Development of a Clinical Pathways Analysis System with Adaptive Bayesian Nets and Data Mining Techniques

D. KOPEC*, G. SHAGAS*, D. REINHARTH**, S. TAMANG*
** Brooklyn College, Department of Computer Science,*
2900 Bedford Avenue, Brooklyn, NY 11210
*** Albert Einstein School of Medicine, Bronx, NY*

Abstract: The use and development of software in the medical field offers tremendous opportunities for making health care delivery more efficient, more effective, and less error-prone. We discuss and explore the use of clinical pathways analysis with Adaptive Bayesian Networks and Data Mining Techniques to perform such analyses. The computation of "lift" (a measure of completed pathways improvement potential) leads us to optimism regarding the potential for this approach.

1 Introduction

1.1 Purpose of the Study

We identify important aspects of patient-related data, such as age, weight, treatment, medication, monitoring process; and clinical outcomes, such as pressure injuries, which are found in patient records. We also create an artificial set of such records and use these for clinical pathways analyses. We use data mining software as a tool to evaluate an action-oriented framework for making decisions directed towards improving care delivery, organizational performance, and patient outcomes. We build a decision support system (DSS) model based on the adaptive Bayesian networks (ABN) methodology and a data mining engine. For clinical pathways analysis we perform (1) installation and tuning of an Oracle database and data mining engine and (2) the following sequence of data mining tasks: collect and preprocess patient data, build and test a model, and compute lift, which is a measure of a completed pathway's improvement potential.

In addition, we make recommendations for a specific configuration based upon the prototype’s capabilities and performance, summarize our findings, and identify opportunities for the future.

1.2 Background and Significance/Justification

Clinical pathways (*CareMaps®* is a name used in some healthcare organizations) define the essential components of the complex care process. For example, a patient with total hip replacement is at special risk for developing pressure ulcers. In the CareMap® developed for total hip replacement, skin care is addressed through specific key interventions such as a consultation with a nutritionist or physical therapist. We believe that there are a number of untapped, yet readily available resources providing basic patient data, which could be used to improve the performance of clinical pathways. We propose the use of ABN's to provide a fast, scalable, non-parametric means of extracting predictive information from data with respect to a target attribute. The target attribute in the application we have developed relates directly to the clinical outcome.

1.3 Hypothesis

We can select a group of patients who are the most responsive to treatment by analyzing a set of medical records, using data mining. Data mining is the process of discovering hidden, previously unknown, and usable information from a large amount of seemingly unrelated data. Data mining can be used as a tool for an action-oriented framework for making decisions directed towards improving health care delivery, organizational performance, and patient outcomes when an action orient framework is implemented. Lift is a measure of how much better prediction results are using a model than could be obtained by chance. Lift indicates how well the model improved the predictions over a random selection given actual results. The lift procedure can be used, for example, to select a group of patients who are the most responsive to a treatment.

1.4 Programming

The research application prototypes were developed on both UNIX and Windows platforms using the following techniques: data mining, on-line transaction and analytical processing.

Clinical Pathways Analysis: The Oracle 9i Database was installed and tuning was performed; the *init.ora* file is presented in Appendix A. Oracle Data Mining (ODM) has been installed; configuration parameters are presented in Appendix A

1.5 Limitations

1. Consumer off-the-shelf (COTS) data mining programs provide two algorithms: Naive Bayesian for Classifications and Predictions, and Association Rules for finding patterns of co-occurring events. The systems knowledge is structured by Bayesian networks. One potential problem that is seldom recognized is the remote possibility that a system's user may wish to violate the distribution of probabilities upon which the system is built for an exceptional case. A DSS that is unable to respond to some previously unforeseen event may end up putting the patient at risk.
2. Another problem is the computational difficulty of exploring a previously unknown network. To calculate the probability of any branch of the network, all branches must be calculated. While the resulting ability to describe the network can be performed in linear time, this process of Naive Bayes network discovery is an extremely complicated task which may well be either be too costly to perform, or impossible, given the number and combinations of variables [2].
3. The third problem centers on the quality and extent of the prior beliefs used in Bayesian inference processing. A Bayesian Network is only as useful as this prior knowledge is

reliable [2]. An assumption of independence among the pieces of evidence “cannot be justified a priori” for Bayesian systems, in general, and especially in areas such as medicine[3].

