American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-5, Issue-10, pp-59-66
www.ajer.org
Research Paper Open Access

Development of National Pathological Data Warehouse (NPDW) For Bangladesh

Vaskor Mostafa[1], F.M.Rahat Hasan Robi[2], Md. Akkas Ali[3]
[1](Lecturer, Department of EEE, Uttara University, Dhaka, Bangladesh)
[2](Lecturer, Department of CSE, Uttara University, Dhaka, Bangladesh)
[3](Assistant Professor, Department of CSE, Bangabandhu Sheikh Mujibur Rahman Science & Technology University, Gopalganj, Bangladesh)

ABSTRACT: *Building a data warehouse for pathological data is a challenging task because data are collected from diverse sources. To capture and integrate these diverse data sources for pathological reporting and analysis, plans have developed data warehouse. A data warehouse is considered as an architecture that is designed for query and analysis over traditional transaction processing. Historical data can be derived from the transaction processing and data warehouse mainly stores these historical data. But it can also include data from other sources. Pathological data warehouse contains historical information of patients and their different test results. The construction of data warehouses involves data cleaning, data integration, and data transformation. In this paper we have developed an architecture of a data warehouse model, a star schema and a development process suitable for integrating data from diverse healthcare sources have been presented. Finally the data warehouse can be viewed as an important preprocessing step for data mining.*

Keywords: *Data warehouse, Pathological data, ODS, OLAP, Data mining, Data preprocessing.*

I. INTRODUCTION

Data warehouses may be relatively new to the health domain but it is an important component of health informatics. It deals with resources and methods needed to optimize the acquisition, storage, retrieval and use of information in medical research and applied to the areas of health care management, diagnosis, clinical care, pharmacy, nursing and public health [1, 2].

A data warehouse provides the ability above and beyond an analog or electronic health record (ECG, ECHO, EMG), text or numbers(e.g., Patient ID, diagnostics data, demography)and images(radiological, ultrasound) to manage and improve the quality of care down to the individual patient level. This can be accomplished through a number of approaches, though all must start with the application of medical informatics. Changes in the pathology interpretation (the diagnosis) can drastically alter the clinician's treatment plan and the patient's prognosis. As in all disciplines of medicine, the goals of pathology are to conform to the ethical principles of beneficence and non-maleficence: the obligation to help and not to harm patients. To this end, pathologists are obligated to provide accurate and timely diagnoses, to protect patients from wrong diagnoses, and to reduce the diagnostic variability that can have a major impact on patient therapy and management.

Pathological Data warehousing is the only viable solution for providing strategic information and vital issue to develop a support system for clinical decision-making process. it is a simple concept for information delivery but traditional operational database does not support critical data analysis tasks of the health care providers. It contains detailed data but do not include important historical data, and since it is highly normalized, it performs poorly for complex queries that need to join many relational tables or to aggregate large volumes of data in order to generate various clinical reports. A health data warehouse is a data store that is different from the hospital's operational databases. It can be used for the analysis of consolidated historical data [3, 4].

In this paper, our main goal is to design a pathological data warehouse. We will describe the data preprocessing using some test value and then we will design a star schema for pathological data. In section 2, some related works on health and pathological data warehouse will be described. In section 3, data preprocessing part will be described. In section 4, the architecture and star schema of our data warehouse will be described. And finally in section 5, there is the conclusion part.

