Calhoun: The NPS Institutional Archive

DSpace Repository

NPS Scholarship | Theses

2001-03

Using On-line Analytical Processing (OLAP) and data mining to estimate emergency room activity in DoD Medical Treatment Facilities in the Tricare Central Region

Ferguson, Cary V.

https://hdl.handle.net/10945/10834

This publication is a work of the U.S. Government as defined in Title 17, United States Code, Section 101. Copyright protection is not available for this work in the United States.

Downloaded from NPS Archive: Calhoun

http://www.nps.edu/library

Calhoun is the Naval Postgraduate School's public access digital repository for research materials and institutional publications created by the NPS community. Calhoun is named for Professor of Mathematics Guy K. Calhoun, NPS's first appointed -- and published -- scholarly author.

Dudley Knox Library / Naval Postgraduate School
411 Dyer Road / 1 University Circle
Monterey, California USA 93943

NAVAL POSTGRADUATE SCHOOL

Monterey, California

THESIS

USING ON-LINE ANALYTICAL PROCESSING (OLAP) AND DATA MINING TO ESTIMATE EMERGENCY ROOM ACTIVITY IN DOD MEDICAL TREATMENT FACILITIES IN THE TRICARE CENTRAL REGION

by

Cary V. Ferguson

March 2001

Thesis Advisor: Daniel R. Dolk
Associate Advisor: Samuel E. Buttrey

Approved for public release; distribution is unlimited.

20010529 038

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 2001	Master's Thesis

4. TITLE AND SUBTITLE
Using On-Line Analytical Processing (OLAP) and Data Mining to Estimate Emergency Room Activity in DoD Medical Treatment Facilities in the TRICARE Central Region

5. FUNDING NUMBERS

6. AUTHOR(S)
Ferguson, Cary V.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(maximum 200 words)***

On-line Analytical Processing (OLAP) and data mining can greatly enhance the ability of the Military Medical Treatment Facility (MTF) emergency room (ER) manager to improve ER staffing and utilization. MTF ER managers use statistical data analysis to help manage the efficient operation and use of ERs. As the size and complexity of databases increase, traditional statistical analysis becomes limited in the amount and type of information it can extract. OLAP tools enable the analysis of multi-dimensional data, which can give the user access to previously undiscovered information. Data mining has the capability to break large sets of data down into groups by classifications, associations, and clusterings to transform previously meaningless data into useful information. This research presents a brief overview of the DoD medical system, OLAP, and data mining. OLAP and data mining tools then analyze a data set containing two years of MTF ER data from the TRICARE Central Region. The results of these analyses provide insight on the predictive capabilities, advantages, and disadvantages of applying OLAP and data mining to MTF ER data.

14. SUBJECT TERMS
On-Line Analytical Processing (OLAP), Data Mining, Medical Treatment Facility (MTF), Emergency Room.

15. NUMBER OF PAGES
122

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFI- CATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

USING ON-LINE ANALYTICAL PROCESSING (OLAP) AND DATA MINING TO ESTIMATE EMERGENCY ROOM ACTIVITY IN DOD MEDICAL TREATMENT FACILITIES IN THE TRICARE CENTRAL REGION

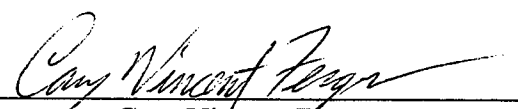
Cary V. Ferguson
Captain, United States Army
B.B.A., University of Notre Dame, 1990

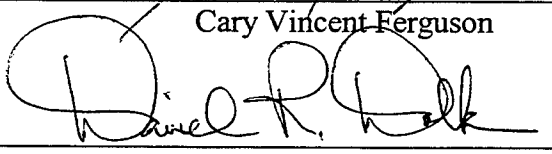
Submitted in partial fulfillment of the
requirements for the degree of

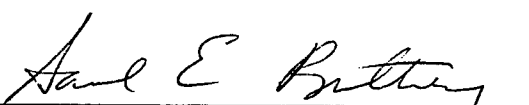
MASTER OF SCIENCE IN INFORMATION SYSTEMS TECHNOLOGY

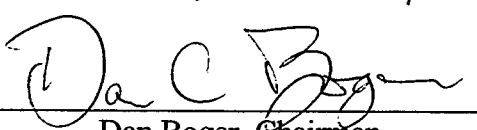
from the

NAVAL POSTGRADUATE SCHOOL
March 2001

Author: 
Cary Vincent Ferguson

Approved by: 
Daniel R. Dolk, Thesis Advisor


Samuel E. Buttrey, Thesis Second Reader


Dan Boger, Chairman
Information Systems Academic Group

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

On-line Analytical Processing (OLAP) and data mining can greatly enhance the ability of the Military Medical Treatment Facility (MTF) emergency room (ER) manager to improve ER staffing and utilization. MTF ER managers use statistical data analysis to help manage the efficient operation and use of ERs. As the size and complexity of databases increase, traditional statistical analysis becomes limited in the amount and type of information it can extract. OLAP tools enable the analysis of multi-dimensional data, which can give the user access to previously undiscovered information. Data mining has the capability to break large sets of data down into groups by classifications, associations, and clusterings to transform previously meaningless data into useful information.

This research presents a brief overview of the DoD medical system, OLAP, and data mining. OLAP and data mining tools then analyze a data set containing two years of MTF ER data from the TRICARE Central Region. The results of these analyses provide insight on the predictive capabilities, advantages, and disadvantages of applying OLAP and data mining to MTF ER data.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

ACRONYMS

AD	Active Duty
ADD	Active Duty Dependent
ADS	Ambulatory Data System
AFB	Air Force Base
ARS	All-Region Server
ASD/HA	Assistant Secretary of Defense for Health Affairs
CART	Classification and Regression Tree
CHCS	Composite Healthcare System
CPT	Current Procedural Terminology
CONUS	Continental United States
DEERS	Defense Enrollment Eligibility Reporting System
DMIS	Defense Medical Information System
DoD	United States Department of Defense
ER	Emergency Room
ER	Entity-Relationship
GRI	Generalized Rule Induction
GUI	Graphical User Interface
HOLAP	Hybrid On-Line Analytical Processing
HSO	Healthcare Support Office
LA	Lead Agent
MAJCOM	Major Command
MB	Megabyte
MEDCOM	Medical Command
MHS	Military Health System
MHZ	Megahertz
MIS	Management Information Systems
MOLAP	Multi-Dimensional On-Line Analytical Processing
MTF	Medical Treatment Facility
OLAP	On-Line Analytical Processing
OLTP	On-Line Transaction Processing
OSD	Office of the Secretary of Defense
RAM	Random Access Memory
RDBMS	Relational Database Management System
RMC	Regional Medical Command
ROLAP	Relational On-Line Analytical Processing
SADR	Standard Ambulatory Data Record
SQL	Standard Query Language
TDMIS	Treatment Defense Medical Information System
TMA	TRICARE Management Activity
USAF	United States Air Force

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGEMENT

I would like to acknowledge the tremendous support I received from the TRICARE Central Region Headquarters. Col Ted McNitt approved the project and provided great insight into the DoD Medical system. I must especially thank Tony Rogers and Terri Cheyney for their responsiveness and support on this project. Many other people took time out of their busy schedules to meet with me on this project. In particular, I would like to thank MAJ Gretchen Cusack, CPT Kevin Watson, Tim Jordan, and Jack Robeda for sharing their time with me. I would also like to thank my advisors for their guidance during the entire thesis process.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. BACKGROUND

Today's volunteer United States military forces face a constant challenge of finding ways to improve the quality of life for their service members. One key area in which the U.S. Department of Defense (DoD) is attempting to improve the quality of life for service members is medical care. In the face of declining resources, including personnel, facilities, and funding, the DoD healthcare system is continuously seeking ways to reduce costs while improving the quality of medical care to all qualified users. The declining number of military medical treatment facilities (MTFs) has caused the DoD to utilize the civilian healthcare system more heavily in order to provide medical services for qualified DoD recipients. The civilian medical system is cost-driven, whereas the military is mission-oriented. Diminishing resources have forced MTFs to closely monitor all aspects of their operations and continuously improve efficiency and reduce costs to justify their existence.

One area that MTFs closely monitor is Emergency Room (ER) management and utilization. The type of emergency medical services an ER provides determines the minimum staffing of emergency medical specialists in the ER. MTF ERs face the challenge of patients using ERs out of convenience as a primary care clinic instead of as an emergency room. Manning and operating an ER is expensive due to the specialists required and improper utilization of ER services by patients. Efficiency gains in ER operation and utilization can deliver significant cost savings for the MTF.

MTF ER managers use statistical data analysis to help manage the efficient operation and use of ERs. As the size and complexity of databases increase, traditional statistical analysis becomes limited in the amount and type of information it can extract. Sophisticated data analysis tools such as on-line analytical processing (OLAP) and data mining can process enormous amounts of historical data and uncover information that classical data analysis methods can not find. OLAP tools enable the analysis of multi-dimensional data, which can give the user access to previously undiscovered information. Data mining has the capability to break large sets of data down into groups by classifications, associations, and clusterings to transform previously meaningless data into useful information. The results of OLAP and data mining could greatly enhance the MTF ER managers ability to improve ER staffing and utilization.

B. OBJECTIVES/RESEARCH QUESTIONS

The objective of this research is to apply OLAP and data mining techniques to historical MTF ER data to produce information that an MTF ER manager can utilize to improve MTF ER staffing and utilization. This research will examine the capabilities of OLAP and data mining, and their potential to predict ER activity in DoD MTFs. Specifically, this research will examine:

- What OLAP and data mining tools are most applicable for estimating future ER activity?
- What is an appropriate database design for storing ER historical data for OLAP and data mining application?
- What levels of data cleansing and data transformation are required to prepare historical ER Data for effective data mining?

- What methods are best for generating and validating ER activity models within a data mining environment?

C. METHODOLOGY

This research first documents the organization of the DoD healthcare system by reviewing available literature and conducting interviews with personnel involved with MTF ER management. This establishes the background of the source of the data used in this thesis. The next section of this research reviews existing literature on the topics associated with OLAP and data mining to develop a context for applying these technologies to the MTF ER application. This research then collects historical MTF ER data and stores it in a relational database for analysis by OLAP and data mining tools. Select OLAP and data mining tools then analyze the historical data. The final step evaluates the predictive ability of the models generated from the OLAP and data mining tools.

D. SCOPE AND LIMITATIONS

This thesis will specifically examine MTF ERs in the TRICARE Central Region of the United States. The Lead Agent Headquarters of the TRICARE Central Region provided the data for this thesis, thus limiting this research to only MTFs within the Central Region. As Chapter II will explain in detail, the TRICARE Central Region has a very large area of responsibility. The combined data set of Central Region's MTF ERs over two years has over 400,000 records. SPSS Clementine™ data mining software and Cognos PowerPlay™ OLAP software will be utilized in depth to analyze the ER data.

E. THESIS ORGANIZATION

This thesis is organized into eight chapters. Chapter I is an overview of the thesis. Chapter II provides an overview of the DoD healthcare system organization to include the flow of data through the system. Chapter III provides an overview of OLAP techniques and the OLAP software used in this research. Chapter IV provides an overview of data mining techniques and the data mining software used in this research. Chapter V gives an overview of data preparation and explains the data preparation used in this research. Chapter VI applies OLAP techniques to the historical MTF ER data and evaluates the results. Chapter VII applies data mining techniques to the historical MTF ER data and evaluates the results. The final chapter provides recommendations and conclusions.

II. DOD HEALTHCARE SYSTEM OVERVIEW

A. DOD HEALTHCARE SYSTEM ORGANIZATION

The DoD healthcare system provides medical care to all eligible members of the Armed Forces, their families, and other qualified beneficiaries throughout the world. The DoD Healthcare System uses a combination of military and civilian medical personnel and facilities to meet these medical needs. The DoD continuously faces the challenge of finding the right balance of military and civilian medical services. Active duty service members maximize the use of military medical facilities. Other eligible users of the DoD Healthcare System use military medical facilities on a space-available basis. Whenever military medical facilities or services are not available, beneficiaries receive medical services in civilian facilities.

1. DoD Medical Organization

The primary assistant and advisor to the Secretary of Defense, the Deputy Secretary of Defense, and the Undersecretary of Defense for Personnel and Readiness for all DoD healthcare policies is the Assistant Secretary of Defense for Health Affairs (ASD/HA). The ASD/HA's primary responsibility is to ensure that the DoD Military Health System (MHS) can provide medical support to the Armed Forces during all military operations. Additionally, the ASD/HA ensures that healthcare services are available for all eligible recipients of DoD Healthcare. The ASD/HA oversees TRICARE, the DoD's primary integrated healthcare management system. Figure 2-1 shows a high-level view of the DoD healthcare system organization. This figure

illustrates how the military medical structure and TRICARE are interconnected ("The Assistant Secretary of Defense for Health Affairs Responsibilities and Functions," 2000).

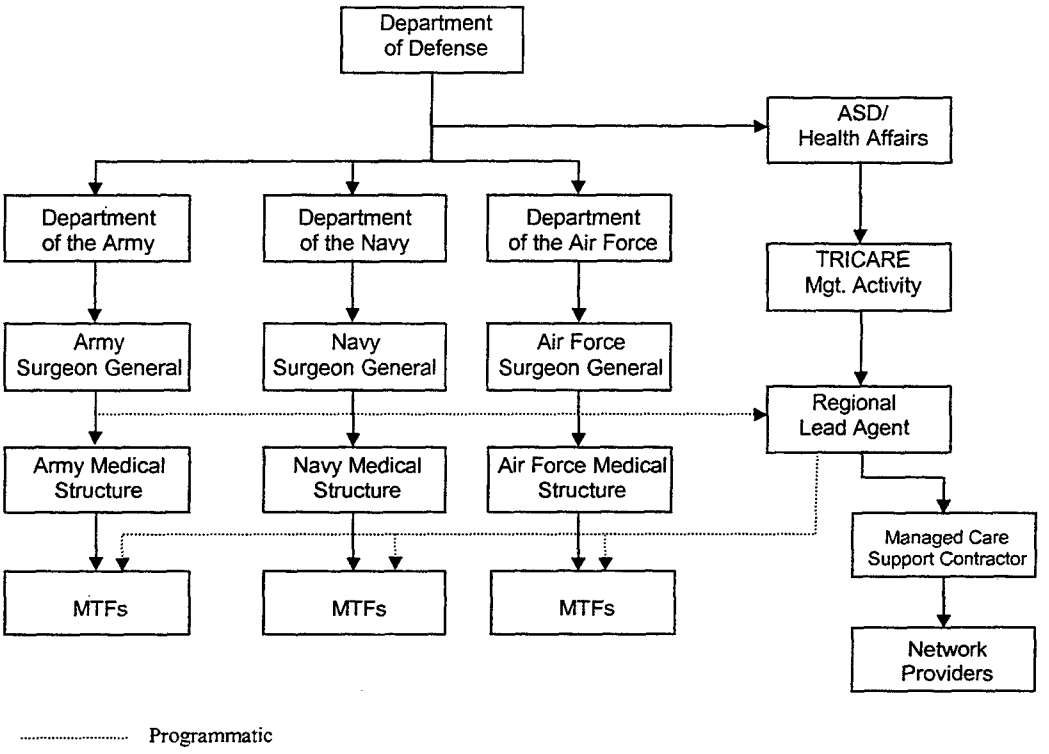


Figure 2-1. DoD Healthcare System Organization

2. U.S. Army Medical Organization

The U.S. Army's largest medical organization is the Medical Command (MEDCOM). The Army Surgeon General also serves as the MEDCOM Commander. The MEDCOM is further subdivided into Regional Medical Commands (RMCs). Each RMC is responsible for all Army MTFs that are located within its region. Figure 2-2 shows a high-level organization chart of the Army medical organization ("Introduction to the U.S. Army Medical Department," 2000).

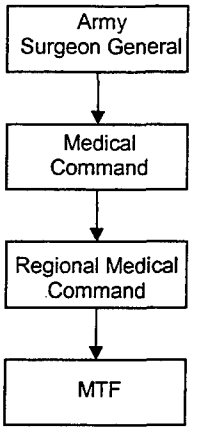


Figure 2-2. Army Medical Organization

3. U.S. Navy Medical Organization

The U.S. Navy's Bureau of Medicine oversees Healthcare Support Offices (HSOs). The HSOs are responsible for the Navy MTFs that are located within their region. The Navy has the unique requirement to provide medical support to its deployed surface fleet. This chapter will not discuss this aspect in detail because the data used in this research does not cover any of the Navy's deployed medical operations. Figure 2-3 shows a high-level organization chart of the Navy's land-based MTF medical organization.

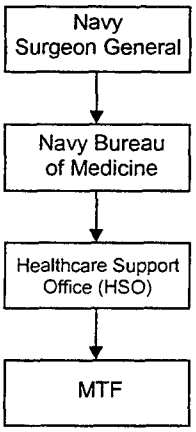


Figure 2-3. Navy Land-Based Medical Organization

4. U.S. Air Force Medical Organization

The U.S. Air Force integrates its medical units with its established Major Commands (MAJCOMs). Most Air Force MTFs are subordinate to an Air Wing, itself subordinate to a MAJCOM. This is different from the Army and Navy, which both align their medical units under a large medical headquarters. One exception to the Air Force medical organization chart that is relevant to this thesis is the 10th Medical Group. The 10th Medical Group operates the MTF at the Air Force Academy near Colorado Springs, Colorado. The Air Force Academy is a direct reporting unit, which means that it functions like a Major Command. Figure 2-4 shows a high-level organization chart for the Air Force medical organization ("The United States Air Force Medical Service Organization," 2001).

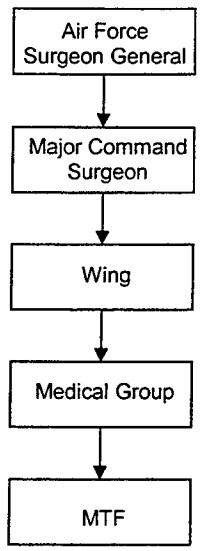


Figure 2-4. Air Force Medical Organization

5. TRICARE

TRICARE is an integral component of the DoD healthcare system. It combines military medical resources with networks of civilian healthcare providers to ensure that all eligible beneficiaries of DoD healthcare services receive quality medical care.

The Assistant Secretary of Defense for Health Affairs (ASD/HA) oversees TRICARE. Figure 2-5 shows the high-level organizational structure of TRICARE. Directly under the ASD/HA is the TRICARE Management Activity (TMA). The TMA administers the TRICARE systems. The TMA manages TRICARE on a regional basis. Figure 2-6 shows the 15 DoD health service regions. A Lead Agent (LA) manages the TRICARE operations in each region.

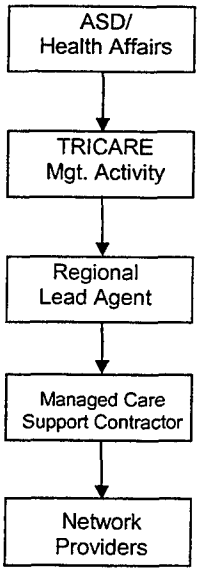


Figure 2-5. TRICARE Organization

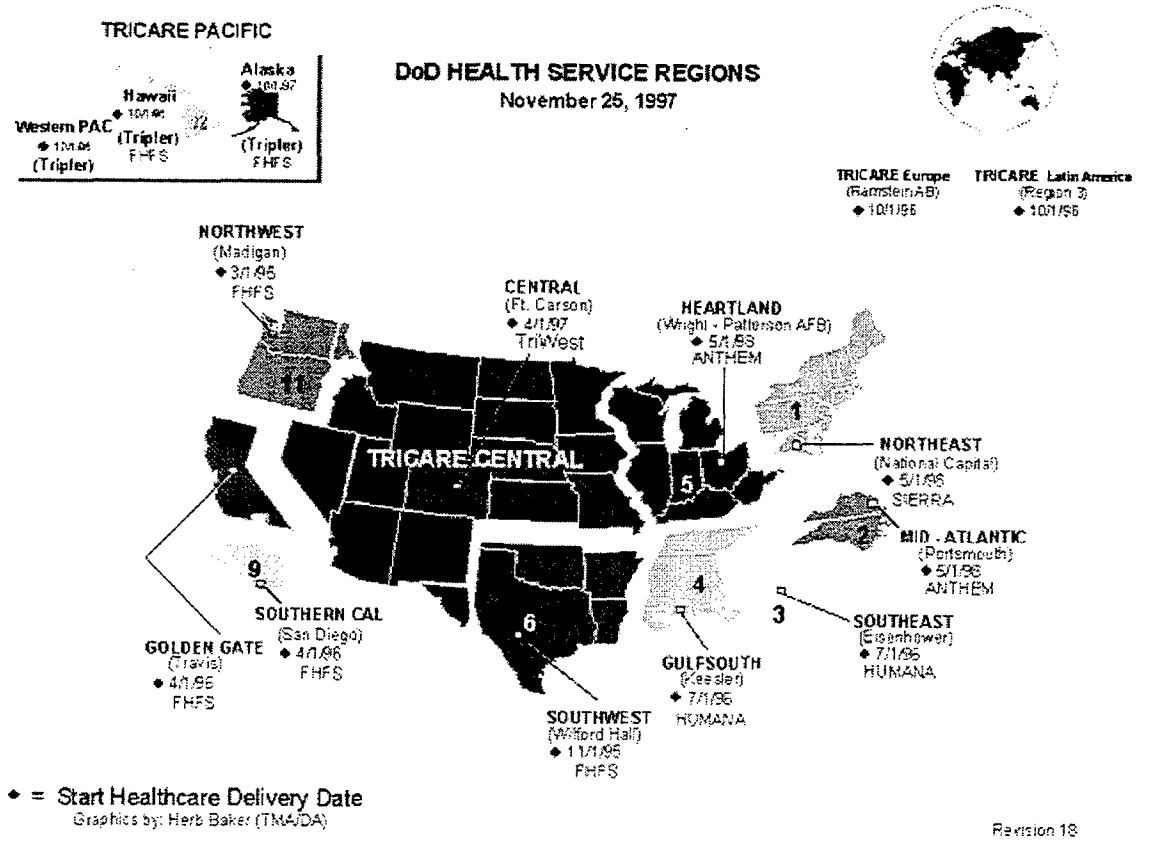


Figure 2-6. DoD Health Service Regions (From McNitt, 2000)

The LA is an active-duty officer from one of the Armed Services. The LA's Armed Service branch varies from region to region. The LA develops, monitors, and implements TRICARE programs within the region. The LA is additionally responsible for integrating TRICARE with the military medical systems within the region.

