumps in values.
Sample Size Indicator - Highlighting	Tumor size has been shown to be a good predictor of survival for certain cancers, including: breast, lung and endocrine. Compare average tumor size in Hillsborough to that of neighboring counties.	How does Hillsborough compare to other counties?	Neighboring counties may have averages based on very small sample sizes.

Table 24 – Confirmatory Focus Group 1 Participants

Gender	Age	Last Degree	Current Position	Course in Statistics?	Years of work	Years of Health-care	Self Reported Comfort with Data
M	34	Ph.D.	Health Economist	Y	7	6	7
M	51	Ph.D.	Assistant Director of Measurement and Evaluation	Y	28	28	7
F	49	Ph.D.	Researcher	Y	28	28	5
F	35	Ph.D.	Project Manager/ Data Manager/ Data Analyst/ Health Science Specialist	Y	9	9	5
M	56	Ph.D.	Health Services Researcher	Y	25	20	6
F	31	MA/ MPH	Program Specialist	Y	8	7	5
F	31	MSPH	Project Manager	Y	8	6	7
F	36	Ph.D.	Health Economist	Y	NR	3	7

Participants

There were a total of eight participants, whose demographics characteristics are summarized in Table 24. This group of participants was different then the first few in that they all held advanced degrees (most had a Ph.D), and they were used to doing the sort of data analysis that these vignettes outlined. In fact, there was some difficulty in having them “buy in” to the vignettes, since they had difficulty abstracting and making a simple decision. There is more discussion about this in the chapters eight and nine. The cases were presented in the same manner as the previous focus groups, and were recorded (sound only) and professionally transcribed.

As explained in Chapter 3, no further changes were made to either the metrics or the methodology to allow comparison across groups.

Coding and Inter-rater Reliability

The final template was used to code the transcripts for CFG1. Cohen's Kappa for CFG1 was 81%. CFG1 was then reconciled between coders. The two independent coders discussed the areas of disagreement, stopping when agreement was reached (King 1998). The transcripts were then recoded based on the reconciliation between the two coders.

Confirmatory Focus Group 2 (CFG2)

The same vignettes described in Table 23 were utilized to conduct the final confirmatory focus group.

Participants

There were a total of seven participants, whose demographics characteristics are summarized in Table 25. This group of participants was the least technical of all groups. Though varied in backgrounds, most of the participants were not heavy users of statistical analysis, but were involved in data intensive positions.

Coding and Inter-rater Reliability

The final template was used to code the transcripts for CFG2. Cohen's Kappa for EFG1 was 78%. CFG2 was then reconciled between coders. The two independent coders discussed the areas of disagreement, stopping when agreement was reached (King 1998). The transcripts were then recoded based on the reconciliation between the two coders.

Table 25 – Confirmatory Focus Group 2 Participants

Gender	Age	Last Degree	Current Position	Course in Statistics?	Years of work Experience	Years of Health-care Experience	Self Reported Comfort with Data Analysis (7 point scale)
F	29	MBA	Marketing Manager – Direct Mail Company	Y	7	0	5
M	32	MBA	Director of Merchandising Analysis	Y	11	7	7
M	44	MBA	Full time student	Y	20	0	7
M	30	MS-MIS	Doctoral Student	Y	3	0	5
F	36	MS	Doctoral Student	Y	9	0	6
M	25	BA	Business Intelligence Lead at Heath-care Consulting Company	Y	3	.5	7
M	26	MS	Doctoral Student	Y	4	0	6

Summary

This chapter describes the procedure, participants, coding, and incremental script changes for the four focus groups that were utilized to evaluate the metrics outlined in chapters four, five and six. The exploratory focus groups were utilized for the refinement of both the focus group procedure and the metrics and the confirmatory focus groups were used for the final evaluation of the metrics. The four focus groups were varied in membership, though each had a predominant type of participant. The focus groups were recorded and transcribed and coded by two independent coders. Cohen's Kappa of inter-rater reliability was calculated for each group. The results are summarized in Table 26.

Table 26 – Summary of Focus Groups

Group	No. of Participants	Average Age	Predominant Job Positions	Last Degree	Average Years of Work Experience	Average Years of Healthcare Experience	Gender	Average Comfort w/ Data Analysis (Low 1 – High 7)	Cohen's Kappa for Inter-Rater reliability
EFG1	3	31.5	Systems Analysts	Master MIS	8	0	M - 2 F - 1	4.5	.77
EFG2	4	32.5	DBA	Master MIS	9	4	M - 3 F - 1	5.75	.64
CFG1	8	40.3	Health Economist, Researcher	Ph.D.	16	13	M - 3 F - 5	6	.81
CFG2	7	33.75	Marketing Manager, Director of Advertising, Doctoral Student	MBA, Master MIS	10	1.75	M – 5 F - 2	6	.78

Chapter Eight: Focus Group Results

Introduction

In design science research, it is imperative to demonstrate evidence of the utility provided by the new artifact (Hevner, March et al. 2004). This chapter describes the evaluation of the three quality metrics proposed in chapters four, five and six. The evaluation is accomplished by interpreting and analyzing the data collected in a series of focus groups, in order to understand the particular implications to research questions two and three (Stewart, Shamdasani et al. 2007), restated below:

- What is the utility of the data quality metrics?
- What is the efficacy of the data quality metrics in altering a decision maker's data analytic strategies?

The identified constructs of utility and efficacy are investigated; looking for rich description of the participants' reactions to the metrics. Utility is defined as "usefulness of the metric" and efficacy as "having the ability to change data analytic strategies". To analyze utility of the metric all passages that were coded as "design feature" were analyzed. Change in data analytic strategies were evaluated contrasting the passages coded as "before" and "after" for each metric.

This chapter has three main sections, one for each of the metrics. For each metric both evidence and counter-evidence of the utility and efficacy of the metrics is presented.

Unallocated data

The unallocated data (UD) metric considers the effects of null values in any of the grouping or filtering variables, providing an operational definition for aspects of incompleteness. The UD metric and presentation methods are intended to highlight the impact of incompleteness on data cubes. Three strategies were presented to the focus groups. The first, showing "unknown groupings", is not a proposed metric, but what generally happens when incomplete data is present. Understanding the usefulness and efficacy of this approach helps to contrast what is currently done to the proposed metrics. The second is Proportional Allocation, and the third is Scenario Based Allocation, which includes: ignoring unknowns, proportionally allocating unknowns and assuming the worst-case for the group under consideration. Participants were asked to make a judgment on whether there was a disparity in care between Hispanics and non Hispanics, with unallocated data in several of the grouping variables and combination of grouping variables.

Utility of the UD Metrics

Table 27 summarizes the utility evaluation by type of UD metric and by focus group. Since there were three UD metrics presented, this discussion contains three separate subsections that summarize the results for each metric.

Unknown Groupings Utility

Not surprisingly, the focus groups did not find showing unknown groupings useful. Mostly this generated distrust in the data, as stated by a participant in EFG1:

"But there are just a lot of unknowns. And I just think in general, from what we've seen, it would be very hard for me to make a decision based on the (this data)."

Table 27 – Summary of Utility of UD Metrics

Focus Group	**UD Approach**	**Evidence of Utility**	**Counter-Evidence of Utility**
EFG1	Unknown Groupings	None, though one participant found it useful – gave ability to make an informed decision.	Participants distrustful of data and unlikely to use it.
EFG2	Unknown Groupings	None	Participants distrustful of data and unlikely to use it.
	Proportional Allocation	Yes	Asked if calculating based on underlying population would be more useful.
	Scenario Based Allocation	Yes	
CFG1	Unknown Groupings	None	Incomplete data is a problem only found with secondary data.
	Proportional Allocation	Yes – Easier to interpret then showing unknown groupings.	Asked if calculating based on underlying population would be more useful.
	Scenario Based Allocation	Useful for sensitivity analysis.	• Worse case Scenario is drastic. • Could be confusing for non-expert user.
CFG2	Unknown Groupings	None	Knowing about unallocated data made participants distrustful of data.
	Proportional Allocation	Yes	
	Scenario Based Allocation	Yes	• Not feasible for every decision. • Would also like to see best case.

One particular participant in EFG2 thought it would be useful to know this information in the formulation of his decision, though the rest of the participants did not agree:

> *"I want information to be available to me, nothing hidden. I'd want to be able to make an intelligent decision with all the information available."*

Interestingly, CFG1 thought that this problem did not apply to their organization, thus they thought the utility was low for the types of data she was accustomed to dealing with:

> *"We're typically more hands on. We are aware of why things are missing in general. So I guess for me it's a little bit harder to talk about something that's completely unknown because I don't have a context."*

The fact that participants did not find this technique useful helps to reinforce the need for different tactics to deal with incomplete data.

Proportional Allocation Utility

All groups that were presented the proportional allocation metric found it to be useful. They liked that the unknowns were eliminated, as stated by a participant in EFG2:

> *"I like the way you used the local data, and then proportioned it."*

Two of the focus groups questioned the proportional allocation approach, wondering if the reallocation should be based on the underlying population (which, as explained in Chapter 4 would be a nearly impossible calculation). For example the following comments from a participant in CFG1 and in EFG2:

"The way I thought about it was to look at the counties that these data are missing in terms of at least ethnicity and then see what percentage of that county's population is either Hispanic or not and then actually maybe emulate from that."

"Wouldn't it make more sense to look at Broward, and look at percentages of your actual population, and then allocate the unallocated data based on that?"

Scenario Based Utility

The scenario based approach was found useful by all the groups in which it was presented. They liked the flexibility, as stated by two of the participants in CFG2:

"I think that's really good. I like having the option to choose which one you use."

"Really, you've got an idea, but you're still guessing, so let me see both, and I'm guessing both ways to see how the data plays out. But I agree all three of the ways that you had mentioned would be useful."

Some of the participants recognized that several scenarios allowed them to consider several options, based on the context of their decision. As stated by a participant in CFG1:

"… it's almost like sensitivity analysis."

Several participants in CFG1 thought that the worst-case scenario approach was drastic, but most of the participants in all groups that were shown this metric agreed that the worst-case, when dealing with something like a disparity in health care, would be more useful than best case. A participant in EFG2 had an interesting example:

> *"Supposing they are building … parking. And you're building it in Davis Island, right next to Tampa General. You would probably want to look at calibrating with the worst-case scenario if the place had a hurricane, starting in a Cat 4 hurricane. You wouldn't want to look at unknowns and say, okay, let's push unknowns proportionately. Because if a Cat 3 or 4 hurricane hit there the weight of the decision would affect your weighing of the unknowns."*

CFG1 and CFG2 were concerned with the cognitive load of several scenarios: A participant in CFG1 stated:

> *"… if you had started having things flip from one to another than it's more difficult for a decision maker to know how to use those data."*

A participant in CFG2 has a similar comment:

> *"I don't know that I would do any kind of sensitivity because then you leave more questions about what do."*

It was probably best summarized by a participant of CFG2, who related it to her present job:

> *"In the line of work that I'm in, we have a lot of unallocated information where we don't know where the demand came from, and so our approach is to weight it equally across everything so that everybody kind of gets a little piece of that pie, and that's our best guess at figuring out something that we truly have no idea of where it came from, how it goes. So I think the three-prong approach that you have is really interesting, but I don't know how feasible it would be to do that every single time you have an analysis."*

Efficacy of UD metrics

In this section we describe the changes in decision making strategies when the participants were asked to make a series of decision with and without the use of the UD metrics. The results are summarized in Table 28.

