The Catch data warehouse: support for community health care decision-making

Donald J. Berndt\textsuperscript{a}, Alan R. Hevner\textsuperscript{a,}\textsuperscript{*}, James Studnicki\textsuperscript{b}

\textsuperscript{a}Information Systems and Decision Sciences Department, College of Business Administration, 4202 Fowler Ave., CIS1040, University of South Florida, Tampa, FL 33620, USA

\textsuperscript{b}College of Public Health, University of South Florida, Tampa, FL 33620, USA

Accepted 1 April 2002

Abstract

The measurement and assessment of health status in communities throughout the world is a massive information technology challenge. Comprehensive Assessment for Tracking Community Health (CATCH) provides systematic methods for community-level assessment that is invaluable for resource allocation and health care policy formulation. CATCH is based on health status indicators from multiple data sources, using an innovative comparative framework and weighted evaluation process to produce a rank-ordered list of critical community health care challenges. The community-level focus is intended to empower local decision-makers by providing a clear methodology for organizing and interpreting relevant health care data. Extensive field experience with the CATCH methods, in combination with expertise in data warehousing technology, has led to an innovative application of information technology in the health care arena. The data warehouse allows a core set of reports to be produced at a reasonable cost for community use. In addition, online analytic processing (OLAP) functionality can be used to gain a deeper understanding of specific health care issues. The data warehouse in conjunction with Web-enabled dissemination methods allows the information to be presented in a variety of formats and to be distributed more widely in the decision-making community. In this paper, we focus on the technical challenges of designing and implementing an effective data warehouse for health care information. Illustrations of actual data designs and reporting formats from the CATCH data warehouse are used throughout the discussion. Ongoing research directions in health care data warehousing and community health care decision-making conclude the paper.

\textsuperscript{2}2002 Elsevier Science B.V. All rights reserved.

Keywords: Health care information systems; Data warehousing; Data staging; Online analytic processing (OLAP); Decision support systems; Community decision-making; Data quality

1. Introduction

The United States spends over a trillion dollars annually on health expenditures. Both as a percentage of national productivity and per capita, health care spending by the United States exceeds that of any other nation in the world. However, this tremendous expenditure has not secured the U.S. a rank among the ‘healthiest’ nations. In fact, for many health indicators, such as infant mortality and measles immunizations, the U.S. ranks below some countries characterized as underdeveloped [23,29]. Prolonged public debates on health care policy in the United States have focused on

\textsuperscript{*}Corresponding author.

E-mail addresses: dberndt@coba.usf.edu (D.J. Berndt), ahevner@coba.usf.edu (A.R. Hevner), jstudnic@hsc.usf.edu (J. Studnicki).
insurance coverage and medical care financing programs without any serious examination of the true health status of the nation.

The need to assess the health status of U.S. communities in a comprehensive and systematic manner has been widely recognized within the health professions. The Institute of Medicine (IOM) of the National Academy of Sciences has acknowledged the importance of a population-based perspective in two influential reports, emphasizing the need for a regular and systematic collection, assemblage, and analysis of the health status of our nation’s communities [16,18]. A community health profile is comprised of socio-demographic characteristics, health status and quality of life indicators, health risk factors, and health resource measures. The intent of such a comprehensive health profile is to assist a community in developing, refining, and monitoring a long-term strategic view of its overall health status. Although there are many sources of health data, there are no standard data definitions, formats, or reports across the health care industry. Thus, health care data are widely used (and misused) in an ad-hoc manner to justify managerial objectives of health institutions and agencies, a maze of mandated categorical funding, and a variety of political agendas. Sound information and accepted analytic techniques are even more important as funding is consolidated in block grants and local community decision-making is emphasized.

As part of the ongoing clarification of the public health role at the community level and the transition from a disease to a health focus and from a treatment to a prevention strategy, there has been recognition that partnerships and collaboration are necessary to support effective action [17,21]. Health organizations, public sector agencies, medical care providers, businesses, the religious community, educational institutions, and other community organizations are interdependent components of a multi-sectoral community health environment. The overall community must be empowered to make the necessary, and sometimes difficult, resource allocation choices to improve health through information, education, behavior change, and social support [7]. Such collaborative action at the community level must be informed by unbiased data describing the community’s health status, needs, and resources. The ability is also needed to track progress over time to meet the community’s health care goals [24].

The gap between current practice in community health care spending and the above goals of collaborative community health care decision-making is vast. The availability and quality of health indicators are problematic. There is little empirical evidence on the use, sharing, or integration of health data into decision-making to provide guidance to community health organizations. While most of the literature on collaborative leadership and community engagement emphasizes the process [4,5], little attention has been focused on the effect of the availability of a common set of data, such as the community health profile, on the quality and inclusiveness of decision-making. There is also scant information about the use of data and information technology to support and monitor the process.

The purpose of this paper is to present an overview of the Comprehensive Assessment for Tracking Community Health (CATCH) methods [25] and then focus on the construction of a comprehensive health care data warehouse that provides automated support for CATCH. The combination of extensive field experience with CATCH and the application of current data warehousing technology make this an innovative interdisciplinary research effort. Section 2 briefly presents the CATCH methods and our motivation for building a data warehouse. In Section 3, we present a detailed discussion of the technical challenges in designing and implementing the data warehouse. Twin star data staging, an effective approach for ensuring quality as data are entered into the warehouse, is highlighted in Section 4. Section 5 discusses the use of the data warehouse for advanced health care applications. The paper concludes with future research directions on data warehousing technical challenges and the use of health profiles to support improved community health care decision-making.

2. The CATCH methods of community assessment

The University of South Florida’s Center for Health Outcomes Research (CHOR) developed CATCH to provide comprehensive and objective health status data for community health planning
purposes. CATCH collects, organizes, analyzes, prioritizes, and reports data on over 250 health and social indicators on a local community basis. The CATCH methods have been tested, refined, and validated in the field over the past 10 years. Reports have been prepared for more than 20 U.S. counties both within and outside of Florida.