2. Requirements Analysis and System Design

2.1 Introduction and Preliminary Investigation Phase

Clinical Pathway Analysis: Using data standardization, health care providers are able to analyze data across clients to determine what interventions make a difference and what outcomes are produced. For example, in the report, "Nursing Report Card for Acute Care Settings," the American Nurses Association states that there are many inputs which influence outcomes, one of which is nursing [11]. Nursing inputs include: health status assessment, surveillance, symptom control, infection control, physical safety, nutrition/hydration, therapeutic procedures, patient education, and discharge planning. One way in which nursing data can be standardized to capture this information is to use recommended classification systems. In the Unites States, the American Nurses Association has designated four classification systems for standardizing nursing data.

2.2 Software Requirements Analysis Phase

Clinical Pathway Analysis The effectiveness of an application may be assessed based on performance, scalability, security and availability. An application should be able to provide reports in the form of human-understandable rules. For example: "*If Patient uses an abduction pillow and patient is turned every 2 hours, the likelihood this Patient’s skin integrity is maintained is YES*." Such an application could be based on an ABN. Reasoning with uncertain knowledge and beliefs has long been recognized as an important research issue in artificial intelligence (AI).

Data mining models are based on one of two kinds of learning: *supervised* and *unsupervised* (sometimes referred to as directed and undirected learning). Supervised learning functions are typically used to predict a value. Unsupervised learning functions are typically used to find the intrinsic structure, relations, or affinities in a body of data but no classes or labels are assigned a priori. Examples of unsupervised learning algorithms include *k-means clustering* and *a priori association* rules. An example of supervised learning algorithms includes the Naive Bayes algorithm for classification.

Most data mining applications support the following functions: *Classification* (supervised), *Clustering* (unsupervised), *Association Rules* (unsupervised) and *Attribute Importance* (supervised). All functions go through an adaptive Bayesian network build process.

A *build result* contains the model details. A *test result*, for classification models, contains the model accuracy and references the confusion matrix, which measures the correctness of predictions made from a text task. A *lift result* of the lift elements is calculated on a per-quantile basis. This measures how much better prediction results are using a model that could be obtained by chance. An *apply* task to data produces scores or predictions with an associated probability. This output may include predicted values, associated probabilities, key values and other associated data. Classification models are also used in a testing phase in which predicted values are compared with actual target values; this model can also be used to compute lift. Classification and Clustering models can both be used to score data tables. Attribute importance supports build since it produces an importance ordering of the attributes[13].

2.3 Decision Analysis Phase

For our study we decided to use Oracle 9i database Enterprise Edition, primarily, due to licensing policy. Database applications were programmed in the Java and PL/SQL languages. The Java programming language was used for most of the programs. Java gives an object-oriented, platform-independent, multithreaded programming environment.

2.4 Design, Construction and Implementation Phases

1. The project required a database and ODM application to be installed (see Appendix A); the typical installation and tuning were performed [17].

3. Methodology

3.1 Data Mining and Clinical Pathways

A *mining result* contains the end products of one of the following mining tasks: *build*, *test*, *compute lift*, or *apply*.

We will perform the following operations of the data mining process:

1. Data preparation (binning and attribute importance),
2. Model creation (corresponding to "build" task),
3. Model testing and Lift computing.