II. RELATED WORKS

This chapter surveys previous work on data warehouse design using pathology or medical health data. Many works have been done before on data warehouse design using pathological data, medical data, clinical data etc. Data warehouse terminology by Ralph Kimball, defines data warehouse as "a copy of transaction data specifically structured for query and analysis" [5]. In [4], the authors provide an operational and technical definition of data warehouses; present examples of data mining projects enabled by existing data warehouses, and describe key issues and challenges related to warehouse development and implementation. In [2], authors mainly focused on medical data and they emphasizes the unique features of medical data mining and they also emphasize on ethical, security and legal aspects of medical data mining. In [3], the authors presented data warehouse architectures using clinical data and gave some solutions to tackle data integration issues using some clinical data warehouse applications. In [6], authors identifies the prospects and complexities of Health data warehousing for Bangladesh perspective and they also propose a data-warehousing model that is suitable for integrating data from different health care sources. In[7], where the architecture of data warehouse is done for Duke University medical center and identify factors compatriots with increased risk for preterm birth for ensuing use of data mining project. In [8], to discover associations between the presence of diseases and clinical findings, authors have used co-occurrence statistics. There are four stages in the medical knowledge development cycle: Planning Tasks, Understanding/learning, Action/internalization, Enforcing/unlearning [9].In [10], author describe the design, development, and implementation of an enterprise-wide data warehouse at the University of Michigan Health System with less specific requirements. In [11] authors presented Main components of KDD and DM and their relationships and Medical Information System development time table. Data mining can be viewed as a process rather than a set of tools, and the acronym SEMMA (sample, explore, modify, model, and assess) refers to a methodology that clarifies this process [12].In [13], the authors use statistical process control charts in hospital epidemiology. In [14], the authors report on the in-house development of an integral part of the data warehouse specifically for the intensive care units (ICU-DWH).In [15], the authors mainly emphasize to use neural network for the diagnosis of diabetes mellitus.

III. DATA PREPROCESSING

Data preprocessing is considered as a major task in building Data Warehouse (DW). It includes outlier detection and removal, handling missing values, transforming data to a suitable format to DW etc. Our first task is about on this step i.e. data preprocessing of building DW.

To do this work, we will take urine test as sample and we will perform the following task on it:

Task 1: Determination of Data Types (nominal, ordinal etc) and Value Range (min-max) for each of the given tests

Task 2: Synthetic random data generation in MS Excel format

Task 3: Missing value handling for each test

Task 4: Transformation analysis (e.g. Min-Max, Z-Score etc)

The following sections will describe the details procedures that have been used to do the tasks mentioned above. Now we will describe the above task:

3.1 Determination of Data Types (nominal, ordinal etc) and Value Range (min-max) for each of the given tests

The task 1 is about determining the data types of the given test samples and also identifying the range of each given tests in which the values can belong to. We have used internet to collect information about each test and thus identified its data type, value range i.e. minimum and maximum values and measuring units. Separate value range has been identified for male and female [27]. An excel file like table 1.1 has been used to store these data.

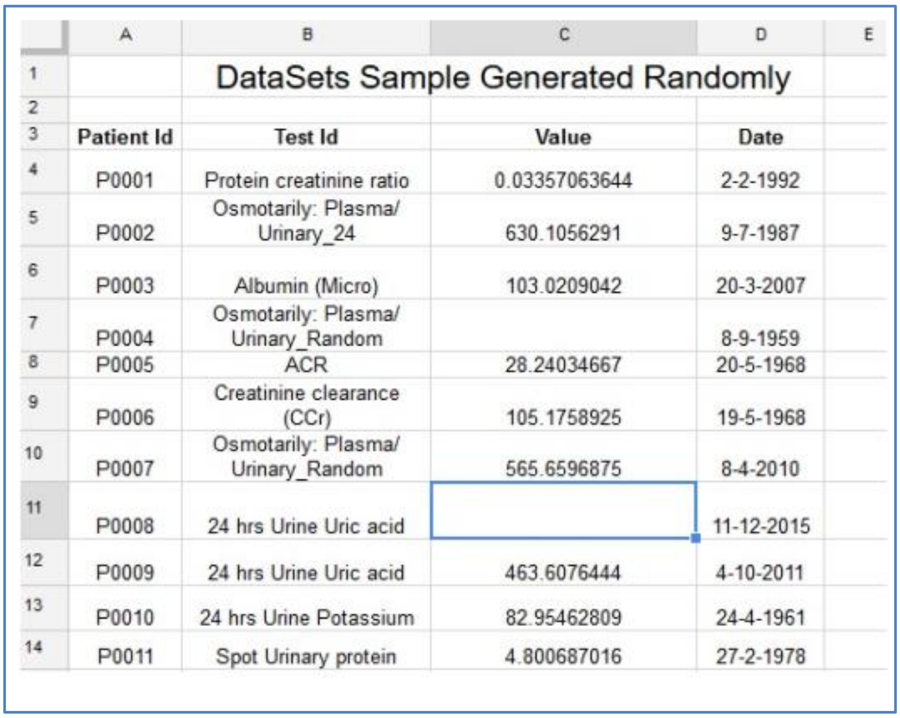
Determination of Data Types for Given Tests							
Test Id	**Test Name**	**Range-Male (Min)**	**Range-Male (Max)**	**Range-Female (Min)**	**Range-Female (Max)**	**Data Type**	**Data Unit**
Albumin (Micro)	Microalbuminuria	30	300	30	300	Numerical	mg/24 hrs
24 hrs Urine Calcium	Urine Calcium Test	100	300	100	300	Numerical	mg/day
ACR	albumin/creatinine ratio (ACR)	0	30	0	30	Numerical	mg/24 hrs
24 hrs Urine Phosphate	Phosphorus 24 hour Urine	360	1600	170	1200	Numerical	mg/24 hrs
Spot Urinary protein	Protein Urine Test (Random urine sample)	0	20	0	20	Numerical	mg/24 hrs
24 hrs Urine Uric	Uric Acid Test	250	750	250	750	Numerical	mg/24 hrs