Each TRICARE region has one managed care support contractor that oversees the network of civilian medical service providers. Figure 2-5 shows the lines of authority in the TRICARE system. This figure does not illustrate the contractual relationships involved in the TRICARE system.

TRICARE offers three choices for healthcare: TRICARE Prime, TRICARE Extra, and TRICARE Standard ("What is TRICARE?," 2000).

a. TRICARE Prime

This option uses military medical treatment facilities (MTFs) as the primary source of healthcare. Civilian medical service providers augment the MTF healthcare. All active-duty service members are enrolled in TRICARE Prime.

b. TRICARE Extra

This option allows the beneficiary the ability to select a healthcare provider from the TRICARE Provider Directory. This option is more expensive than TRICARE Prime, but gives the beneficiary more flexibility.

c. TRICARE Standard

This option offers the greatest flexibility. It allows the beneficiary to select the authorized healthcare provider of his or her choice. It is normally the most expensive of the three options.

B. TRICARE CENTRAL REGION

TRICARE Central Region is the largest geographic health service region in the continental United States (CONUS). Figure 2-6 shows the geographic size of the Central Region. The Central Region covers all or part of 16 states and has military medical facilities from the Army, Navy, and Air Force. As of 30 SEP 99, the Central Region's eligible beneficiary population was 1,097,740.

1. Organizational Structure

The Lead Agent (LA) of TRICARE Central Region is an Army officer and is linked to the Army Surgeon General. The Central Region LA takes directives from the Army Surgeon General, but receives his or her direction from the TRICARE Management Activity (TMA) (McNitt Telephone Conversation, 2000). Figure 2-7 shows the relationship of the LA with the Army, TMA, and the 26 MTFs that are in the region.

2. Central Region Military Medical Treatment Facilities (MTFs)

Central Region contains one Navy MTF, seven Army MTFs, and 18 Air Force MTFs. Figure 2-8 shows the Central Region MTF locations. Of the 26 MTFs, 15 are designated as clinics and 10 MTFs are bedded facilities with emergency rooms.

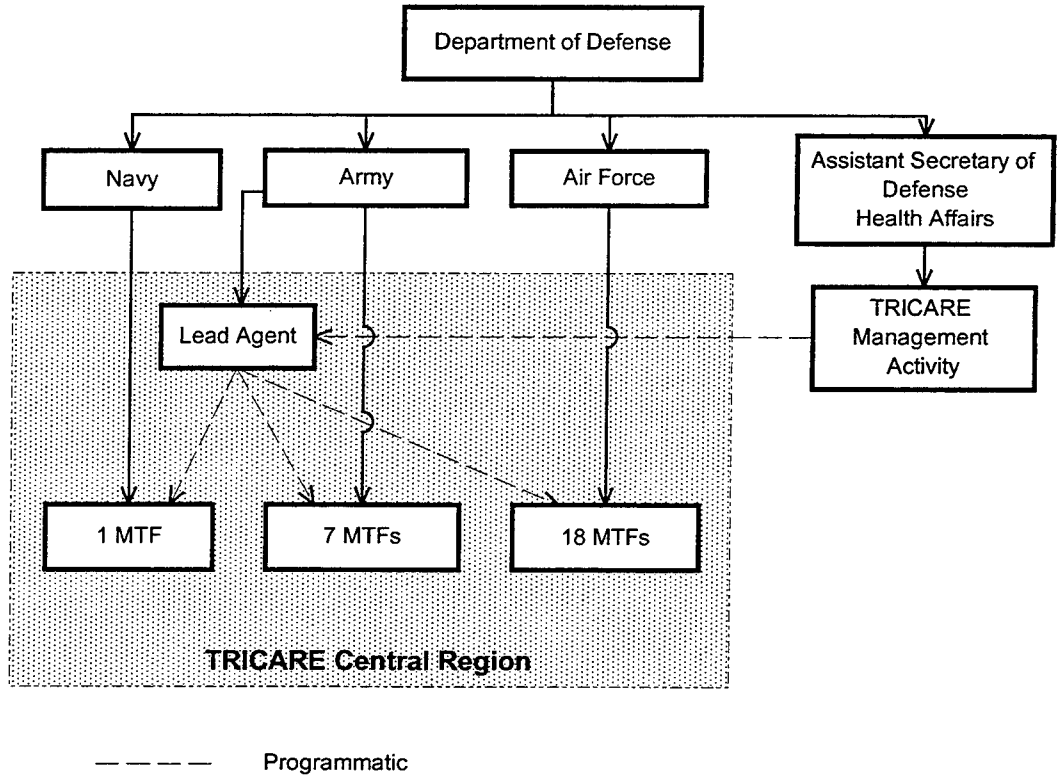


Figure 2-7. TRICARE Central Region Organization (From McNitt, 2000)

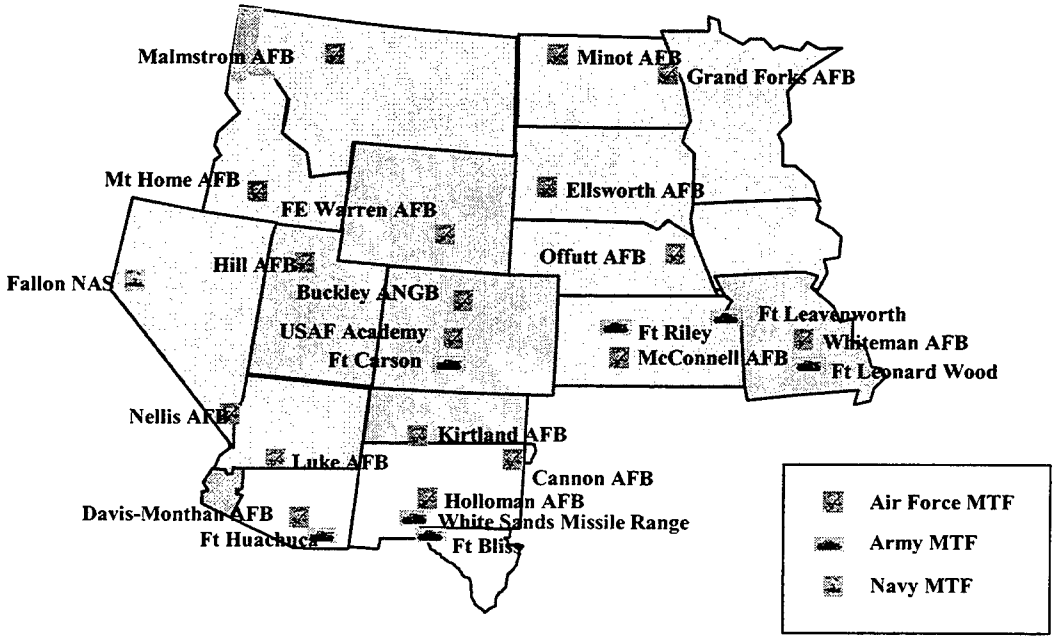


Figure 2-8. TRICARE Central Region MTF Locations (After McNitt, 2000)

3. Eligible Beneficiary Population

The Central Region's eligible DoD health service beneficiary population spans all of the Armed Services. Table 2-1 shows the eligible beneficiary population by service. Table 2-2 shows the breakdown of beneficiaries in Central Region enrolled in TRICARE Prime. These enrollment numbers are important because they include all active-duty service members and a large number of their dependents in Central Region. TRICARE Prime beneficiaries account for the majority of the MTF emergency room visits.

	Active Duty	All Others	Total	
Air Force	**87,707**	**429,873**	**517,580**	**47%**
Army	**71,156**	**335,132**	**406,288**	**37%**
Navy/Marine	**10,759**	**149,927**	**160,686**	**15%**
Coast Guard	**553**	**5,920**	**6,473**	**1%**
Others	**1,562**	**5,151**	**6,713**	**1%**
Total	**171,737**	**926,003**	**1,097,740**	

Table 2-1. TRICARE Central Region Eligible Beneficiary Population as of 30 September 1999 (From McNitt, 2000)

		Civilian	
	MTF	Network	Total
AD	**129,728**	**6,118**	**135,846**
ADD	**200,316**	**16,522**	**216,838**
NADD <65	**107,458**	**25,433**	**132,891**
NADD 65+	**3,733**	**0**	**3,733**
TOTAL	**441,235**	**48,073**	**489,308**

AD - Active Duty
ADD- Active Duty Dependent
NADD- Non-Active Duty

Table 2-2. Central Region TRICARE Prime Enrollment as of 31 March 2000 (After McNitt, 2000)

C. MILITARY MEDICAL TREATMENT FACILITIES (MTFs)

MTFs are the backbone of the Military Health System (MHS). Military medical professionals operate the MTFs, which vary in size and available medical services based on the military medical needs in an MTF's area of coverage. The controlling branch of service of the MTF is normally the same as the commander of the MTF's military installation. Although controlled by a single branch of service, each MTF provides medical care to all eligible beneficiaries based on priority and availability.

1. Clinics

Small MTFs operate as clinics. These MTFs are normally on small military installations, or function as a part of a network of clinics that feed into a larger MTF in the area. MTF clinics have a limited range of outpatient medical services and do not provide inpatient medical care.

2. Hospitals

Larger MTFs are hospitals that provide both inpatient and outpatient services. The range of medical services provided by the MTF is based on the military medical needs of the MTF's area of coverage.

3. Emergency Rooms

Many of the larger MTFs have emergency rooms (ERs). The MTF's area military medical needs determine the level of emergency medical service provided by the ER. The level of emergency medical services provided determines the organization, staffing, and equipping of the ER. ERs normally operate in conjunction with an ambulance

service. The extent of ambulance service is related to the level of emergency medical care provided by the ER.

MTF managers closely manage ER utilization. Too often, beneficiaries use MTF ERs out of convenience rather than necessity. This is a costly problem because ERs are more expensive to operate than primary care clinics. Many MTF managers have had great success in shifting non-emergent medical patients into primary care clinics, but continue to face the challenge of reducing the ER workload.

D. DOD AUTOMATED INFORMATION SYSTEMS

The MTF ER data used in this thesis is based on outpatient records. This section describes the primary DoD automated systems that feed outpatient data into the All Region Server Bridge (ARS Bridge). The ARS Bridge is the source of the data used in this research. Figure 2-9 shows the outpatient data flow from the patient visit to the ARS Bridge.

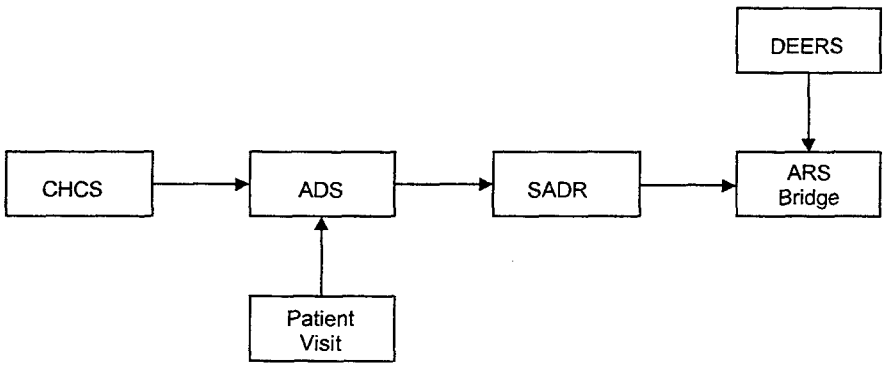


Figure 2-9. Outpatient Data Flow

1. Composite Healthcare System (CHCS)

CHCS supports clinical management in MTFs. CHCS is installed in more than 700 DoD Hospitals and clinics. CHCS supports medical appointment scheduling, inpatient records, outpatient administrative data, laboratory functions, radiology work, pharmacy activities, and other MTF activities. CHCS has ad hoc reporting capabilities and it interfaces with 40 other clinical and administrative systems. Through one interface, it supplies outpatient demographic and appointment data to the Ambulatory Data System.

2. Ambulatory Data System (ADS)

ADS provides detailed outpatient ambulatory data. ADS receives patient demographic and appointment data from CHCS. ADS combines this data with diagnostic and procedural data for each outpatient encounter. MTF database servers maintain the ADS data, providing medical professionals at the MTF with access to real time data. ADS transmits outpatient data to the DoD via the Standard Ambulatory Data Record.

3. Standard Ambulatory Data Record (SADR)

ADS generates and transmits an SADR to the DoD for each outpatient encounter. The SADR is a 353-character record containing patient demographics and clinical data. The DoD uses the SADR to populate outpatient data in multiple automated information systems.

4. All Region Server Bridge (ARS Bridge)

The ARS Bridge serves as an access point to DoD medical data. It contains inpatient records, outpatient records, TRICARE data, DEERS data, and other DoD

healthcare system data. Authorized DoD agencies download medical data from the ARS Bridge to perform data analyses.

5. Defense Enrollment Eligibility Reporting System (DEERS)

DEERS is a database of military sponsors, their dependents, and others who are covered by TRICARE. DoD information systems use DEERS to establish an individual's eligibility to receive DoD medical services. Some patient visit data extracted from the ARS Bridge originated in DEERS.

E. SUMMARY

The DoD healthcare system is a very complex organization. It has components from both civilian and military medical systems. One challenge the DoD faces is to find the right balance of civilian and military medical services to provide the best quality medical care to all beneficiaries of the DoD healthcare system. Optimizing the management of MTF ERs is a critical factor in achieving this balance.

Efficient ER management is not an easy task. Standard data analysis to improve MTF ER management is limited due to the complexity of multiple databases and the large amount of data available. Chapter III will examine on-line analytical processing and data mining, which are powerful data analysis tools that can uncover information that is beyond the capabilities of standard data analysis. This information could greatly enhance the ER manager's ability to optimize MTF ERs.

III. ON-LINE ANALYTICAL PROCESSING (OLAP)

A. INTRODUCTION TO OLAP

OLAP enables end users to analyze large amounts of multi-dimensional data on their desktop computers. OLAP uses custom-designed multi-dimensional data cubes to allow quick access to large amounts of data and to facilitate multiple views of the data set. The OLAP user interface is easy to use and designed to allow average computer users to quickly become productive using the software. OLAP processes data from many different sources, including relational databases, data warehouses, and data marts.

OLAP requires the end user to have a clear understanding of the business model of the environment in which he or she is working. The end user must also understand the data that will be explored using OLAP tools. The end user does not necessarily need to know how to construct a data cube. With input from the end users, a database professional can design and create data cubes that are tailored to meet an individual or group of users' needs. OLAP tools are designed to work with any data cube that meets that OLAP tool's supported formats. When multiple data cubes are available, the end user can select the data cube that best fits the question that he/she is trying to answer. For example, if a user wanted to analyze data only from a specific medical facility, then he/she would not need to use a data cube that contained data on all medical facilities. On large data sets, this could provide the user with a noticeable increase in performance.

B. ON-LINE TRANSACTION PROCESSING (OLTP) VERSUS OLAP

OLTP encompasses the daily operational database functions of an organization such as transaction and query processing, whereas OLAP accommodates the data analysis needs of an organization. OLTP and OLAP differ from each other in orientation, design, and function.

1. OLTP Characteristics

OLTP is customer-oriented and focuses on operational processing. An OLTP system can have thousands of users that include clerks, database administrators, and data analysts. OLTP systems maintain current data and concentrate on day-to-day operations. OLTP users can have read and write access and normally work with a small number of records. OLTP's database design is application-based and uses an entity-relationship (ER) model. Users view detailed data in two dimensions. A primary metric for OLTP is transaction throughput.

2. OLAP Characteristics

OLAP is market-oriented and focuses on information processing. An OLAP system typically accommodates hundreds of users that include managers, executives, and analysts. OLAP systems maintain historical data and focus on long-term information and decision support. Users can access millions of records and have mainly read-only privileges. Users view data in summarized, multi-dimensional views. OLAP's database design is subject-oriented and normally follows a multi-dimensional data model. Primary metrics for OLAP systems are query throughput and response time. (Han and Kamber, 2001, pp. 42-43).

C. DATA CUBES

OLAP accesses data through data cubes. A data cube is a data structure that stores data in arrays to produce multiple dimensions. Data cubes are tailored to a particular business model in order to meet an individual user's or group of users' needs. A data cube uses multiple dimensions and measures to form its data structure. Figure 3-1 shows a sample three-dimension data cube for emergency room (ER) visits. The data cube's three dimensions are time, facility, and patient age. It uses the number of ER visits as a measure of performance. The number '150' in this data cube, for example, represents the number of ER visits by 11-21 year-old patients to the medical facility at the U.S. Air Force Academy during the first quarter of 1998. Data cubes are normally based on one of three data structure schemas: the star schema, the snowflake schema, or the fact constellation schema.

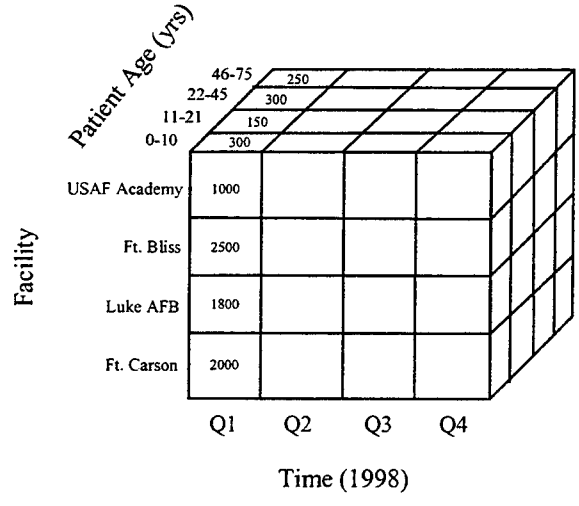


Figure 3-1. Emergency Room Visits Data Cube (Values for Demonstration Only)

1. Star Schema

The star schema is the least complex of the three data structure schemas. Figure 3-2 shows an example of the star schema. The star schema uses a central fact table, which is non-redundant and contains keys to each dimension table. The fact table maintains measure data that is used in OLAP data views to analyze performance. Each dimension can have only one dimension table. This restriction can introduce data redundancy. For example in Figure 3-2, data redundancy will occur in dimension tables if the addresses of two facilities contain the same state.

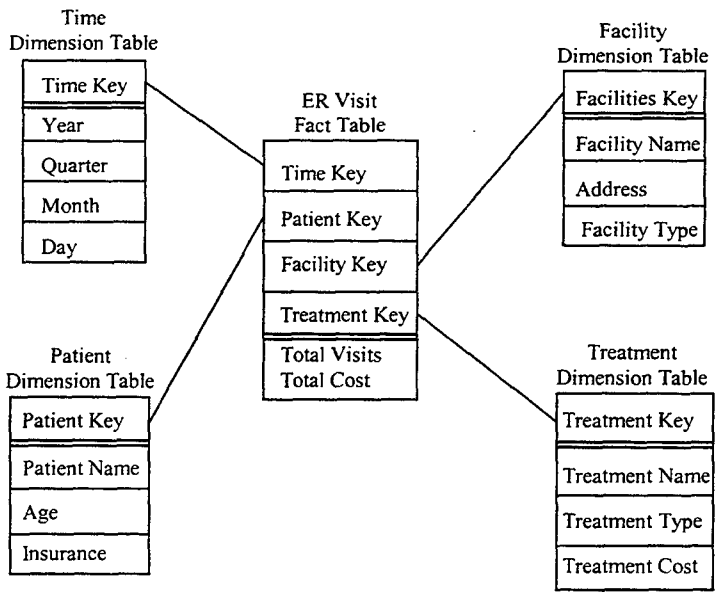


Figure 3-2. Star Schema

2. Snowflake Schema

The snowflake schema is an extension of the star schema. Dimension tables are added to the star schema to reduce data redundancy. In Figure 3-2, an address dimension

table and an insurance table are added to the star model in Figure 3-1 to reduce data redundancy in the facility and patient dimension tables. Reducing data redundancy saves storage space and reduces the size of individual dimension tables. Although it saves storage space, the snowflake schema reduces system performance because of the additional joins that must be navigated to access data. Most users are more concerned about performance than saving storage space, so the star schema is normally favored over the snowflake schema.