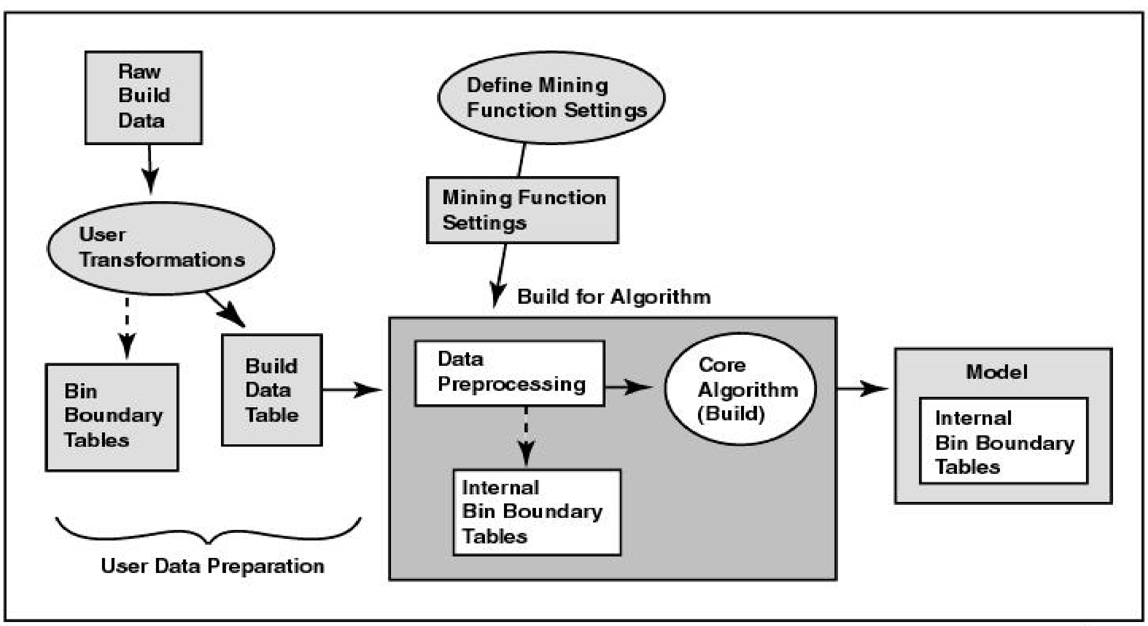

Figure 1 The Adaptive Bayesian Network Build Process (adopted from Oracle)

The left side of the diagram shows the steps to perform before data can be supplied to the algorithm that will build the model. These "pre-processing" steps are termed in the figure "user data preparation." The pre-processing result is the build data table. The build data table and the mining function settings are shown as boxes with arrows pointing to the algorithm, indicating that they are supplied (as input) to the algorithm.

The user may bin the data manually, which produces (in addition to binned data) bin boundary tables (see Appendix A, Tables A.1-A.4). In cases where there is no clear way to determine optimal or preferred bin boundaries, ODM can perform the binning automatically. This approach to automatic binning (not used in this research) also results in bin boundary tables that are associated with the model.

The mining function settings are parameters that determine the type of model to build. These settings can be specified by the user or, if the user elects not to specify certain parameters, by ODM (see Appendix A). The algorithm then builds the model, using the pre-processed data, the mining function settings, any internal bin boundary tables, and the core algorithm.

A total of 55 samples were obtained by simple random sampling from a Palliative Health Care Center. We identified important aspects of a patient's treatment, such as age, length of stay, weight, treatment, medication, monitoring process, and clinical outcomes (such as pressure injuries) from this data. This data was used as prototype patient data for creation of binning tables (see Appendix A, Tables A.1-A.4). Samples were used for *discretization.* This groups related values together under a single value (or bin). *Bin* boundaries for one or more attributes, both for categorical data - a list of categorical values to be contained in each bin - and for numerical data, a set of upper and lower boundaries, for each of the bins, were calculated.

We also created an artificial set of such records and used them for evaluation of the Bayesian network's performance. A total of 5000 data sets were obtained from the Oracle 9i database, ODM Schema, CENSUS_2D_TEST_BINNED Table. The data was slightly altered and was employed as hypothetical medical data.