The CATCH process can be briefly described as shown in Fig. 1. Community health indicator data are gathered from a variety of sources. Secondary data sources include health care data reported by hospitals, local, state, and federal health agencies, and national health care groups. Primary data sources would involve data gathered from door-to-door or mail-in surveys. All health care data are translated into common formats and integrated with other data warehouse components to support the production of health care report cards.

Over 250 indicators are used within CATCH and are organized into 10 indicator categories. These indicators and categories represent a wide spectrum of health care issues and have evolved through both research and field practice. Table 1 lists the 10 indicator categories and presents a few representative indicators to lend a sense of perspective to the level of detail provided in CATCH reports. These indicators are collected from a variety of sources.

Each indicator value is compared against the state average, an average from a peer group of counties, and other interesting values (e.g., a national goal for that indicator) [26]. The results of these comparisons are organized into a multi-dimensional matrix based on favorable or unfavorable comparisons against each comparison dimension. Fig. 1 shows a 2-by-2 comparison matrix based on state averages and peer

Table 1
Ten indicator groups with representative indicators

<table>
<thead>
<tr>
<th>Demographic Characteristics</th>
<th>Health Status: Morbidity and Mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Population</td>
<td>Breast Cancer</td>
</tr>
<tr>
<td>Racial Composition</td>
<td>Cardiovascular Disease</td>
</tr>
<tr>
<td>Net Migration</td>
<td>Stroke</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Socioeconomic Characteristics</td>
<td>Sentinel Events</td>
</tr>
<tr>
<td>Employment</td>
<td>Rubella</td>
</tr>
<tr>
<td>High School Dropouts</td>
<td>Measles</td>
</tr>
<tr>
<td>Per Capita Income</td>
<td>Late Stage Cancer</td>
</tr>
<tr>
<td></td>
<td>Avoidable Hospitalizations</td>
</tr>
<tr>
<td>Maternal and Child Health</td>
<td>Health Resource Availability</td>
</tr>
<tr>
<td>Infant Mortality</td>
<td>Licensed Hospital Beds</td>
</tr>
<tr>
<td>Low Birthweight</td>
<td>Licensed Medical Doctors</td>
</tr>
<tr>
<td>Birth Defects Mortality</td>
<td>Licensed Registered Nurses</td>
</tr>
<tr>
<td>Social and Mental Health</td>
<td>Infectious Disease</td>
</tr>
<tr>
<td>Domestic Violence</td>
<td>Syphilis</td>
</tr>
<tr>
<td>Homicide Rate</td>
<td>AIDS</td>
</tr>
<tr>
<td>Psychiatric Admissions</td>
<td>Hepatitis</td>
</tr>
<tr>
<td>Physical Environmental Health</td>
<td>Behavioral Risk Factors</td>
</tr>
<tr>
<td>Foodborne Outbreaks</td>
<td>Smoking</td>
</tr>
<tr>
<td>Contaminated Wells</td>
<td>Obesity</td>
</tr>
<tr>
<td>Lead Poisoning</td>
<td>Mammograms</td>
</tr>
</tbody>
</table>

![Fig. 1. The CATCH process.](image-url)
averages. Community indicators that demonstrate unfavorable comparisons on all dimensions are highlighted as community health challenges. After this simple comparison, the health care challenges are prioritized using a set of five filters.

- **Number Affected**—number of persons in the community affected by the indicator.
- **Economic Impact**—an estimate of the direct cost per case for individuals affected by the indicator.
- **Availability of Efficacious Intervention**—an estimate of the relative degree to which treatment or prevention is likely to be effective.
- **Magnitude of Difference**—the degree to which the community indicator is worse than the dimensional comparisons.
- **Trend Analysis**—for a 5-year period is the trend favorable or unfavorable and what is the magnitude of change in the trend direction?

The community stakeholders are given an opportunity to weight the importance of each of the above factors. The final product of the CATCH methodology is a comprehensive, prioritized listing of community health care challenges. A more detailed description of the CATCH methodology with a complete listing of health care indicators can be found in Ref. [25].

### 2.1. Limitations

While the value of CATCH is incontrovertible, the ultimate deployment of CATCH throughout Florida and the nation has been constrained by several serious limitations:

- The handcrafted process is labor-intensive and slow. Hundreds of individual sources of data must be identified and contacted. Data are often provided in hard copy formats and must be manually checked, validated, and entered into spreadsheets. With manual methods, it takes 3–4 months to complete a CATCH report for a single county.
- Longitudinal trend analyses over many years are cost prohibitive for most communities. Since each application is expensive and time-consuming, the capability to fund and produce annual assessments in a single community is limited.
- Most public health funding comes from state and federal governments. A statewide CATCH assessment would help to prioritize funding and serve to enable effective program evaluation based on quantifiable outcome assessment. Since nearly all data indicators available in Florida are available in most other states, there is reason to be confident that CATCH will be expanded nationally and even internationally.

- With the massive amount of health care data involved, many interesting relationships and correlations between health status indicators can be found and investigated. In the manual system, such discovery was not feasible. A comprehensive and integrated data warehouse provides the infrastructure for such data mining efforts.

### 2.2. CATCH data warehouse challenges

The application of data warehousing technologies for the automated support of CATCH holds tremendous promise. The remainder of this paper describes our work to construct an effective and efficient data warehouse solution, enabling both cost-effective report generation and ad-hoc analyses of critical health care issues. The construction of a data warehouse for public health care data poses major challenges beyond those required for the construction of a commercial data warehouse (e.g., retail sales). Such challenges include the following.

- Data come from a very diverse set of sources. Health care data are published in a wide variety of formats with differing semantics. There are currently few standards in the health care field for such data. The data integration task to build the data warehouse requires significant effort.
- CATCH reports are disseminated to a diverse and geographically distributed set of stakeholders.
- The data warehouse is required to support the activities of public policy formulation. The socio-political issues of health care planning impact design features such as security, availability, data quality, and performance.