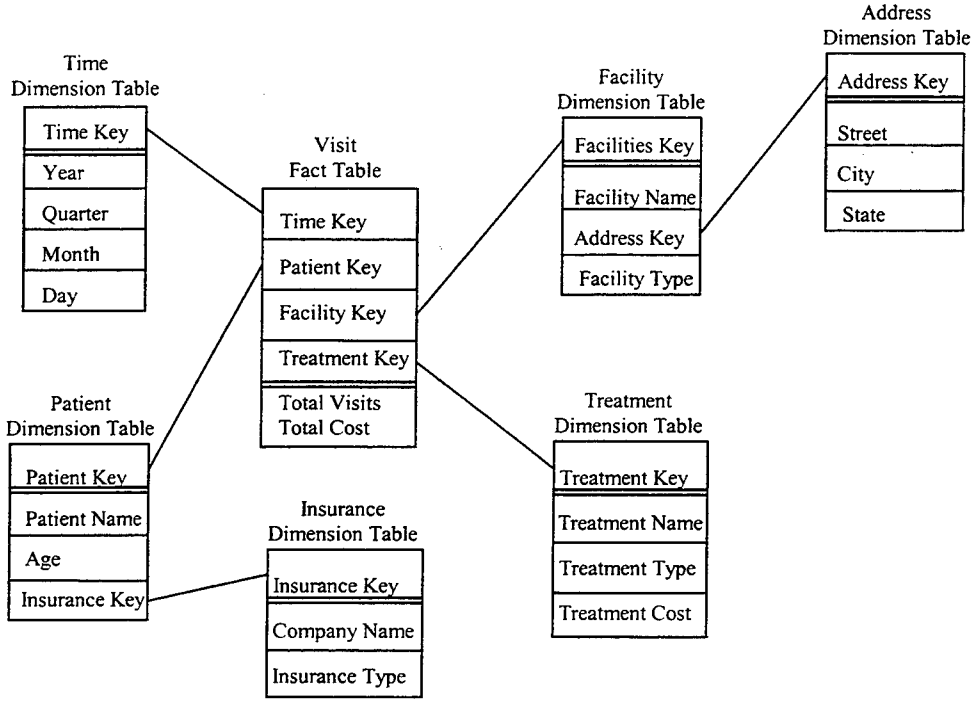


Figure 3-3. Snowflake Schema

3. Fact Constellation

The fact constellation schema is a collection of star schemas. Figure 3-4 shows an example of the fact constellation schema. It uses multiple fact tables that can share

dimension tables. The fact constellation schema's complexity can slow down performance; however, some advanced applications may require multiple fact tables to share dimension tables.

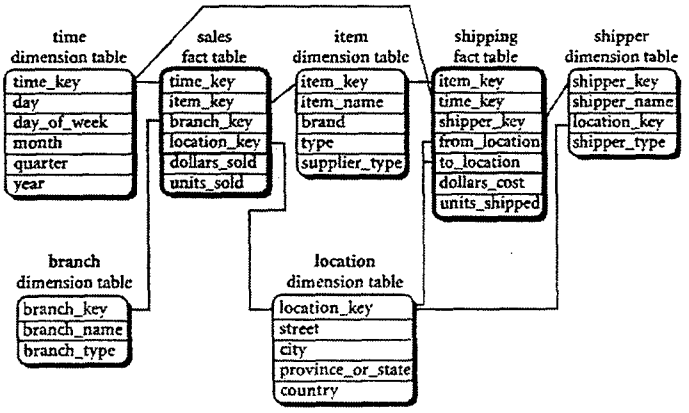


Figure 3-4. Fact Constellation Schema (From Han and Kamper, 2001, p. 51)

D. OLAP OPERATIONS

1. Drill-Down

Drill-down allows the user to navigate from summary data to the more detailed data that were aggregated to produce the summary data. The drill-down operation steps down to a lower level category of a particular dimension. For example, a drill-down operation of a time dimension may navigate through year, quarter, and month categories. More dimensions can be added to the data cube to facilitate drill-down needs of the user.

2. Drill-Up (Roll-Up)

The drill-up operation is the reverse of the drill-down operation. Drill-up operations produce different levels of aggregation of detailed data by navigating up category levels of a given dimension or by removing one or more dimensions from a data

view. A drill-up operation for a time dimension may navigate through month, quarter, and year. Another drill-up example involves an emergency room (ER) visit data cube that contains a time and a facility dimension. If the facility dimension were removed, the cube would show total ER visits for all facilities in a given time category.

3. Slice and Dice

Slicing singles out a specific category level of one dimension, creating a 'slice' of that dimension. Figure 3-5 shows a slice of Figure 3-1 for all 1998 ER visits to the USAF Academy. Dicing selects two or more dimensions to create a subcube based on those dimensions. An example of applying dicing to the data cube in Figure 3-1 is a subcube that contains all 1998 ER visits and age groups for the facilities at the USAF Academy and Fort Bliss.

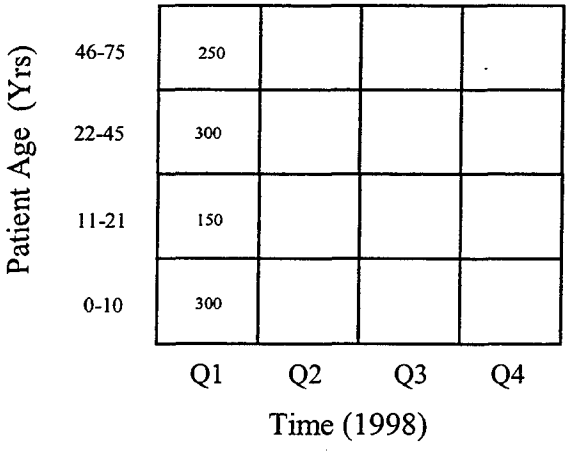


Figure 3-5. Data cube slice for USAF Academy ER visits (Values are arbitrary)

4. Pivot

The pivot operation presents an alternate view of the data by rotating the data axes. For example, in a 2-D table view, row and column data would exchange positions. The pivot operation can also be applied to 3-D cubes and other views.

5. Drill-Across

The drill-across operation allows the user to view data across more than one fact table. A drill across example using the fact constellation schema in Figure 3-4 would be a query that extracts data through the sales and shipping fact tables.

6. Drill-Through

The drill-through operation drills through a data cube to view detailed data that resides in a relational database. Drill-through uses standard query language (SQL) to access the data in the relational database (Han and Kamper, 2001, pp. 60-61). For example, if an ER manager wanted to see the exact demographics of patient ER visits for broken arms, he or she could drill through a data cube to look at the individual records in the relational database that were aggregated to form the summary OLAP data.

E. TYPES OF OLAP

1. Relational OLAP (ROLAP)

ROLAP systems are designed to work with relational database management systems (RDBMSs). ROLAP eliminates the need to create a multi-dimensional data structure because it uses metadata, which is data about data, to create the data structure. It enables the user to view two-dimensional data in multi-dimensional views through the use of the Star Schema data structure. Figure 3-6 shows an example of a ROLAP system.

ROLAP uses standard query language (SQL) to communicate with the RDBMS and extract the data that it needs to create multi-dimensional data views. ROLAP tends to be more scalable than other OLAP technologies because of the lower complexity of its data structures. However, ROLAP's lower complexity data structure results in slower data access times due to the series of joins that must be navigated to process data (Berson and Smith, 1997, p. 254).

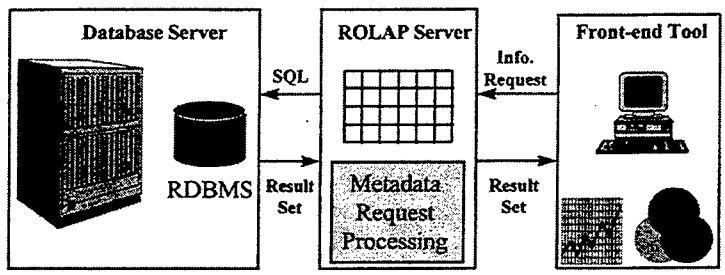


Figure 3-6. ROLAP System (From Berson and Smith, 1997, p. 254)

2. Multi-Dimensional OLAP (MOLAP)

MOLAP systems interface with array-based multi-dimensional data. MOLAP can map multi-dimensional views directly to a data cube array structure. This allows fast data access, greatly increasing MOLAP's performance over ROLAP. Figure 3-7 shows an example of a MOLAP system. The ability to handle arrayed data enables MOLAP to maximize data compression technology. MOLAP stores dense data subcubes in quick access arrays, while it compresses sparsely populated subcubes to save storage space. Figure 3-8 shows MOLAP's storage space advantage over ROLAP. The areas of the circles in figure 3-8 represent the data set sizes required for each system. This figure also displays how MOLAP outperforms ROLAP at the expense of greater complexity.

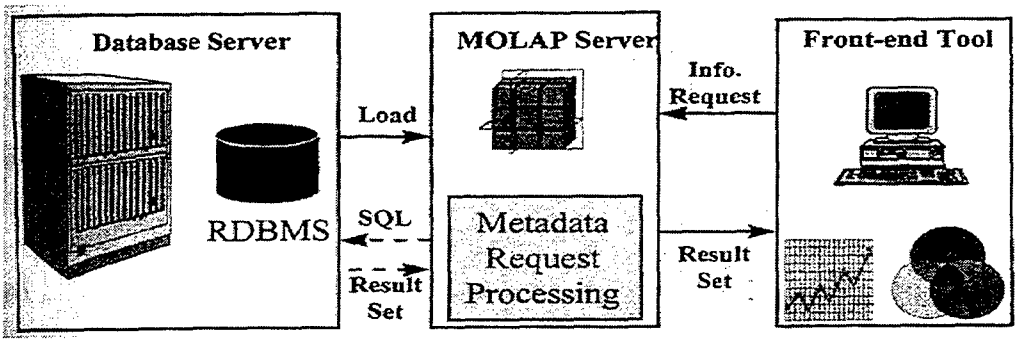


Figure 3-7. MOLAP System (From Berson and Smith, 1997, p. 253)

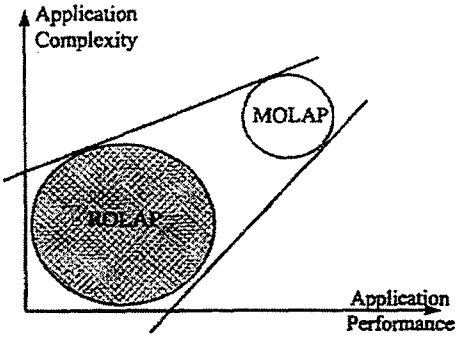


Figure 3-8. Performance, Complexity, and Data Set Size for MOLAP and ROLAP (From Berson and Smith, 1997, p. 252)

3. Hybrid OLAP (HOLAP)

HOLAP systems use a combination of ROLAP and MOLAP technologies to exploit the advantages of each. A HOLAP system can store large volumes of detailed data in a relational database environment, while maintaining pre-calculated summary data in a MOLAP environment. This structure would allow the user to access detailed data without requiring a complex multi-dimensional data storage structure. This lower complexity gives the HOLAP system increased scalability to more easily handle larger

data sets than an MOLAP system. The HOLAP system maintains MOLAP's rapid access to summary data, giving the user quick response to data inquiries. Figure 3-9 shows an example of an HOLAP system.

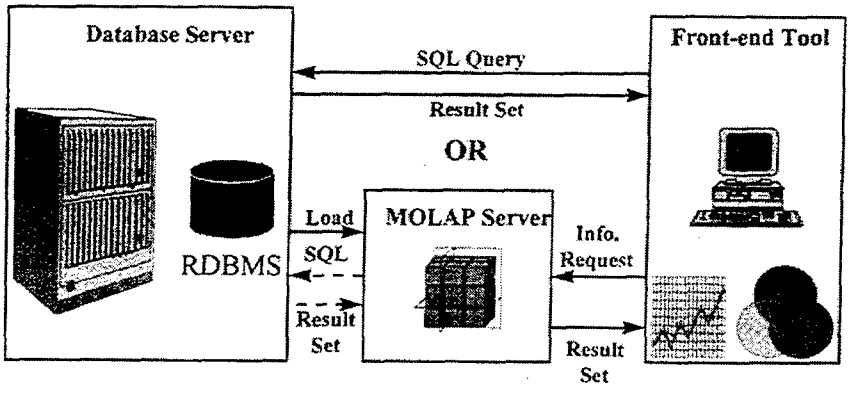


Figure 3-9. HOLAP System (From Berson and Smith, 1997, p. 255)

F. COGNOS POWERPLAY™ OLAP SOFTWARE

This research uses Cognos PowerPlay™ Version 6.6 OLAP software in a Microsoft NT Workstation operating environment. The primary reason for selecting this software is its availability. PowerPlay™ is a recognized leader in OLAP software. It is one of many products in the Cognos' line of data management and analysis software. PowerPlay™ is an HOLAP tool, meaning that it can interface directly with relational databases and with multi-dimensional data cubes. The PowerPlay™ OLAP software package includes PowerPlay™ for Windows™, PowerPlay™ for Excel™, PowerPlay™ Web, and PowerPlay™ Transformer. This research uses PowerPlay™ for Windows™ and PowerPlay™ Transformer.

PowerPlay™ Transformer creates the custom data cubes called PowerCubes that enable PowerPlay™ to view multi-dimensional data. Transformer can input data from ASCII files, relational databases, spreadsheets, and custom Cognos data formats. Transformer allows the user to structure the PowerCubes to meet the organization's data analysis requirements. The user selects the dimensions, performance measures, and data sources for the PowerCube. Transformer allows the user to create special categories within the dimensions, design custom drill-down paths, and establish drill-through capabilities to external data sources. Transformer has visual tools that display the hierarchy of dimensions and drill-down paths. PowerPlay™ for Windows™ is the end user tool that enables data exploration and analysis. It interfaces with Transformer PowerCubes as well as data cubes from other major vendors. PowerPlay™ for Windows has two main components: Explorer and Reporter. The main difference between these components is that Explorer views are dynamic while Reporter views are normally static. In Explorer, the data view changes as the user changes categories; however, each row and column in a data view may only contain one category from a single level of a single dimension. In Reporter, unless the user creates subsets, the data views do not change after the user selects different categories. The user has more freedom to create custom data views in Reporter than in Explorer. (Cognos, 2000, p.35-36).

PowerPlay™ for Windows™ supports many different OLAP techniques. While exploring the data sets, the user can drill-down and roll-up different levels of a dimension as well as drill-through to other data sources such as a query, a report, or another data cube. Other supported OLAP techniques include the slice and dice, and pivot operations.

PowerPlay™ for Windows™ allows the user to view data in pie charts, bar charts, line charts, and multiple nested charts. The user can select dimension levels to filter the data view. For example, the user could select the dimension levels that would only show the active duty Army service members that visited the MTF ER at Ft. Carson during February 1999. PowerPlay™ data views display one measure at a time, such as cost, revenue, or margin percent. The user can view different numbers in the data view by selecting a different measure.

PowerPlay™ users can examine local data cubes or access remote data cubes. Users can access remote data cubes from servers within their organization's local area network (LAN) or from the Internet if the data cube is stored on a web-enabled server running software like PowerPlay™ Web. This means that users could access remote cubes from anywhere that has Internet access (see Berndt, 2001, for more information on client and server-based OLAP services).

G. SUMMARY

OLAP is designed to enable managers, executives, and analysts to quickly explore large quantities of data and extract useful information. The end user OLAP interface is not complex, reducing the learning curve required to effectively use the software. However, all users of OLAP tools must understand the business model and the data of their organization in order to realize the benefits of OLAP. Chapter VI will detail the application of PowerPlay™ OLAP software to the MTF ER data, showing the

Transformer PowerCube design and PowerPlay™ for Windows techniques used to explore the data set.

IV. DATA MINING

A. INTRODUCTION TO DATA MINING

Today's information technology allows organizations to rapidly accumulate enormous volumes of data. However, much of the value of this data will go unrealized without a set of tools that can navigate through vast amounts of data to uncover its hidden knowledge. Data mining provides organizations with the capability to find and extract the knowledge that is buried in their massive databases and data warehouses. Data mining not only discovers trends and patterns in data, but also gives meaning to the discoveries and predicts what future data values will be. One author compares data mining to database processing, stating that database processing organizes and stores data according to semantics about the data, while data mining discovers the semantics of the data (Lin, 1999).

Data mining must be closely tied to the organization's business model. A common misconception is that data mining is a magic software package into which the user dumps huge amounts of data and which extracts patterns and solutions to business problems on its own. Data mining is an interactive process that requires the input and guidance from a user that fully understands the organization's business process (Introduction to Clementine™, 1999). With the guidance of the user, data mining can use its many tools to transform formerly meaningless data and patterns into knowledge. The results of data mining could be discoveries about the data, rules, or predictive models (Kasif, and others, 1999).

Data mining is not just a single discipline; it is a combination of disciplines that work together to extract knowledge. These disciplines include, but are not limited to, database technology, statistics, visualization, and machine learning. Data mining systems are often categorized by: (Han and Kamber, 2001, pp. 29-30)

- The kinds of databases mined, such as relational or object-oriented;
- The kinds of knowledge mined based on the granularity and level of abstraction of the knowledge;
- The kinds of techniques used, such as statistics or neural networks;
- The functional area applied to, such as finance or telecommunications.

B. DATA MINING COMPARED TO OTHER TECHNOLOGIES

1. Data Mining Versus Traditional Statistics

Traditional statistics is a key component of a data mining system. An area where data mining and traditional statistics differ is that traditional statistics are limited to manually created hypotheses, while data mining can both use manually created hypotheses and also create its own hypotheses to test automatically (Hogl, and others, 2001). In other words, traditional statistics can only test questions that are supported by the statistics model built by the user. Data mining is not limited to a single statistics model and can discover and test questions that are beyond the scope of the established statistics models.

Traditional statistical analysis normally requires a professional statistician to perform and understand. Data mining gives the user a means to use powerful tools of statistical analysis and visualize and understand the results. Data mining automates much

of the statistical analysis process and allows non-statisticians to analyze data and understand the results. This is not to say that any user can successfully use and understand data mining. Because of the complexity of data mining tools, users can require significant training before becoming productive.

2. Data Mining Versus On-Line Analytical Processing (OLAP)

OLAP tells the user what has been going on in a business. Data mining can tell the user what is going to happen next (Berson and Smith, 1997, p. 334). Data mining has the ability to learn from historical data and make predictions about future data. OLAP is normally limited to summarizing historical data. OLAP reveals larger patterns in data, whereas data mining can sometimes turn even a minor pattern into a significant finding. With both tools, the user must understand the organization's data and business model.

The user ability requirements of a data mining system are much greater than for an OLAP system. If the user does not have to create a data cube, most OLAP tools have short train-up times for the average user to become productive. Data mining systems have a much steeper user learning curve than OLAP systems because they require more user training and user understanding of data mining methods.

C. DATA MINING METHODS

Data mining methods can be characterized as supervised or unsupervised (learning methods). Supervised learning methods receive input from the user and provide a numeric or categorical response as an end state. Unsupervised learning methods do not

receive direction from the user and do not produce a response as an end state. Data mining is an evolving field with new methods continuously under development. This section of Chapter IV describes four of the common data mining methods used today.

1. Clustering

Clustering methods group data objects with similar attributes together. Clustering is an intuitive process for humans, who naturally group like items together. Clustering is an unsupervised learning method because there is no end state to guide the algorithm. Clusters are formed such that data objects within a cluster are similar to the other objects in the cluster, but dissimilar to the objects in other clusters. Data clusters seek to maximize intraclass similarities and to minimize interclass similarities. Clustering can, for example, reveal sparse and dense areas within a data set. Figure 4-1 shows an example of this 'density-based' clustering. For more details on different types of clustering see Han and Kamber, p.346 (Han and Kamber, 2001, pp.25, 335).