<u>Unknown Groupings Efficacy</u>

For all groups when they were presented with figures on unknown groupings, if the unknown percentage was low, participants felt comfortable drawing conclusions. However, as the percentage of unknown increased they were less confident with the data, and less willing to make a decision based on that data alone. One of the strategies frequently seen was to speculate on the reason for the unknown data, which certainly could lead to poor decision making. For example one participant in CFG1, assumed unknowns were people that did not self-report that they smoked:

> *"I don't think people want to admit that they smoked pot or other drugs."*

Other participants in the same group recognized that this would not be a good strategy.

> *"What the real answer is. I mean so you really can't make any assumptions about that unknown data*."

Table 28 – Changes in Data Analytic Strategies

Focus Group	**UD Approach**	**Change in Data Analytic Strategies?**	**Comments/Observed Changes**
EFG1	Unknown Groupings	N/A	When unknown percentages were low, ignored, but once number grew speculation /stakeholder issues surfaced.
EFG2	Unknown Groupings	Yes	When unknown percentages were low, ignored, but once number grew speculation /stakeholder issues surfaced.
	Proportional Allocation	Yes	Preferred using this approach in combination of scenario based reasoning.
	Scenario Based Allocation	Yes	
CFG1	Unknown Groupings	N/A	Rejected task, group disliked low realism of the vignettes, refused to make decision.
	Proportional Allocation	No	
	Scenario Based Allocation	No	
CFG2	Unknown Groupings	N/A	When unknown percentages were low, ignored, but once number grew speculation /stakeholder issues surfaced.
	Proportional Allocation	Yes	Preferred using this approach in combination of scenario based reasoning.
	Scenario Based Allocation	Yes	

Some participants said that it depended on who was paying for the data collection. This may influence how unallocated data would be considered. For example a participant in EFG1 stated for the smoking/cancer correlation example:

> *"If I was on the side of the case…smoking is bad, I would say only 27% of them had never smoked, whereas, if I were …on the side saying smoking isn't part of the problem, only 44%."*

Most groups said that they would just ignore it, as a participant in EFG2 stated:

> *"Well I think traditionally what we would do is just to ignore it, not include it in your analysis."*

Overwhelmingly most of the participants agreed that it would be difficult to make a decision because of the uncertainty that the unknown data caused. As a participant in EFG2 states:

> *"I guess the one way to think about it is you waffle."*

Proportional Allocation Efficacy

All groups but one, CFG1, altered their decision and data analysis strategies when given the ability to consider proportional and scenario based allocation. For example, in CFG2 participants noticed that the disparity was not as severe as it seemed when making a decision and ignoring the unallocated data.

> *"You are starting to see some non-Hispanics also not receiving the treatment."*

Scenario Based Efficacy

When examining several versions of the data, again all groups but one altered their decision making. Most participants stated that depending on the context of the decision, their decision making would change.

Summary of UD Metrics Utility and Efficacy

There was enough evidence to indicate that focus groups found the UD metrics of proportional allocation and scenario based allocation useful. Efficacy was also demonstrated, in all but one group. CFG1 did not find the task realistic, since they were not able to make a simple decision without considering external circumstances. They refused to make any decision, thus, it was hard to demonstrate a change in decision making strategies due to the metric.

Information Volatility

The vignette used for the information volatility metric asked participants to compare several series of numbers and a graph (with a linear trend line) that described a trend. A statement was made that counties neighboring Miami-Dade were better at early detection/prevention of breast cancer based on trend on volumes of cases, which was flat for Miami-Dade but decreasing for neighboring counties. The problem with several of the counties selected for comparisons was that the volumes exhibited large jumps in values, thus were probably less reliable and an unrealistic comparison from which to draw any conclusions. The participants were asked to make a judgment by observing a trend line on a graph, as well as the actual numbers, and then again, after introducing and presenting the information volatility metric for each series of numbers.

Utility of Information Volatility Metric

The results for the utility of the Information Volatility Metric are summarized in Table 29. In general all the groups found this to be a useful metric. For the exploratory focus groups, benchmarking was not used, and the participants voiced that just a numeric representation was difficult to understand, and that they had a difficult time interpreting what this number meant, as demonstrated by the following comments from EFG2.

> *"I think it would depend on the user. I mean, I think we would be able to figure it out, and I think a lot of people would, but I think a lot wouldn't."*

> *"Yeah, the volatility one is a little – I think a little more –difficult. Well, because people don't have a lot of background in what that means."*

Table 29 – Utility of Information Volatility Metric

Focus Group	**Evidence of Utility**	**Counter-Evidence of Utility**
EFG1	Yes	Difficulty Interpreting
EFG2	Yes	Difficulty Interpreting
CFG1	Yes - Saw several instances where this would be useful in their daily data analysis	None
CFG2	Yes	None

In fact, the benchmarking idea was a design feature that was added after the exploratory focus groups. For example, one participant in EFG2 stated:

> *"You (need to) draw a line in the sand and say, this is a problem, this is not. And maybe if it goes over that line, it pops up and says, 'Hey, check this out.'"*

This was corroborated by the confirmatory focus groups. For example a participant of CFG2 stated the:

> *"…benchmarking is a necessary component of it."*

Two of the groups EFG2, CFG1 were extremely enthusiastic about the utility of this metric. For example a participant in CFG1 stated:

> *"I like that calculation and the idea of having a metric or measuring and giving you this kind of information."*

In fact, most of the focus group participants made very similar comments, and discussed several ways that this metric would be useful in their current jobs. For example a participant in CFG2 related:

> *"So this – this applied to an example of the VA where you have some nursing homes that have less than 30 beds, small n's can make it 5 percent – 10 percent change as opposed to a 300 bed plus facility where it takes 25 people to get the same kind of, you know, impact. So we can use this."*

CFG2 found the benchmarking information useful, also relating it to issues in their workplace, as one participant commented:

> *"It's keeping the institutional memory to what those numbers really mean because you know we – we've sat over at this end and don't see much. Okay, let's compare that example of like Tampa to Miami and or you're looking at costs or you're looking at clinical wait times or something and then you have some sort of huge variation between the two and you can make a conclusion like they don't know what they're doing. … and then you get down to the numbers and the nitty gritty and you talk to someone over there and say oh, we're in a transition period and we've got some issues with our data."*

Efficacy of Information Volatility Metric

The counties provided for comparison were highly volatile, but in general, most focus groups prior to receiving the information volatility metric thought Miami-Dade was not declining as rapidly as other counties. For example, prior to seeing the IV metric a participant in CFG2 noted:

> *"No, they're not doing as well, because they have a straight across line, and there's no decrease; whereas the other two counties that you showed had a decrease."*

Most focus groups changed their decisions once they were informed about the volatility in the data. The results are summarized in Table 30.

Table 30 – Efficacy of Information Volatility Metric

Focus Group	**Change in Data Analytic Strategies?**	**Comments/Observed Changes**
EFG1	Yes	
EFG2	Yes	
CFG1	Slight	Rejected task, group disliked low realism of the vignettes, refused to make decision.
CFG2	Yes	

When information on volatility was available, the participants were less likely to compare trends if one of the trends were labeled as highly volatile. In the case of this vignette, they reversed their prior decision, since the counties that were being compared to Miami-Dade had high IV numbers.

CFG1 had difficulty "buying into" the reality of such a scenario, though they found this metric useful and saw the potential for its use in their daily tasks. This time they did show some changes in data analytic strategies when they decided they would "think like a manager":

> *"… if I were a manager and I'm looking at these trend lines and one looks flat and one looks down and the variability looks about the same, you know it's not huge on one or huge on another, I think I'd be asking what's going on. I think I'd*

say you know what are they doing right and what are we doing wrong here or whatever."

Summary of Information Volatility Utility and Efficacy

There was enough evidence to indicate that focus groups found the IV metrics useful. Efficacy was also demonstrated, although in one group thought this evidence was not as strong because the task was rejected. Since CFG1 refused to make a decision, it was hard to demonstrate a change in decision making strategies due to the metric, though they did display a possible change in data analytic strategies by role-playing "a manager", but detaching themselves from the example.

Sample Size Indicator

This metric differs from the other two, in that it is not a calculation, but rather it is a simple approach to draw the attention of a decision maker to aggregated data based on small sample sizes. This research investigates how in a BI tool, the data presentation can be utilized to mitigate the bias of insensitivity to sample size by drawing attention to sample sizes. The focus groups were used to identify if participants were indeed insensitive to sample size as shown by precious studies (Bar-Hillel 1982; Tversky and Kahneman 1982; Hastie 2001). The demonstration that participants are insensitive to sample size reveals the necessity of any form of metric that highlights potential sample size problem. Table 31 summarizes the insensitivity to sample size by focus group. The majority of the focus groups did indeed ignore the volume when comparing averages, until they were made aware of the small sample sizes. Each of the focus groups did indicate that in fact that they were surprised to have made such a mistake, since they were well aware that sample size should have been taken into consideration when comparing averages.

Table 31 – Insensitivity to Sample Size by Focus Group

Focus Group	**Insensitivity to Sample Size**	**Comments/ Reaction**
EFG1	Yes	Questioned definition of small sample size
EFG2	No	Pointed out possible effect of outlier
CFG1	Yes	Screen shown too quickly for consideration
CFG2	Yes, but only initially	After several examples questioned the effect of outliers and eventually pointed out small sample size

One EFG2 did question sample size immediately. One participant noticed the difference in sample size:

"Well, you know it's – on (an) average, though, (and) the sizes are different."

This informed the rest of the group, and the rest of the observations took sample size into consideration, with phrases like:

"That would be good data – because you're getting a good, strong sample."

There was a similar comment in CFG1:

"You know one of the problems – well one of the reasons those – those small counties are volatile is that they have relatively few cases."

Interestingly enough, even after this comment was made, this group continued to ignore sample size. Several counties were judged as better at cancer detection then other counties though the volumes were quite small.

A simple design feature (highlighting) is suggested and prototyped to the focus groups and its utility and efficacy is evaluated.