### 3. The CATCH data warehouse

The goals of the CATCH data warehouse include the support and enhancement of the CATCH methods, the provision of cost-effective and thorough reports to communities, and the creation of a rich environment
for more detailed research into critical health care issues. In addition, a focus on data quality makes the data warehouse an especially valuable asset over time as a rich and trustworthy historical repository is built. Lastly, the data warehouse lends itself to a variety of dissemination strategies based on hardcopy reports, interactive access, and Web-enabled information delivery. The different access technologies allow a diverse group of community planners and stakeholders to investigate important health care issues using comparable data. All of these characteristics make the CATCH data warehouse a unique application of technology in the field of public health. In fact, the implementation of this type of data warehouse and its use in monitoring, as well as improving health status, will become a primary role of public health agencies in the future.

The CATCH data warehouse includes a variety of components arranged in three broad categories: reporting tables for direct support of the CATCH methods, aggregated dimensional structures, and fine-grained or transaction-oriented dimensional structures. In the sections that follow, examples of these data warehouse components are presented. All of the components draw on the dimensional model or star schema, some components with more than a dozen dimensions and some with a few simple dimensions.

3.1. The dimensional model

Important missions of a data warehouse include the support of decision-making activities and the creation of an infrastructure for ad-hoc exploration of very large collections of data. Decision-makers should be able to pursue many of their investigations using browsing tools, without relying on database programmers to construct queries. The emphasis on end-user data access places a premium on an understandable database design that provides an intuitive basis for navigating through the data. The star schema or dimensional model has been recognized as an effective structure for organizing many data warehouse components [12,15,19]. The star schema is characterized by a center fact table, which usually contains numeric information that can be used in summary reports. Radiating from the fact table are dimension tables that provide a rich query environment. This structure provides a logical data cube, with dimensions such as time and location identifying a set of numeric measurements within the cube. Fig. 2 contains a fragment from the hospital discharge transaction-oriented star schema discussed in this paper.

3.1.1. Fact tables

The most appropriate facts are additive numeric data items that can be summed, averaged, or combined in other ways across the dimensions to form summary statistics. The only way to compress the millions of data points and produce a reasonably sized answer set is to present some mathematical summarization. No human will want thousands, let alone millions, of items in answer to their queries. As Kimball [19] points out, “the best and most useful facts are numeric, continuously valued, and additive.” The CATCH data warehouse includes facts such as counts of hundreds of different health events, population-based rates, age-adjusted rates, and even fine-grained financial data in the case of the hospital discharge data depicted in Fig. 2. For example, using the hospital discharge star it is possible to focus on a single hospital (using the hospital dimension), select a single disease (using the ICD DIAGNOSIS dimension), and investigate how the length of stay has varied over a specified time period. Using the hierarchical nature of the dimensions, it is also possible to ‘roll-up’ to compare types of hospitals, disease categories, or even patient age bands. While the dimensional structure is simple and readily understandable, it supports a large and very useful universe of queries.

3.1.2. Dimension tables

The dimensions define the query environment, the richer the set of dimensions the more ways the data can be accessed via queries. Two of the important characteristics of dimensions are the richness of the attributes that describe the dimension and the hierarchical nature of the dimension. For example, the COUNTY dimension in the CATCH data warehouse includes attributes that describe whether a county is coastal, wealthy, urban, dense, large in area, or includes a military base. Therefore, the counties can be organized by any value in this attribute set. Some of the attributes lend themselves to hierarchical organization. In the case of COUNTY, there is natural geographic hierarchy that includes groups of counties that form regions within the state and the state itself. The county is also composed of finer geographic units
such as communities, ZIP codes, and census tracts. The dimension hierarchies enable roll-up and drill-down operations that control the level of detail in queries. These formally defined hierarchies also provide the framework for navigation or data browsing.

In order to describe the dimension hierarchies succinctly to both end-users and developers, *dimension hierarchy diagrams* have been utilized in the CATCH data warehouse design process. These diagrams show the hierarchical nature so that end-users have an uncluttered view of how they can navigate and designers can easily understand the dimensional structures. Fig. 3 illustrates an important health care dimension based on the International Classification of Disease (ICD) codes. Currently, we are using versions 9 and 10 of the ICD codes. These codes are divided into chapters and sections, which provides a natural hierarchy for the codes. Fig. 3 shows the hierarchical structure using separate tables, but these tables can be easily denormalized to enhance query performance. In addition, there are several other tables that provide alternative hierarchies for this important dimension. This ICD PROCEDURE dimension is combined with many other dimensions such as patient age, gender, mortality risk, and severity of illness to form star schemas (see Fig. 2) with rich query environments.

### 3.2. Data warehouse design: the data access pyramid

The mission of the CATCH data warehouse is to support the automated and cost-effective application of CATCH, as well as to enable more detailed
analyses that were not possible using the coarse-grained data that typified past CATCH reports. In order to meet these goals, the data warehouse design includes several levels of data granularity, from the coarse-grained data used in generic report production to actual event-level data, such as hospital discharges. The data warehouse design includes major components at all three levels of granularity as illustrated in the data access pyramid found in Fig. 4.

**Report indicators**—Reporting tables with derived or highly aggregated data are used to support the core CATCH reports, including comparisons between a target county and peer counties. These tables also provide fast response for interactive access via data browsing tools and can provide the foundation for simple community-wide Internet access. In addition, the metadata play an important role at the reporting level, providing indicator definitions, state or federal goals, and expert domain knowledge for priority filters (e.g., economic impact and treatment availability). This report level of the data warehouse may not be needed in all data warehouse applications but provides important support for rapid generation of community CATCH reports.