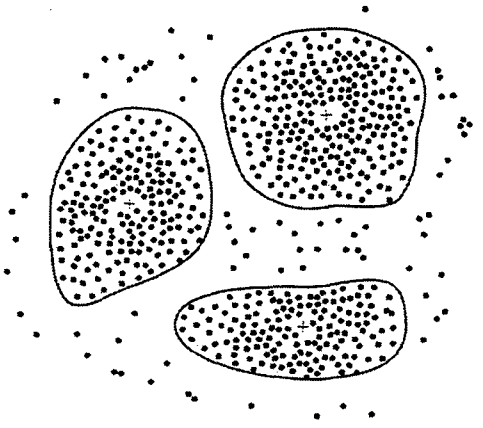


Figure 4-1. Density-based Clustering Example (From Han and Kamber, 2001, p. 26)

2. Classification

Classification takes observations, each of which have a class attribute, and constructs a model to predict the class of the new data. Classification is a supervised learning method because the classification model is told to what classes the observations belong. Classification uses the predefined class information to classify records for which the data class is unknown. For example, a classification model may be designed to classify a loan applicant as high, medium, or low risk. The classification process works by creating a model to classify data based on the predefined data classes. A sample data set for which the classes are known is used to train the model. The learned model is then tested to verify its accuracy using another data set with known data classes. The test data set is different from the training data set to ensure that the model applies to a general range of data sets and is not overfit to the training data set (Han and Kamber, 2001, pp. 279-281).

3. Association Rules

Association rules seek to develop rules that produce a conclusion based on a set of existing conditions. Discovering association rules is an unsupervised learning method because the user does not specify what to look for. Association rules uncover which items, such as attributes, events, purchases, etc, tend to occur together (Introduction to Clementine™, 1999, p. 7-6). For example, an association rule could state that if a military officer is a student at the Naval Postgraduate School and studies Information Systems Technology, there is a 98% chance that he or she owns a personal computer. Discovering association rules from large data sets is especially useful to organizations

trying to determine customer shopping habits. This is often referred to as shopping basket analysis.

Many of the rules that association discovers could be determined manually by using data visualization techniques. The advantage of using association rule algorithms is that they automatically pull out association relationships with great speed and they can search for patterns anywhere in the data set. A drawback of association rules is that they have the potential to create very large search spaces, causing a very slow search. (Introduction to Clementine™, 1999, p. 7-7).

4. Outlier Analysis

Outlier analysis identifies data that are inconsistent or grossly different from the other data within a data set. Outliers can be caused by data entry errors or by natural data variability. Some data mining algorithms simply eliminate outlier data, possibly losing important information. Outlier analysis can be beneficial in many areas such as fraud detection. Detecting outliers can be challenging when the data is hidden in seasonal or other cyclic changes. Outliers in time-series data could appear normal in time snapshots, when in fact they are not consistent with the other data over time (Han and Kamber, 2001, pp. 381-382).

D. DATA MINING ALGORITHMS

This section describes four of the common algorithms used in data mining methods today. Data mining practices use many variations of these algorithms as well as

other algorithms. The data miner selects the algorithm or group of algorithms that best match the organization's data mining goal.

1. Decision Trees

Decision trees use the attributes of known classes to predict the value of a specified class. The example in Figure 4-2 classifies customers in the cellular telephone industry into those who renew their phone contracts and those who do not renew their phone contracts (Berson and Smith, 1997, pp. 351-352). As this example shows, decision tree models are simple in design, making them easy to understand.

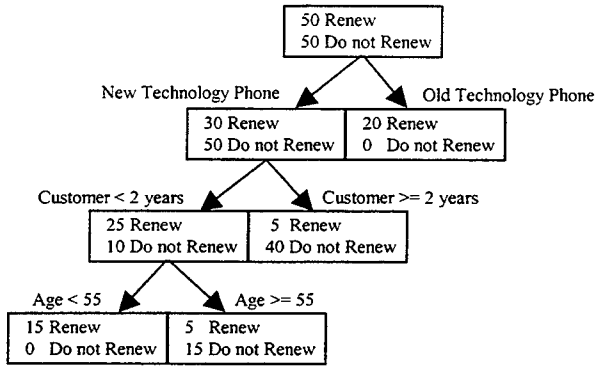


Figure 4-2. Cellular Phone Customer Decision Tree
(After Berson and Smith, 1997, p. 351)

Decision tree algorithms normally automate hypothesis generation and validation more than other data mining algorithms. Decision trees can be used for both data mining exploration and prediction applications. Although simple and effective, decision trees do have limitations. Some predictive applications, such as a simple regression prediction, can be solved more easily and quickly using other methods. However, decision trees are very useful for more complex problems.

Decision trees are a supervised learning technique. The quality of decision tree results is highly dependent on the decision tree asking the right question at the right branch of the tree. Decision tree algorithms examine the different possibilities at each branch of classifying the data set into partitions that best fits the design of the algorithm. For example, the classification tree algorithm built into the S-Plus statistical package tries all possible questions and selects the question with the best results (S-PLUS 2000 User's Guide, 2000). Other decision trees may use heuristics or random selection to determine which questions to use (Berson and Smith, 1997, pp. 355-357).

2. Neural Networks

A neural network is a computer-based model originally designed to mimic the human brain. It uses a series of nodes with adaptive weights to learn and improve the network over time. Figure 4-3 shows an example of a neural network. Neural networks are very powerful prediction techniques, but they produce models that are so complex that even experts have difficulty understanding them. Neural networks, unlike some data mining techniques, require preprocessing of data to be effective. They also require long training times, which would not make them a good choice for time-critical applications. Neural networks can process and adapt to noisy data. They also can classify patterns that the network has not been trained on (Han and Kamber, 2001, p. 303).

A common algorithm by which the values of the weights are determined is back-propagation. Back-propagation learns from a multi-layer feed-forward network. Figure 4-3 shows an example of a multi-layer feed-forward network. The system feeds inputs into a set of nodes creating the input layer. The weighted outputs of these nodes are then

fed into the hidden layer. The network can use multiple hidden layers. The output layer consists of the weighted outputs of the last hidden layer. The output layer sends the values it receives to the system as the resultant predictive values. The network learns through the iterative processing of samples, adjusting the node weights and comparing the results with samples with known predictive values. The neural network continues to adjust its node weights to reduce, for example, the mean-squared error between the network's prediction and the sample. It makes these adjustments backwards from the output layer through the hidden layers (Han and Kamber, 2001, pp. 303-305).

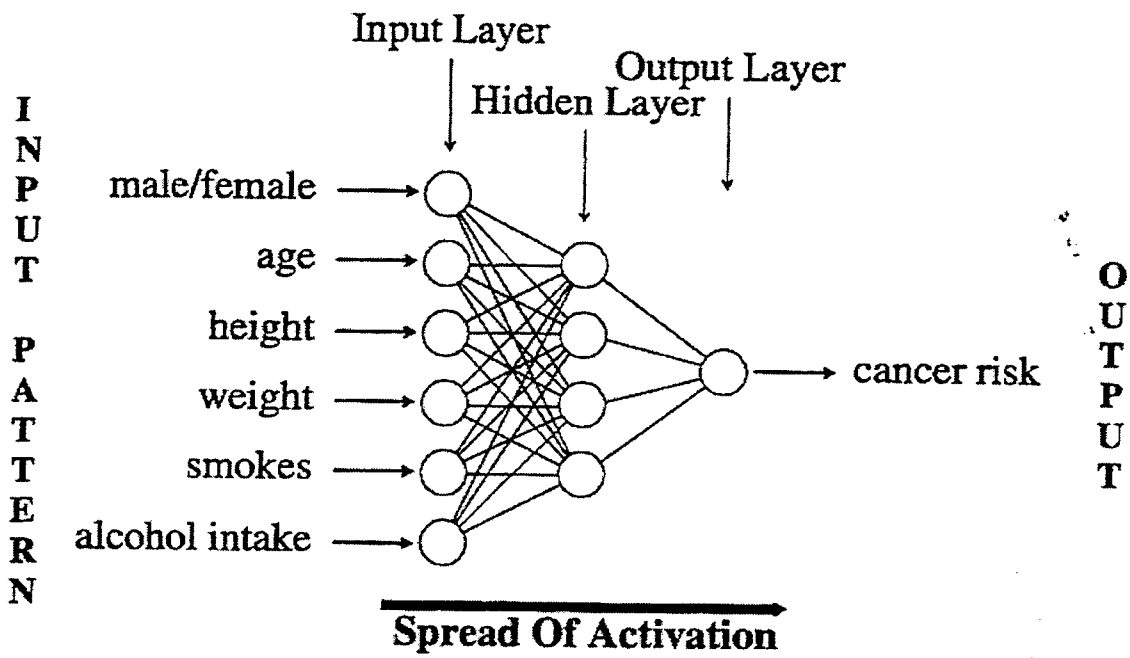


Figure 4-3. Neural Network that Predicts the Risk of Cancer
(From Clementine™ User's Guide Version 5, 1998 , p. 7)

3. Genetic Algorithms

Genetic algorithms are based on the principles of natural evolution and survival of the fittest. In the context of neural networks, genetic algorithms start with random node weights to develop different networks. The networks that perform poorly are weeded out, while the networks that perform better survive. The algorithm will then make multiple reproductions of the best-performing networks applying a slight modification to each reproduction. The best-modified network's weights will then evolve as the basis of the surviving network (Han and Kamber, 2001, p. 315).

4. Nearest-Neighbor

Nearest-neighbor algorithms are similar to those used in clustering. These algorithms predict a data value by using the known predictive value of the historical data that most resembles the unknown data. For example, if the algorithm determines that the individuals whose attributes are closest to the test subject are 1990 graduates of the Management Information Systems (MIS) program at the University of Notre Dame, and if most of those had starting salaries in the $30-50,000 range, it would predict the unknown starting salary of the test subject to be in the $30-50,000 range.

E. SPSS CLEMENTINE™ DATA MINING SOFTWARE

This research uses SPSS Clementine™ Data Mining system software in a Microsoft NT™ workstation operating environment. The primary reason for this software's selection was its availability at the Naval Postgraduate School. Integral Solutions Limited (ISL) originally developed Clementine™; however, SPSS Inc. has now

acquired ISL (Han & Kamber, 2001, p. 461). Clementine™ is one of the leading data mining software packages today.

Clementine™ contains a range of data manipulation, statistical analysis, data visualization, machine learning, and modeling tools. Its machine learning and modeling techniques include association rule detection, rule induction (decision trees), clustering, and neural networks (Introduction to Clementine™, 1999, p. 1-2). It can import data from ASCII files, ODBC linked databases, or Clementine™-created cache files. Clementine™'s data manipulation tools can clean and format fields and records in preparation for data mining. Clementine™'s data visualization tools feature data tables, XY plots, histograms, data distribution, and web diagrams. Clementine™ can send output to flat files, ODBC linked databases, MS Excel spreadsheets, and custom SPSS files.

Clementine™ uses two different algorithms to discover association rules: Generalized Rule Induction (GRI) and APRIORI. GRI searches data to find the most interesting independent rules. GRI can search numeric and symbolic fields. APRIORI is slightly more efficient than GRI, however APRIORI can only search symbolic fields. Both algorithms produce a rule set that can be used to predict data values (Introduction to Clementine™, 1999, p. 12-2).

Clementine™ uses Kohonen networks to find clustering in a data set. Kohonen networks find patterns that share similar attributes and groups similar patterns together. Kohonen networks organize neurons into one or two dimensional grids or arrays. The difference between Kohonen networks and neural networks is that Kohonen networks do

not have an output layer. Instead, Kohonen networks produce an output grid of neurons called a Kohonen map. The Kohonen networks then place each data record into the neuron that is most similar to the record's attributes. While this method does not directly produce a rule set that can be applied to data, Clementine™ can generate a rule set from the resulting Kohonen network (Introduction to Clementine™, 1999, p. 7-5).

Clementine™ has two rule induction options: Build Rule and C5.0. Both of these algorithms enable the user to attempt to predict the value of a specific field by using the other fields of the data set. The result of these algorithms is a set of decision tree-type rules. Build Rule can predict symbolic or numeric fields, while C5.0 can only predict symbolic fields. C5.0 is more sophisticated than Build Rule and C5.0 will normally produce more accurate and simpler rules. C5.0 can adjust to mismatched data types and missing data, while Build Rule, by default, will generate an error when it encounters data noise (Introduction to Clementine™, 1999, p. 9-2).

F. SUMMARY

Data mining techniques are extremely powerful tools that can sometimes discover valuable information that is hidden in large data sets. Data mining enables users to perform statistical analyses without needing the statistical expertise of a professional statistician. However, data mining tools are not simple to use and require more training than OLAP systems for users to become effective. As is the case with OLAP, users of data mining tools must understand the business model and the data of their organization in order to realize the benefits of data mining.

V. DATA PREPARATION

With the number of human factors involved with the data entry process, there are almost always inconsistencies somewhere in a data set. These inconsistencies, when uncorrected or unaccounted for, can cause OLAP and data mining systems to fail or to output incorrect information. To obtain accurate results, it is critical to clean data sets and prepare the data in the required format before performing OLAP and data mining operations.

A. PROBLEMS WITH DATA SETS

1. Data Noise

Data sets usually contain imperfections such as mismatched data types and missing data. Data entry errors and data system processing errors account for much of this data noise. The analyst must evaluate the noise of a data set to determine its impact on the effectiveness of the data analysis. For example, a key field that is only 20% populated could have a significant impact on the effectiveness of the data analysis. In this case, people may reach inappropriate conclusions about certain associations, thinking the relationship holds for the entire data set when indeed it applies to a sample of only 20%. Fields specified as numeric cannot be fully processed if some of the data in the fields contain symbols instead of numeric data. In some cases, data analysis tools cannot handle data noise. When this is the case, the data noise must be corrected before the data set is useable by the data analysis tool.

2. Data Manipulation

Sometimes the format of the data in a data set will not facilitate the level of detail of data analysis that the user wishes to achieve. For example, if the user wants to analyze a data field by special groups that cannot be created by the data analysis tool, then the user must create and add the special groups to the data set first. Data manipulation can include changing data values, changing field data types, deleting selected records, and creating new fields. For example, a symbolic data field that contains only numbers can be converted to a numeric data type to enable calculations using that field. Another example of data manipulation is to delete records that contain too many empty data fields to be considered reliable.

B. MTF ER DATA PREPARATION

The raw MTF ER data set contains 28 fields and 415,424 records. (See Appendix A for a complete list of the data fields.) Each record reflects an individual patient's visit to an MTF ER in the TRICARE Central Region from the two-year period of 1 April 1998 to 31 March 2000. A database professional at the Lead Agent (LA) Headquarters of the TRICARE Central Region queried the ARS-Bridge to retrieve the raw MTF ER data. (See Chapter II for more information on TRICARE Central Region and the ARS-Bridge). Due to bandwidth limitations, the database professional had to prepare more than one query to retrieve the entire raw data set. A data analyst at the LA Headquarters then combined the raw data queries into one data file and used an algorithm to scramble the patient ID into an unrecognizable, but unique character string. The data analyst then

imported the resulting data file into a Microsoft Access™ database, which was used for this research.

1. Data Noise

The first step taken to assess the amount of noise in the data set was to run a quality check in the Clementine™ software to determine the degree of population of the data fields. Fifteen of the fields were populated in 99.9% or more of the records. Four fields were populated between 73-98%. The remaining fields were populated below 40%. Fortunately, the critical fields needed for this research fell into the 99.9% and above category. Appendix B contains the populations of the data fields of the raw MTF ER data set. All fields in the data mining models were populated at 99.9% or above because the data mining machine learning techniques required clean data. All but two fields used in the OLAP tool were populated at 99.9% or above. These two exception fields were populated at 89% and 98% and were added to the OLAP model to enable determination of the reason that these fields were less populated. Both data mining and OLAP tools can create a 'Blank' or 'Empty' field category that indicates the number of null values in a symbolic field. However, null values in numeric fields in the data set had to be replaced with a predetermined numeric value that identified it as null before applying either the data mining or the OLAP tools. Otherwise the tools generate a data mismatch error.

2. Data Manipulation

The 'Visit Date,' 'Age,' and 'Diagnosis Code' fields in the raw data did not meet the format needed to perform the analyses in this research. The 'Visit Day of the Week'

and 'Visit Month' fields were derived from the 'Visit Date' field to facilitate analyzing the day of the week and the month of the visit independent of the year and specified date. For example, the date '14 December 99' would generate 'Tuesday' and 'December' field values. These fields help reveal visit trends during specific days of the week and seasonal trends during certain months. The 'Age' field contains ages from 0-99. To reduce the challenge of analyzing a set of 100 different possibilities, the individual ages were placed into 10 age groups. The data set contains thousands of different diagnosis codes. To bring this number down to a reasonable level, the diagnosis codes were placed into the standard classes established by the International Classification of Disease Version 9 (ICD-9) (Rogers, 1994).

The 'Visits' field for each record should equal '1' because each record represents one ER visit. A relatively small number of records had a value of '2' or '3' in the 'Visits' field. There could be circumstances where the system set the number of visits to be greater than '1', but these circumstances were not known at the time of this research. Therefore, the 'Visits' field was changed to 'Raw Visits', maintaining the original 'Visits' value, and all values in the 'Visits' field were changed to '1'. Leaving both fields in the data set enables subsequent exploration of the reason that there is more than one visit in these single records.

Both PowerPlay™ and Clementine™ have functions that can easily handle adding the age groups to the data set. Extracting the additional date groups and diagnosis codes, changing the 'Visits' field to '1', and replacing the null numeric values was accomplished using the Make-Table-Query function within Microsoft Access™. This function allows

the user to reproduce an existing table with modification to existing fields and with the addition of new fields. Upon execution the Make-Table-Query function converts the results of the query into a new table in the database. The added fields are the result of built-in MS Access functions and Visual Basic subroutines. Figure 5-1 shows part of the design view of the Make-Query-Table used to create the new data set. Notice that the 'ICD9 Group' field expression calls a subroutine named 'func_ICD9_Group' that looks at the disposition code field, 'D1,' and determines which ICD-9 group it belongs to. Appendix C contains the Visual Basic code used to determine ICD-9 groups.

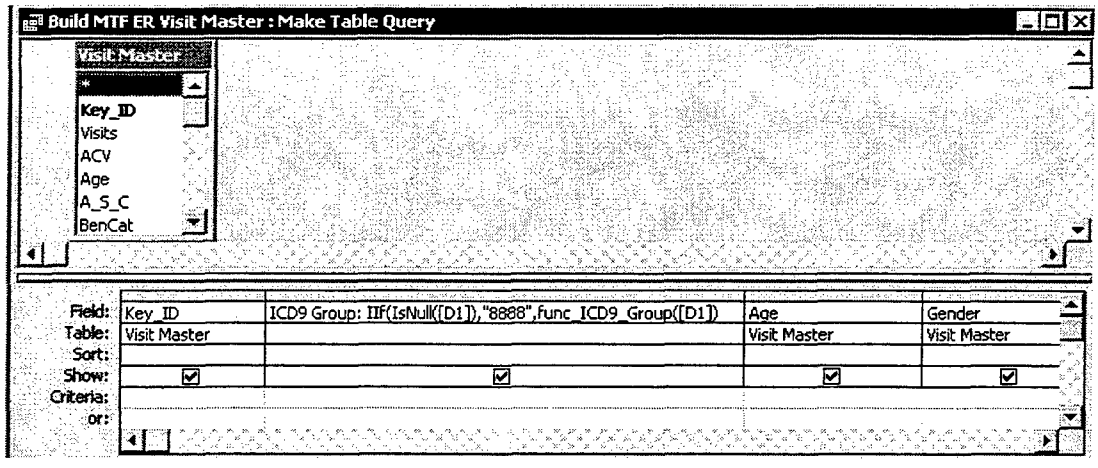


Figure 5-1. Microsoft Access™ Make Table Query

C. SUMMARY

Data preparation can be a very time-consuming task. Many data sets are not fully instantiated in every field of every record and are not free from data type mismatches. Sometimes data must be modified or new fields added to a data set in order to effectively

use a data analysis tool. The data analyst must determine what data preparation is necessary for the data set before performing data mining or OLAP data analysis.

VI. OLAP ANALYSIS OF MTF ER DATA

This chapter examines the capabilities of PowerPlay™ OLAP software and applies them to the MTF ER data set. The OLAP techniques mentioned here have been described in detail in Chapter III. Appendix A contains detailed descriptions of the data fields in the MTF ER data set mentioned in this chapter.

