International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:1, No:8, 2007

Data Mining for Cancer Management in Egypt Case Study: Childhood Acute Lymphoblastic Leukemia

Nevine M. Labib, and Michael N. Malek

Abstract**—***Data Mining* aims at discovering knowledge out of data and presenting it in a form that is easily comprehensible to humans. One of the useful applications in Egypt is the Cancer management, especially the management of Acute Lymphoblastic Leukemia or ALL, which is the most common type of cancer in children.

This paper discusses the process of designing a prototype that can help in the management of childhood ALL, which has a great significance in the health care field. Besides, it has a social impact on decreasing the rate of infection in children in Egypt. It also provides valubale information about the distribution and segmentation of ALL in Egypt, which may be linked to the possible risk factors.

Undirected Knowledge Discovery is used since, in the case of this research project, there is no target field as the data provided is mainly subjective. This is done in order to quantify the subjective variables. Therefore, the computer will be asked to identify significant patterns in the provided medical data about ALL. This may be achieved through collecting the data necessary for the system, determimng the data mining technique to be used for the system, and choosing the most suitable implementation tool for the domain.

The research makes use of a data mining tool, Clementine, so as to apply Decision Trees technique. We feed it with data extracted from real-life cases taken from specialized Cancer Institutes. Relevant medical cases details such as patient medical history and diagnosis are analyzed, classified, and clustered in order to improve the disease management.

Keywords**—**Data Mining, Decision Trees, Knowledge Discovery, Leukemia.

I. Introduction

DATA Mining or "the efficient discovery of valuable, non-obvious information from a large collection of data" [1] has a goal to discover knowledge out of data and present it in a form that is easily comprehensible to humans.

There are several *data mining techniques*, such as Market Basket Analysis, Memory-Based Reasoning (MBR), Cluster Detection, Link Analysis, Decision Trees, Artificial Neural Networks (ANNs), Genetic Algorithms, and On-Line Analytic Processing (OLAP).

Michael N. Malek is with Department of Computers and Information Systems, Faculty of Management, Sadat Academy for Management Sciences, Cairo, Egypt.

Decision Trees may be used for classification, clustering, prediction, or estimation.One of the useful medical applications in Egypt is the management of *Leukemia* as it accounts for about 33% of pediatric malignancies. [2]

Childhood Acute Lymphoblastic Leukemia (also called acute lymphocytic leukemia or ALL) is a cancer of the blood and bone marrow. This type of cancer usually gets worse quickly if it is not treated. It is the most common type of cancer in children.

There are different approaches in Data Mining, namely *hypothesis testing* where a database recording past behavior is used to verify or disprove preconceived notions, ideas, and hunches concerning relationships in the data, and *knowledge discovery where* no prior assumptions are made and the data is allowed to speak for itself. As for knowledge discovery, it may be directed or undirected. *Directed knowledge discovery* tries to explain or categorize some particular data field while *undirected knowledge discovery* aims at finding patterns or similarities among groups of records without the use of a particular target field or collection of predefined classes.

The remainder of this paper is organized as follows. In Section 2, we give a brief explanation of the data mining concept. In Section 3, we give a recent review of similar work in the field. In Section 4, we describe the implemented approach in detail. As for the results and conclusions, they are provided in Section 5.

II. Data Mining Concepts

A. Definition

Data mining may be defined as "the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules" [3].

Hence, it may be considered mining knowledge from large amounts of data since it involves knowledge extraction, as well as data/pattern analysis [4].

B. Tasks

Some of the tasks suitable for the application of data mining are c*lassification, estimation, prediction, affinity grouping, clustering,* and *description. Some* of them are best approached in a top-down manner or *hypothesis testing while* others are best approached in a bottom-up manner called *knowledge discovery either* directed or undirected.

As for Classification, it is the most common data mining task and it consists of examining the features of a newly presented object in order to assign it to one of a predefined set of classes.

While classification deals with discrete outcomes, estimation deals with continuously-valued outcomes. In real-life cases, estimation is often used to perform a classification task.

Prediction deals with the classification of records according to some predicted future behavior or estimated future value.

Both Affinity grouping and market basket analysis have as an objective to determine the things that can go together.

Clustering aims at segmenting a heterogeneous population into a number of more homogeneous subgroups or clusters that are not predefined.