**Aggregate data**—There are families of star schemas that provide true dimensional data warehouse capabilities, such as interactive roll-up and drill-down operations. These components have carefully designed dimensions that can be utilized by more sophisticated data browsing tools. The star schemas are populated using thorough data staging and quality procedures that usually involve processing detailed data sets extracted by various health care agencies and organizations. Typically, the data are aggregated and transformed for loading into a family of related star schemas—a constellation—that share important dimensions and support interactive online analytic processing (OLAP) techniques.

**Transaction data**—For certain types of information, the design calls for retaining very fine-grained or even event level data. An example is the hospital discharge data that includes each hospital discharge event for the more than 200 hospitals that are mandated to report such information in Florida. These data are retained at the transaction level because of the rich set of facts and dimensions available for analysis and the density of potential aggregations that result in negligible space savings.

These three levels of aggregation within the data warehouse combine to meet a wide range of reporting requirements and performance goals, thus providing a flexible basis for disseminating health care information to community decision-makers. The following two sections (Sections 3.3 and 3.4) provide some examples of the major data warehouse components. At the aggregate data level, a coarse-grained component based on the Public Health Information Data System (PHIDS) is used to support CATCH report production and high-level browsing. A second exam-

---

**Fig. 3.** ICD Procedure dimension hierarchy.

**Fig. 4.** Data access pyramid.
ple aggregate is procedure volume information formed from the underlying hospital discharge data. The original hospital discharge data provide an example of transaction-oriented data that supports detailed analyses, along with other data such as vital statistics (e.g., births and deaths) and specific disease registries.

3.3. Aggregated Florida Department of Health Data

An example of a highly aggregated data warehouse component is the Public Health Information Data System (PHIDS) star schema. The Florida Department of Health collects, analyzes, and reports a large number of public health indicators. These items have always provided critical assessment measures within CATCH. The importance of the PHIDS indicators made them obvious candidates for inclusion in the data warehouse and a natural resource for automation of the traditional CATCH report.

The PHIDS indicators are clearly not the fine-grained data that support a detailed OLAP environment. The data are highly aggregated and provided annually at the county level. Therefore, the data set is suitable for generating the traditional CATCH report, but unsuitable for more specific analyses. Essentially, the construction of the data warehouse has been a search for both fine and coarse data that can provide synergies through integration. The simple star schema used to implement the PHIDS-based data warehouse component has only the year reported and the county as explicit dimensions. Currently, many of the PHIDS indicators are maintained using spreadsheets at the Florida Department of Health. For use in the data warehouse, the data are first extracted from the spreadsheets, reformatted using custom staging programs, and then loaded via a bulk loader utility. The twin star staging process, as described in Section 4, is used to ensure data quality. Data correctness is verified by sampling the data and comparing the data warehouse values with published PHIDS reports.

3.4. Transaction-oriented hospital discharge data

Florida hospital discharge transactions are collected by the Agency for Health Care Administration (AHCA) from the more than 200 short-term acute care hospitals in the state. These hospitals report every discharge transaction, regardless of payer, throughout the state. Hospital discharge data are used to derive several CATCH indicators such as avoidable hospitalizations due to diabetes and other chronic diseases. Typically, the large volume of hospital discharge transactions is scanned to form derived or aggregated data for CATCH indicators. However, the broader mission of the CATCH data warehouse is both to support the CATCH methods and enable more detailed investigations of critical local health care issues. It is the ability to fully explore issues at appropriate levels of detail that make the fine-grained components so important. While first staging and preprocessing the hospital discharge data for use in forming CATCH indicators, the value of the discharge transactions themselves became very apparent. The hospital discharge transactions provide an interesting set of numeric data items, such as length of stay and a breakdown of revenues, which are very well suited for a data warehousing approach. In addition, the transactions include a rich set of attributes that provide many natural dimensions for use in formulating queries.

Transaction-based star schemas can provide very useful functionality within a data warehouse framework, making the hospital discharge star an important component of the CATCH data warehouse. The hospital discharge data includes over 20 interesting dimensions such as the discharging hospital characteristics, admission criteria, diagnostic codes, procedure codes, reimbursement categories, time, geographic location, and many others. Furthermore, many of these dimensions are hierarchical in nature, easily supporting important roll-up/drill-down operations. Fig. 2 is a partial representation of the discharge star schema. The discharge star is equally rich in additive numeric facts. For instance, length of patient stay is a particularly important measurement for analysis. There is also a measurement indicating elapsed days until the medical procedure. Finally, there is a total revenue item that provides important cost information. In fact, there is also a large text field with embedded revenue items that provides a breakdown of the various costs from room charges to laboratory fees. Procedures to parse this text field have been developed as part of the data staging activities and are used to extract revenue items, providing nearly 30 interesting numeric facts for each transaction. It is not uncommon to have useful information buried in text
fields that must be preprocessed using data staging tools or customized procedures. This can be a challenging task since the source database has no understanding of the structure embedded in such text fields and therefore, simple query access is impossible. In this case, the rich set of facts and highly dimensional structure of the hospital discharge data make it a powerful warehouse component for detailed investigations and customized analyses.

The hospital discharge star has repeating groups for diagnoses (ICD D X 1–10) and procedures (ICD Procedure 1–10). This design mirrors the underlying data and simplifies the data staging process for the millions of discharge records used in the project. An alternative design without repeating groups might simplify some queries, but this fine-grained data is at the bottom of the data access pyramid and is typically aggregated for most query processing. The original positional representation also conveys information relevant to health care coding practitioners and is used in several ancillary algorithms. For many purposes, the primary diagnosis or procedure is used in calculating higher-level health care indicators, so this structure is maintained in the transaction-oriented data [28].