Utility of the Sample Size Indicator

All four groups found that highlighting was useful in drawing their attention to small sample sizes, thus warning them about comparing aggregated data from these highlighted fields (data are summarized in Table 32). For example, a participant in CFG1 stated:

> *"Yeah, it just leaves that. You get – well if you see a chart with red and yellow all over it, you're thinking okay, you know be really careful."*

One participant was able to relate the usefulness of this method to his job in telecommunications:

> *"In my job, we use a lot of these networking tools, and you'll see a significant average failures- you'll see a huge number. But there was 12 observations compared to the 300 at another. So it's imperative in our case, because we've got to compare apples to apples before we spend a couple grand to upgrade something."*

In another group, there was some counter-evidence of the utility of this method. CFG2, thought highlighting would actually draw their attention to those numbers:

> *"I think it's distracting."*
>
> *"I agree, because I'm looking at all the colors, and it's hard for me to actually look at the values that are colored. The ones to focus on are the ones that aren't highlighted."*

Even when the moderator explained that the focus of the study was to see if drawing attention to these cells would minimize this bias (so highlighting could be replaced with a small flag, or a slightly different gradient in the text), several members in CFG2 insisted anything would be distracting (thus not useful) and they would prefer if the tool did not report these numbers at all:

> *"Just don't show me anything that I can't get caught up with. Keep it simple. This other stuff's just distracting. You have those yellow gradients, so they may or may not be significant?"*

Table 32 – Utility of Sample Size Indicator

Focus Group	**Evidence of Utility**	**Counter-Evidence of Utility**
EFG1	Yes	No
EFG2	Yes	No
CFG1	Yes	No
CFG2	Yes	Yes

Efficacy of Sample Size Indicator

In three groups, once highlighting was introduced, the participants were careful to utilize those averages that had small volumes. They immediately changed their decision making strategies. As participants in CFG2 stated:

> *"We'll take these with a grain of salt."*

> "(*Or) go get more data."*

Participants in EFG1 even discussed strategies for different contexts, with one participant questioning the definition of a small sample size and another discussing some approaches:

> *"Well, for a very common cancer, you would want a larger sample size, for an uncommon cancer, a smaller one."*

CFG2 did not exhibit a change in data analytic strategies. The group realized that the volumes were small for neighboring counties after only a few comparisons, thus highlighting was unnecessary. In fact, as previously described, this group disliked the highlighting. The results are summarized in Table 33.

Table 33 – Efficacy of Sample Size Indicator

Focus Group	**Change in Data Analytic Strategies?**	**Comments/Observed Changes**
EFG1	Yes	
EFG2	Yes	
CFG1	Yes	
CFG2	No	Was already aware of small sample sizes

Summary of Sample Size Utility and Efficacy

There was enough evidence to indicate that most of the participants were insensitive to sample size and that drawing attention to these small sample sizes was a useful technique, though there was some counter-evidence that highlighting was the best method. Once highlighting was introduced, in most cases data analysis strategies changed, since the participants were less willing to compare averages with small sample sizes.

Summary and Discussion

There was enough evidence to indicate that focus groups found the metrics useful and that the metrics were efficient in altering a decision maker's data analytic strategies. An interesting finding is that for all the metrics participants preferred some sort of comparison, such as the scenario based approach for UD, benchmarking for the IV metric and even for sample size, participants questioned the definition of “small”.

Another interesting finding is that most groups showed some form of the insensitivity to sample size bias, even though all had extensive statistics training. This highlights the need to consider judgment biases when designing BI tools, and is an interesting area for further development and research.

Chapter Nine: Conclusions

Concluding Remarks

This work is one of the first to investigate issues of data quality in the information supply chain. It proposes three result-driven data quality metrics that inform and aid decision makers with incomplete and inconsistent data and help mitigate insensitivity to sample size, a well known decision bias. The Unallocated Data metrics consider the effects of null values in any of the grouping or filtering variables. Information volatility describes the rate of change in the values of stored data. For both of these metrics it was found that comparative techniques, such as benchmarking or scenario based approaches are promising approaches in data quality. In addition, results from this research indicate that decision making literature should be considered in the design of BI tools.

This research is also one of the first to propose the use of focus groups as a technique to evaluate design science research. It outlines a methodology for planning, selecting participants, conducting, analyzing and reporting the results of the focus groups to demonstrate utility and efficacy of the artifacts.

This research provides practitioners three implementable result-driven data quality metrics that allow the consideration of the context in decision making and consider decision making biases.

Limitations

The defined data quality metrics should be useful in an environment where decision makers utilize aggregated data. This research took a simplistic approach, by implementing these metrics for straightforward decisions in a controlled environment. Further thought needs to be given to how and when to present these metrics, and whether the decision maker will have some control on setting the sensitivity. In several of the focus groups, participants noted that it would be cognitively taxing to receive this information for every single cell or sheet.

The Focus Group technique has several limitations. Firstly, the participants were not randomly selected, but rather a "convenience sample" was used, which could limit generazability (though the goal was to find people with a certain skill set). Secondly, the moderator had control of the interface in which these metrics were presented. The results could be different if the decision maker had been able to access them directly. Thirdly, the context was very important to participants. Careful care has to be taken to design tasks that are relevant to the group, as was shown by the resistance shown in CFG2. Finally, the two confirmatory focus groups were quite contrasting when evaluating utility. Focus groups should continue until nothing new is learned (Krueger and Casey 2000).   Having contrasting results for the evaluation of efficacy for the two confirmatory groups is a limitation and indicates that further focus groups should be conducted.

Future Research

Certainly, the proposed metrics can be further refined. Moreover, other evaluation techniques need to be investigated. For example, controlled experimentation may help to clearly understand impact of metrics on decision making, or simulations may help in evaluating the quality of the metrics. Furthermore, these metrics need to be compared and contrasted to existing metrics in the literature. Finally, the utility and efficacy of these metrics should be extended to other decision making contexts.

The in intent of the focus groups was to evaluation of the proposed data quality metrics. However, several other “user views" of data quality emerged that merit serious consideration. The focus group technique allowed the researcher to observe data quality in action (in decision making). Three crucial aspects of this user view emerged:

1. Participants were skeptical of the data in the examples (which for the most part was from a real ISC), but were not skeptical about their own data (data that they utilized in their jobs), perhaps because they have very high ownership of that data and believe it to be of high quality.

2. The user model of data quality must recognize the finding of research on behavioral decision making that is relevant to issues of data quality. This research took a step in this direction by identifying ways to mitigate the bias of insensitivity to sample size. There are however several other behavioral decision making issues that were noticed in the focus groups. For example, when faced with uncertainty, several approaches were taken to analyze the data. These included speculating on the reasons for poor data quality (Hispanics don’t go to the doctor as much), or bringing in stake holder issues (who’s side am I on? I can use poor data quality to sway the decision depending on what you want the answer to be).

3. Several data quality attributes that are conceptually separate in the technical model and interrelated in the user model. For example volatility and seemed to

communicate problems with small sample size as well. Identifying interrelated data quality metrics is important for measurement purposes and may reveal some attributes that are most interconnected with other data quality attributes. This should identify to the research community the most crucial data quality dimensions to identify to the user.

Summary

This work is one of the first to propose three result-driven data quality metrics designed for use in an environment where managers and decision makers utilize aggregated data (summary information) retrieved from several data sources in the information supply chain to make tactical decisions. The study is based in a rich environment, health planning which provides relevance, yet these metrics can easily be extended to other context where aggregated data is utilized.

References

"http://hcecf.org/hcabout.html."

Al-Shorbaji, N. (2001). Health and Medical Informatics:Technical Paper. Health Information Support, Regional Office for the Eastern Mediterranean,World Health Organization. Cairo,Egypt.

Anonymous (2006). "Information Systems and Health Care XI: Public Health Knowledge Management Architecture Design: A Case Study." Communications of the Association for Information Systems **18**: 15.

Baker, T. and D. A. Collier (2005). "The Economic Payout Model for Service Guarantees." Decision Sciences **36**(2): 197.

Ballou, D., R. Wang, et al. (1998). "Modeling Information Manufacturing Systems to Determine Information Product Quality." Management Science **44**(4): 462-484.

Ballou, D. P. and H. L. Pazer (1985). "Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems " Management Science **31**(2): 150-162.

Ballou, D. P. and H. L. Pazer (2003). "Modeling completeness versus consistency tradeoffs in information decision contexts." Knowledge and Data Engineering, IEEE Transactions on **15**(1): 240-243.

Bar-Hillel, M. (1982). Studies of Representativeness. Judgment Under Uncertainty:Heuristics and Biases. D. Kahneman, P. Slovic and A. Tversky. Cambridge, Cambridge University Press.

Basili, V. R. (1996). The role of experimentation in software engineering: past, current, and future Proceedings of the 18th international conference on Software engineering Berlin, Germany IEEE Computer Society.

Baskerville, R. and M. D. Myers (2004). "Special Issue on Action Research in Information Systems: Making Is Research Relevant To Practice-Foreword." MIS Quarterly **28**(3): 329.

Benbasat, I. and R. Weber (1996). "Research commentary: Rethinking "diversity" in information systems research." Information Systems Research **7**(4): 389.

Berndt, D. J., A. R. Hevner, et al. (2003). "The CATCH data warehouse: Support for community health care decision-making." Decision Support Systems **35**(3): 367.

Burdick, D., P. M. Deshpande, et al. (2006). "OLAP over uncertain and imprecise data " The VLDB Journal **16**(1): 123-144.

Chapman, G. B. and A. S. Elstein (2000). Cognitive Processes and Biases in Medical Decision Making. Decision Making in Healthcare:Theory, Psychology and Applications. G. B. Chapman and F. A. Sonnenberg, Cambridge University Press.

Chen, P. P.-S. (1976). "The entity-relationship model—toward a unified view of data " ACM Trans. Database Syst. **1**(1): 9-36.

Cohen, J. (1960). "A coefficient of agreement for nominal scales." Educational and Psychological Measurement **20**: 37-46.

Cole, R., S. Purao, et al. (2005). Being Proactive: Where Action Research Meets Design Research. Twenty-Sixth International Conference on Information Systems, Las Vegas.

Croome, S. (2003). "Understanding Volatility Measurements." http://www.investopedia.com/articles/mutualfund/03/072303.asp Retrieved March 26, 2007, 2007, from http://www.investopedia.com/articles/mutualfund/03/072303.asp.

Cui, Y. and J. Widom (2000). Practical Lineage Tracing in Data Warehouses. 16th International Conference on Data Engineering (ICDE'00), San Diego, California.

Davenport, T. H., D. Cohen, et al. (2005). Competing on Analytics, Babson Executive Education Working Knowledge Research Center.

Davenport, T. H., S. L. Jarvenpaa, et al. (1996). "Improving knowledge work processes." Sloan Management Review **37**(4): 53.

Davis, G. B. (2002). "Anytime/anyplace computing and the future of knowledge work " Commun. ACM **45**(12): 67-43.

Debreceny, R., M. Putterill, et al. (2003). "New tools for the determination of e-commerce inhibitors." Decision Support Systems **34**(2): 177.

Derose, S. F. and D. B. Petitti (2003). "Measuring Quality of Care and Performance from a Population Health Care Perspective." Annual Review of Public Health **24**(1): 363-384.

Derose, S. F., M. A. Schuster, et al. (2002). "Public Health Quality Measurement: Concepts and Challenges." Annual Review of Public Health **23**(1): 1-21.

Desanctis, G. and G. R. (1987). "A Foundation for the Study of Group Decision Support Systems." Management Science **33**(4): 589-609.

Drucker, P. F. (1993). Post-Capitalist Society. New York, NY, Harper Collins.

Drucker, P. F. (1999). "Knowledge-worker productivity: The biggest challenge." California Management Review **41**(2): 79-+.

Einhorn, H. J. and R. M. Hogarth (1981). "Behavioral Decision Theory: Processes of Judgment and Choice." Annual Review of Psychology **32**(1): 53-88.