Description is concerned with describing and explaining what is going in a complicated database so as to provide a better understanding of the available data.

C. The Virtuous Cycle of Data Mining

The four stages of the virtuous cycle of data mining are:

1. Identifying the problem: where the goal is to identify areas where patterns in data have the potential of providing value.
2. Using data mining techniques to transform the data into actionable information: for this purpose, the produced results need to be understood in order to make the virtuous cycle successful. Numerous pitfalls can interfere with the ability to use the results of data mining. Some of the pitfalls are bad data formats, confusing data fields, and lack of functionality.

In addition, identifying the right source of data is crucial to the results of the analysis, as well as bringing the right data together on the computing system used for analysis.

3. Acting on the information: where the results from data mining are acted upon then fed into the measurement stage.
4. Measuring the results: this measurement provides the feedback for continuously improving results. These measurements make the virtuous cycle of data mining *virtuous*. Even though the value of measurement and continuous improvement is widely acknowledged, it is usually given less attention than it deserves.

III. Data Mining Techniques

Some of the mostly used techniques are the following:

A. Neural Networks

A Neural network may be defined as "a model of reasoning based on the human brain" [5]. It is probably the most common data mining technique, since it is a simple model of neural interconnections in brains, adapted for use on digital computers. It learns from a training set, generalizing patterns inside it for classification and prediction. Neural networks can also be applied to undirected data mining and time-series prediction.

B. Decision Trees

Decision trees are a way of representing a series of rules that lead to a class or value. Therefore, they are used for directed data mining, particularly classification. One of the important advantages of decision trees is that the model is quite explainable since it takes the form of explicit rules. This allows the evaluation of results and the identification of key attributes in the process. The rules, which can be expressed easily as logic statements, in a language such as SQL, can be applied directly to new records.

C. Cluster Detection

Cluster detection consists of building models that find data records similar to each other. This is inherently undirected data mining, since the goal is to find previously unknown similarities in the data. Clustering data may be considered a very good way to start any analysis on the data. Self-similar clusters can provide the starting point for knowing what is in the data and for figuring out how to best make *use* of it.

D. Genetic Algorithms

Genetic algorithms (GA), which apply the mechanics of genetics and natural selection to a search, are used for finding the optimal set of parameters that describe a predictive function. Hence, they are mainly used for directed data mining. Genetic algorithms use many operators such as the selection, crossover, and mutation to evolve successive generations of solutions. As these generations evolve, only the most predictive survive, until the functions converge on an optimal solution.

IV. Related Work

A. Classification using Partial Least Squares with Penalized Logistic Regression [6]

In this paper, the classification problem is viewed as a regression one with few observations and many predictor variables. A new method is proposed combining partial least squares (PLS) and Ridge penalized logistic regression. The basic methods are then reviewed based on PLS and/or penalized likelihood techniques, their interest in some cases are outlined and their sometimes poor behavior is theoretically explained. This procedure is compared with these other classifiers. The predictive performance of the resulting classification rule is illustrated on three data sets:
Leukemia, Colon and Prostate.

B. Marker Identification and Classification of Cancer Types using Gene Expression Data and SIMCA [7]

The objective of this research was the development of a computational procedure for feature extraction and classification of gene expression data. The Soft Independent Modeling of Class Analogy (SIMCA) approach was implemented in a data mining scheme in order to allow the identification of those genes that are most likely to confer robust and accurate classification of samples from multiple tumor types. The proposed method was tested on two different microarray data sets where the identified features represent a

rational and dimensionally reduced base for understanding the biology of diseases, defining targets of therapeutic intervention, and developing diagnostic tools for classification of pathological states. The analysis of the SIMCA model residuals was able to identify specific phenotype markers. On the other hand, the class analogy approach allowed the assignment to multiple classes, such as different pathological conditions or tissue samples, for previously unseen instances.