It is sometimes preferable to store the actual transactions rather than lightly aggregated data that has been derived from the underlying transactions. Kimball [19] uses the term sparsity failure to describe the size explosion that can occur when creating aggregate data from a sparsely populated fact table. Detailed fact data such as hospital discharge transactions will probably not have all combinations of the dimensions present in the actual data. In other words, not all diseases occur in all hospitals during a particular year and therefore the effect with regard to size is not multiplicative. If we consider only the cardinality of the actual dimensions then the possible combinations of dimension key values is very large for the hospital discharge data. For example, consider the following four dimensions with approximate cardinalities, hospitals (250), ICD codes (15,000), severity ratings (5), and payers (10). This could result in 187.5 million dimension key combinations. Further, we can define density as the actual number of records (roughly 2 million/year for discharges) divided by the potential combinations of dimension keys, yielding a density of 2/187.5 or roughly 1.07%. This remarkably low density makes intuitive sense since the very fine ICD distinctions lead to sparse usage. Imagine that we decide to construct an aggregate table by creating 150 disease categories that summarize the 15,000 ICD codes, reducing the dimension size by a factor of 100. In this case, all 150 categories may appear for each hospital (a reasonable assumption) giving a density of 100% and roughly 1.9 million rows. This rather insignificant space savings comes at the expense of losing the richness of the original ICD codes and the flexibility of having individual cost data for each transaction. Therefore, in the CATCH data warehouse and many other applications, transaction-oriented components make good sense. In fact, to really understand the implications for tasks such as data warehouse capacity planning it is often necessary to sample the data to discover the actual distribution of dimension values. The design challenge is to carefully consider the number of fine-grained items that are summarized to form the aggregate data and look for a factor of 10 or more as a reasonable compression ratio [19].

### 3.5. Performance issues

The large volumes of data contained in the CATCH data warehouse coupled with demanding queries can conspire to produce some truly awful performance. As in any database project, good design is the most effective tool for enhancing performance. The CATCH data warehouse design continues to evolve in response to new challenges. In addition to design changes, three other techniques offer avenues for improving performance: aggregate tables, star schema indexing strategies, and physical table partitions.

#### 3.5.1. Aggregates

Many data warehouse designers identify aggregates as one of the most effective strategies for improving performance. Kimball [19] notes that “aggregates can have a very significant effect on performance, in some cases speeding queries by a factor of 100 or even 1000.” If the aggregate data are useful, having the data physically ready and waiting will certainly improve query speeds. In addition, if sparsity failure is avoided, then the amount of data required may also be substantially reduced. That is, benefits from both reduced space and previously
handled computations can accrue through the use of aggregates. In addition, many data warehousing navigation tools are aggregate-aware, making the aggregate structures transparent to the end user. However, there are a potentially large number of aggregates that are possible given a rich set of dimensions. The choice of which aggregate tables to build is based on the type of queries being executed and will naturally change over time [14].

Aggregates play an important role in the CATCH data warehouse. Some data are extracted and loaded in aggregate form, such as the PHIDS indicators discussed above, and other aggregates are derived from more detailed data warehouse components. For instance, vital statistics such as death and birth certificates are used to derive a collection of aggregated mortality and birth-related indicators. There are two somewhat different purposes for aggregates. Highly aggregated data are used to directly support traditional CATCH report production, while lightly aggregated data are used to improve query performance. The continual re-evaluation of aggregates is an important task in data warehouse administration.

### 3.5.2. Indexing

Many database management systems intended for data warehousing support **bitmap index** structures. Bitmap indexes are especially suited to low cardinality dimensions such as admission quarter, day of the week, gender, and others. These indexes are space efficient and speed the star queries that characterize access to fine-grained structures such as the hospital discharge data. Another technique is to cache the smaller dimension tables in memory for improved query performance. All of these techniques have been employed and performance tuning continues to be an ongoing activity as the user community grows and explores new uses for the data warehouse.

### 3.5.3. Partitioning

The third important performance tuning technique is the use of physical table partitioning [6]. The use of table partitions is important both for query performance, as well as data warehouse management. Since the data are loaded or staged at different times, these activities can be isolated through partitioning. This also allows preprocessing and data quality procedures to be run on separate partitions. In addition, partitioned indexes can also be used. One of the most important benefits of partitioned tables is the opportunity for the optimizer to exclude large portions of the data when queries include restrictions on partitioning attributes. An excellent example of partitioned tables in the CATCH data warehouse is the hospital discharge data. In recent years, there have been roughly 2 million discharge transactions/year. The goal is to keep at least 10 years of discharge data or 20 million transactions available for analysis, but often only a few years are necessary for any given query, thereby creating an ideal parameter for partitioning. The hospital discharge data is partitioned by year, with roughly 1.5–2 million rows per partition. If a query specifies a single year or a small range of years, the optimizer can create an execution plan that only searches the required partitions, leaving the vast majority of data untouched. Since most of the detailed interactive analyses fit this mold, the performance tends to be quite good. However, the entire collection of data is still available for queries that cover a wide range of years, it just takes more time.

### 4. Data staging and quality assurance

The extraction, transformation, and loading (ETL) functions in a data warehouse are considered the most time-consuming and expensive portion of the development lifecycle [22]. These processes are concerned with the extraction of data from legacy systems, transformation and preprocessing requirements to produce useful, integrated data, and the transportation of the data into the actual data warehouse structures. The CATCH data warehouse involves somewhat unusual challenges with regard to data staging activities. The data are drawn from multiple organizations, which usually apply in-house transformations to data collected by yet another layer of organizations. For instance, the hospital discharge data are originally collected by hospitals and reported to the Florida Department of Health. These data are then integrated, preprocessed, and provided to other interested organizations, including the CATCH data warehouse project. In the case of demographic data, population levels are extracted from the Florida Governor’s Office and the Census Bureau. Overall, the data warehouse has continued to grow without the need for a data purging
strategy. However, as the size continues to increase and finer geographic levels are used, a purging strategy will become necessary in the near term. The design is already multi-level, as described in Fig. 4, and it is the base of the data pyramid that accounts for most of the space. As space becomes an issue, the earlier years of fine-grained data will be purged and retained offline. These structures are maintained as physical partitions, so the purging operations can be conducted without disrupting data access and data can easily be re-introduced.

Two innovative techniques, twin star data staging and data quality filters, have been developed to manage the ETL processing required in the CATCH data warehouse.