Friede, A., H. L. Blum, et al. (1995). "Public Health Informatics: How Information-Age Technology Can Strengthen Public Health." Annual Review of Public Health **16**(1): 239-252.

Galhardas, H., D. Florescu, et al. (2001). Improving data cleaning quality using a data lineage facility. . Workshop on Design and Management of Data Warehouses (DMDW), Interlaken, Switzerland.

Galliers, R. D. (1991). Choosing Appropriate Information Systems Research Approaches: A Revised Taxonomy. Information Systems Research: Contemporary Approaches and Emergent Traditions. H. E. Nissen, H. K. Klein and R. Hirschheim. Amsterdam, The Netherlands, Elsevier Science Pub.

Harries, C. (2003). "Correspondence to what? Coherence to what? What is good scenario-based decision making?" Technological Forecasting and Social Change **70**(8): 797-817.

Hastie, R. (2001). "PROBLEMS FOR JUDGMENT AND DECISION MAKING." Annual Review of Psychology **52**(1): 653-683.

Heath, D. and E. Platen (2006). "Local volatility function models under a benchmark approach." Quantitative Finance **6**(3): 197 - 206.

Hevner, A., S. March, et al. (2004). "Design Science Research in Information Systems." Management Information Systems Quarterly **28**(1): 75-105.

Hotopp, S. (1997). Practical Issues Concerning Volatility and Its Measurement Past and Predicted. Volatility in the capital markets : state-of-the-art techniques for modeling, managing, and trading volatility. I. Nelken. Chicago ; London ; New Delhi, Glenlake/Fitzroy Dearborn**:** xii, 224.

Imielinski, T. and W. Lipski (1984). "Incomplete Information in Relational Databases " Journal of the ACM **31**(4): 761-791.

Jakob, H. I., M. Lars, et al. (2004). "Managing Risk in Software Process Improvement: An Action Research Approach." MIS Quarterly **28**(3): 395.

Jarke, M., M. A. Jeusfeld, et al. (1999). "Architecture and quality in data warehouses: An extended repository approach." 10th International Conference on Advanced Information Systems Engineering **24**(3): 229-253.

Jarvenpaa, S. L. and K. R. Lang (2005). "Managing the Paradoxes of Mobile Technology." Information Systems Management **22**(4): 7.

Kahn, B. K., D. M. Strong, et al. (2002). "Information quality benchmarks: product and service performance " Communications of the ACM **45**(4): 184-192

Kahneman, D. and A. Tversky (1979). "Prospect Theory: An Analysis of Decision under Risk." Econometrica: Journal of the Econometric Society **47**(2).

Kimball, R. and M. Ross (2002). The data warehouse toolkit: the complete guide to dimensional modeling. New York, Wiley.

King, N. (1998). Template Analysis. Qualitative Methods and Analysis in Organizational Research. G. Symon and C. Cassell. London, Sage Publications.

Klar, Y. (1990). "Linking Structures and Sensitivity to Judgment-Relevant Information in Statistical and Logical Reasoning Tasks." Journal of Personality and Social Psychology **59**(5): 841-858.

Klein, B. D., D. L. Goodhue, et al. (1997). "Can humans detect errors in data? Impact of base rates, incentives, and goals." MIS Quarterly **21**(2): 169-194.

Kontio, J., L. Lehtola, et al. (2004). Using the focus group method in software engineering: obtaining practitioner and user experiences.

Kotze, A. A. (2007) "An Introduction to Volatility and how it can be calculated in Excel." **Volume**, DOI:

Krueger, R. A. and M. A. Casey (2000). Focus groups : a practical guide for applied research. Thousand Oaks, Calif., Sage Publications.

Maibach, E. and D. R. Holtgrave (1995). "Advances in Public Health Communication." Annual Review of Public Health **16**(1): 219-238.

Manning, P. K. (1996). "Information technology in the police context: The "sailor" phone." Information Systems Research **7**(1): 52.

Mantei, M. M. and T. J. Teorey (1989). "Incorporating Behavioral Techniques Into The Systems Development." MIS Quarterly **13**(3): 257.

Markus, M. L., A. Majchrzak, et al. (2002). "A Design Theory for Systems that Support Emergent Knowledge Processes." MIS Quarterly **26**(3): 179-212.

Massey, A. P. and W. A. Wallace (1991). "Focus groups as a knowledge elicitation technique: an exploratory study." Knowledge and Data Engineering, IEEE Transactions on **3**(2): 193-200.

Matchar, D. B. and G. P. Samsa (2000). Linking Modeling with Health Policy Formation and Implementation. Decision Making in Healthcare:Theory, Psychology and Applications. G. B. Chapman and F. A. Sonnenberg, Cambridge University Press.

McClave, J., P. G. Benson, et al. (2005). Statistics for Business and Economics. Upper Saddle River, NJ, Pearson Prentice Hall.

McMillan, L. G. (1996). McMillan on options. New York, J. Wiley.

Morey, R. C. (1982). "Estimating and improving the quality of information in a MIS " Commun. ACM **25**(5): 337-342

Morgan, D. L. (1988). Focus groups as qualitative research. Newbury Park, Calif., Sage Publications.

Mun, J. (2006). Real options analysis : tools and techniques for valuing strategic investments and decisions. Hoboken, N.J., John Wiley & Sons.

Nielsen, J. (1997). "The use and misuse of focus groups." Software, IEEE **14**(1): 94-95.

Oliver, T. R. (2006). "The Politics of Public Health Policy." Annual Review of Public Health **27**(1): 195-233.

Parssian, A. (2006). "Managerial decision support with knowledge of accuracy and completeness of the relational aggregate functions." Decision Support Systems **42**(3): 1494-1502.

Parssian, A., S. Sarkar, et al. (2004). "Assessing Data Quality for Information Products:Impact for Selection, Projection, and Cartesian Product." Management Science **50**(7): 967-982.

Pipino, L. L., Y. W. Lee, et al. (2002). "Data Quality Assessment." Communications of the ACM **45**(4ve): 211-218.

Redman, T. C. (1996). Data quality : the Field Guide. Boston,Ma, Digital Press.

Redman, T. C. (1996). Data Quality for the Information Age. Boston, Artech House.

Rikard, L., H. Ola, et al. (2004). "Design Principles for Competence Management Systems: A Synthesis of an Action Research Study." MIS Quarterly **28**(3): 435.

Sackett, D. L., W. M. C. Rosenberg, et al. (1996). "Evidence based medicine: what it is and what it isn't." BMJ **312**(7023): 71-72.

Seiner, R. S. (2000) "Questions Metadata Can Answer." The Data Administration Newsletter (TDAN.com)

Shankaranarayan, G. and Y. Cai (2006). "Supporting data quality management in decision-making." Decision Support Systems **42**(1): 302-317.

Shankaranarayan, G., M. Ziad, et al. (2003). "Managing Data Quality in Dynamic Decision Environments: An Information Product Approach." J. Database Management **14**(4).

Shankaranarayanan, G. and Y. Cai (2006). "Supporting data quality management in decision-making." Decision Support Systems **42**(1): 302-317.

Simon, H. A. (1996). The Sciences of the Artificial Cambridge,MA, MIT Press.

Smith, H. J., S. J. Milberg, et al. (1996). "Information privacy: Measuring individuals' concerns about organizational practices." MIS Quarterly **20**(2): 167.

Stewart, D. W., P. N. Shamdasani, et al. (2007). Focus groups : theory and practice. Newbury Park, Calif., Sage Publications.

Strong, D. M., Y. W. Lee, et al. (1997). "10 Potholes in the Road to Information Quality." IEEE Computer **30**(8): 38-46

Sun, S. and J. Yen (2005). Information Supply Chain: A Unified Framework for Information-Sharing Intelligence and Security Informatics. SpringerLink, Springer Berlin / Heidelberg. **3495/2005:** 422-428.

Timko, I., C. E. Dyreson, et al. (2006). Pre-aggregation with probability distributions Proceedings of the 9th ACM international workshop on Data warehousing and OLAP Arlington, Virginia, USA, ACM Press.

Torkzadeh, G., J. C.-J. Chang, et al. (2006). "Identifying issues in customer relationship management at Merck-Medco." Decision Support Systems **42**(2): 1116.

Tracey, T. J. and J. Rounds (1999). Inference and Attribution Errors in Test Interpretation. Test interpretation: Integrating science and practice. R. K. Goodyear and J. W. Lichtenberg. Boston, Allyn & Bacon.

Tremblay, M. C., R. Fuller, et al. (2006). "Doing More with More Information: Changing Healthcare Planning with OLAP Tools." Decision Support Systems **In Press**.

Trembly, A. C. (2002). Poor Data Quality: A $600 Billion Issue. National Underwriter (Life & Health/Financial Services Edition). **11**.

Tversky, A. and D. Kahneman (1982). Judgment Under Uncertainty:Heuristics and Biases. Judgment under Uncertainty: Heuristics and Biases. D. Kahneman, P. Slovic and A. Tversky. Cambridge, Cambridge University Press.

U.S., C. (1995). Bringing Health Care Online: The Role of Information Technologies. O. o. T. Assessment, U.S. Government Printing Office

Wand, Y. and R. Y. Wang (1996). "Anchoring data quality dimensions in ontological foundations " Commun. ACM **39**(11): 86-95.

Wang, R., M. P. Reddy, et al. (1993). An Object-Oriented Implementation of Quality Data Products. WITS-93, Orlando, Florida.

Wang, R. Y., M. P. Reddy, et al. (1995). "Toward quality data: An attribute-based approach." Decision Support Systems **13**(3-4): 349-372.

Wang, R. Y., V. C. Storey, et al. (1995). "A Framework for Analysis of Data Quality Research." IEEE Transactions on Knowledge and Data Engineering **7**(4): 623-640.

Wang, R. Y. and D. M. Strong (1996). "Beyond accuracy: What data quality means to data consumers." Journal of Management Information Systems **12**(4): 5-34.

Xia, W. and G. Lee (2005). "Complexity of Information Systems Development Projects: Conceptualization and Measurement Development." Journal of Management Information Systems **22**(1): 45.

Bibliography

"http://changingminds.org." (November 2006). "Using the Data:Quality Measures." U.S. Census Bureau Retrieved November, 2006, from http://www.census.gov/acs/www/UseData/sse/ita/ita_def.htm.

Abate, M. L., K. V. Diegert, et al. (1998). "A Hierarchical Approach to Improving Data Quality " Data Quality **4**(1).

Alavi, M. and D. E. Leidner (2001). "Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues." MIS Quarterly **25**(1): 107.

Ashton, R. H. (1992). "Effects of justification and a mechanical aid on judgment performance." Organizational Behavior and Human Decision Processes **52**(2): 292-306.

Benbasat, I., D. K. Goldstein, et al. (1987). "The Case Research Strategy in Studies of Information Systems." MIS Quarterly: 369-385.

Berndt, D. (2006). Next-Generation Software Engineering: Challenges in Data and Knowledge Management. Next Generation Software Engineering Workshop at HICSS, Hawaii.

Berndt, D. J. and J. W. Fisher (2001). Understanding Dimension Volatility in Data Warehouses (or Bin There Done That). Proceedings of the Sixth INFORMS Conference on Information Systems and Technology (CIST 2001), Miami, Florida.