During the next step, data should be separated among *bins*. ODM provides four ways to bin data:

1. For categorical data, a list of categorical values to be contained in each bin.
2. For numerical data, a set of upper and lower boundaries for the bins.
3. Top N most frequent items: For categorical attributes only, the user selects the value N and the name of the "other" category. ODM determines the N most frequent values and puts all other values in the "other" category.
4. Quantile binning: For numerical attributes only, the values are sorted, and the values are divided into the number of user-specified quantiles. ODM determines which values are in which bins. (The *quantile* of a distribution of values is a number x_p such that a proportion p of the population values are less than or equal to x_p.)

Clinical pathways' or CareMaps' values and interventions are binned as shown in Tables A.1-A.4 (see Appendix A). Table A.1 displays the original data, before binning. Table A.2 shows the bin boundaries for numeric data; Table A.3 shows bin boundaries for categorical data. Table A.4 shows the results of binning.

4. Findings

4.1 Results

The following results were produced by the critical pathways analysis and data mining system.

4.1.1 Clinical Pathways Analysis and Data Mining.

We analyzed the hypothetical medical data by data mining operations such as building and testing a model, and to computing lift. We also conducted performance analysis of the mining procedure as a function of the Java pool size and the maximum ABN depth. We examined several ways of improving software performance, including tuning database buffer cache size (See Appendix A, Section 1).

We identified six independent variables or input, such as age, weight, treatment, medication, monitoring process, length of stay, and one dependent variable or output, which is the pressure injury (ulcer). A pressure ulcer is a lesion caused by unrelieved pressure, generally occurring over bony prominences [18]. The data mining engine splits the dataset into two subsets: the *training set* and the *test set*. The data mining algorithms used the training set while generating the Bayesian network, and after training we used a test set to test the accuracy of the classifiers on a new set of examples.

The data mining results were obtained by executing the adaptive Bayesian network "build" and "lift and test" ODM programs (see above and Appendix D). A *build result* contains the model details, a *test* result contains the model accuracy and references the confusion matrix. A *lift result* of the lift elements is calculated on a per-quantile basis. An *apply* task to data produces scores or predictions with an associated probability. For this study, a patient is classified as having positive outcomes if his/her posterior probability for a positive outcome is greater than 0.5. This parameter is arbitrary; values less than 0.2 generate large and often conclusive changes from pre-test to post-test probabilities; 0.2 - 0.5 generate small, but sometimes important changes, and values 0.5 - 1 rarely make significant changes.

The result of a test task for classification models is a *confusion matrix*. A confusion matrix (Table 1) provides a quick understanding of the model's accuracy and the types of errors the model makes when scoring records. The row indices of a confusion matrix correspond to actual values observed and used for model building; the column indices correspond to predicted values produced by applying the model. For any pair of actual/predicted indices, the values in the table indicate the number of records classified in that pairing.

Table 1 Confusion Matrix for the Naive Bayes Classification Method
("s" denotes positive outcome while "!s" denotes negative outcome.)

	Predicted s	Predicted !s
Actual s	464	165
Actual !s	38	167

In the case of Table 1, the predictions were correct 464+167= 631 times, and incorrect 165 + 38 = 203 times. The sum of the values (631+203=834) in the matrix is equal to the number of scored records in the input data table. The error rate is 203/834=0.243; the accuracy rate, therefore, is = 631/834=0.757. The average classification accuracy of a similar Bayesian network for dementia severity determination was 0.72±0.07 [19] our result is within the window of common values.

Lift is a measure of how much better predicted results are using a model than would be obtained by chance alone; in other words, how well does the model improve predictions over a random selection given the actual results? Positive cases with highest confidence come first, followed by positive cases with lowest confidence. Based on this ordering, records were partitioned into quantiles (bins), and the following statistics were calculated: (1) Cumulative part of records, (2) Bin Lift and (3) Lift cumulative. Applying the "model compute lift" procedure to an adaptive Bayesian network model and the data set leads to the results presented in the Table 2:

Table 2 Compute Lift Procedure Output
(Assumes the number of elements=10, ABN maximum depth =10.)