A. OLAP MODEL

1. Data Cube Design

The data structure for the MTF ER data cube follows the star schema (see Figure 6-1). The star schema for this model contains one fact table and eight dimension tables. The fact table is the MTF ER Visit Master Table, which contains detailed records of each ER visit, keys to each of the dimension tables, and measure data. For better clarity, the fact table in Figure 6-1 only shows the keys to the dimension tables and measure data. Each dimension table provides a drill down path within the data cube. This model contains two measures: 'Visits' and 'Visits Raw'. 'Visits' is a modified field that represents one visit for each record in the MTF ER Visit Master Table. 'Visits Raw' is the original, unmodified 'Visits' data in which a small number of records contain a value of '2' or '3'. Chapter III gives more details on measure data. The years used in this OLAP tool are as follows:

- 1999 = 1 April 1998 - 31 March 1999 (e.g. Q2 1999 = April 1998-June 1998)
- 2000 = 1 April 1999 - 31 March 2000

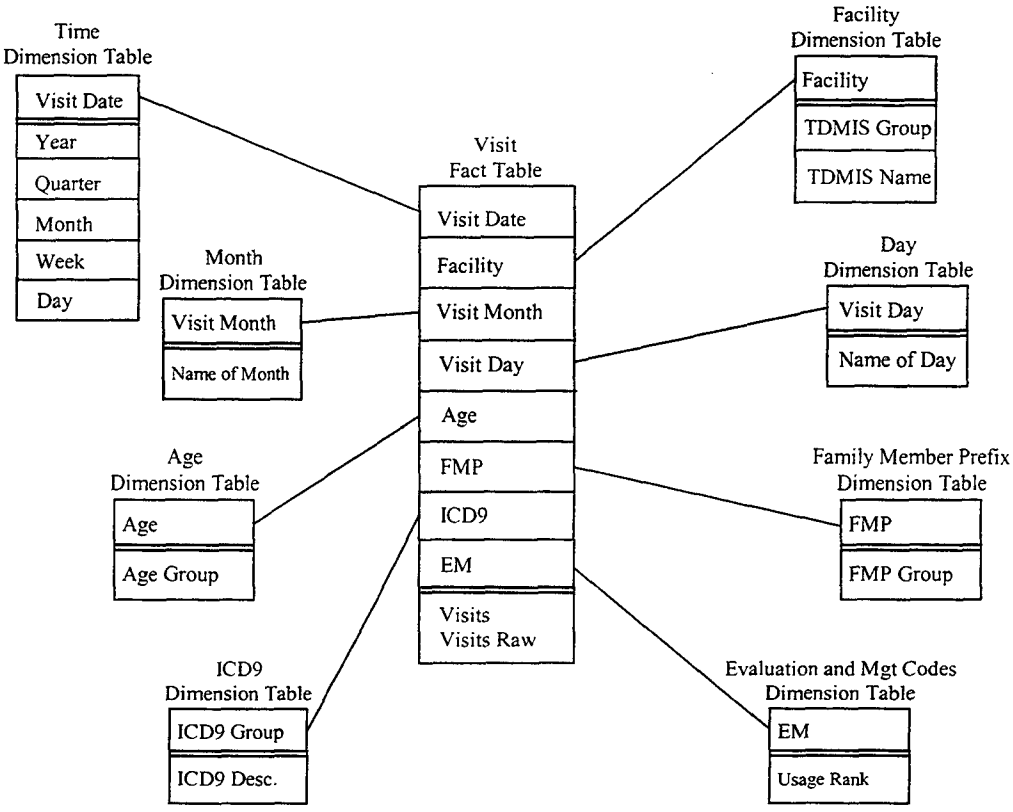


Figure 6-1. MTF ER Star Schema

PowerPlay™ Transformer allows users to construct data cubes with a graphical user interface (GUI) that is very user-friendly. Most tasks in the design of the data cubes are "point and click" or "drag and drop" operations. Figure 6-2 shows the Transformer GUI for the MTF ER data cube. The dimension map in the center of the Transformer display shows all of the dimensions of the data cube. The dimensions with only a single level such as 'Gender' were not included in the star schema in Figure 6-1 because they are resident in the MTF ER Visit Master Table. These single level dimensions do not have drill-down paths unless the user adds special drill-down categories. These single level dimensions function as filters during OLAP data exploration. The extra vertical line

in the facility dimension in the dimension map represents an alternate drill-down path. The normal drill-down path for the facility dimension passes from the TDMIS Group through the TDMIS name to the TDMIS, which is the individual facility number. The alternate drill-down path allows the user to bypass the TDMIS Groups and drill-down directly to the individual facilities in the TDMIS field. The bottom section of the Transformer display shows the data sources, which are the fact table and dimension tables of the star schema, the measures, and the generated data cube (PowerPlay™ refers to its data cubes as PowerCubes).

The measures in this research are simply the number of MTF ER visits. The visit data is an attribute of the MTF ER Visit Master Table. In other applications, such as a retail business, the measure data can be more complex and include calculated measures such as ratios or profit margins. For example, PowerPlay™ Transformer allows the user to create calculated measures such as foreign currency conversion.

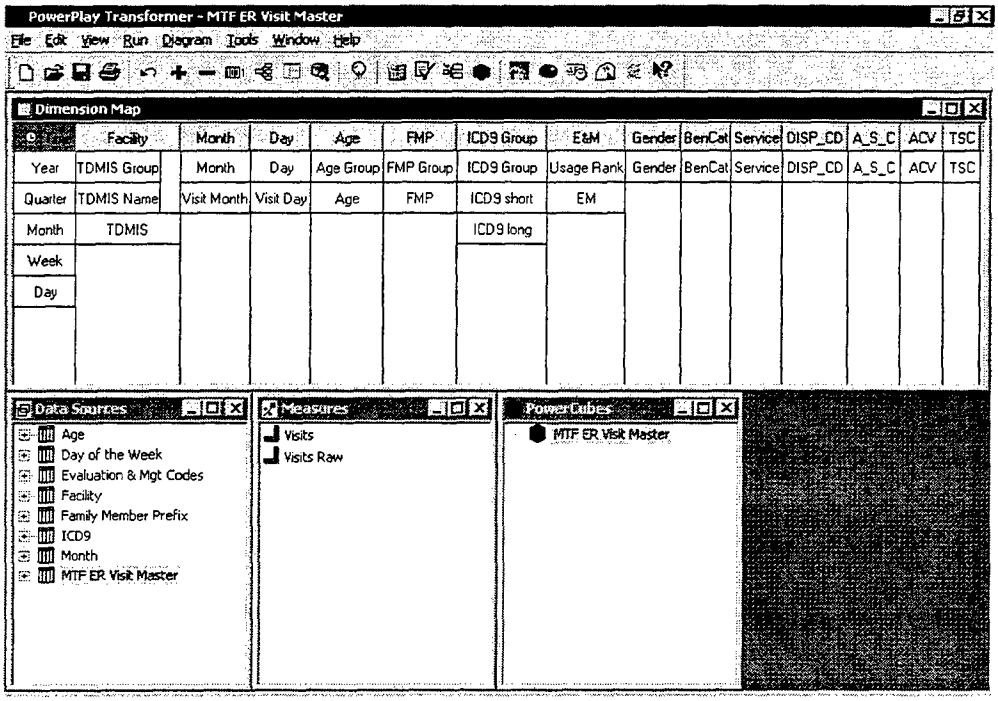


Figure 6-2. MTF ER Data Cube Model in PowerPlay™ Transformer

As mentioned in section E of Chapter III, Transformer provides a function that displays the categories that represent drill-down options within a dimension. Figure 6-3 shows the categories display for the month dimension. Each item with '+' next to it contains subordinate levels that are not displayed. The 'By Month' category is automatically generated from the dimension map. All categories generated from the dimension map, in this case, 'By Month' and all of its subordinates; are enclosed in a box. Any categories outside and below this box are user-created categories. The 'Seasons' category was added to the month dimension to provide an alternate view of data to facilitate the detection of seasonal patterns. Dimension level and category adjustments must be completed before generating the data cube. Any changes made to

categories or dimension levels for an existing data cube require data cube regeneration to incorporate the changes.

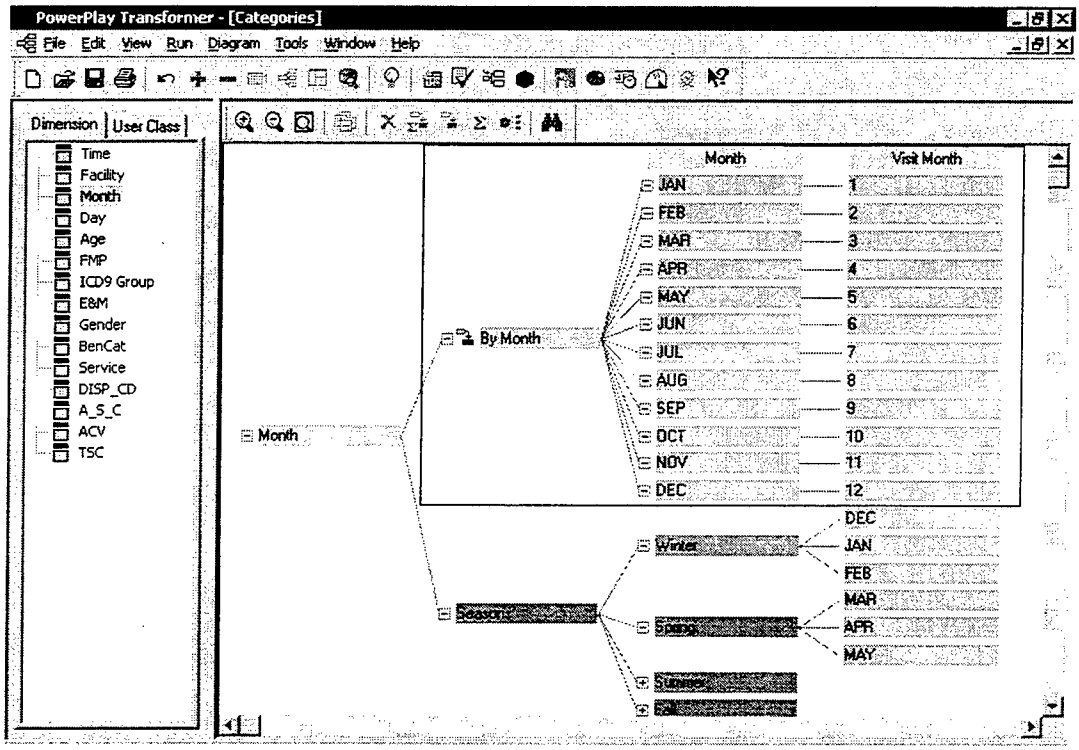


Figure 6-3. Month Categories for the MTF ER Data Cube in PowerPlay ™ Transformer

2. PowerPlay™ for Windows™ Interface

PowerPlay™ for Windows allows the user to view a data cube with a variety of customizable charts and tables, and to view a hierarchical map of the dimensions of the data cube. Figure 6-4 shows several of the 'Nested Charts' data view option of PowerPlay™ for Windows™ GUI. Notice that the dimension tab names in this display (located at the top of the workspace window) match the dimension names in the Transformer dimension map shown in Figure 6-2. Each dimension tab contains all of the

possible drill-down paths within that dimension. The dimension tabs allow the user to select a drill-down level for that dimension. The user selects the level of each dimension needed to create the desired data view. For example, Figure 6-5 shows the table view for all 24-35 year old patients who visited Army ER facilities during February 1999. The user can also drill-down or roll-up dimension levels in a data view by clicking on the dimension level in the data view window. PowerPlay™ gives the user an alternate means to view the data cube through the dimension map viewer. Figure 6-6 shows the dimension map for the MTF ER data cube. As in the categories display in Transformer, each item with a '+' next to it contains subordinate levels that can be expanded by clicking on the '+'.

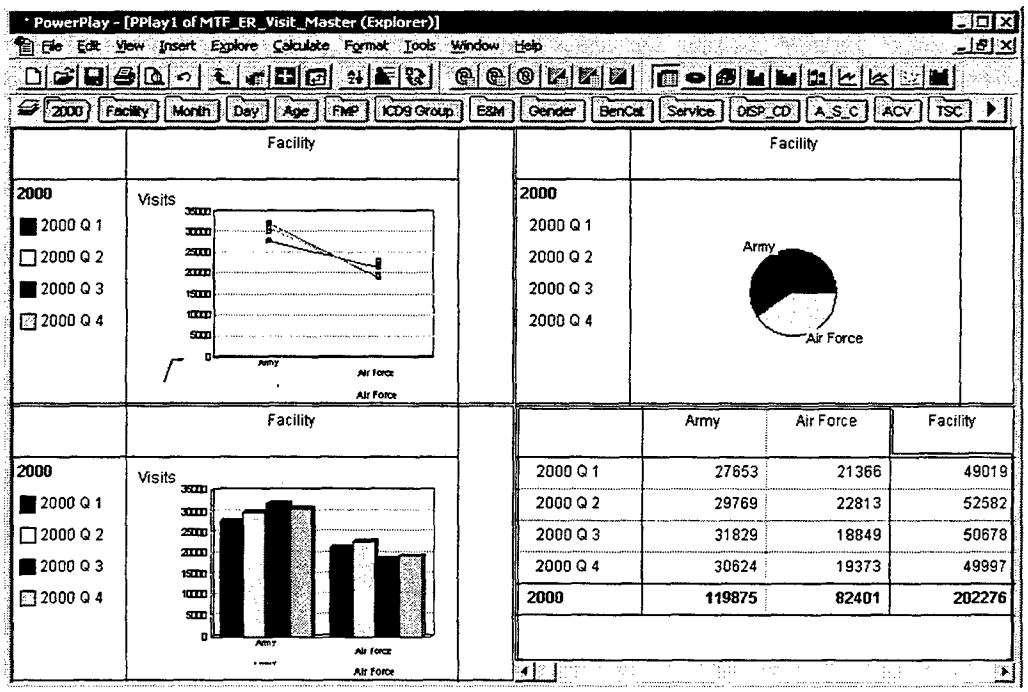


Figure 6-4. PowerPlay™ for Windows™ Explorer Display

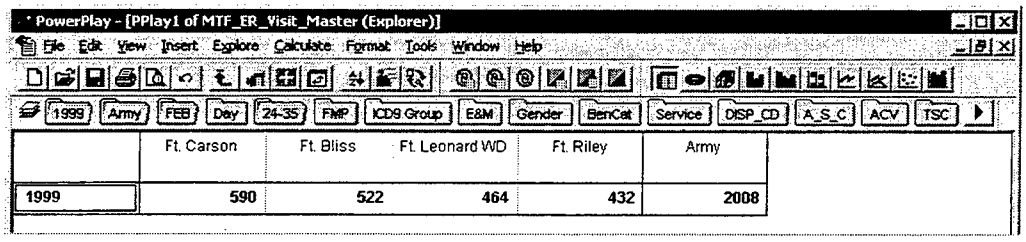


Figure 6-5. PowerPlay™ Explorer Data View

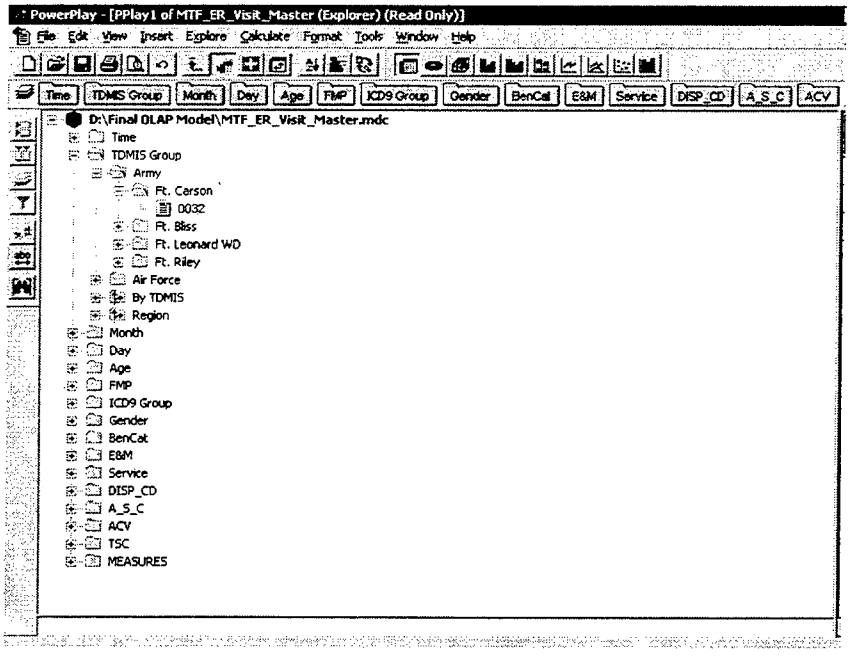


Figure 6-6. PowerPlay™ Data Cube Dimension Map

B. DATA ANALYSIS

1. Data Exploration

The first step in data exploration is to understand the organization's business model and the data itself. As mentioned in Chapter V, the MTF ER was analyzed using the statistical functions of the Clementine™ data mining software to determine the

distributions of each data item in each field and to determine the relationships between the data fields. In terms of an OLAP application, this means the user needs to understand all of the dimensions and categories within the data cube. In the case of the MTF ER data set, the dimensions are set up in PowerPlay™ from left to right to facilitate answering the questions:

- How many patients over time are going to the ER?
- To which facilities are they going?
- What day of the week and month are they going to the ER?
- Who is going to the ER?
- Why are they going to the ER?

The next step is to start exploring the data set. The dimensions were first examined at high aggregate levels using the different multiple views available in PowerPlay™. This method can reveal large inconsistencies in the data set. Once an inconsistency appears, the slice and dice operation can be used to isolate the discrepancy. Now within this smaller search space the drill-down operation can further isolate the inconsistency and possibly reveal the source of the problem. At any point during any of these operations, the user can invoke the pivot operation, change the numbers to percentages, add calculated fields, or add graphical statistical lines to the data views in order to create an alternate perspective.

2. Data Discoveries

a. 18-Year Old Active Duty Service Members

While exploring the age group dimension by facility over time, a large concentration of the 18-23 year old group appeared at one facility. Using the methods described in the previous section, this age group spike was narrowed down to an 18-year old age group for active duty service members. The facility turned out to be Ft. Leonard Wood, which is one of the Army's basic training posts, and, as such, would have a disproportionate number of soldiers in this age category when compared to a non-basic training post.

b. Two Quarters with No ER Visits

While looking at yearly ER visits for each facility, facility 0094 had nearly a 25% decrease in ER visits from the years 1999 to 2000. Isolating that facility and drilling-down into yearly quarters revealed that there was over a 200% increase in ER visits from quarter one to quarter two and no ER visits in quarters three and four of the year 2000 (See Figure 6-6). Changing the data view to a table and comparing this facility to ICD-9 codes revealed that the ICD-9 code group 'V01-V82' accounted for 97.9% of all of the year 2000 ER visits to this facility. This is significant because the 'V01-V82' ICD-9 code group was only used 390 times at this facility in 1999 versus 4581 times for quarter two of 2000 (see Figure 6-7). Looking at individual age groups for quarter two of the year 2000 revealed that the '0-4,' '18-23,' and '24-35' age groups experienced the largest increase in number of ER visits (see Figure 6-8). Viewing the Evaluation and

Management codes used by this facility during the year 2000 quarter two revealed that 65.6% of the ER visits for that quarter were coded as Physician telephone consultations.