Butler, S. A. (1985). "Application of a Decision Aid in the Judgmental Evaluation of Substantive Test of Details Samples." Journal of Accounting Research **23**(2): 513-526.

Cappiello, C., C. Francalanci, et al. (2003-2004). "Time-Related Factors of Data Quality in Multichannel Information Systems." Journal of Management Information Systems **20**(3): 71-92.

Chen, P. P.-S. (1976). "The entity-relationship model—toward a unified view of data " ACM Trans. Database Syst. **1**(1): 9-36.

Chengalur-Smith, I. N., D. P. Ballou, et al. (1999). "The impact of data quality information on decision making: An exploratory analysis." Ieee Transactions on Knowledge and Data Engineering **11**(6): 853-864.

Chung, W., H. Chen, et al. (2005). "A Visual Framework for Knowledge Discovery on the Web: An Empirical Study of Business Intelligence Exploration." Journal of Management Information Systems **21**(4): 57.

Cooper, B. L., H. J. Watson, et al. (2000). "Data warehousing supports corporate strategy at First American Corporation." MIS Quarterly **24**(4): 547.

Damianos, L., S. Wohlever, et al. (2002). MiTAP: A Case Study of Integrated Knowledge Discovery Tools. 36th Hawaii International Conference on System Sciences(HICSS'03), Hawaii, IEEE.

Datta, A. and H. Thomas (1999). "The cube data model: A conceptual model and algebra for on-line analytical processing in data warehouses." Decision Support Systems **27**(3): 289.

Davenport, T. H. and R. J. Thomas (2002). "The mysterious art and science of knowledge-worker performance." MIT Sloan Management Review **44**(1): 23-30.

Davis, G. B. and R. W. Collins (1991). "Conceptual Model for Research on Knowledge Work."

Dennis, A. R. and J. S. Valacich (2001). "Conducting Research in Information Systems." Communications of AIS **7**(7).

Doherty, N. F. and G. Doig (2003). "An analysis of the anticipated cultural impacts of the implementation of data warehouses." IEEE Transactions on Engineering Management **50**(1): 78.

Einhorn, H. J. and R. M. Hogarth (1981). "Behavioral Decision Theory: Processes of Judgment and Choice." Annual Review of Psychology **32**(1): 53-88.

Ekblad, S., A. Marttila, et al. (2000). "Cultural challenges in end-of-life care: reflections from focus groups' interviews with hospice staff in Stockholm." Journal of Advanced Nursing **31**(3): 623-630.

Even, A., G. Shankaranarayanan, et al. (2006). Enhancing Decision Making with Process Metadata: Theoretical Framework, Research Tool, and Exploratory Examination.

Fisher, C. W., I. Chengalur-Smith, et al. (2003). "The Impact of Experience and Time on the Use of Data Quality Information in Decision Making." Information Systems Research **14**(2): 170-188.

Fisher, J. W. (2002). Creating False Memories: Temporal Reconstruction Errors in Data Warehouses. Department of Information and Decision Sciences. Tampa, University of South Florida.

Fontanills, G., T. Gentile, et al. (2003). The volatility course workbook step-by-step exercises to help you master The volatility course. New York, John Wiley & Sons.

Frenk, J. (1993). "The New Public Health." Annual Review of Public Health **14**(1): 469-490.

Fuller, S. (1992). Knowledge as Product and Property. The Culture and Power of Knowledge: Inquiries into Contemporary Societies. S. N. and R. V. Ericson. Berlin, de Gruyter**:** 177-190.

Gardyn, E. (1997). A data quality handbook for a data warehouse. Conference on Information Quality. Cambridge, MA.

Glassey, K. (1998). "Seducing the end user." Association for Computing Machinery. Communications of the ACM **41**(9): 62.

Gnatovich, R. (2006). BI Versus BA--What's the Difference? CIO.

Goodhue, D. L. (1995). "Understanding user evaluations of information systems." Management Science **41**(12): 1827-1844.

Goodhue, D. L. and R. L. Thompson (1995). "Task-Technology Fit And Individual-Performance." MIS Quarterly **19**(2): 213-236.

Hammond, J. S., R. L. Keeney, et al. (2006). " The hidden traps in decision making." Harvard Business Review (Decision Making)(Best of HBR)(Reprint) **84**(1): 118-126.

Han, J. and M. Kamber (2001). Data mining : concepts and techniques. San Francisco, Morgan Kaufmann Publishers.

Hastie, R. (2001). "PROBLEMS FOR JUDGMENT AND DECISION MAKING." Annual Review of Psychology **52**(1): 653-683.

Herbert, M. and I. Benbasat (1994). "Adopting information technology in hospitals: the relationship between attitudes/expectations and behavior." Hospital Health Service Administration **39**(3): 369-383.

Hoffman, T. (2001). "Conference attendees: CRM initiatives may miss their marks." Computerworld **35**(9): 7.

Huber, G. P. (1983). "Cognitive Style as a Basis for MIS and DSS Design: Much Ado About Nothing?" Management Science (pre-1986) **29**(5): 567.

Iezzoni, L. I. (1997). "Assessing Quality Using Administrative Data." Ann Intern Med **127**(8_Part_2): 666-674.

Imhoff, C. (2006). Poor-Quality Data...Can your Company Afford the Risk? Terradata.

Jarvenpaa, S. L. (1989). "The Effect of Task Demands and Graphical Format on Information Processing Strategies." Management Science **35**(3): 285-303.

Kaeppel, J. and NetLibrary Inc. (2002). The option trader's guide to probability, volatility, and timing. New York, John Wiley & Sons.

Kaplan, R. M. and D. L. Frosch (2005). "Decision Making in Medicine and Health Care." Annual Review of Clinical Psychology **1**(1): 525-556.

Keathley, D., A. Kearney, et al. (2001). ESCAP II: Analysis of Missing Data Alternatives for the Accuracy and Coverage Evaluation, US Census Bureau.

Klar, Y. (1990). "Linking Structures and Sensitivity to Judgment-Relevant Information in Statistical and Logical Reasoning Tasks." Journal of Personality and Social Psychology **59**(5): 841-858.

Kohli, R. and W. J. Kettinger (2004). "Informating the clan: Controlling physicians' costs and outcomes." Mis Quarterly **28**(3): 363-394.

Kon, H., J. Lee, et al. (1993). A process view of data quality. Working paper TDQM-93-01, MIT TDQM Research Program, E53-320, 50 Memorial Drive, Cambridge, Ma. 02139.

Lee, Y. W. (2003-2004). "Crafting Rules: Context-Reflective Data Quality Problem Solving." Journal of Management Information Systems **20**(3): 93-119.

Little, R. G., Jr and M. L. Gibson (1999). Identification of Factors Affecting the Implemementation of Data Warehousing. 32nd Hawaii International Conference on System Sciences, Hawaii.

Louie, T. A. (2005). "Hindsight bias and outcome-consistent thoughts when observing and making service provider decisions." Organizational Behavior and Human Decision Processes **98**(1): 88-95.

Maibach, E. and D. R. Holtgrave (1995). "Advances in Public Health Communication." Annual Review of Public Health **16**(1): 219-238.

March, S. and A. Hevner (2006). "Integrated Decision Support Systems: A Data Warehousing Perspective." Decision Support Systems.

McDonald, C. J., J. M. Overhage, et al. (1997). "A Framework for Capturing Clinical Data Sets from Computerized Sources." Ann Intern Med **127**(8_Part_2): 675-682.

McKeen, J. D., H. A. Smith, et al. (2005). "Developments in Practice XX - Digital Dashboards: Keep Your Eyes on the Road." Communications of the Association for Information Systems **16**: 1.

Mike, W. C. and D. Elizabeth (2005). "Taking Industry Seriously in Information Systems Research." MIS Quarterly **29**(4): 591.

Moon, B., I. F. V. Lopez, et al. (2003). "Efficient Algorithms for Large-Scale Temporal Aggregation." IEEE Transactions on Knowledge and Data Engineering **15**(3): 744-759.

Nelken, I. (1997). Volatility in the capital markets : state-of-the-art techniques for modeling, managing, and trading volatility. Chicago ; London ; New Delhi, Glenlake/Fitzroy Dearborn.

Payne, J. W., J. R. Bettman, et al. (1992). "Behavioral Decision Research: A Constructive Processing Perspective." Annual Review of Psychology **43**(1): 87-131.

Ramaprasad, A. (1987). "Cognitive Process As a Basis for MIS and DSS Design." Management Science **33**(2): 139-148.

Rebonato, R. (2004). Volatility and correlation : the perfect hedger and the fox. Chichester, West Sussex, England ; Hoboken, NJ, J. Wiley.

Reed, J. and V. R. Payton (1997). "Focus groups: issues of analysis and interpretation." Journal of Advanced Nursing **26**(4): 765-771.

Reneau, J. H. and C. Blanthorne (2001). "Effects of information sequence and irrelevant distractor information when using a computer-based decision aid." Decision Sciences **32**(1): 145.

Rossi, P. E. (1996). Modelling stock market volatility : bridging the gap to continuous time. San Diego, Academic Press.

Sarawagi, S., R. Agrawal, et al. (1998). Discovery-Driven Exploration of OLAP Data Cubes. I. A. R. Center. San Jose, California.

Seiner, R. S. (2000) "Questions Metadata Can Answer." The Data Administration Newsletter (TDAN.com) **Volume**, DOI:

Shaw, M. J., C. Subramaniam, et al. (2001). "Knowledge management and data mining for marketing." Decision Support Systems **31**(1): 127.

Shiller, R. J. (1989). Market volatility. Cambridge, Mass., MIT Press.

Shim, J. P., M. Warkentin, et al. (2002). "Past, present, and future of decision support technology." Decision Support Systems **33**(2): 111.

Shin, B. (2003). "An Exploratory Investigation of System Success Factors in Data Warehousing." Journal of the Association for Information Systems **4**: 141-170.

Starr, P. (1982). The Social Transformation of American Medicine. New York, Basic Books.

Strong, D. M., Y. W. Lee, et al. (1997). "Data Quality in Context." Communications of the ACM **40**(5): 104-110.

Studnicki, J., B. Steverson, et al. (1997). "Comprehensive Assessment for Tracking Community Health (CATCH)." Best Practices and Benchmarking in Healthcare **2**(5): 196-207.

Thomas, H. and A. Datta (2001). "A Conceptual Model and Algebra for On-Line Analytical Processing in Decision Support Databases." INFORMATION SYSTEMS RESEARCH **12**(1): 83-102.

Todd, P. and I. Benbasat (1992). "The Use of Information in Decision Making: An Experimental Investigation of the Impact of Computer-Based Decision Aids." MIS Quarterly **16**(3): 373.

Todd, P. and I. Benbasat (1999). "Evaluating the impact of DSS, cognitive effort, and incentives on strategy selection." Information Systems Research **10**(4): 356.

Van de Van, A. (2005). "Running in Packs to Develop Knowledge-Intensive Technologies." MIS Quarterly **29**(2).

Venable, J. R. (2006). The Role of Theory and Theorizing in Design Science Research. DESRIST, Claremont, California.