Bin	Cumulative part of records	Bin Lift	Lift cumulative
0	0.1	3.0835	3.0835
1	0.2	1.9050	2.4943
2	0.3	1.7436	2.2440
3	0.4	1.3561	2.0220
4	0.5	1.0783	1.8351
5	0.6	0.4901	1.6127
6	0.7	0.1471	1.4048
7	0.8	0.0980	1.2424
8	0.9	0.0490	1.1105
9	1.0	0	1

The graphical data representation can be found in Figure 2. The data mining procedure separates all the records into 10 equal bins. For example, suppose that 20% of the patients who received treatment without using the model would have a positive response. From a randomly selected 10% of the patients only 2% of the entire set (0.2 of 10%) will respond positively. However, using the model to separate patients, we can calculate that for bin number 0 the probability of a positive outcome will be 3.0835 times higher, or 0.1*20%*3.0835 = 6.167%.

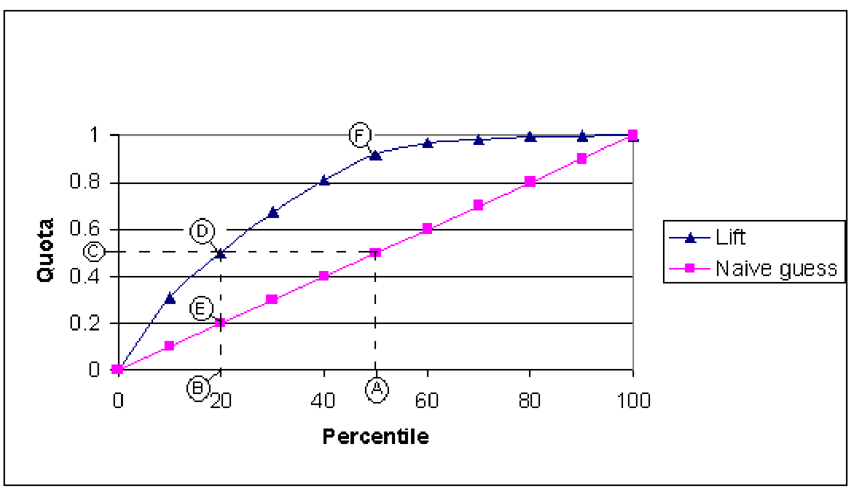

Figure 2 Lift (incremental improvement) Over the Naïve Guess

The lift quota was obtained by multiplying the cumulative part of the records by the lift cumulative. For example, bin 4 has the cumulative part of its record equal to 0.5 and its lift cumulative equal to 1.8351. The product of these values is equal to 0.5*1.8351=0.918 (point “F”, Figure 2).

For example, let’s assume a hospital expects to have a 15% success rate with a specific treatment. The probability of a positive outcome after a patient’s treatment is equal to 30%. The patient’s treatment cost is $1,000 per day. The ABN model was applied to 100 patients with similar diagnosis and treatment, and thus each percentage point is equal to $1,000. Randomly distributed treatment should cover 50% (that is 15% if 50% of 30%, points “A” and “C”, Figure 2) of the patients. The application of the “lift” procedure can separate 20% (point “B”) of the patients with cumulative lift of 2.4943, who are the most responsive to the treatment (points “D” and “E”, Figure 2). The treatment will be 2.5 times more effective for the selected patients (that is 20% equals 50%/2.5). The savings, therefore, equals 100*((0.5 * $1,000)-(0.2 * $1,000)), or $30,000 daily.

4.2 Discussion and Recommendations: Clinical Pathways and Data Mining

We conducted performance analysis of the mining procedure as a function of the Java pool size and the maximum ABN depth. We also analyzed the hypothetical medical data by data mining operations such as to build and test a model, and compute lift. Database tuning can significantly improve data mining performance. Increasing the database cache size leads to a high cache hit ratio. Increasing the Java pool size leads to a decrease in the time required for all data mining procedures.

Increasing the maximum ABN depth slightly increases a procedure’s accuracy, but significantly increases the computation build time. We do recommend changing the default value from 10 to 15, because this could increase accuracy. Additional increasing of the maximum ABN depth would not bring overall benefits.