4.1. Twin star staging

Fig. 5 outlines the twin star staging process and its three component stages. The approach is designed to utilize the power of commercial database systems, especially referential integrity constraints and exception processing. The various stages use a combination of scripting languages, bulk-loading tools, and database procedures.

4.1.1. Stage 1

The process begins with file-based preprocessing and cleansing activities. These procedures can be written in any programming language, but AWK and Perl have been especially useful in the CATCH project with their built-in parsing and pattern matching capabilities. Data transformation, quality checks, and simple reports can all be performed on the initial data file. Even though many checks will be repeated throughout the data staging process, the presence of redundant checks is an asset with regard to data quality. Stage 1 of the twin star strategy involves using a bulk loader to move the data into a staging table within the database system. The staging table is designed for maximum flexibility in storing data, minimizing data type conflicts, and providing a workbench for database-resident transformation procedures. Typically this includes additional attributes that are created as part of the preprocessing and cleansing tasks. Bulk loading utilities are used to quickly populate the staging table and capture problematic data in a series of log files. Data type, uniqueness, and “not null” checks for critical staging table attributes can be used to control the thoroughness of this data staging step. With care, many simple data quality issues can be resolved at this early stage.

4.1.2. Stage 2

The temporary star shares the critical data dimensions with the permanent star, and is essentially a ‘twin’ of the permanent star (though there may be different supporting dimensions for particular tasks). The fact table attributes and important dimensions should be exact duplicates so that any operations or referential integrity checks will be consistent between the stars. Stage 2 entails moving the data from the staging table to the temporary star. Attribute data types should be compatible and referential integrity constraints can be used to check for valid dimension keys. The referential integrity constraints are disabled and later re-enabled sequentially after the load to perform the checks in one sweep, thereby improving processing time. Most database systems provide a method of capturing invalid rows and it is important to make use of such capabilities during both the Stage
2 and 3 transfers. Since the temporary star is the functional equivalent of the permanent star, just much smaller, the interface and data browsing tools developed for the actual data warehouse can be used to exercise the temporary star. Test reports, browsing by power users, and sanity checks based on comparisons with previously loaded data in the permanent star are all useful methods of ensuring high quality data in the temporary star.

4.1.3. Stage 3

The **permanent star** is the long-term storage area for the data warehouse. This star must be carefully indexed, distributed across storage devices to avoid I/O bottlenecks, and possibly partitioned. As noted earlier, partitioned tables can provide performance improvements by distributing information across physical devices and by allowing the query optimizer to select only the relevant partitions. In addition, partitioned tables ease data warehouse management tasks through creation, loading, and archiving of independent partitions. The Stage 3 transfer from the temporary star to the permanent star should be fast and free of data type and referential integrity violations. The simple transfer will allow large volumes to be processed within most load windows. Redundant referential integrity constraints can be used as a final check (again disabling and re-enabling for efficiency). The resulting exception tables should be empty, but any offending rows are a clear sign that somehow problems survived Stages 1 and 2. This provides a last opportunity to postpone releasing or publishing the data.

4.2. Data quality filters

The data quality issues that surface while initially constructing a data warehouse are among the most challenging obstacles, contributing significantly to the time spent in data staging activities. As noted above, the ETL processes and quality assurance procedures can account for the majority of time and resource commitments in a data warehouse project. This has been the case in the CATCH data warehouse project, where there are a large number of data sources and many intermediate stages for errors to be introduced. In addition, the challenge of producing a truly integrated design requires translations to common definitions and shared dimensions. Rather than any “magic bullet,” a long-term effort to develop a comprehensive set of preprocessing procedures will produce the best data quality. The procedures under development on the CATCH project include a measure of redundancy to provide added insurance against quality problems surviving various phases of the ETL process. As more procedures have been added to the quality assurance arsenal, an interesting structure has emerged, mirroring the natural structure of the data warehouse, with procedures falling into the following categories of **quality filters**.

- **Fact filters** are the quality procedures used to check the fine-grained data, such as hospital discharge transactions. For example, any discrepancies between itemized fees and total charges should be flagged. Quality procedures at this level compare attributes within a fact table row, or may compare between two rows, but the focus is on fine-grained data.

- **Aggregate filters** include quality checks that become possible only when the focus is on summaries of fact-level answer sets. As we have seen, aggregates are important for boosting performance, but they also present data quality assurance opportunities. At this level, ‘roll-up’ operations over important dimensions allow aggregate averages, maximums, or other summaries to be compared. With regard to hospital discharge transactions, comparisons of average lengths of stay, maximum costs, or diagnostic volumes can all be usefully compared by hospital and by year. That is, large hospitals can be verified against each other and new data can be compared against previous years. Aggregate filters can be the basis for some very powerful data quality procedures, effectively using the capabilities existing in the data warehouse.

- **Dimension filters** are the procedures used to investigate ‘dirty’ dimensions. For instance, many business-oriented data warehouses include a customer dimension that can be very large and may have severe data quality problems [19]. Duplicate customer entries, household matching, and data obsolescence issues are among the problems inherent in such dimensions [3,10]. In the CATCH data warehouse, dimensions that must be carefully monitored include hospitals, practitioners, and geographic entities such as counties and communities. Dimension filters can be used to monitor many problems with regard to changing dimensions.
These three broad categories of quality filters can be further refined based on the type of comparison being used. For instance, the intratuple filters involve comparisons between attributes within a single record. Of course, the record itself may be at an aggregate level and represent a summary of a fact-level answer set. For example, average pharmacy costs may have a fairly predictable relationship with total charges for a given disease. This type of comparison could be used as a quality check within a given aggregate hospital discharge record, an example of an intratuple aggregate filter.