C. Data Mining the NCI Cancer Cell Line Compound GI(50) Values: Identifying Quinone Subtypes Effective Against Melanoma and Leukemia Cell Classes [8]

Using data mining techniques, a subset (1400) of compounds from the large public National Cancer Institute (NCI) compounds data repository has been studied. First, a functional class identity assignment for the 60 NCI cancer testing cell lines was carried out via hierarchical clustering of gene expression data. Comprised of nine clinical tissue types, the 60 cell lines were placed into six classes-melanoma, leukemia, renal, lung, and colorectal, and the sixth class was comprised of mixed tissue cell lines not found in any of the other five classes. Then, a supervised machine learning was carried out using the GI(50) values tested on a panel of 60 NCI cancer cell lines. With this approach, identified two small sets of compounds that were most effective in carrying out complete class separation of the melanoma, non-melanoma classes and leukemia, non-leukemia classes, were identified. As for attempts to subclassify melanoma or leukemia cell lines based upon their clinical cancer subtype, they met with limited success.

D. Cancer Surveillance using Data Warehousing, Data Mining, and Decision Support Systems [9]

This study discussed how data warehousing, data mining, and decision support systems can reduce the national cancer burden or the oral complications of cancer therapies. For this goal to be achieved, it first will be necessary to monitor populations; collect relevant cancer screening, incidence, treatment, and outcomes data; identify cancer patterns; explain the patterns, and translate the explanations into effective diagnoses and treatments.

Such data collection, processing, and analysis are time consuming and costly. Success is highly dependent on the abilities, skills, and domain knowledge of the interested parties. Even the most talented and skilled parties have incomplete knowledge about the study domain, pertinent information technology (IT), and relevant analytical tools. Data sharing across interested groups is limited. Consequently, much useful information may be lost in the cancer surveillance effort.

E. Data Mining with Decision Trees for Diagnosis of Breast Tumor in Medical Ultrasonic Images [10]

In this study, breast masses were evaluated in a series of pathologically proven tumors using data mining with decision tree model for classification of breast tumors.

Accuracy, sensitivity, specificity, positive predictive value and negative predictive value are the five most generally used objective indices to estimate the performance of diagnosis results. Sensitivity and specificity are the most two important indices that a doctor concerned about. With sensitivity 93.33% and specificity 96.67%, the proposed method provides objective evidences for good diagnoses of breast tumors.

V. Research Methodology

A. What kind of data are we working on?

Childhood Acute Lymphoblastic Leukemia [13]

Childhood acute lymphoblastic leukemia (also called Acute Lymphocytic Leukemia or ALL) is a cancer of the blood and bone marrow. This type of cancer usually gets worse quickly if it is not treated. It is the most common type of cancer in children.

Normally, the bone marrow produces stem cells (immature cells) that develop into mature blood cells.
In ALL, too many stem cells develop into a type of white blood cell called lymphocytes. These lymphocytes may also be called lymphoblasts or leukemic cells.
In ALL, the lymphocytes are not able to fight infection very well. Also, as the number of lymphocytes increases in the blood and bone marrow, there is less room for healthy white blood cells, red blood cells, and platelets. This may lead to infection, anemia, and easy bleeding.

The following tests and procedures may be used:

- Physical exam and history.
- Complete blood count(CBC)
- Bone marrow aspiration and biopsy
- Cytogenetic analysis
- Immunophenotyping
- Blood chemistry studies
- Chest x-ray

In childhood ALL, risk groups are used instead of stages.Risk groups are described as:

- Standard (low) risk: Includes children aged 1 to 9 years who have a white blood cell count of less than 50,000 μ/L at diagnosis.
- High risk: Includes children younger than 1 year or older than 9 years and children who have a white blood cell count of 50,000/μL or more at diagnosis.

It is important to know the risk group in order to plan treatment.

B. Data Collection

Sources of Data

-National Cancer Institute's database.
-Real patients' cases from patients' profiles (tickets).
-Medical reviews (geographical divisions and disease categories).
-International Cancer Resources (NCI USA) from official website. [11]
-Doctors, Professors and Biostatisticians from NCI.

- No. of cases: 172

Methods of Data Collection

-Data acquisition from the NCI database using digital media such as Flash Memories, diskettes and CDs.
-Capturing data from the NCI network from various spots.
-Collecting data from patients' tickets (hard copies) and digitizing them (feeding the data into preformatted database).
-Note taking
-Using published reviews from NCI containing percentages and distributions [12].
-Structured interviews with experts.

C. Data Cleaning

-Real world data, like data acquired from NCI, tend to be incomplete, noisy and inconsistent. Data Cleaning routines attempt to fill on missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

1) Missing Values

Many methods were applied to solve this issue depending on the importance of the missing value and its relation to the search domain.