The build and test a model and compute lift data mining tasks have been performed on hypothetical medical data. We have to test the application on a real data set. Using the robust Bayes classifier could increase the accuracy of outcome prediction for intensive-care patients up to 78-84% [1].

5. Summary

Data mining can be used as a tool for an action-oriented framework for making decisions directed toward improving health care delivery, organizational performance, and patient outcomes. A decision support system model, based on the ABN methodology, and the Oracle data mining engine, was developed to conduct performance measurements of ABN, such as accuracy and lift. We conducted an analysis of the mining procedure and recommend allocating 24 megabytes for the database buffer cache size and 32 megabytes for the Java pool size, which could be increased in the case of a switch to medium or large systems. Increasing the maximum ABN depth from value 10 slightly increases a procedure’s accuracy but significantly increases the computation build time. The build and test a model, and compute lift data mining tasks have been performed on hypothetical medical data. The results are promising, and we have to test a model on real data.

Developments and changes in software applications today present incredible opportunities for improving health care delivery, organizational performance, and patient outcomes. Hence we are optimistic that great progress can and will be made in this field.

References

[1] Ramoni, M., Sebastiani, P., Dybowski, R. *Robust Outcome Prediction for Intensive-Care Patients. Methods of Information in Medicine* 2001; 40: 39-45.

[2] Niedermayer, D., I.S.P., B.Sc., B.A., M.Div. *An Introduction to Bayesian Networks and their Contemporary Applications* (1998). Available: http://www.niedermayer.ca/papers/bayesian/bayes.html.

[3] Luger, G. *Artificial Intelligence. Structural and Strategies for Complex Problem Solving*, 4th edition. Harlow, England, Addison Wesley, 2002.

[4] Belluck, P. *Prosecutors Say Greed Drove Pharmacist to Dilute Drugs*. The New York Times Newspaper. August 18, 2001.

[5] Henderson, M. *Health Care Fraud and Abuse: Diagnosis Related Group (DRG) Creep*. 2003 PricewaterhouseCoopers. Available: http://www.pwcglobal.com/extweb/manissue.nsf/.

[6] Fowler, M. and Scott, K. *UML Distilled: A Brief Guide to the Standard Object Modeling Language*. 2nd edition. Boston, MA, Addison-Wesley Pub Co., 1999.

[7] Intelligent Business Machines Corp. *Software. DB2 Information Management. DB2 Product Family. DB2 Intelligent Miner* (2003). Available: http://www-3.ibm.com/software/data/iminer/

[8] Dlugacz, Y., Stier, L., Greenwood, A. *Changing the System: A Quality Management Approach to Pressure Injuries.* Article: Journal of Health Care Quality 107 (September/October 2001). Available: http://www.allenpress.com/jhq/107/107.htm

[9] Leo, J., M.D. APACHE Medical Systems, Inc. *Mission Critical: Shopping for the Right Outcomes Management Systems.* Available: http://www.apache-msi.com/ttdx.shtml. [August 20, 2003].

[10]Cerner Corp. *Outcomes Management: Committed to Critical Care* . Available: http://www.cerner.com/products/products_4a.asp?id=277 [August 20, 2003].

[11] American Nurses Association. (1995). *Nursing report card for acute care settings.* Fairfax, VA: Lewin-VHI, Inc..

[12] Westra, B. *Standardized Data: The Foundation for Home Care Report Cards*. The Quality Messenger, 1(3), 5, 7, 8, 1994. Available: http://www.carefacts.com/art11.htm.

[13] 2002b *Data Mining Concepts*. Release 9.2.0.2. Part # A95961-02 (March 2002).

[14] JCAHO (Joint Commission on Accreditation of Healthcare Organizations), *Performance Measurement in HealthCare.* Available: http://www.jcaho.org/pms/index.htm [August 20, 2003].

[15] Microsoft Corp. *Microsoft SQL Server. Powerful and Accessible Analysis for Business* White Paper (May 31, 2001). Available: http://www.microsoft.com/sql/evaluation/BI/AnalysisSvcWP.asp.