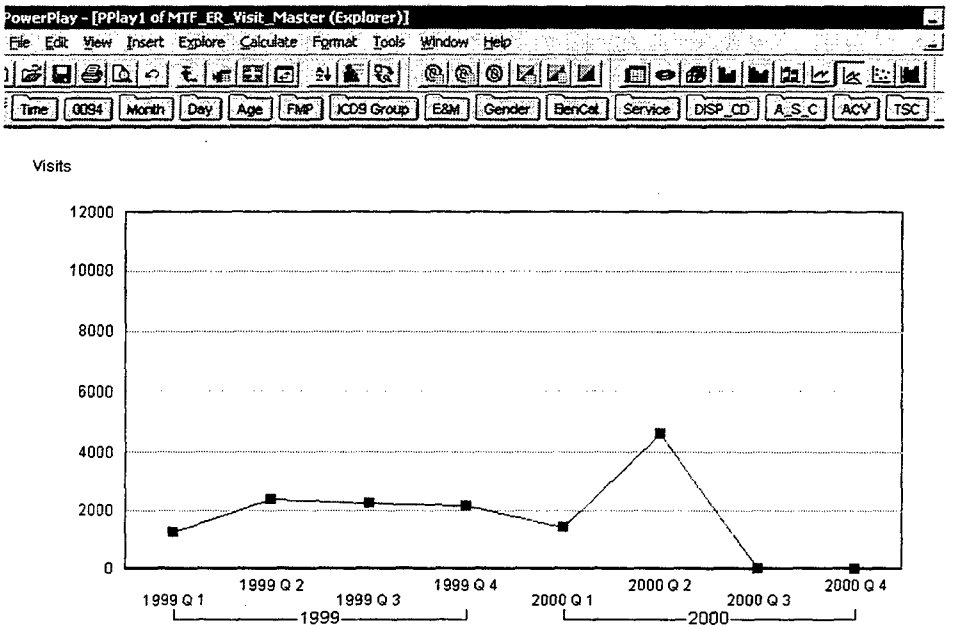
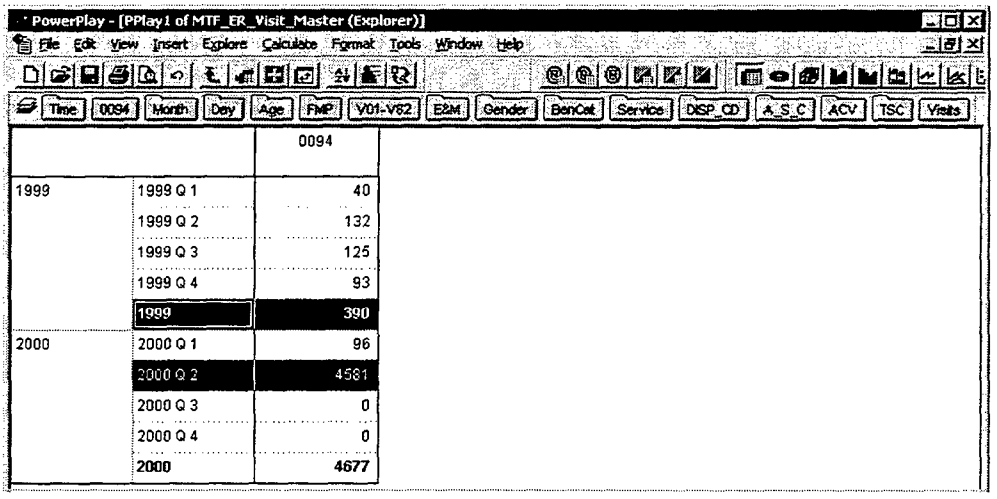


Figure 6-7. ER Visits to Facility 0094



		0094
1999	1999 Q 1	40
	1999 Q 2	132
	1999 Q 3	125
	1999 Q 4	93
	1999	**390**
2000	2000 Q 1	96
	2000 Q 2	4581
	2000 Q 3	0
	2000 Q 4	0
	2000	**4677**

Figure 6-8. ICD-9 Group 'V01-V82' Coded ER Visits for Facility 0094

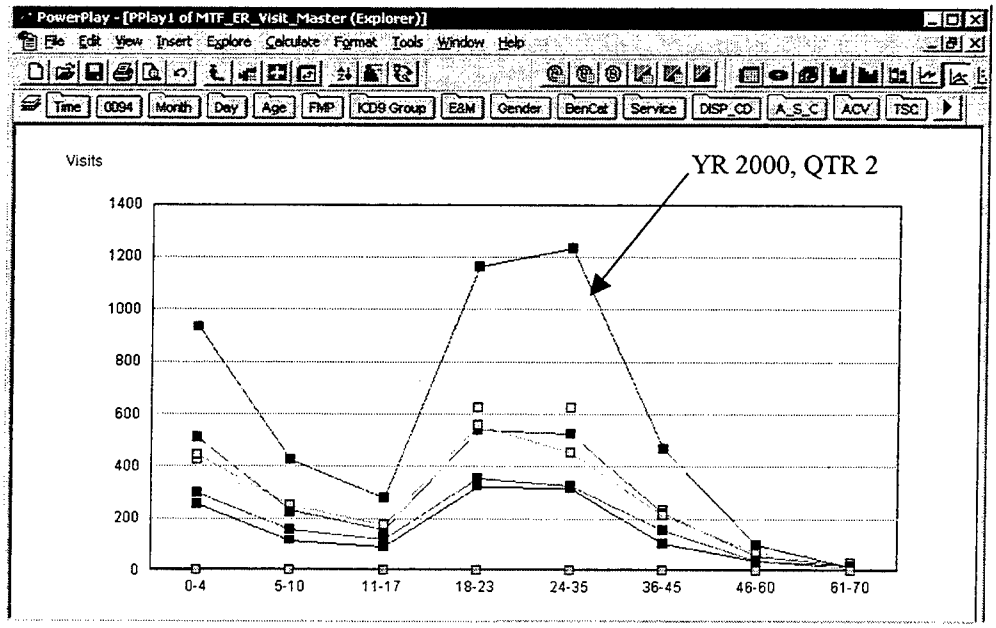


Figure 6-9. Number of ER Visits in Quarter Two of the Year 2000

The following is a summary of the inconsistencies found in the data for facility 0094 during quarter two of the year 2000:

- The number of 'V01-V82' ICD-9 coded ER visits represented 97.9% of all of the year 2000 ER visits to this facility.
- The number of 'V01-V82' ICD-9 coded ER visits for the previous year was 390 for this facility compared to 4581 for quarter two of the year 2000.
- The age groups '0-4,' '18-23,' and '24-35' experienced the greatest increase of the 'V01-V82' ICD-9 coding.
- The evaluation codes '99371' and '99372', which are both physician telephone consultation codes, represent 65.6% of all ER visits for that quarter.

After uncovering these inconsistencies, an inquiry to the TRICARE Central Region Lead Agent revealed that this facility, Minot AFB, closed its ER some time around January 1999. At that point, the ER functioned as a walk-in clinic and no longer provided ER-level services. This explains why there were so many physician phone consultations, which are common in clinics.

C. EVALUATION OF OLAP

This OLAP tool is an excellent means for an average computer user to rapidly explore a data set. The response time to data dimension and data view adjustments is very fast so the user does not have to wait for the software to run a large query as with traditional databases. Designing the data cubes is not difficult if the user has a good understanding of the organization's data and business model. Generating the data cubes is not as time-intensive as one might expect. A PentiumII 400MHz machine with 192 MB of RAM required just under six minutes to generate the data cube that originated from over 415,000 records and over 1300 categories. OLAP systems can be web-enabled to allow remote users to access data cubes through the Internet. This is a significant capability, especially if the organization is geographically dispersed or the organization's personnel conduct business travel.

OLAP is a very good tool with which to discover trends and anomalies quickly in historical data sets. Without any prior knowledge of the Minot AFB ER closure, this research discovered that the facility had shut down its ER services and operated as a walk-in clinic. This helps to validate the effectiveness of OLAP tools. However, OLAP

does not possess the ability to predict future activity in the sense that it generates a model that can be applied to a data set to predict future values. OLAP discoveries can be further exploited by other data analysis tools such as data mining which can produce predictive models.

D. SUMMARY

The OLAP tool used in this research was not difficult to use. The user interface for both the data cube creation software and the data exploration software was user friendly and facilitated a short train-up time for the user to become productive. OLAP can be web-enabled to allow users to access data cubes from anywhere that has Internet access. OLAP is an excellent tool to explore historical data for trends, patterns, and anomalies, but by itself cannot make predictions of future data values.

THIS PAGE INTENTIONALLY LEFT BLANK

VII. DATA MINING ANALYSIS OF MTF ER DATA

This chapter applies data mining methods to the MTF ER data set using SPSS Clementine™ data mining software. The data mining methods used in this research have been described in more detail in Chapter IV. Descriptions of the data mentioned in this chapter can be found in Appendix A.

A. USING CLEMENTINE™ WITH MTF ER DATA

1. User Interface

Clementine™ has a graphical user interface (GUI) that allows the user to visually build data mining models. Although not as intuitive as the OLAP tool used in this research, the GUI was not difficult to use after the initial learning process. Understanding the data mining methods used by the software required more time than the OLAP tool.

2. Non-Data Mining Specific Tools

Clementine™ has a wide range of data manipulation, statistical analysis, and data visualization tools that were instrumental in this research. As mentioned in Chapter V, Clementine™'s statistical analysis and data visualization tools reveal characteristics of the data and relationships between different data fields in preparation for the application of OLAP and data mining operations. Specifically, the data distribution and quality check functions give a clear picture of the population and range of the data in each field from the data set. Appendix B shows the results of the quality check function. The results of these functions clearly identify the strengths and weaknesses of this data set to allow

optimization of data analysis techniques. For example, the 'Marital' status field in the data set is only 39.9% populated, which meant that it is not useful for the data analysis in this research. The 'Marital' field is immediately eliminated, reducing the size of the data set. The Clementine™ statistics function analyzes each data field and generates the minimum, maximum, and mean values, number of occurrences, standard deviation, and correlations between data fields. Figure 7-1 shows the statistics generated for the 'Age' field.

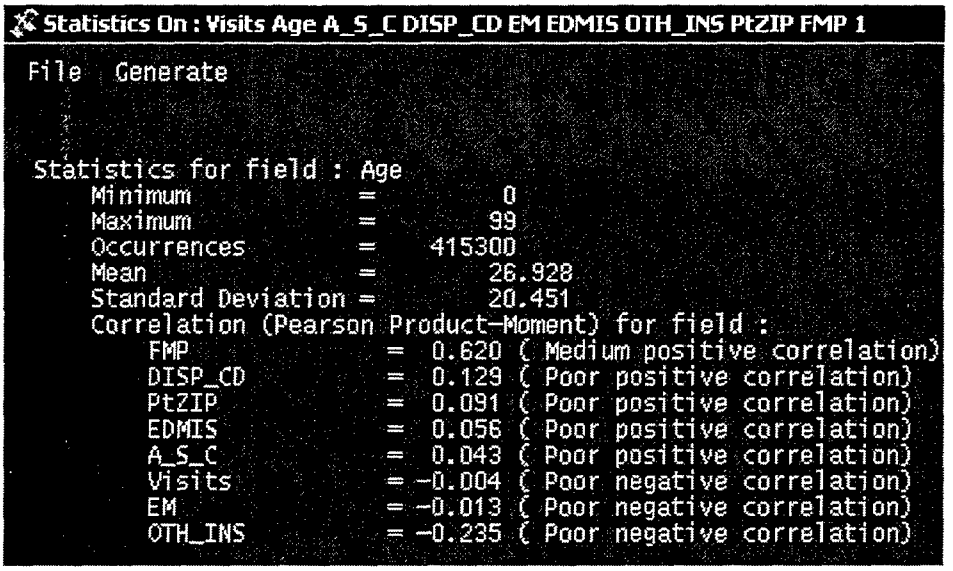


Figure 7-1. Clementine™ Generated Statistics for the 'Age' Field

B. RULE INDUCTION

1. Rule Induction Model

This research used Clementine™'s C5.0 rule induction algorithm to create a model to build a rule set that predicts a specified field in the MTF ER data set. The first step before executing any of the data mining methods in Clementine™ is to run the data

through a type node. This process allows the type node to check the data type of each field. This is necessary because the data mining algorithms in Clementine™ will generate an error if mistyped data enters a model-generating node. The type node allows the user to preset the data types or it can automatically determine the data types. After the data typing is complete, the user sets the data direction for each field, which determines if the data field is an input, output, both input and output, or is not included in the model. In the case of rule induction, there are multiple input fields and one output field. The output field is what the rule set will try to predict. Figure 7-2 shows the properties of a type node that is set up to use 'Gender,' 'FMP,' 'EM,' and 'Age Group' as inputs and 'ICD9 Group' as the output of the model. The 'BenCat,' 'Service,' and 'TDMIS' fields will not be used in the model. The input fields were selected on this particular model because the combination of ages, evaluation and management codes (which represent the intensity of service), and the gender of the patient seemed to have the characteristics needed to predict the diagnosis (ICD-9) of the patient.

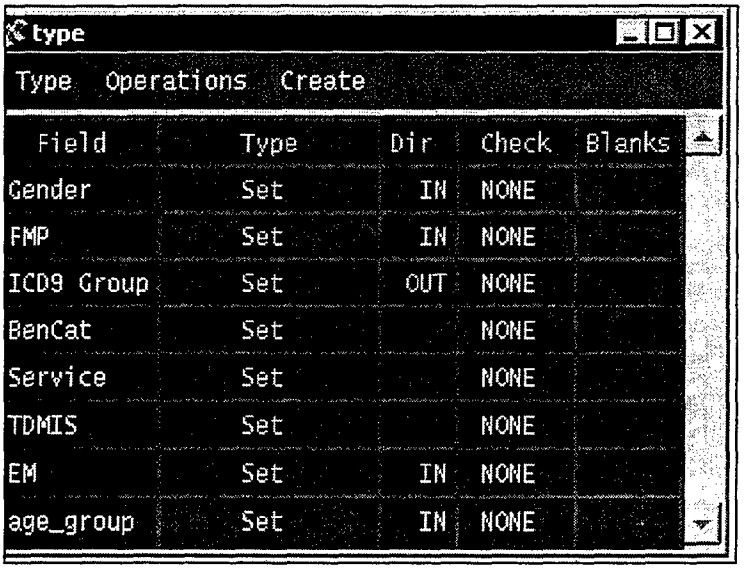


Figure 7-2. Clementine™ Data Type Node Properties

After the type node is initialized, the next step of rule induction is to train the C5.0 node. During this process, the C5.0 node processes the data set through its algorithm to develop a model to build rules in the form of a rule set or decision tree (The user selects the output form). This results in a generated rule node. Figure 7-3 shows the data stream used to generate the C5.0 rule node, which is the node located in the 'Generated Models' pallet on the right side of the screen shown in the figure. The generated C5.0 rule node is now placed into the design window and tested by running a data stream through the C5.0 rule node and into a matrix that compares the predicted values with the actual values. After testing, if the accuracy of the results is acceptable to the user, then the rule set may be used to predict unknown data. Otherwise the user must generate and test another rule set using a data set that varies from the first data set, or change the inputs of the model.

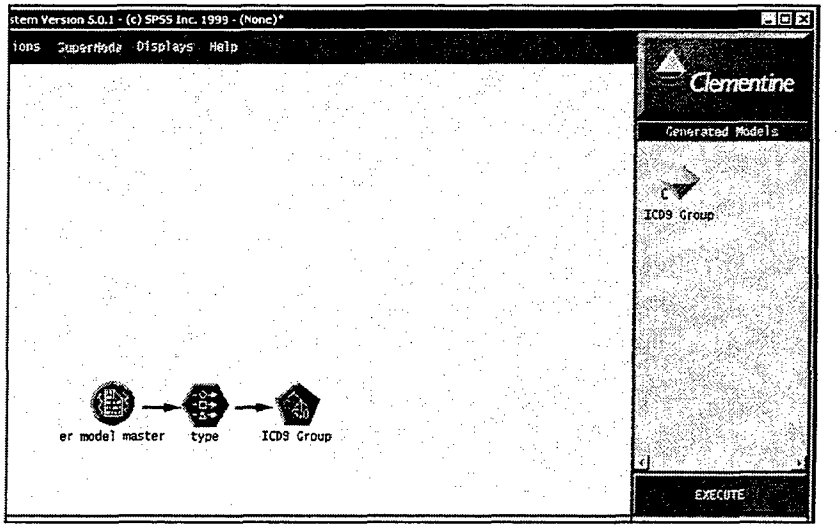


Figure 7-3. Clementine™ Data Stream Model to Generate C5.0 Rule Set

2. Analysis Results

This research uses the C5.0 algorithm to create rule sets to predict several different fields in the MTF ER data set. Figure 7-4 shows the data stream model used to pass data through a generated C5.0 rule node to predict the 'ICD9 Group.' In this case the value of the 'ICD9 Group' is known so the output of this model is a matrix that compares the predicted value with the actual value. Table 7-1 shows the resulting matrix. The labels of the X-axis of the matrix represent the actual 'ICD9 Group' and the labels of the Y-axis represent the predicted 'ICD9 Group'. The numbers in the matrix are the percentage of the times that the predicted group matched the actual group. For example in Table 7-1, the value of the cell where the actual '320-389' group and the predicted '320-389' group intersect is '21.427'. This means that the model predicted the correct '320-389' group 21.427% of the time.

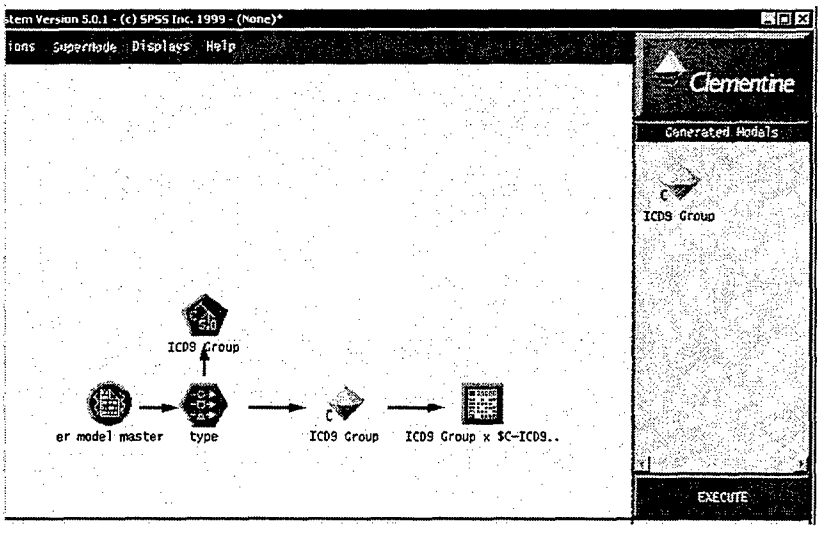


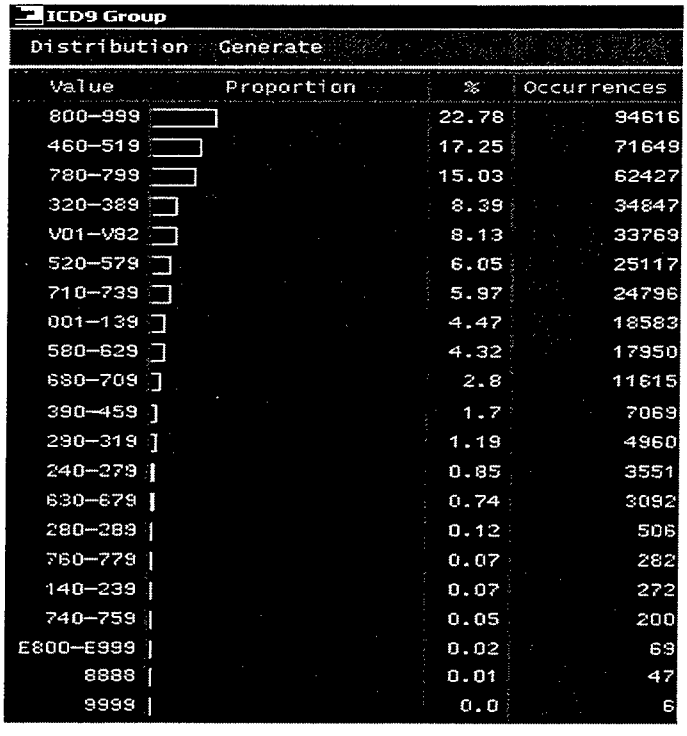
Figure 7-4. Clementine™ Data Stream Model to Predict the 'ICD9 Group'

ICD9 Group x $C-ICD9.. : Percentages Across

Matrix	Generate											
	320–389	390–459	460–519	520–579	580–629	630–679	680–709	710–739	740–759	760–779	800–999	V01–V82
001–139	15.479	0.027	16.534	0.0	0.016	0.0	0.0	0.011	0.0	13.454	53.823	0.555
140–239	1.103	0.0	5.147	0.0	0.0	0.0	0.0	0.0	0.0	44.853	47.426	1.471
240–279	3.305	0.226	8.559	0.0	0.056	0.0	0.0	0.0	0.0	35.48	51.893	0.424
280–289	3.38	0.199	10.736	0.0	0.0	0.0	0.0	0.0	0.0	39.96	45.129	0.199
290–319	0.707	0.263	3.942	0.0	0.0	0.0	0.0	0.0	0.0	32.262	62.503	0.202
320–389	21.427	0.028	19.619	0.0	0.003	0.0	0.0	0.0	0.0	12.204	46.153	0.545
390–459	0.681	2.908	3.858	0.0	0.0	0.0	0.0	0.028	0.0	50.255	41.901	0.369
460–519	8.795	0.066	15.696	0.0	0.006	0.0	0.0	0.01	0.0	18.135	56.649	0.63
520–579	7.676	0.068	12.522	0.036	0.004	0.0	0.0	0.02	0.0	22.486	56.773	0.391
580–629	2.364	0.039	11.368	0.011	0.089	0.0	0.0	0.022	0.0	36.396	49.091	0.619
630–679	0.194	0.0	7.867	0.032	0.065	0.032	0.0	0.0	0.0	58.174	33.053	0.583
680–709	6.016	0.026	9.506	0.0	0.0	0.0	0.052	0.009	0.0	19.512	64.423	0.431
710–739	0.977	0.028	5.631	0.004	0.004	0.0	0.004	0.081	0.0	24.273	68.279	0.71
740–759	11.5	0.0	13.5	0.0	0.0	0.0	0.0	0.0	0.5	24.5	50.0	0.0
760–779	27.66	0.0	47.163	0.0	0.0	0.0	0.0	0.0	0.0	8.865	15.603	0.709
780–799	7.288	0.128	10.291	0.003	0.014	0.002	0.002	0.008	0.0	30.207	50.966	1.081
800–999	5.793	0.043	8.501	0.003	0.006	0.0	0.002	0.005	0.0	13.89	71.512	0.239
8000	10.87	0.0	15.217	0.0	0.0	0.0	0.0	0.0	0.0	17.391	56.522	0.0
9999	16.667	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	83.333	0.0
E800–E999	8.696	0.0	10.145	0.0	0.0	0.0	0.0	0.0	0.0	27.536	53.623	0.0
V01–V82	8.727	0.033	7.48	0.0	0.003	0.0	0.003	0.006	0.0	12.887	42.656	28.199

Table 7-1. Clementine™ Matrix that Compares Predicted with Actual ICD9 Groups

All but three groups of predictions fall below 20%. Of note is the '800-899' group that is predicted correctly 71.512% of the time. This initially looks like a promising number until the entire '800-899' column is examined (See Table 7-1). The rule set predicted 19 of the 21 'ICD9 Groups' as '800-899' over 40% of the time. At 22.78% of the population, the '800-899' ICD-9 Group is the largest ICD-9 Group in the data set, which could have contributed to the frequent '800-899' predictions of the rule set (see Figure 7-5).