Comparisons across records, or intertuple filters, provide a rich set of quality assurance opportunities that examine relationships between fact table rows or aggregates of these rows. An example of this type of filter would be comparisons between disease volumes by year. Unlikely disease distributions, after accounting for population growth, might indicate a data quality problem with new data. The distinction between intertuple and intratuple comparisons, combined with the major filter categories, leads to six interesting filter categories that seem to naturally describe the many types of quality procedures being built into our data warehouse.

An additional quality assurance strategy involves comparing data warehouse aggregates with known summaries published by outside sources. This process can best be described as a quality benchmark, where externally derived data is used to check internal data warehouse procedures. This type of quality procedure usually includes permanent data quality tables populated with externally produced data summaries based on published reports or spreadsheet calculations. For instance, state-level reports on the number of specific disease occurrences provide a benchmark for data warehouse aggregates based on the underlying hospital data. Automated comparison procedures report only reasonably large departures based on user-defined thresholds. Quality benchmarks are particularly important as ongoing development activities yield both larger volumes of data, as well as new aggregation procedures. Before new versions of procedures or interface tools are deployed, historical quality benchmarks can be used to evaluate their performance. Both quality benchmarks and filters are part of the substantial infrastructure necessary to meet data quality goals. These tools account for a significant portion of the CATCH data warehouse development effort [2].

5. CATCH data warehouse applications

The data warehouse is used to support a variety of activities, from automating the original CATCH reports to supporting current health care research initiatives. The CATCH methods provided a solid foundation for the initial implementation efforts, but as components have been added the synergies have opened new application opportunities. Clearly, the human–computer interface is of paramount importance in the data warehouse environment and the primary determinant of success from the end-user perspective [1]. In order to support analysis and reporting tasks, the data warehouse must have high quality data and make that data accessible through effective interface technologies. The act of releasing data in a warehouse is in a very real sense the same as publishing that data in printed form—retractions in both media can be very painful.

5.1. Producing CATCH reports

CATCH reports have been refined over the past decade in the field. The field expertise available in the interdisciplinary research team infused the requirements process and provided a clearly identifiable goal as the first step in data warehouse construction. Hundreds of stored procedures, as well as the design itself, implement many aspects of this domain expertise. The stored procedures generate the health status indicators and move them upward in the data access pyramid for final report production. The reports allow quick and easy access to comprehensive summaries and more detailed collections of information from the data warehouse using standard report writing technologies. This type of pre-defined and thorough reporting is critical for implementing a more automated CATCH report and will probably be the preferred format for many users. For example, the comparison between target counties and peer counties, as well as state averages, are fundamental components of the original CATCH reports and important tools for community health care planners. In addition, current and historical trend information is provided on fact sheets for each
health indicator. The final reports are really reference books (numbering over 300 pages) with several major parts, such as comparisons, fact sheets, and prioritized lists.

New features that move beyond the original CATCH reports include components that enable user-defined communities to supplement the traditional county-level perspective. Users can define smaller communities based on geographic or demographic criteria, with community fact sheets providing an exploded view of selected health status indicators. While CATCH has traditionally focused on large hardcopy reports, the reports are now produced directly in Web-friendly formats for electronic distribution. The advantage of this approach is that a strong methodological structure can be retained as the reports are much more widely distributed. The interested reader can refer to the Center for Health Outcomes Research (CHOR) Web site for examples of current reports [chor.hsc.usf.edu].

In addition to static reports, the high-level components of the data warehouse can be accessed dynamically using data browsing tools. It is usually possible to constrain the navigation, while still providing enough freedom to explore many more perspectives than can be accommodated in a traditional report. Fig. 6 shows an online analytic processing (OLAP) tool being used to browse through trend information for specific indicators at the county level. Most of these tools can support both desktop and Web browser access, making this an important new avenue for data dissemination.

5.2. Investigating health care issues

Data warehouse browsing tools provide star query-like access through a flexible menu-based interface, with pull-down menus representing important dimensions. These types of tools are easy to use and support some ad-hoc exploration, but are usually controlled
through an administrative layer that determines the data available to end-users. In developing a flexible interface, there is a tradeoff between the ability to express ad-hoc queries and the ease-of-use that results from pre-defined constructs implemented by data warehouse designers and administrators. Of course, SQL can provide an ad-hoc query facility, but requires some care in the data warehouse environment with very large tables and ill-formed queries conspiring to sharply degrade performance. In addition, use of SQL by casual users often produces incorrect queries resulting in erroneous results from the data warehouse. As noted above, OLAP tools can be used to empower standard report users and allow simple navigation through many more views than can be produced using traditional reporting tools, yet still curtail unwanted operations.

A second, and in some ways more important role for the browsing tools is to provide a flexible interface for more customized analysis. Health care issues highlighted by preliminary reports can be investigated more fully using the finer levels of detail maintained in the data warehouse. These tasks might entail querying the true dimensional star schemas that include age, gender, race, and other dimensions, or even the event-oriented data, such as hospital discharges. These data warehouse resources support much more detailed analyses, allowing the user to focus on issues such as differences in age or race with regard to specific health status indicators. Once decision-makers review the CATCH report, they may have community-specific issues that relate to the diverse population groupings that inevitably fall within somewhat arbitrary political boundaries. Dealing effectively with such important

![Fig. 7. Browsing screen for hospital disease indicators.](image-url)
issues requires a more careful and focused analysis that is precluded at the higher levels of aggregation that make up the generic CATCH reports.

A current research initiative involves the exploration of volume and cost-related issues in health outcomes. Data browsing tools are used for exploratory analysis. Used in this manner, OLAP tools provide a first step in the data mining process [13]. Fig. 7 illustrates a browsing screen in which detailed volume, length of stay, and cost data are presented for specific hospitals, or groups of hospitals. In addition to tabular representations, these tools provide graphic capabilities that support simple data visualization.

5.3. Security issues

Currently, dynamic access to the CATCH data warehouse is restricted to the development team and associated health care researchers. Obviously, some of the data may be sensitive in nature. Data security is an important issue in the health care environment and the CATCH data warehouse attempts to balance the information requirements for local health care planning with critical security issues. It is important to note that no patient identifiers of any kind are incorporated within the data warehouse. Security issues are mostly concerned with reporting rare events in geographic areas that might allow a person to be identified through other data sources. Detailed security policies provide guidance for the manipulation and reporting of health care data in the data warehouse.