- Fill in the missing value manually
- Use a global constant to fill in the missing value

25,000	free	90	L2	C-All	1,26
39,000	free	90	L1-L2	C-All	1,167
13,000		96	L2	C-All	0,99
1,230	free	94	L2		1,18
74,000	free	95	L2	C-All	1,147
16,100	free	75	Disseminated L	C-All	1,29
143,000	free	90	L2	C-All	1,0

Fig. 1 Missing values

2) Noisy Data

Noise is a random error or variance in a measured variable. Many techniques were used to smooth out the data and remove the noise.

- Clustering

Outliers were detected by clustering, where similar values are organized into groups, or clusters, values that fall outside of the set of clusters may be considered outliers.

- Combined computer and human inspection

Using clustering techniques and constructing groups of data sets, human can then sort through the patterns in the list to identify the actual garbage ones. This is much faster than having to manually search through the entire database.

F	251000
F	254000
M	256000
F	280000
M	281000
M	2000
M	282000
F	315000
M	317000
F	325000

Fig. 2 Outliers

3) Inconsistent Data

There may be inconsistencies in the data recorded for some transactions. Some data inconsistency may be corrected manually using external references, for example errors made at data entry may be corrected by performing a paper trace (the most used technique in our search, to guarantee the maximum data quality possible, by reducing prediction factors).

Other inconsistency forms are due to data integration, where a given attribute can have different names in different databases. Redundancies may also exist.

Giza	11	F
Cairo	3	F
Ismailia	3	M
El-Menya	10	M
El-Minya	3	F
Sohag	156	M
Cairo	1	M
Giza	10	F
Cairo	15	M
Giza	17	M
El-Qalyoubeyya	15	M
6-Oct	11	F
Giza	15	F

Fig. 3 Data Inconsistency (El-Menya and El-Minya represent the same value)

D. Data Integration

Data Mining often requires data integration, the merging of data from multiple data sources into one coherent data store. These sources include in our case NCI database, flat files, and data entry values. Equivalent real-world entities from multiple data sources must be matched up, for example, *patient_id* in one database must be matched up with *patient_number* in another database.

Careful integration of the data from multiple sources helped reducing and avoiding redundancies and inconsistencies in the resulting data set. This helped improving the accuracy and speed of the subsequent mining process.

E. Data Selection

Selecting fields of data of special interest for the search domain is the best way to obtain results relevant to the search criteria. In this research Acute Lymphoblastic Leukemia clustering was the aim, so data concerning the diagnosis of ALL and data concerning the patients of ALL were carefully selected from the overall data sets, and mining techniques were applied to these specific data groups in order to reduce the interesting patterns reached to the ones that represent an interest for the domain.

F. Data Transformation

In Data Transformation, the data is transformed or consolidated into forms appropriate for mining.

- **Smoothing:** which works to remove the noise form data. Such techniques include binning, clustering, and regression.
- **Aggregation:** where summary or aggregation operations are applied to the data.

- **Generalization of the data:** where low-level data are replaced by higher-level concepts through concept hierarchies.
- **Normalization:** where the attribute data are scaled so as to fall within a small specified range.
- **Attribute construction:** where new attributes are constructed and added from the given set of attributes to help the mining process.

G. Data Mining

1) Choosing the Tool

SPSS Clementine 8.1

As a data mining application, Clementine offers a strategic approach to finding useful relationships in large data sets. In contrast to more traditional statistical methods, you do not necessarily need to know what you are looking for when you start. You can explore your data, fitting different models and investigating different relationships, until you find useful information.

Working in Clementine is working with data. In its simplest form, working with Clementine is a three-step process. First, you read data into Clementine, then run the data through a series of manipulations, and finally send the data to a destination. This sequence of operations is known as a **data stream** because the data flows record by record from the source through each manipulation and, finally, to the destination--either a model or type of data output. Most of your work in Clementine will involve creating and modifying data streams.

At each point in the data mining process, Clementine's visual interface invites your specific business expertise. Modeling algorithms, such as prediction, classification, segmentation, and association detection, ensure powerful and accurate models. Model results can easily be deployed and read into databases, SPSS, and a wide variety of other applications. You can also use the add-on component, Clementine Solution Publisher, to deploy entire data streams that read data into a model and deploy results without a full version of Clementine. This brings important data closer to decision makers who need it.