Wells, J. D. and T. J. Hess (2002). "Understanding decision-making in data warehousing and related decision support systems: An explanatory study of a customer relationship management application." Information Resources Management Journal **15**(4): 16.

Wixom, B. H. and H. J. Watson (2001). "An empirical investigation of the factors affecting data warehousing success." MIS Quarterly **25**(1): 17.

Yin, R. (2003). Case Study Research - Design and Methods. Thousand Oaks, Sage Publications.

Appendices

Appendix A: Online Analytical Processing and Dimensional Modeling

OLAP (Online Analytical Processing) is used to describe decision support software that allows the user to analyze information that has been summarized into multidimensional views and hierarchies. The data cubes are formed from data in a dimensional model. In a relational database, the dimensional model is often realized as a star schema, with data stored in two types of tables: dimensional tables and fact tables. The term *fact* is used to represent a business measure (Kimball and Ross 2002). Fact data represent the measurable, quantitative, and additive results of a business event. Dimensional data contain descriptive information about those events, and defines the *grain* of the fact table. The fact table contains a composite primary key which is made up of a set of foreign keys to the dimension tables.

The dimension tables contain the textual descriptors of the business (Kimball and Ross 2002). Dimension tables usually have a fairly low number of rows, but contain a large number of attributes. The attributes in a dimension table serve as the query constraints, groupings and report labels, providing the structure for a large number of possible information products. The fact table is joined to a set of dimension tables with a star join schema. Figure 39 is an example of a star schema for daily sales of a product at a certain store .

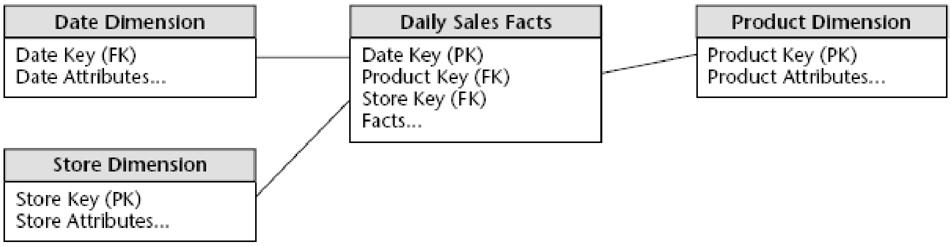
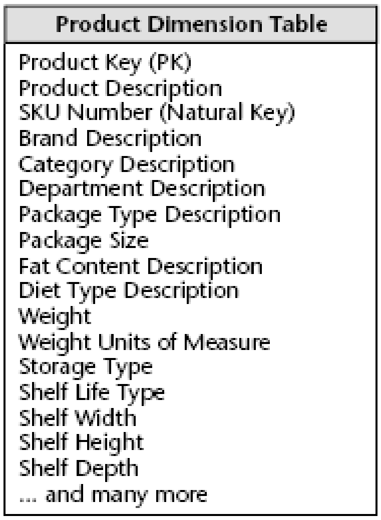


Figure 39 - Fact and Dimension Tables In A Dimensional Model

One of the strengths of dimensional modeling is that the dimensional tables often contain hierarchies, which allow the data to be displayed at different granularities. For example, the sample product dimension table shown in Figure 40 allows products to roll up into brands and then into categories. Similarly, a date hierarchy would allow for sales facts to be aggregated a daily, weekly, monthly, quarterly or yearly levels.



Product Dimension Table
Product Key (PK)
Product Description
SKU Number (Natural Key)
Brand Description
Category Description
Department Description
Package Type Description
Package Size
Fat Content Description
Diet Type Description
Weight
Weight Units of Measure
Storage Type
Shelf Life Type
Shelf Width
Shelf Height
Shelf Depth
... and many more

Figure 40 – Sample Dimension Table (Kimball and Ross 2002)

To provide a more concrete example of a data cube created from a healthcare information supply chain, we consider a simple health-care cube created with data extracted from the Florida Cancer Registry. The Florida Cancer Data System (FCDS) is Florida's statewide, population-based cancer registry and has been collecting incidence data since 1981 when it was contracted by the State of Florida Department of Health in 1978 to design and implement the registry. The University Of Miami Miller School Of Medicine has been maintaining FCDS (fcds.med.miami.edu) since that time.

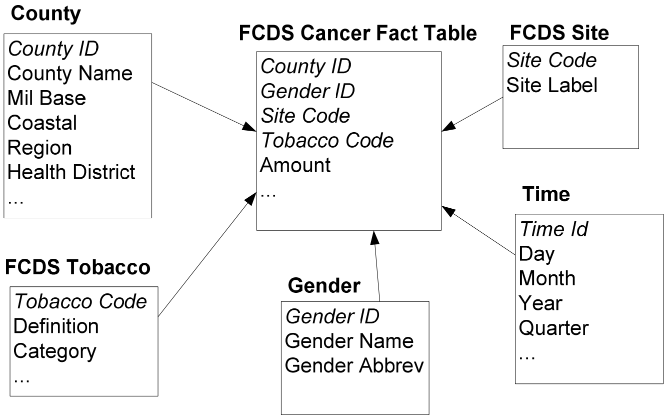


Figure 41 – Star Schema for Smoking/ Cancer Data

The dimensions of interest for our example include COUNTY, SITE, TOBACCO, GENDER, and TIME. Figure 41 shows the star schema for this example. Notice there are several dimensions with hierarchies: the COUNTY dimension contains county-region hierarchy, and the TIME dimension contains a day-month-quarter-year hierarchy. This allows for different levels of aggregation granularity. For example, monthly data would be a lower level of aggregation along the TIME dimension, then daily, or region would be a higher level of aggregation along the COUNTY dimension then county.

Suppose we are to count occurrences corresponding to GENDER, TIME, and whether the patient smoked, TOBACCO (the query is shown in

Figure 42). For the sake of simplicity we hold the type of cancer and the county static in this cube illustration, since it this would be difficult to illustrate more then three dimensions.

```
SELECT      tobacco_category, cat_year, county_name, gender_name, site_label,
            sum(cat_count)
FROM        counties, fcds_tobacco, fcds_genders , fcds_sites
WHERE       (tobacco_code = fcds_tobacco ) AND
            (county_id = cat_county ) AND
            (gender_code = gender ) AND
            (code = fcds_site_grp ) ) AND
            cat_year >= 2000 AND
            cat_year <= 2002 )
GROUP BY    tobacco_category, cat_year, county_name, gender_name, site_label;
```

Figure 42 – Sample Cube Query

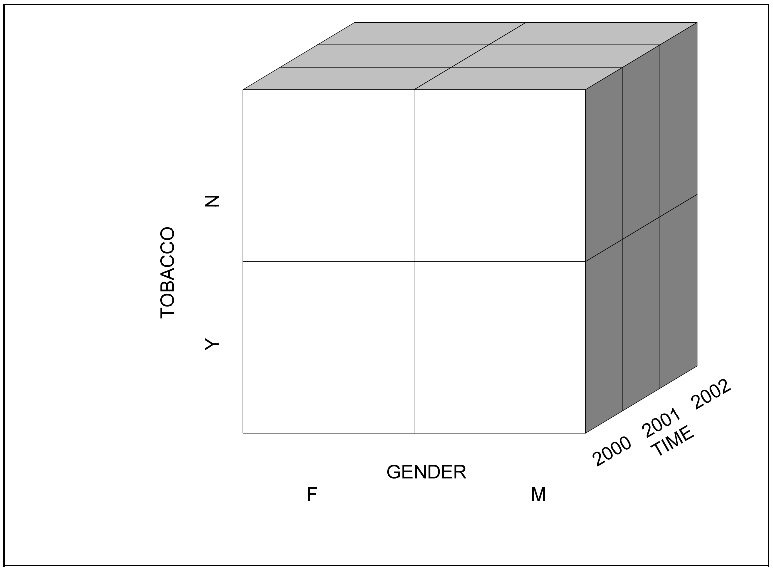


Figure 43 illustrates the resulting data cube, with three of the dimensions: TOBACCO, GENDER and TIME. Each of the cells corresponds for the count for the corresponding TOBACCO, GENDER and TIME. The lowest right hand cell contains the count of male patients that were tobacco users in the year 2000 (in a certain county, for a certain cancer). Figure 44 illustrates how this data cube would be presented using an OLAP tool (in this case Oracle Discoverer).

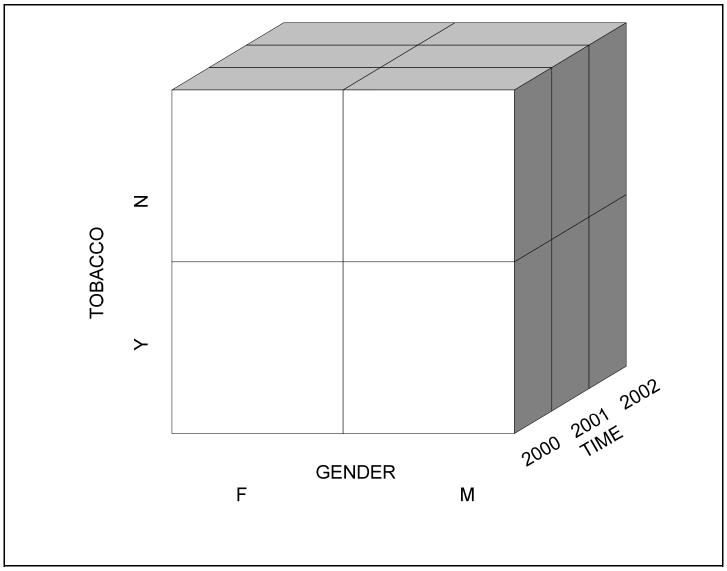
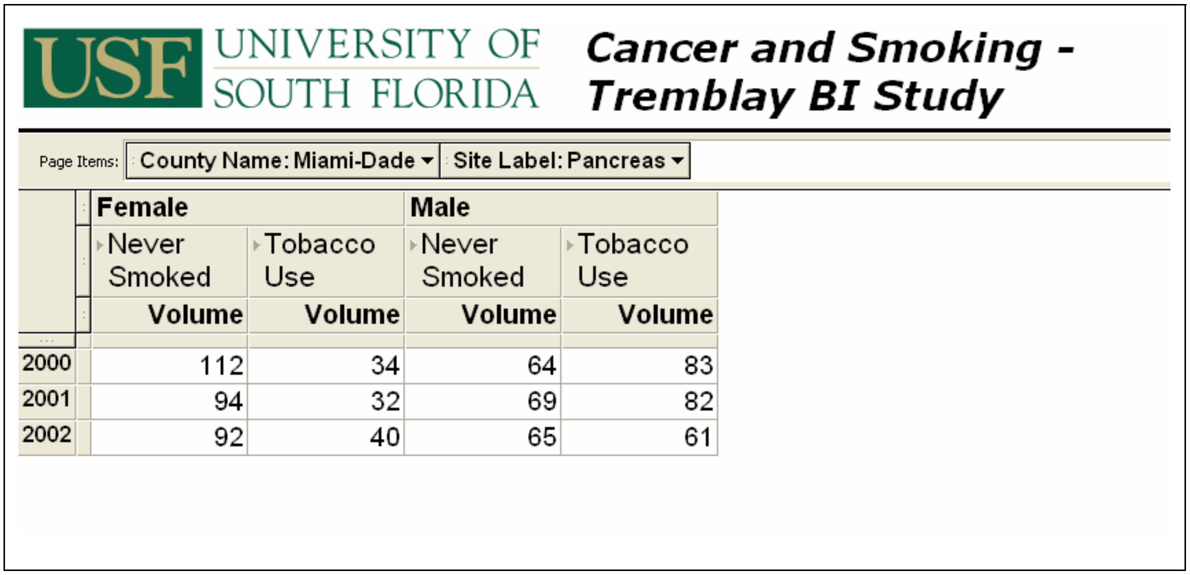


Figure 43 – Data Cube for Cancer/ Smoking Information



Cancer and Smoking - Tremblay BI Study

Page Items: County Name: Miami-Dade ; Site Label: Pancreas

	Female		Male	
	Never Smoked	Tobacco Use	Never Smoked	Tobacco Use
	Volume	Volume	Volume	Volume
2000	112	34	64	83
2001	94	32	69	82
2002	92	40	65	61

Figure 44 – OLAP Interface for Smoking/Cancer Data Cube

Appendix B: Focus Group Scripts

Arrange furniture for focus group

Set up tape recorder & test

Set out pencils, questionnaire

Open all examples:

- PPT presentation
- Cancer and Smoking
- New Chemo Treatment
- Volatility in Cancer Volumes Trend

Greet and chat with people as they come in.