6. Conclusions

The CATCH data warehouse can have an important impact on our health status by making rigorous, quantitative information available to health care decision makers in local, state, national, and international communities. In this paper, we have described some of the technical challenges faced in designing and implementing a data warehouse for health care information. We have presented innovative research contributions in the areas of data warehouse design, data staging for ETL processing, data quality assurance, and health care data warehouse applications.

The CATCH data warehouse is now fully functional. For example, it has been recently used to produce a comprehensive CATCH report for Miami–Dade COUNTY, Florida’s largest county. As part of this report we were asked to provide more detailed assessments of the eight commission districts within the county. The flexibility of the data warehouse to provide customized reporting allowed us to provide these analyses rapidly and effectively. Because this report was the first to be fully automated, we verified the accuracy of the report with a complete hand check of every data table. The discrepancies between the automated data tables and the manually derived tables were minimal and easily reconciled.

The CATCH data warehouse remains a work in progress. We are pursuing an active research agenda to enhance the technical data warehousing capabilities and community health care applications. We invite the reader to follow the progress of the data warehouse at our CHOR web site [chor.hsc.usf.edu]. In the next sections (Sections 6.1 and 6.2), we briefly review our current research directions.

6.1. Data warehouse research directions

The CATCH data warehouse provides a rich research environment for focused investigation in the following areas.

- Data Warehouse Design—The variety and volatility of health care data sources make the maintenance of the data warehouse design a true challenge. Changes to source data formats frequently require the updating of dimension table schemas. Often historical data cannot be placed into the new format without information loss. Finding solutions for maintaining historical accuracy while providing efficient use of all data in current applications is difficult. We are researching design techniques to minimize the impact of dimension table changes on the maintenance and operations of the data warehouse [2].

- Data Staging—As presented in Section 3, we have implemented an innovative twin star data staging procedure. Ongoing research will study the performance of twin star data staging on various data loads. Enhancements to the procedure will be proposed and implemented.

- Data Quality—Issues of data quality dominate our research agenda. The health care field places particular emphasis on data accuracy, timeliness, privacy, and ease of use [27]. We are in close contact
with the communities who receive the CATCH reports and have interviewed a number of users to elicit data quality requirements. This information will be used to drive our research to improve data quality in the CATCH data warehouse.

- Data Dissemination—The technologies for disseminating CATCH reports to communities are rapidly evolving. The requirements of the receiving communities and the capabilities of the data warehouse system will drive future research directions.
- Data Mining—We are aggressively investigating several areas of knowledge discovery in the CATCH data warehouse [11]. We have a unique capability to perform detailed studies in such areas as physician and hospital volume, racial disparities in health care, and environmental impacts on community health status.

6.2. Community decision-making with CATCH data

The CATCH data warehouse will result in widespread distribution of data previously unavailable to most communities, as well as online access for specialized inquiry. Many issues arise as to how the communities will make the most effective use of the CATCH data for health care decision-making. This is an area with considerable research potential.

There is a rich literature on the decision-making process both with and without information technology. The study of group decision support systems and environments has a strong tradition in the management information systems field [8]. In many ways, this important body of work is appropriate to health care decision-making, which is usually group-oriented. For example, Dennis et al. [9] study the effects of minority influence on decision-making and find that the presence or absence of technology has very different effects. Another important contributing area would be the political process and its ramifications to decision-making [20]. Certainly, policy making in health care is very much a political process.

The use of the CATCH methodology and state-of-the-art data warehousing technology across many Florida communities will provide a rich research opportunity for studying interesting issues on group decision-making in community health care organizations. Such issues would include the composition of the decision-making group, the community stakeholders and their political influence, the decision-making process, and dissemination patterns of health care information in the community. The complexities and the interrelationships among these issues make the design of research studies both a challenge and an opportunity. As the automated CATCH reports are produced for various communities in Florida, we will study how effectively the CATCH information is used for health care planning.

Acknowledgements

The authors gratefully recognize the U.S. Department of Commerce, which has provided funding through a Technology Opportunities Program (TOP) grant. The Florida Department of Health has been a research partner. Research collaborators in the College of Public Health include R. Campbell, E. Gilbert, S. Luther, B. Myers, and B. Steverson. Contributing graduate students in the College of Business Administration include S. Hedge-Desai, R. Marsh, D. McCorkel, M. Nevrekar, M. Pearl, R. Rajendrababu, and J. Slayton. The authors also thank Oracle Corporation for making their state-of-the-art development tools available through the Oracle Academic Initiative.

References

making: you can lead a group to information but you can’t make it think, MIS Quarterly 20 (4) (1996) 433–458.


Donald J. Berndt is an Assistant Professor in the Information Systems and Decision Sciences Department in the College of Business Administration at the University of South Florida. His research interests include data warehousing, knowledge discovery, and data mining. Dr. Berndt received a PhD in Information Systems from the Stern School of Business at New York University. He is a member of Beta Gamma Sigma, ACM, AIS, and INFORMS.

Alan R. Hevner is an Eminent Scholar and Professor in the Information Systems and Decision Sciences Department in the College of Business Administration at the University of South Florida. He holds the Salomon Brothers/HRCP Chair of Distributed Technology. His research interests include software engineering, software testing, distributed database systems, and health care information systems. He received a PhD in Computer Science from Purdue University. Dr. Hevner is a member of ACM, IEEE, AIS, and INFORMS.

James Studnicki is a Professor of Health Policy and Management at the University of South Florida College of Public Health. His research interests include measuring the health status of communities, evaluating alternative treatment outcomes, and studying the influence of managed care penetration on the utilization and quality of health services. Dr. Studnicki received a ScD from Johns Hopkins University.