[16] 2002a *Oracle9i Database Administrator's Guide*. Release 2 (9.2). Part # A96521-01 (March 2002).

[17] 2002e *Oracle9i Database Installation Guide*. Release 2 (9.2.0.1.0) for Windows. Part # A95493-01 (May 2002).

[18] Posey, M. *Quality Indicators: Skin Care.* American Society of Consultant Pharmacists, (Feb 2004). Available: http://www.ascp.com/public/pubs/tcp/1996/jul/mds2.html.

[19] Shankle, R., Mani, S., Dick, M and Pazzani, M. *Simple Models for Estimating Dementia Severity Using Machine Learning.* MedInfo'98: 9th World Congress on Medical Informatics, Seoul, Korea, August 1998 Available: http://www.ics.uci.edu/~pazzani/Publications/medinfo98.pdf

Appendices

Appendix A. Oracle Database and Data Mining Processes

Section 1. Data before Binning and Bin Boundaries for Numeric Data

Table A.1 Binning Clinical Pathways Analysis Data (before binning)

PT_ID	AGE	LOS[1]	WEIGHT	TREATMENT	MEDICATION	MONITOR[2]
2	57	4	150	Abduction pillow	None	Admission Assessment
8	66	10	120	None	Spray	None
10	74	3	180	Abduction pillow	Cream	Daily Assessment
11	63	5	220	Trapeze	Cream	Daily Assessment
41	80	7	190	Trapeze	Spray	Admission Assessment

Note: [1]LOS stands for Length of Stay
[2]MONITOR stands for Monitoring Process

Table A.2 Binning Clinical Pathways Analysis Data: Bin boundaries for Numeric Data

COLUMN_NAME	LOWER_ BOUNDARY	UPPER_ BOUNDARY	BIN_ID	DISPLAY_NAME
AGE	0	49	1	0-49
AGE	50	59	2	50-59
AGE	60	69	3	60-69
AGE	70	Up	4	70-Up
LOS	0	3	1	0-3
LOS	4	7	2	4-7
LOS	8	Up	3	8-Up
WEIGHT	0	79	1	0-79
WEIGHT	80	149	2	80-149
WEIGHT	150	Up	3	150-Up

Table A.3 Binning Clinical Pathways Analysis Data***:*** Bin Boundaries for Categorical Data

COLUMN_NAME	CATEGORY	BIN_ID	DISPLAY_NAME
TREATMENT	Abduction Pillow	1	Abduction Pillow
TREATMENT	Trapeze	2	Trapeze
TREATMENT	None	3	None
MEDICATION	Spray	1	Spray

COLUMN_NAME	CATEGORY	BIN_ID	DISPLAY_NAME
MEDICATION	Cream	2	Cream
MEDICATION	None	3	None
MONITOR	Admission Assessment	1	Admission Assessment
MONITOR	Daily Assessment	2	Daily Assessment
MONITOR	None	3	None

Table A.4 Binning Clinical Pathways Analysis Data***:*** Assignment of Original Data to Bins (fragment)

PT_ID	AGE	LOS	WEIGHT	TREATMENT	MEDICATION	MONITOR
2	2	2	3	1	3	1
8	3	3	2	3	1	3
10	4	1	3	1	2	2
11	3	2	3	2	2	2
41	4	2	3	2	1	1

Appendix B. Abbreviations

Oracle is a registered trademark, and Oracle9*i*, *SQL*Net*, and *SQL*Plus* are trademarks or registered trademarks of the Oracle Corporation. *Java*™ and all *Java*-based marks are a trademark or registered trademark of Sun Microsystems, Inc. Windows refers to the Windows NT, Windows 2000, and Windows XP operating systems of Microsoft Corp.

(a) Abbreviations

ABN	Adaptive Bayesian Network
COTS	Consumer Off-The Shelf (product)
DSS	Decision Support System
ODM	Oracle9*i* Data Mining
PL/SQL	Procedural Language / Structured Query Language