The numerous features of Clementine's data mining workbench are integrated by a visual programming interface. You can use this interface to draw diagrams of data operations relevant to your business. Each operation is represented by an icon or node, and the nodes are linked together in a stream representing the flow of data through each operation.

2) Using the Tool

Creating the Data Source to use:

In our case an Oracle 9i Database was set and data fed into it in one table, ODBC was used to link the data source with the Clementine engine.

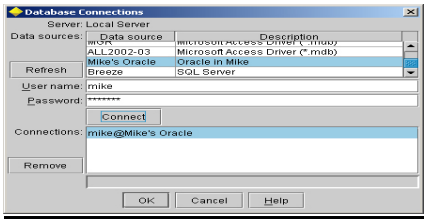

Fig. 4 Database Connection

Viewing the Data in a Tabular Form

After linking the software with the data source, we view the data from the database in a tabular form by linking the data source to a "table" output.

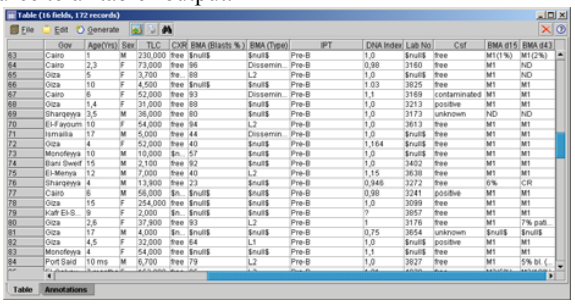

Fig. 5 Viewing data in tabular form

This enables us to assure that the link is successfully built and let us take a look on the form of data read by the software to detect any loss, inconsistency or noise that may have occurred in the linking process.

Manipulating the Data

By using the record Ops, operation concerning the records as a whole can be applied and used to operate on the data, using sampling, aggregation, sorting etc…

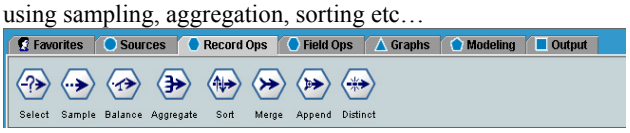

Fig. 6 Record Operations

Record Ops are linked to the data source directly and their output can be in any "output" means or can be directly fed as an input to other functions.

Using the Field Ops on specific fields allows us to explore the data deeper, by using type selecting and filtering some fields as input or output fields, deriving new fields and binning fields.

Fig. 7 Field Operations

H. Data Evaluation

After applying the data mining techniques comes the job of identifying the obtained results, in form of interesting patterns representing knowledge depending on interestingness measures. These measures are essential for the efficient

discovery of patterns of value to the given user. Such measures can be used after the data mining step in order to rank the discovered patterns according to their interestingness, filtering out the uninteresting ones. More importantly, such measures can be used to guide and constrain the discovery process, improving the search efficiency by pruning away subsets of the pattern space that do not satisfy pre-specified interestingness constraints.

I. Knowledge Representation (outcome)

In this step visualization and knowledge representation techniques are used to present the mined knowledge to the user. All the operations applied on the records and fields, and the mining process itself are represented in the form visualizations and graphics in this step.

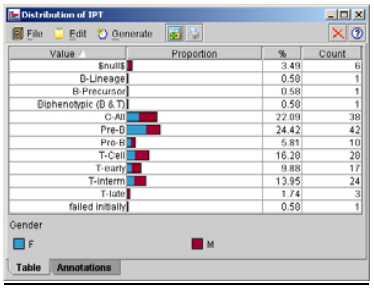

Fig. 8 Distribution of IPT

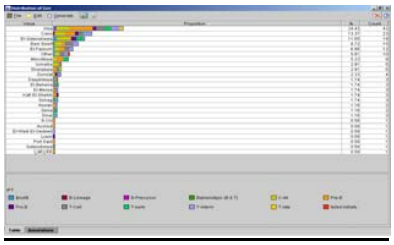

Fig. 9 Distribution of Gov and IPT

VI. Conclusions

Based on the previous work, the following conclusions were drawn:

1. Decision Trees, as a data mining technique, is very useful in the process of knowledge discovery in the medical field, especially in the domains where available data have many limitations like inconsistent and missing values.