Encourage them to fill out questionnaire and consent forms while they wait.

Thanks for filling out the questionnaires and forms, again. Please hand them all in now.

We will be showing you a typical OLAP interface. OLAP interfaces (which are usually embedded in Business Intelligence tools) are increasingly being used for decision support. They display the results in a tabular form, and allow flexibility to reconfigure the results depending on the task. Often there are some data quality problems that are hidden from the decision maker. In fact, most decision makers assume that the information is 100% accurate. You are being asked to participate because:

1. We want to understand how including information about data quality in a business intelligence tool will affect your decision-making process
2. To get your opinion on the way it is presented
3. To get your suggestions on how you would improve it.

Our goal is to eventually automate these data quality calculations so they match the information shown to you by the OLAP tool. Every time you reconfigure, these calculations will match the information on the screen. Keep in mind this is not the final tool. We are at a “prototype” stage, and we seek to understand how to present this information in a useful and understandable way.

The data we are going to look at is real data from the cancer registry. This data comes from several data sources (all Florida hospitals). I am going to take several cases and demonstrate three different ways to tell you about the quality of the data. After each case we will discuss for each case what decision you would make for each scenario. This should take about 1 ½ hours, after which we will go to dinner (have lunch**).** Imagine yourself in a position that helps define public policy. For example, making decisions about where in the state you may open a cancer center, or whether a certain ethnicity or race is underserved…

These are the three cases we are considering:

1. The amount of data that we could not place because of missing information
2. The variability in the data, for example if we observing a trend across a time period, how much does it fluctuate?
3. When comparing data, do we know enough about the data to make valid comparisons?

The tape recorder is here to allow us to tape the discussion so that we can listen and study the conversation later – "rigorous qualitative analysis". Everything you say is strictly confidential – your real names will not be used in any report. **Please try to speak one at a time** so that we can all hear what is being said and so that we'll be able to follow the conversation on the tape.

Let's begin with introductions. Please tell us your name and a brief description on you current job. I'll start then we can go around the room.

I'm Monica Tremblay, a doctoral candidate in Information Systems and Decision Sciences and this work is part of my dissertation. I am interested in business intelligence and data analytics. Prior to pursuing my PhD I worked in industry as a systems analyst.

Participant introductions

Vignette Script

The participants are shown several vignettes

- *One will highlight how information about missing data is illustrated*
- *One will illustrate information about variability in the data*
- *One will illustrate how information on sample size will be shown*

For each vignette:

1. Ask participants to discuss how this extra information on data quality and on sample size would impact their decision
    - *Allow conversation to flow – the goal of focus groups is to stimulate conversation from comments of other participants*
2. Ask participants to make a final decision, allow the use of scrap paper – have them write down their choice(s) before discussing them as a group

Start with Powerpoint, Start with Anna Nicole example…

Case 1 – Cancer and Smoking – Lung – unallocated data

The participants are shown the data – a part of the chart shows a percentage for which we do not have information on smoking status. This is not pointed out – the goal is to see if anyone comments on this. Show information for different years/counties.

Imagine if you were asked to make some sort of blanket statement – about the correlation between cancer and each particular cancer. I realize there are probably other causes, and it is unrealistic tot think you can do this with just this data, but acts as if you had no other data available.

1. Start with Lung Cancer Numbers:
   - PPT slide : "Cigarette smoking causes 87 percent of **lung** cancer deaths. These are the numbers for the state of Florida. We can navigate and see individual counties and break it down by year."
   - Demonstrate Lung Cancer Numbers, use this explain OLAP tools
2. Move on to other cancers:
   - "Smoking is also responsible for most cancers of the larynx, oral cavity and pharynx, esophagus, and bladder. In addition, it is a cause of kidney, pancreatic, cervical, and stomach cancers, as well as acute myeloid leukemia.
   - Work with Cancer of the Pancreas

Allow discussion

1. Does smoking cause this kind of cancer?
2. How confident would you be in saying that there is a relationship?

- Compare Females and Males

Show male/female example, see if anyone makes comment on women less likely to report smoking (we don't know why data is unknown, should not pass judgment here)- use pancreatic cancer as example

Case 2 – Hispanic Disparity with Cancer Treatment

The participants are shown the data – a part of the chart shows a percentage for which we do not have information on Hispanic/not Hispanic or whether they received treatment.

Start with ppt slide on Hispanic and Cancer Risks and “Ignotus” cancer

We are looking at a particular cancer and I should point out that – that this is a fake cancer, okay? The premise here is that when Hispanics are diagnosed with a certain cancer, this cancer which I called Ignotus (ignotus : unknown, obscure, ignorant, ignoble), they’re less likely to receive chemotherapy than non Hispanics.

- Explain all the types of unallocated data.

- Start with ppt slide on Hispanic and Cancer Risks and “Ignotus” cancer
- Show all counties together, then break down by counties (Miami Dade is a good example)

1. Is there a disparity based on this data (alone)?
   - Allow discussion
   - What might be some of the approaches you might take to consider this data (allow them to talk! Even if they do not have suggestions!)
2. If not brought up, illustrate how this is more complex since there are unallocated data amounts in several fields
3. Discuss approaches with ppt (3 slides) -
4. Describe worst-case scenario
   - How confident would you be using this data
   - How would you explain your answer
5. Show distributed approach (use ppt slide)
   - How confident would you be using this data
   - How would you explain your answer
6. How do you feel about the three scenario approach (show without nulls, worse, distributed)?
   - Does it change your opinion?
   - How about your confidence?
   - How would you change it?

Case 3 – Cancer Volumes and Data Volatility

Use EXCEL Volatility in Cancer Volumes. Start with First Sheet , 1 slide in ppt

- *Examining Breast Cancer Trends*
- *Claim is that your neighboring counties are better then you at early detection/prevention – thus are seeing decreasing volumes*

You are Miami Dade: Compare to Collier, Broward, and Palm Beach - have them make a decision

Note that Collier County exhibits a much steeper decrease – are they doing better?

Don't bring up volatility – see if they notice it – if they don't after a pause, bring it up (keep comparing the three counties)

Introducing Volatility…

How reliable are these trends – how much variability is there in this data?

- *Explanation of volatility in stock market and how we apply it to data.*

Show PPT slide of stock market volatility

"As you can see from the chart, volatility soared during the Crash of 1987. It jumped when Iraq invaded Kuwait a few years later. It jumped during the Asian crisis in late 1997, and after the crash of the LTCM hedge fund in 1988. It jumped up after

September 11th, 2001. You get the idea - volatility in the stock market soars after major uncertainty appears"

- For each of the counties show the volatility:
    - ➢ With metric
    - ➢ With Chart
    - ➢ With Benchmark Numbers
    - ➢ With Benchmark Chart
    - ➢ With Indicator

What is the decision now?

- Ask which ones look stable judging from the metric
- Judging from these what volatility metrics would concern you
- Time permitting play with some smaller counties or infrequent diseases
- Which do you prefer?
- How would you approach problem solving?
- Benchmarking data – One approach suggested by another group
- Show Breast Broward vs Larynx – Broward
- Ask for comments

Case 4 - Average Tumor Size and Age

- Use Oracle Discoverer – Tumor Size – sample size
- Start with Tumor 2 sheet (no highlighting)

This is the average tumor size in mm for counties in Central Florida, for several years. This measure has been shown to be a good predictor of survival for certain cancers, including: breast, lung and endocrine.

1. Let's look at breast. How does Hillsborough compare to other counties?
    - Any particularly bad year?

- Point out 1996 and drill – compare monthly numbers to other counties…DO NOT BRING UP SAMPLE SIZE – but see if any one does…

2. Lets look at Lung – 40 mm or bigger usually is bad news according to a certain study
    - Again, any bad years?
    - Drill 1996, compare May Hillsborough and May Osceola – see if anyone comments..
3. Finally lets look at Endocrine

    - Compare Several Numbers

If no one has yet brought up sample size explain insensitivity to sample size with the ppt slide

Run through the same exercise with Tumor Size – aid

- See if people understand why data is highlighted aid, explain highlighting for any sample size smaller then 30 (Central Limit theorem)

Show Tumor-Size – aid 2 -> Highlighting – Red for more severe cases

General Questions on Sample Size:

1. Were you aware that you have to consider sample size when comparing averages?
2. If so, do you often consider sample size?
3. Does highlighting help?
4. Could you suggest other ways that you might show this?
5. What about the cases where there is a lot of highlighting, for example a rare cancer?
6. If I had just shown you the highlighting would you have understood why?

Wrap up - Have them fill this out first and then discuss …

1. Do you think receiving data quality information would be beneficial to you?
2. Do you find this information useful?
3. Would data quality information improve the way you do your work?
4. Given data quality information, how many of you would utilize it?
5. Do you think you could figure out how to use the tool?
6. Do you understand what the data quality metrics mean?
7. Do you think data quality information may complicate your work?
8. Do you think the tool may make you waste time on mechanical operations?
9. Take too long to learn or understand?
10. Do you think the decisions you would make would be more/less effective if you had data quality information?

Appendix B: (Continued)

Hand out final questionnaire – I realize these are the same questions, but I am interested in individual opinions

Thank everybody for participating – invite everyone to join for lunch/dinner.

Appendix C: Telephone Screening Questionnaire

(This is a guide – not to be read verbatim, the goal is to sound conversational)

Name of Person ______________________________

Phone Number ______________________________

Time Called ______________________________

Better Time to Call ___________________________

Hi, this is <name> and I am with the Information Systems and Decision Sciences Department, College of Business at USF. You may remember me as your instructor in the data warehousing/data mining class **or** from <past projects>**or** we got your name from <name> who said you may be interested in participating. We want to talk to people that do a lot of data analysis as part of their job. You are being asked to participate because we want to observe how information about data quality in a business intelligence tool would affect your decision-making process and also to get your opinion on how it is presented and suggestions on how to improve it. We plan to get together

Date, day

Time (1 ½ hrs to 2 hrs)

Place

We will meet in the ISDS conference room. After the meeting, if you wish we will go to dinner, our treat.