ICD9 Group

Distribution Generate

Value	Proportion	%	Occurrences
800-999		22.78	94616
460-519		17.25	71649
780-799		15.03	62427
320-389		8.39	34847
V01-V82		8.13	33769
520-579		6.05	25117
710-739		5.97	24796
001-139		4.47	18583
580-629		4.32	17950
680-709		2.8	11615
390-459		1.7	7069
290-319		1.19	4960
240-279		0.85	3551
630-679		0.74	3092
280-289		0.12	506
760-779		0.07	282
140-239		0.07	272
740-759		0.05	200
E800-E999		0.02	69
8888		0.01	47
9999		0.0	6

Figure 7-5. Clementine™ Distribution of the 'ICD9 Group' Field

The '8888' and '9999' groups are the groups created to replace null and invalid data fields, respectively. Had the incorrect '800-899' predictions for the other non '800-

899' groups been low and the '800-899' predictions for the '8888' and '9999' groups been high, this could have indicated the approximate percentage of the null and invalid ICD-9 Group entries that should have been coded '800-899'. In this case, the 83.33% incorrect '800-899' predictions for ICD-9 Group '9999' are not relevant since there were only a total of six occurrences of the'9999' field (see Figure 7-5).

C. ASSOCIATION RULE DETECTION

1. Association Rule Detection Model

This research uses the APRIORI algorithm in Clementine™ to build a model for association rule detection. As with the rule induction process, the first step in building the association rule detection model is to run the data through the type node. In this case all data fields used in the model are set to the data direction 'BOTH' so the association rule algorithm can examine every field as an input and as an output of association rules. Next, the data stream is run through the APRIORI node to develop a model. The result is an unrefined model that can be browsed but not applied to data. However, a rule set similar to the one produced by rule induction can be generated from the unrefined model. To do this, the user selects a single field that the rule set will try to predict. Once the software generates the rule set, it can be applied to a data set and tested using the same procedures used in rule induction outlined in the preceding section.

2. Analysis Results

This research used the APRIORI association rule algorithm to detect association rules and generate rule sets to predict MTF ER data fields. Figure 7-6 shows the data

stream model used to generate the unrefined APRIORI association rule node. This node cannot be placed in a data stream, but can be used to generate a rule set. The unrefined association rule node contains association rules that the user can browse. Figure 7-7 shows the results of one of the unrefined APRIORI nodes generated from the MTF ER data set. The numbers in parentheses represent the number of occurrences in the data set, the percent coverage of the data set, and the prediction accuracy of the rule.

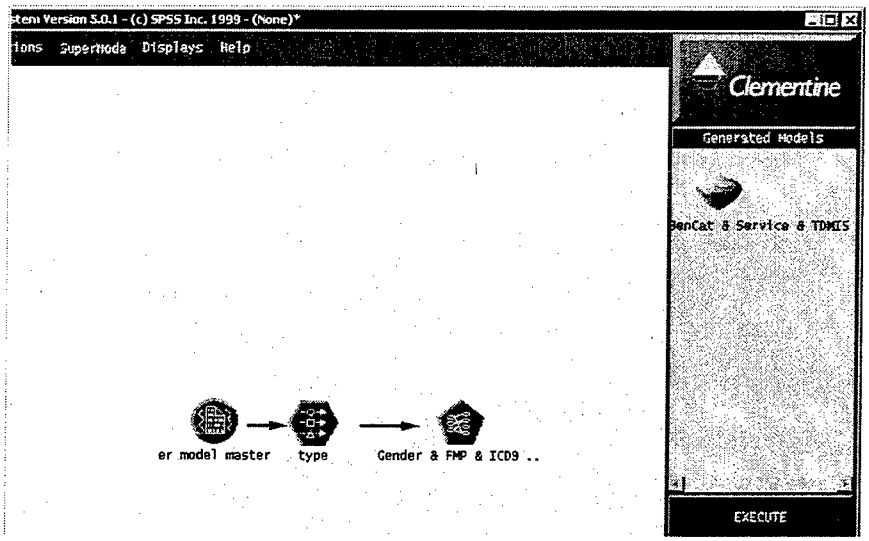


Figure 7-6. Clementine™ Data Stream Model that Generates the KNet Node

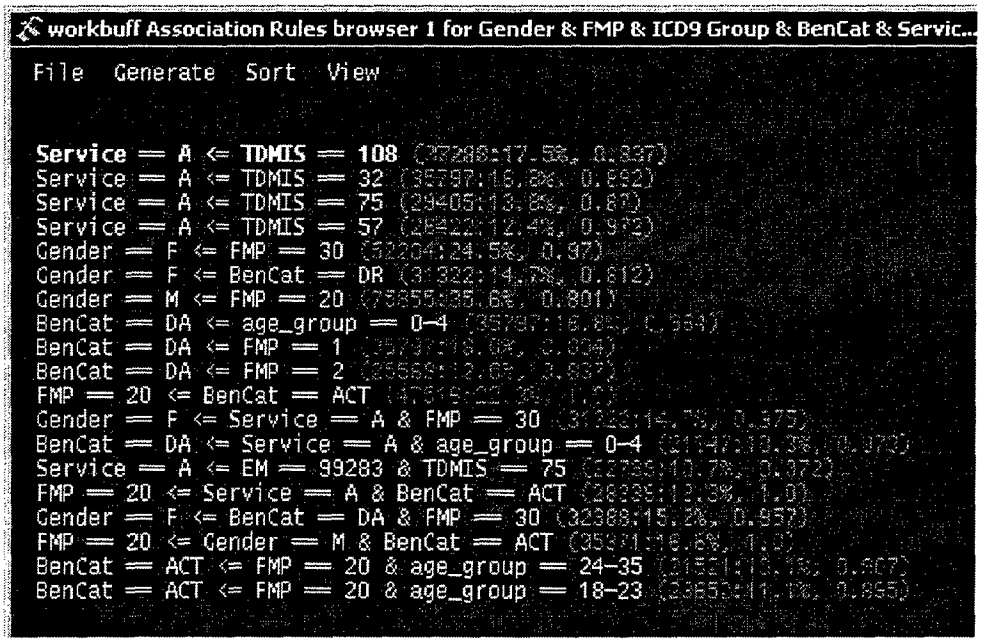


Figure 7-7. Clementine™ KNet Generated Association Rule Set

Close examination of the resulting rule sets did not reveal any rules that were of significant predictive value. For example, one association rule states that if a patient is on active duty, is male, and is in the Army, he has a family member prefix of 20. This rule does not reveal anything new because most active duty service members, regardless of gender and branch of service, have a family member prefix of 20, which is the prefix for the military sponsor. Again, this demonstrates the need to understand the data semantics well before undertaking any data mining activity. Additional association rule nodes were generated using different combinations of input fields with no better results.

D. CLUSTERING

1. Clustering Model

Clementine™ uses the Kohonen algorithm described in Chapter V, Section E to detect clustering in a data set. A fundamental difference with this method compared to the previous two methods, is that the Kohonen network does not make a prediction. Instead, it attempts to determine relationships between the data.

As with the other methods, the type node is initialized. In the type node, all of the data fields that are used in the model are set to 'IN' and no fields are set to 'OUT' since there is no output. The next step is to send data through the 'Train Kohonen' node. This results in a model called the KNet node. The KNet node produces an X and Y coordinate for each data record that passes through it. These X and Y coordinates can be plotted using the plot node to produce an XY graph. The resulting XY graph shows the clusters found in the data set.

The next step is to concatenate the X and Y coordinates of each record to form a cluster number. For example, if record 20 has an X coordinate of '1' and a Y coordinate of '2,' its cluster number would be '12.' Now that each record has a cluster number, the data set can be run through a distribution node to show the proportion of the population that each cluster represents. At this point, the user must decide which clusters are most interesting. The user then selects the records in the interesting clusters for further analysis. At this point, the user can view the data in web plots and distribution graphs to determine the attributes of each cluster. It is hoped that these attributes will describe each cluster. For example, the resulting description for cluster '12' may be that this cluster is

associated with male, active-duty Air Force service members between the ages of 18 and 23.

2. Analysis Results

The most noticeable difference between this method and the previous two data mining methods is that building, training, and analyzing the Kohonen network is much more time-consuming than the other two methods. The Kohonen network data stream in Figure 7-8 shows the complexity of the Kohonen network method compared to the previous two methods. The MTF ER data set had to be greatly reduced because most of the fields had many different unique data points. For example, the 'ICD9 Group' field contained 21 different groups (19 ICD-9 Groups, one group for null values, and one group for invalid data). Each additional group in a data field significantly increases the search space of the Kohonen network because the network checks each data field group against all other data field groups in the data set. This method requires a computer with large amounts of RAM to generate a model in any reasonable amount of time.

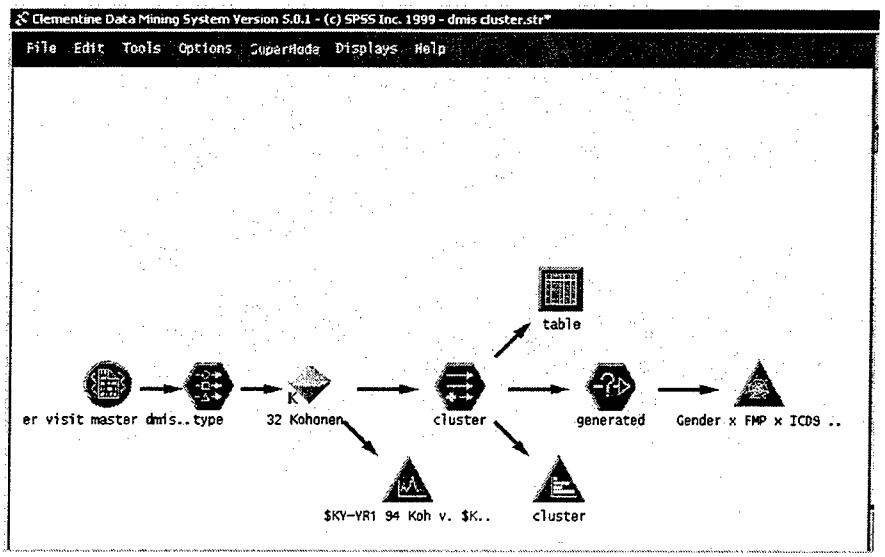


Figure 7-8. Clementine™ XY Plot of Kohonen Network Coordinates for Facility 0032

To adjust to the capabilities of the computer hardware available to run the Clementine™ software, a Pentium III 500 MHz machine with 128 MB of RAM, the data set was scaled back to focus on one facility. After initializing the type node, the 'Train Kohonen' node generated the 'KNet' node model. Even with the scaled back data set, it took five hours and twenty-two minutes to train the Kohonen Network. The data set was then passed through the KNet node and the resulting coordinates were plotted on the graph in Figure 7-9. The KNet node placed each record in one of 25 derived clusters. A derive node then concatenated the XY coordinates to give each record a cluster number. In this case the cluster numbers ranged from '00' to '44.' '00' is at the intersection of the X- and Y-axes and '44' is in the upper right corner. The resulting clusters were then run through a distribution node. Using the results of the XY plot and the cluster distributions, the top four populated clusters were selected to further explore with web and distribution

plots. Figure 7-10 shows the modified web plot for cluster '40', which indicates that it has strong ties to female patients, active duty dependents, and first spouses; and medium ties to age groups '18-23' and '24-35', and ICD-9 Group '800-999'. The dark lines are strong ties and the lighter lines are medium ties. The distribution plots for cluster '40' confirm the findings of the web plot. Table 7-2 is a summary of the attributes of the four most populated clusters for facility '0032,' which is the MTF located at Ft. Carson, Colorado.

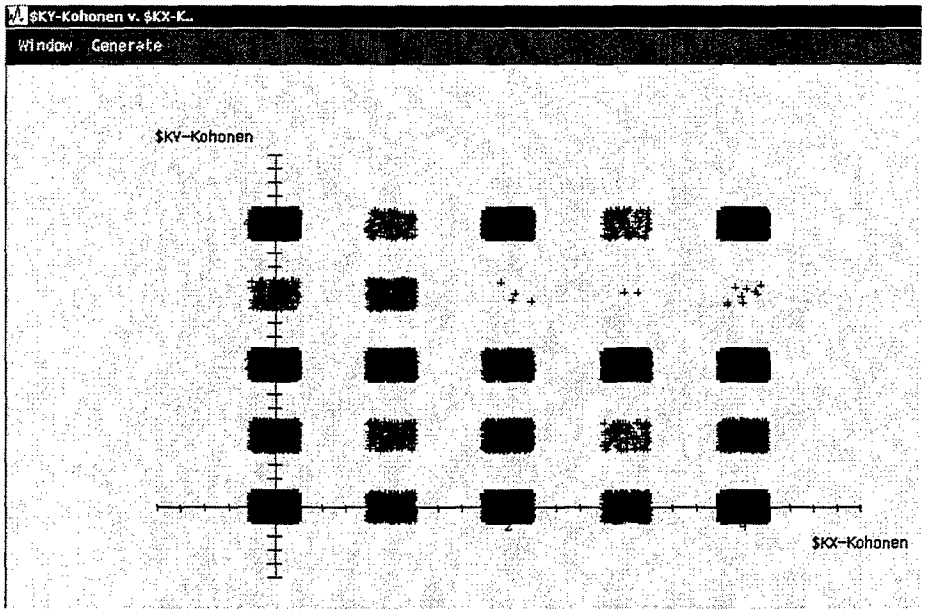


Figure 7-9. Clementine™ XY Plot of Kohonen Network Coordinates for Facility 0032

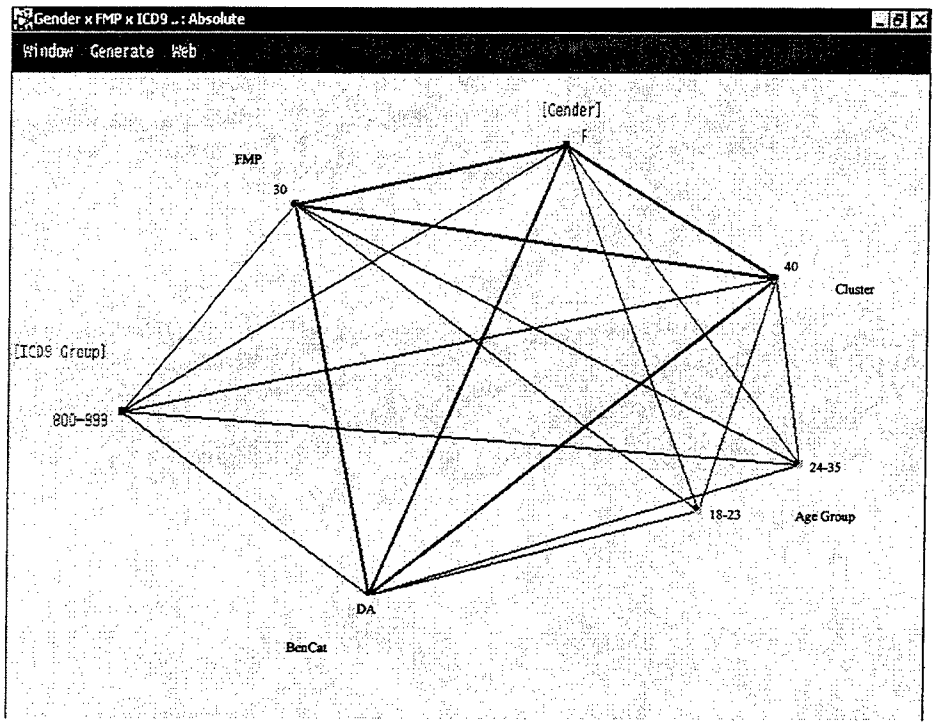


Figure 7-10. Clementine™ Web Plot for Cluster 40 of Facility 0032

Cluster	Gender	Family Member Prefix	Beneficiary Category	Age Group	ICD-9 Group
00	Male	20- Sponsor	Active Duty	24-35	800-899
40	Female	30- 1st Spouse	Active Duty Dependent	18-23 24-35	800-899
42	Female	1- 1st Child 2- 2nd Child	Active Duty Dependent	0-4	No Common ICD-9 Group
44	Male	1- 1st Child 2- 2nd Child	Active Duty Dependent	0-4	No Common ICD-9 Group

Table 7-2. Attributes of the Most Populated Clusters for Facility 0032

Data mining is an iterative process and one data mining method's results may serve as a starting point for another method. The significance of the results shown in

Table 7-2 is that they give the data miner a direction to apply other data mining methods. For example, the rule sets that were built in the earlier section in this chapter were based on the entire data set. With the information on clustering in the data, the user could select records out of a particular cluster to train a rule set node. This might result in a more accurate rule set for predicting fields of records in the same cluster, since the rule set algorithm is using a data set that has many characteristics in common.

E. EVALUATION OF DATA MINING

1. Non-Data Mining Specific Tools

The data manipulation, statistical analysis, and data visualization tools provided by Clementine™ were instrumental in the data preparation phase of this research. Although many software packages exist that can perform these functions, it was convenient and time-saving to have these tools and the data mining tools in the same package. Especially useful were the data field quality check, data distribution, and histogram functions.

2. Rule Induction and Association Rule Detection

The rule induction and association rules models are not difficult to set up and their algorithms quickly process data to generate a rule set or set of association rules. They have great potential to predict unknown data given the right conditions. Two conditions have a significant influence on the effectiveness of rule induction and association rule detection. First, the data set must have characteristics that are conducive to rule building. For example, relationships and patterns among the data of the MTF ER data set that

support building rules that apply to the whole data set may not exist. In other words, the reason no significant rules are discovered may be because no or very few significant rules exist in the data set. The other significant factor in building rules is having a complete and accurate data set. In the MTF ER data set, many fields were unusable because they were too sparsely populated (See Appendix B for more information on the MTF ER data set population). Some of the under- populated fields may have influenced the prediction accuracy of the rule sets.

3. Clustering

Clustering is a good tool to find out what attributes individuals in a group within a data set have in common and to find outliers in a data set. For example, an MTF may want to send information packets that explain the available healthcare options to 25% of its enrolled patients in an effort to reduce ER visits. Clustering can sometimes find groups that were not previously discovered by traditional statistical analysis. This would give the MTF more options on which groups to send information packets.

F. SUMMARY

Data mining has many powerful tools that can sometimes extract significant findings previously unattainable by other methods of data analysis. Data mining is an iterative process that can sometimes use the results of one data mining method as the starting point for another method. Although the data mining analysis in this research did not produce any major discoveries, it did reveal the potential benefits of data mining.

The characteristics of the data and the relationships between the data in a data set, as well as the quality of the data, play a significant role in the effectiveness of data mining.

VIII. CONCLUSIONS AND RECOMMENDATIONS

The objective of this research was to evaluate the capabilities of OLAP and data mining tools and determine their ability to predict MTF ER data. The user of OLAP and data mining tools must understand the organization's data and business model. The DoD medical system is very complex with multiple interdependent organizations and information systems which will require careful training of OLAP and data mining users to ensure that they understand how their organization fits into the DoD medical structure and what information systems maintain their data.

A. OLAP SUMMARY

OLAP is a very effective tool to explore historical data without the requirement of extensive user training. The application of an OLAP tool to the MTF ER data set quickly revealed anomalies and trends in the data set, validating the tool's effectiveness. User understanding of the organization's data and business model are imperative to the success of OLAP data analysis. The following subsections summarize the advantages and disadvantages of OLAP.

1. Advantages of OLAP

- OLAP has low train-up time for a user to become effective due to the simplicity of the design of its user interface.
- OLAP queries have very fast response times due to OLAP's ability to interface with multi-dimensional data cubes that contain pre-calculated summary data.

- OLAP can quickly switch between multiple data views such as tables, charts, and graphs, allowing the user to view the data set from different perspectives.

- OLAP allows the user to easily change the performance measures in a data view, giving the user a different perspective of the data.

- OLAP can discover trends, patterns, and anomalies in an historical data set. OLAP can enable the MTF ER manger to explore a data set to answer specific questions without having to go through a database professional.

- OLAP can feed findings into another system, such as data mining, for further data analysis.

- OLAP data cubes can be web-enabled, allowing access to the data cubes from anywhere that there is an Internet connection. If web-enabled, data cubes on a server at the Lead Agent Headquarters in Colorado Springs, Colorado, could be accessed from any of the MTFs in Central Region through an Internet connection.

- OLAP can provide, on-line, on demand reports to personnel, which may not be possible form the underlying operational databases.

2. Disadvantages of OLAP

- OLAP by itself does not produce a model to predict future data; it is primarily an historical data analysis tool.

- The OLAP tool used in this research had limited data preparation capabilities, requiring other software packages to cleanse and manipulate the data set before the application of OLAP.