In addition, using this technique is very convenient since the Decision Tree is simple to understand, works with mixed data types, models non-linear functions, handles classification, and most of the readily available tools use it.

2. SPSS Clementine is very suitable as a mining engine with its interface and manipulating modules that allow data exploration, manipulation and exploration of any interesting knowledge patterns

3. Using better quality of data influences the whole process of knowledge discovery, takes less time in cleaning and integration, and assures better results from the mining process.

4. Using the same data sets with different mining techniques and comparing results of each technique in order to construct a full view of the resulted patterns and levels of accuracy of each technique may be very useful for this application.

Acknowledgment

Thanks and gratitude to supervisors from NCI Egypt,
Department of Biostatistics & Cancer Epidemiology
Prof. Dr. Inas El-Attar, Head of Department
Dr. Nelly Hassan, Assistant Professor
Dr. Manar Mounir, Lecturer
Pediatrics Department : Dr. Wael Zekri, Specialist
Special Thanks to May Nabil El-Shaarawy and Mr. Kamal Amin.

References

[1] J. P. Bigus, *"Data Mining with Neural Networks"*, New York: McGraw-Hill, 1996
[2] NCI Egypt website (www.nci.edu.eg) viewed on 1st August 2005
[3] M. Berry and S. Gordon, *"Data Mining Techniques: For Marketing, Sales, and Customer Support"*, May 1997
[4] Han, J. and Kamber, M., *"Data Mining Concepts and Techniques"*, 2001
[5] M. Negnevitsky, *"Artificial Intelligence, A Guide to Intelligent Systems"*, England: Pearson Education Limited, 2002.
[6] G. Fort, S. Lambert Lacroix, *"Classification using partial least squares with penalized logistic regression"*, England: Bioinformatics-Oxford, 2005.
[7] S. Bicciato, A. Luchini, C. Di-Bello, *"Marker identification and classification of cancer types using gene expression data and SIMCA"*, Germany: Methods-of-information-in-medicine, 2004.
[8] K. A. Marx, P. O'Neil, P. Hoffman, M. L. Ujwal, *"Data mining the NCI cancer cell line compound GI(50) values: identifying quinone subtypes effective against melanoma and* ***leukemia*** *cell classes"*, United-States: Journal-of-chemical-information-and-computer-sciences, 2003.
[9] G. A Forgionne, A. Gagopadhyay, and M. Adya, *"Cancer Surveillance Using Data Warehousing, Data Mining, and Decision Support Systems"*, Topics in Health Information Management, vol. 21(1); Proquest Medical Library, August 2000
[10] W. Kuo, R. Chang, D. Chen and C. C. Lee, "Data Mining with Decision Trees for Diagnosis of Breast Tumor in Medical Ultrasonic Images", Breast Cancer Research and Treatment, Dordrecht, vol. 66, Iss. 1, Mar 2001.
[11] National Cancer Institute official website (www.nci.nih.gov) viewed on 1st August 2005
[12] Periodicals of NCI Egypt (2001)
[13] "Introduction to Data Mining and Knowledge Discovery - Third Edition", Two Crows Corporation (pdf)

Nevine Makram Labib, a full timer Lecturer at the Sadat Academy for Management Sciences, Department of Computer & Information System and Vice-Dean of the Training Center Alexandria Branch in addition to being an International Trainer in Information Technology and Management-related disciplines.

Dr. Makram has over ten years experience in the Information Technology field. She attained both her Doctoral Degree and Masters Degree specializing in this very important area and has her academic background into practice, gaining vast hands on experience.

Dr. Makram is known for the wealth of real-life expertise and experience she brings to the courses she runs leaving an outstanding impression for all her attendees and is always asked to return for more training. She has also served as a speaker at numerous international preferences focusing on Artificial Intelligence, Expert Systems, and Information Technology.

Michael Nabil Malek, Researcher graduated from Sadat Academy for Management Sciences, department of Computers and Information systems. One year experience in IT field including training as System Administrator in Mena House Oberoi Hotel and Casino, and as Communication Engineer in ALCATEL. He has many projects concerning Data Mining, especially in the fields of hospitality and medicine. In addition, he has a working knowledge of Oracle Systems (Development and Administration).