No___ Ok. Thanks for your time

Yes___ Great! I will send you an e-mail to confirm as well as a reminder e-mail

e-mail___________

Can you suggest someone else that may be interested?

Name____________

Phone_________________

E-mail _________________

Great! Thanks so much.

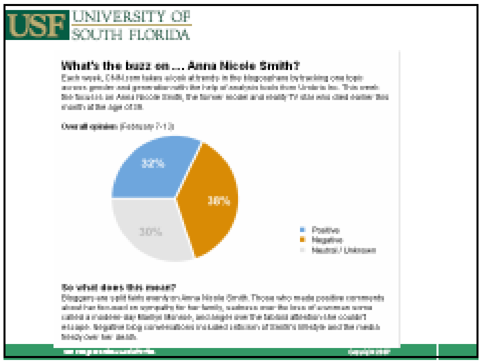
Appendix D: PowerPoint Presentation

UNIVERSITY OF SOUTH FLORIDA

Data Quality in Healthcare

Monica Chiarini Tremblay

Monica Chiarini Tremblay



UNIVERSITY OF SOUTH FLORIDA

What's the buzz on .... Anna Nicole Smith?

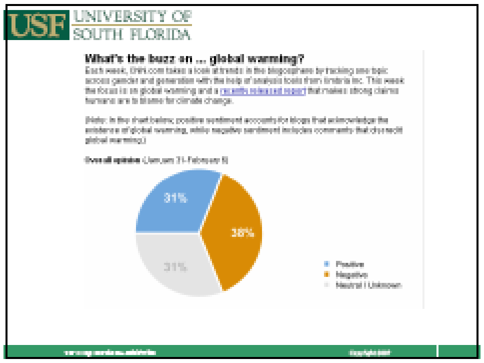
Each week, CNN.com takes a look at trends in the blogosphere by tracking one topic across gender and generation with the help of analysis tools from Umbria Inc. This week the focus is on Anna Nicole Smith, the former model and reality TV star who died earlier this month at the age of 39.

Overall opinion (February 7-13)

So what does this mean?

Bloggers are split fairly evenly on Anna Nicole Smith. Those who made positive comments about her focused on sympathy for her family, sadness over the loss of a woman some called a modern-day Marilyn Monroe, and anger over the tabloid attention she couldn't escape. Negative blog conversations included criticism of Smith's lifestyle and the media frenzy over her death.

Copyright 2007



UNIVERSITY OF SOUTH FLORIDA

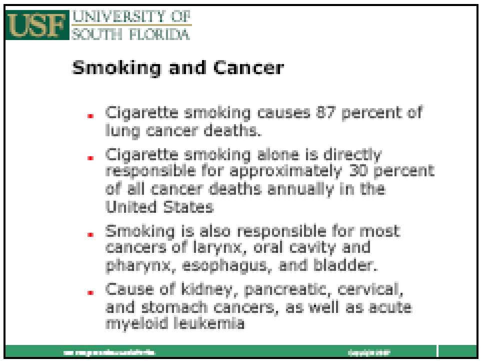
What's the buzz on ... global warming?

Each week, CNN.com takes a look at trends in the blogosphere by tracking one topic across gender and generation with the help of analysis tools from Umbria Inc. This week the focus is on global warming and a recently released report that makes strong claims humans are to blame for climate change.

(Note: In the chart below, positive sentiment accounts for blogs that acknowledge the existence of global warming, while negative sentiment includes comments that discredit global warming.)

Overall opinion (January 31-February 6)

Copyright 2007

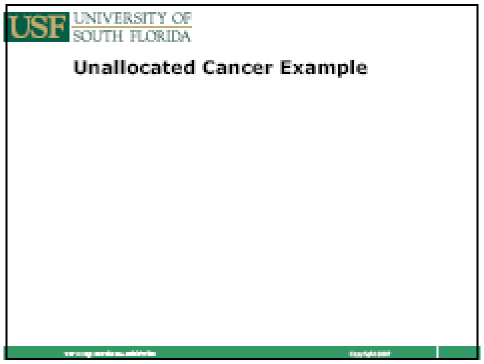


UNIVERSITY OF SOUTH FLORIDA

Smoking and Cancer

- Cigarette smoking causes 87 percent of lung cancer deaths.
- Cigarette smoking alone is directly responsible for approximately 30 percent of all cancer deaths annually in the United States
- Smoking is also responsible for most cancers of larynx, oral cavity and pharynx, esophagus, and bladder.
- Cause of kidney, pancreatic, cervical, and stomach cancers, as well as acute myeloid leukemia

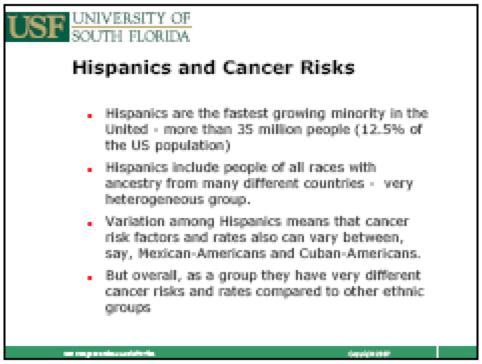
Copyright 2007



UNIVERSITY OF SOUTH FLORIDA

Unallocated Cancer Example

Copyright 2007



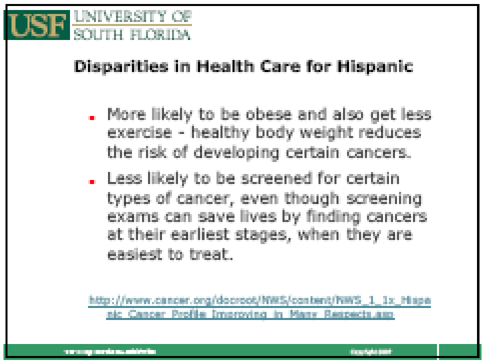
UNIVERSITY OF SOUTH FLORIDA

Hispanics and Cancer Risks

- Hispanics are the fastest growing minority in the United - more than 35 million people (12.5% of the US population)
- Hispanics include people of all races with ancestry from many different countries - very heterogeneous group.
- Variation among Hispanics means that cancer risk factors and rates also can vary between, say, Mexican-Americans and Cuban-Americans.
- But overall, as a group they have very different cancer risks and rates compared to other ethnic groups

Copyright 2007

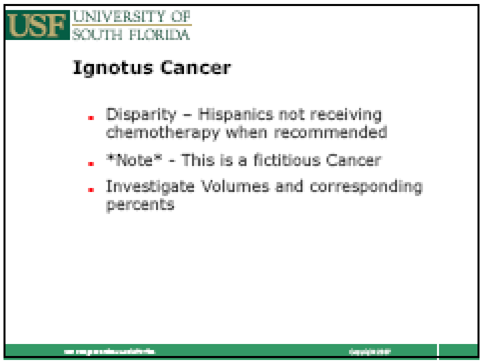
Appendix D: (Continued)



Disparities in Health Care for Hispanic

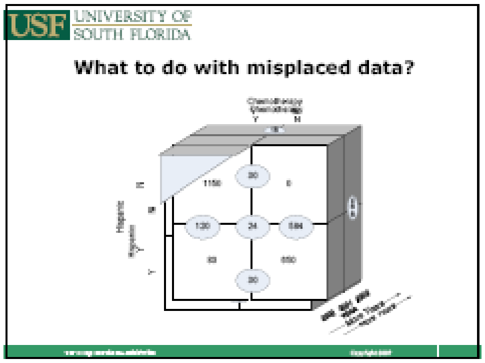
- More likely to be obese and also get less exercise - healthy body weight reduces the risk of developing certain cancers.
- Less likely to be screened for certain types of cancer, even though screening exams can save lives by finding cancers at their earliest stages, when they are easiest to treat.

http://www.cancer.org/docroot/NWS/content/NWS_1_1x_Hispanic_Cancer_Profile_Improving_in_Many_Respects.asp

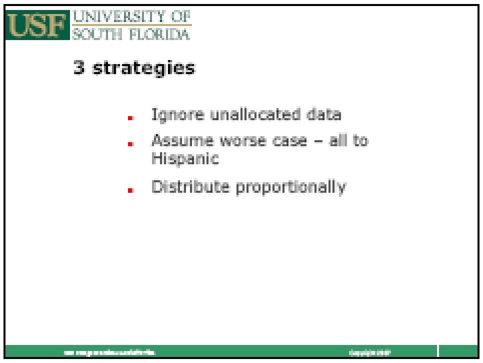


Ignotus Cancer

- Disparity – Hispanics not receiving chemotherapy when recommended
- *Note* - This is a fictitious Cancer
- Investigate Volumes and corresponding percents

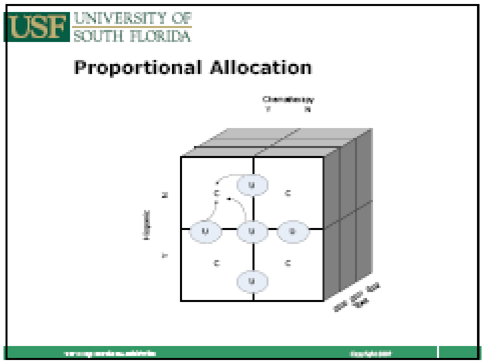


What to do with misplaced data?

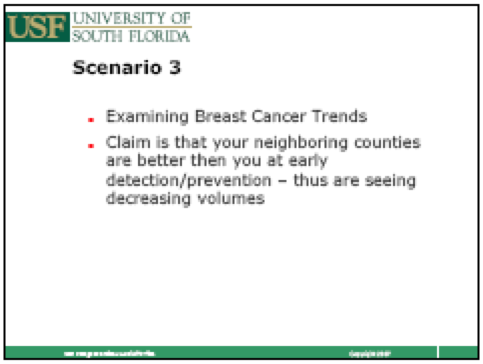


3 strategies

- Ignore unallocated data
- Assume worse case – all to Hispanic
- Distribute proportionally



Proportional Allocation



Scenario 3

- Examining Breast Cancer Trends
- Claim is that your neighboring counties are better then you at early detection/prevention – thus are seeing decreasing volumes

About the Author

Monica Chiarini Tremblay received her M.S. in Information Systems degree from the University of South Florida and a B.S. degree in Industrial and Systems Engineering from the University of Florida. Prior to her academic career, Monica worked as a senior systems analyst at a multi-national company where she was involved in the design, implementation and support several ERP, material handling, and accounting information systems. Her research interests focus on data analytics and business intelligence, data and text mining, data quality, and data warehousing. Her research has been published in the Journal of Computer Information Systems and Decision Support Systems. Monica was selected to participate in the International Conference on Information Systems Doctoral Consortium and received the College of Business' Doctoral Research Award and the University Graduate Fellowship.

Mrs. Tremblay will be joining the faculty at Florida International University, in Miami, Florida as an Assistant Professor.