B. DATA MINING SUMMARY

Data mining is a very powerful tool that can discover hidden information in a data set that other data analysis methods overlook. However, data mining is not a magical solution to data analysis. If relationships, rules, and clusterings do not exist in a data set, then data mining cannot find them. Even though this research did not make any major

discoveries in the MTF ER data set, it did demonstrate the beneficial capabilities of data mining. As with OLAP, it is essential that the user of the data mining system have a good understanding of the organization's data and business model. The following subsections summarize the advantages and disadvantages of data mining.

1. Advantages of Data Mining

- Data mining gives the user the ability to use powerful statistical methods without having to be a professional statistician.
- Data mining has powerful data prediction capabilities. After the user sets the parameters, data mining can automatically find rule sets and association rules.
- Data mining methods can sometimes infer values for missing or invalid data. Many data fields in the MTF data set were essentially unusable because of missing data. Data mining prediction methods could possibly populate enough missing data in a field to make it useable for data analysis.
- Data mining can find patterns and relationships automatically, whereas in OLAP, the user has to search for them.
- Data mining methods use the power of machine learning.
- The data mining tool used in this research has a wide range of data preparation tools, so it can do most of the data preparation without having to use other software packages.

2. Disadvantages of Data Mining

- Data mining can require significant user training. The user must understand the data mining methods before using the software. This typically requires someone with a background in statistical analysis and inference. This limitation would make it difficult to implement data mining at the MTF level because the individual performing the data mining would essentially have to be dedicated full-time to data mining tasks. This would mean adding another salary to the MTF payroll.
- Data mining can be time-consuming, both in building models and waiting for models to execute.

- Some data mining methods are resource intensive, requiring powerful workstations with large RAM capacities.

C. RECOMMENDATIONS

OLAP and data mining tools can greatly enhance the MTF ER manager's ability to improve ER staffing and utilization. The Lead Agent Headquarters must assess its data operations and determine what the organization's data analysis needs are in order to decide if OLAP and data mining tools are appropriate. If OLAP and data mining tools are selected, the next step is to determine the proper mix of OLAP and data mining tools and at what levels to place these tools. The following are recommendations for selecting OLAP and data mining tools for use in the TRICARE Central Region.

- The organization must focus on data quality to fully benefit from OLAP and data mining, and to have confidence in the findings of these tools. If the organization's data is inaccurate or sparse, then these tools will have limited capabilities. Most of the critical fields in the MTF ER data set were well populated; however, the OLAP and data mining tools were excluded from exploring many data fields because the fields were not populated to an acceptable level.

- OLAP could provide the TRICARE Central Region with an excellent set of tools for managers to explore historical data from a desktop computer. This research focused on MTF ER data, but the same OLAP principles and benefits apply to the data set from all areas of the MTF. The MTF ER data set is relatively small compared to the MTF's other data sets, so the time to design and create the data cubes will increase. However, the performance of the end user OLAP tool should still be acceptable using a larger data cube because OLAP takes advantage of multi-dimensional data access speed and pre-calculated summary data..

- To accommodate the large data sets produced by MTF visits, custom data cubes and subcubes can be created to meet the needs of specific users and to reduce the size of the data cube by excluding unneeded data. For example, a subcube for each MTF with an ER in Central Region could be generated from the MTF ER data cube created in this research. These subcubes would include all of the detailed data for a particular facility, and only summary data for the other facilities. This allows the MTF to

compare its summary data with other facilities in the region, and it greatly reduces the size of the data cube compared to the master data cube. This makes the data cube much easier to work with from a data handling and data storage perspective.

- The web capabilities of OLAP could be very beneficial to Central Region because the MTFs of the region are so geographically dispersed. The Lead Agent Headquarters could maintain data cubes on servers that individual MTFs could access through an Internet connection. OLAP's security features, like most network products, would allow the Lead Agent Headquarters to establish MTF user privileges to control access to the data cubes. This concept also applies to an MTF and its area clinics. An MTF could maintain data cubes on a server that its area clinics could access through a web connection.

- Not every individual in the MTF needs to have an OLAP tool on his or her desk, but certainly each department within the MTF could benefit from access to OLAP tools. Anyone in the MTF who is performing data queries on a regular basis should have access to OLAP tools.

- Most OLAP tools fall short of providing forecasting and predictive capabilities. Data mining complements this shortcoming; however, it requires sophisticated users who are comfortable with the principles of statistical inference. Data mining operations performed by users that do not understand the data mining methods used by the tool may produce erroneous information and make false predictions.

- Data mining tools could greatly enhance the TRICARE Central Region's ability to predict future MTF ER activity. However, data mining's user training and understanding requirements may limit its use to the Lead Agent Headquarters. Data mining would require one or more experienced individuals to operate the system full-time, which the Lead Agent Headquarters may be better able to accommodate than most MTFs.

- If the organization selects both OLAP and data mining tools, it should plan to share information between the two systems to take advantage of the strong points of each tool. Findings in OLAP can be passed to a data mining system for further exploitation and predictive model building.

- Before implementing either OLAP or data mining, the MTFs must ensure that all users of these systems understand the organization's data and its business model.

D. CLOSING REMARKS

OLAP and data mining can greatly enhance the MTF ER manager's ability to improve ER staffing and utilization. An organization that understands its data and its business model as well as the capabilities and limitations of OLAP and data mining can maximize the benefit of these data analysis tools.

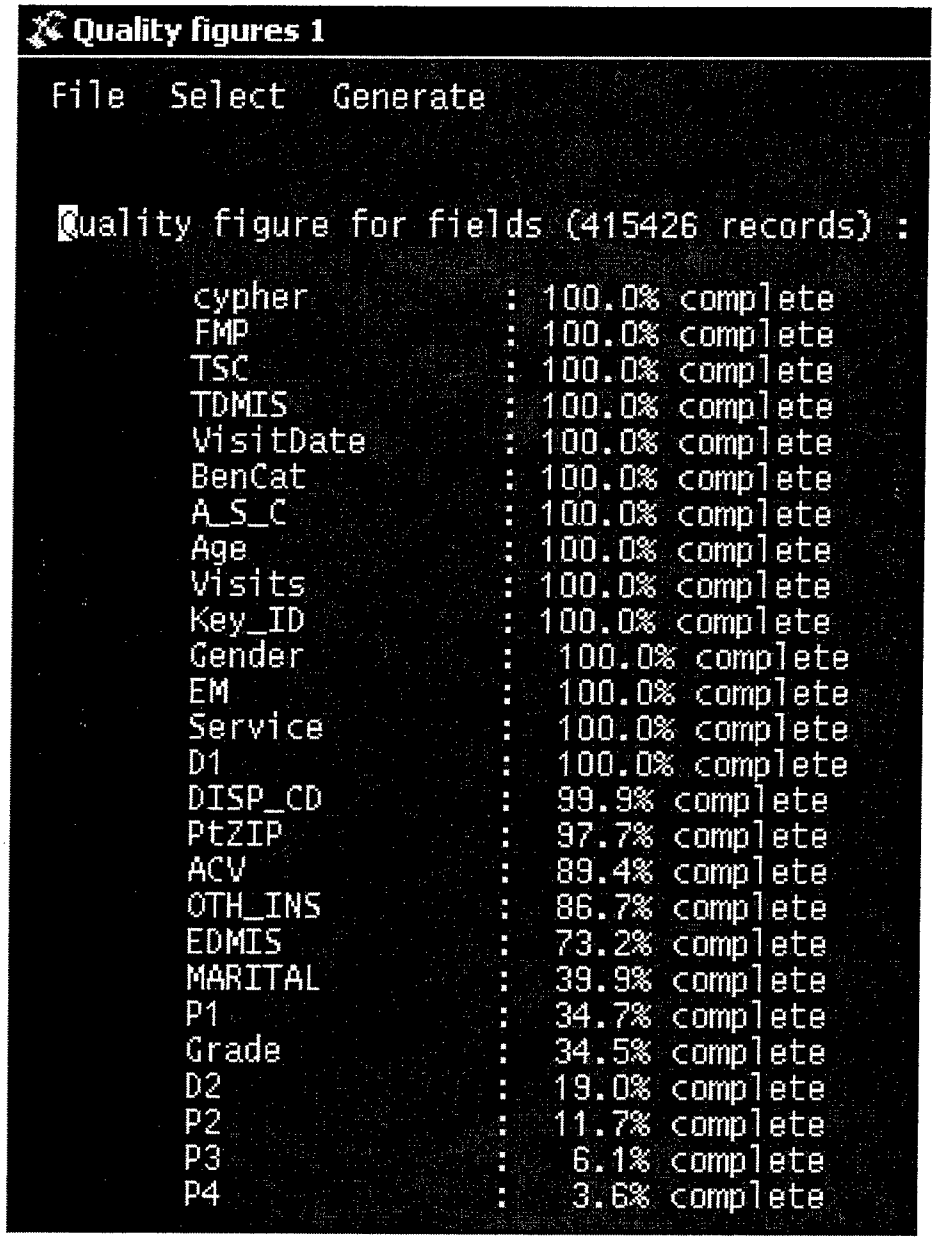
APPENDIX A. MTF ER DATA DESCRIPTIONS

Field Name	Field Description	Example Value	Data Type
A_S_C	Appointment Status Code - identifies the status of the patient's appointment	1 = Scheduled Appointment	String
ACV	Alternate Care Value - code that indicates the enrollment status of the beneficiary	A = TRICARE Prime	String
Age	Age of Patient - patient age in years	Patient age < 1 = 0	Numeric
BenCat	Beneficiary Category of patient	ACT = Active Duty	String
D1	First Diagnosis (ICD-9 Code)	ICD-9 Code without Decimals	String
D2	Second Diagnosis- (ICD-9 Code)	ICD-9 Code without Decimals	String
DISP_CD	Disposition Code - represents the outcome of the visit	1 = Released without limitations	String
EDMIS	Enrollment DMIS of Patient - treatment facility that patient is enrolled	0032 = Evans Army Hospital at Ft. Carson, CO	String
EM	Evaluation and Management Code - indicates the level of patient care provided at that particular visit	99201 = Physician Phone Consult	String
FMP	Family Member Prefix identifies the relationship between the beneficiary and the sponsor	01 = 1st Dependent Child	String
Gender	Gender of Patient	F = Female	String
Grade	Grade of Sponsor - pay grade of sponsor at time of treatment	E1-E9 = Enlisted O1-10 = Officer	String
Key_ID	Sort Key for Data Table	Auto-Number Field	Numeric
MARITAL	Marital Status of Patient	S = Single, never married	String

Field Name	Field Description	Example Value	Data Type
OTH_INS	Other Insurance - indicates if patient has other health insurance	1 = Yes 2 = No	String
P1	First Procedure - (CPT4 Code)		String
P2	Second Procedure - (CPT4 Code)		String
P3	Third Procedure - (CPT4 Code)		String
P4	Fourth Procedure - (CPT4 Code)		String
Patient ID	Patient Identification - a unique identifier for each patient in which the SSN is not recognizable for privacy reasons		String
PtZIP	Patient Zip Code	5-digit ZIP Code	String
Service	Branch of Military Service	A = Army	String
TDMIS	Treatment DMIS - facility where treatment took place		String
TSC	Treatment Service Clinic - identifies type of clinic in which treatment took place	0032 = Evans Army Hospital at Ft. Carson, CO	String
Visit Date	Visit Date to MTF ER		Numeric
Visits	Visit to ER, All Values = 1		Numeric
Visits Raw	Visits to MTF ER - original data in 'Visits' field, most fields = 1, some fields = 2 or 3		Numeric
Age Group	Age Group of Patient - created for this research	18 - 23	String
ICD-9 Group	ICD-9 Group of First Diagnosis - created for this research using the established standard ICD-9 groupings (Rogers, 1994)	800-899	String
Visit Day	Day of Visit - created for this research	Monday	String
Visit Month	Month of Visit - created for this research	January	String

APPENDIX B. MTF ER DATA POPULATION

The following is the output from Clementine™'s 'Quality' node which shows the percent each field is populated.



THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX C. CODE TO CREATE ICD-9 GROUPS

This appendix contains the Microsoft Access Module Visual Basic code that reads ICD-9 code and places the code into an ICD-9 group.

```
Option Explicit
Public Function func_ICD9_Group(strPassICD9 As String) As String
   On Error Resume Next

' Declare variables

   Dim strICD9_Group As String
   Dim bol_ICD9_Check As Boolean


' Check for "?....." ICD9 Code


bol_ICD9_Check = strPassICD9 Like "0*"

If bol_ICD9_Check = True Then
  strICD9_Group = "001-139"
  Else

bol_ICD9_Check = strPassICD9 Like "1[0-3]*"

If bol_ICD9_Check = True Then
  strICD9_Group = "001-139"
  Else

  bol_ICD9_Check = strPassICD9 Like "1[4-9]*"
  If bol_ICD9_Check = True Then
        strICD9_Group = "140-239"
        Else

  bol_ICD9_Check = strPassICD9 Like "2[0-3]*"
  If bol_ICD9_Check = True Then
        strICD9_Group = "140-239"
        Else
    bol_ICD9_Check = strPassICD9 Like "2[4-7]*"
    If bol_ICD9_Check = True Then
     strICD9_Group = "240-279"
     Else

     bol_ICD9_Check = strPassICD9 Like "28*"
```

```
If bol_ICD9_Check = True Then
 strICD9_Group = "280-289"
 Else

 bol_ICD9_Check = strPassICD9 Like "29*"
 If bol_ICD9_Check = True Then
  strICD9_Group = "290-319"
  Else

 bol_ICD9_Check = strPassICD9 Like "3[0-1]*"
 If bol_ICD9_Check = True Then
  strICD9_Group = "290-319"
  Else

  bol_ICD9_Check = strPassICD9 Like "3[2-8]*"
  If bol_ICD9_Check = True Then
   strICD9_Group = "320-389"
   Else

   bol_ICD9_Check = strPassICD9 Like "39*"
   If bol_ICD9_Check = True Then
    strICD9_Group = "390-459"
    Else

   bol_ICD9_Check = strPassICD9 Like "4[0-5]*"
   If bol_ICD9_Check = True Then
    strICD9_Group = "390-459"
    Else

    bol_ICD9_Check = strPassICD9 Like "4[6-9]*"
    If bol_ICD9_Check = True Then
     strICD9_Group = "460-519"
     Else

    bol_ICD9_Check = strPassICD9 Like "5[0-1]*"
    If bol_ICD9_Check = True Then
     strICD9_Group = "460-519"
     Else

     bol_ICD9_Check = strPassICD9 Like "5[2-7]*"
     If bol_ICD9_Check = True Then
      strICD9_Group = "520-579"
      Else

      bol_ICD9_Check = strPassICD9 Like "5[8-9]*"
      If bol_ICD9_Check = True Then
       strICD9_Group = "580-629"
       Else
```

```
bol_ICD9_Check = strPassICD9 Like "6[0-2]*"
If bol_ICD9_Check = True Then
 strICD9_Group = "580-629"
 Else

 bol_ICD9_Check = strPassICD9 Like "6[3-7]*"
 If bol_ICD9_Check = True Then
  strICD9_Group = "630-679"
  Else

 bol_ICD9_Check = strPassICD9 Like "6[8-9]*"
 If bol_ICD9_Check = True Then
  strICD9_Group = "680-709"
  Else

 bol_ICD9_Check = strPassICD9 Like "70*"
 If bol_ICD9_Check = True Then
  strICD9_Group = "680-709"
  Else

 bol_ICD9_Check = strPassICD9 Like "7[1-3]*"
 If bol_ICD9_Check = True Then
  strICD9_Group = "710-739"
  Else

 bol_ICD9_Check = strPassICD9 Like "7[4-5]*"
 If bol_ICD9_Check = True Then
  strICD9_Group = "740-759"
  Else

 bol_ICD9_Check = strPassICD9 Like "7[6-7]*"
 If bol_ICD9_Check = True Then
  strICD9_Group = "760-779"
  Else

 bol_ICD9_Check = strPassICD9 Like "7[8-9]*"
 If bol_ICD9_Check = True Then
  strICD9_Group = "780-799"
  Else

 bol_ICD9_Check = strPassICD9 Like "[8-9]*"
 If bol_ICD9_Check = True Then
  strICD9_Group = "800-999"
  Else

  bol_ICD9_Check = strPassICD9 Like "E*"
  If bol_ICD9_Check = True Then
```

```
                strICD9_Group = "E800-E999"
                Else

               bol_ICD9_Check = strPassICD9 Like "V*"
               If bol_ICD9_Check = True Then
                strICD9_Group = "V01-V82"
                Else
                strICD9_Group = "9999"    'Default value

               End If          ' strICD9_Group = "V01-V82"
               End If          ' strICD9_Group = "E800-E999"
              End If           ' strICD9_Group = "800-999"
              End If           ' strICD9_Group = "780-799"
              End If           ' strICD9_Group = "760-779"
              End If           ' strICD9_Group = "740-759"
              End If           ' strICD9_Group = "710-739"
              End If           ' strICD9_Group = "680-709"
              End If
              End If           ' strICD9_Group = "630-679"
             End If          ' strICD9_Group = "580-629"
             End If
            End If          ' strICD9_Group = "520-579"
          End If          ' strICD9_Group = "460-519"
          End If
         End If          ' strICD9_Group = "390-459"
         End If
        End If          ' strICD9_Group = "320-389"
       End If          ' strICD9_Group = "290-319"
       End If
      End If         ' strICD9_Group = "280-289"
     End If        ' strICD9_Group = "240-279"
    End If       ' strICD9_Group = "140-239"
    End If
End If     ' strICD9_Group = "001-139"
End If

func_ICD9_Group = strICD9_Group

End Function
```

LIST OF REFERENCES

[Citation by name: in alphabetical order:]

Berndt, Donald J., "Consumer Decision Support Systems: A Health Care Case Study," *Proceedings of the Thirty-Fourth Annual Hawaii International Conference on System Sciences,* IEEE Computer Society Press, 2000.

Berson, Alex, and Smith, Stephen J., *Data Warehousing, Data Mining, and OLAP*, McGraw-Hill, 1997.

Clementine™ User Guide Version 5, Integral Solutions Limited, 1998.

Cognos PowerPlay™ Discovering PowerPlay™, Cognos Inc., 2000.

Han, Jiawei, and Kamber, Micheline, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

Hogle, Oliver, and others, "On Supporting Medical Quality with Intelligent Data Mining," *Proceedings of the Thirty-Fourth Annual Hawaii International Conference on System Sciences,* IEEE Computer Society Press 2000.

Introduction to Clementine™, SPSS Inc., 1999.

"Introduction to the U.S. Army Medical Department," [www.armymedicine.army.mil]. 2000.

Kasif, Simon, and others, "Data Mining Research: Opportunities and Challenges," ed. Grossman, Robert, [www.ncdm.uic.edu/dataminingresearch.htm]. January 1999.

Lin, T.Y., "Data Mining: Granular Computing Approach," *Methodologies for Knowledge Discovery and Data Mining: Third Pacific Asia Conference, PAKDD-99, Beijing, China, April 26-28, 1999 Proceedings,* Springer-Verlay, 1999.

McNitt, T. R., Colonel, Medical Corps, U.S. Army, "Overview Briefing, TRICARE Central Region," 2000.

Rogers, Gregg, ed., *ICD-9 CM, 1995: International Classification of Diseases, 9th Revision: Clinical Modification,* 4th Edition, vol. 1-3, Practice Management Information Corporation, 1994.

S-PLUS 2000 User's Guide, Data Analysis Products Division, MathSoft, 2000.

Telephone Conversation between McNitt, T. R., Colonel, Medical Corps, U.S. Army, Lead Agent, TRICARE Central Region, and the author, 31 October 2000.

"The Assistant Secretary of Defense for Health Affairs Responsibilities and Functions," [www.ha.osd.mil/ha.htm]. 23 October 2000.

"The United States Air Force Medical Service Organizations," [sg-www.satx.disa.mil/af/sg/orgs/orgs.cfm]. 15 February 2001.

"What is TRICARE?," Military Health System Web Site, [www.tricare.osd.mil/tricare/beneficiary/whatistricare.html]. 18 September 2000.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center 2
   8725 John J. Kingman Road, STE 0944
   Fort Belvoir, VA 22060-6218

2. Dudley Knox Library 2
   Naval Postgraduate School
   411 Dyer Road
   Monterey, CA 93943-5101

3. CPT Cary V. Ferguson 2
   3638 Honolulu Ave
   Eugene, OR 97404

4. Professor Daniel R. Dolk, Code IS/Dk 1
   Naval Postgraduate School
   Monterey, CA 93943-5118

5. Professor Samuel E. Buttrey, Code OR/Sb 1
   Naval Postgraduate School
   Monterey, CA 93943-5118

6. Lead Agent, TRICARE Central Region 3
   Regional Health Service Ops
   Speker Ave
   Bldg. 1011
   Colorado Springs, CO 80913

7. Regional Health Service Ops 1
   Attn: Tony Rogers
   Speker Ave
   Bldg. 1011
   Colorado Springs, CO 80913

8. Regional Health Service Ops 1
   Attn: Terri Cheyney
   Speker Ave
   Bldg. 1011
   Colorado Springs, CO 80913

9. Chair, IS Academic Group, Code IS 1
   Naval Postgraduate School
   Monterey, CA 93943-5118