acid	(Urine Analysis)						
Protein creatinine ratio	protein/creatinine ratio	0	0.11	0	0.11	Numerical	mg/mg creatinine
24 hrs Urine Sodium	Urine Sodium Level Test	40	220	40	220	Numerical	mmol/day
Creatinine clearance (CCr)	Creatinine clearance test	97	137	88	128	Numerical	ml/min
24 hrs Urine Potassium	Potassium Urine Test	25	125	25	125	Numerical	mEq/L
24 hrs UTP	Total protein, 24-hour urine	0	0.2	0	0.2	Numerical	g/V
Osmotarily: Plasma/ Urinary_Random	Urine Osmolality, Random	300	900	300	900	Numerical	mOsm/kg of water
Osmotarily: Plasma/ Urinary_24	Urine Osmolality, 24 hrs	500	800	500	800	Numerical	mOsm/kg of water

Table 3.1: Data Type and Value Range Determination

3.2 Synthetic random data generation in MS Excel format

In task 2, we have to generate a random dataset in excel format from the data collected in task 1. To fulfill the randomness criteria, we have written a java program to generate data in excel format. Our java program randomly generates patient test data and store it in a excel file. During the generation of dataset, we also generate some missing value randomly. For generating missing values, numerical attribute i.e. test value is only considered. A sample dataset is shown in figure 3.1



DataSets Sample Generated Randomly

Patient Id	Test Id	Value	Date
P0001	Protein creatinine ratio	0.03357063644	2-2-1992
P0002	Osmotarily: Plasma/ Urinary_24	630.1056291	9-7-1987
P0003	Albumin (Micro)	103.0209042	20-3-2007
P0004	Osmotarily: Plasma/ Urinary_Random		8-9-1959
P0005	ACR	28.24034667	20-5-1968
P0006	Creatinine clearance (CCr)	105.1758925	19-5-1968
P0007	Osmotarily: Plasma/ Urinary_Random	565.6596875	8-4-2010
P0008	24 hrs Urine Uric acid		11-12-2015
P0009	24 hrs Urine Uric acid	463.6076444	4-10-2011
P0010	24 hrs Urine Potassium	82.95462809	24-4-1961
P0011	Spot Urinary protein	4.800687016	27-2-1978

Figure 3.1: Randomly generated dataset with missing values

3.3 Missing value handling for each test

The task 3 asked us to handle missing values generated in task 2. Handling missing value is very important step in producing consistent and clean data. There are several techniques to handle missing values. Some of the common techniques are discussed in [1]-

a. Ignore the tuple: This is usually done when the class label is missing (assuming the mining task involves classification).

b. Fill in the missing value manually: In general, this approach is time consuming and may not be feasible given a large data set with many missing values.

c. Use a global constant to fill in the missing value: Replace all missing attribute values by the same constant such as a label like Unknown.

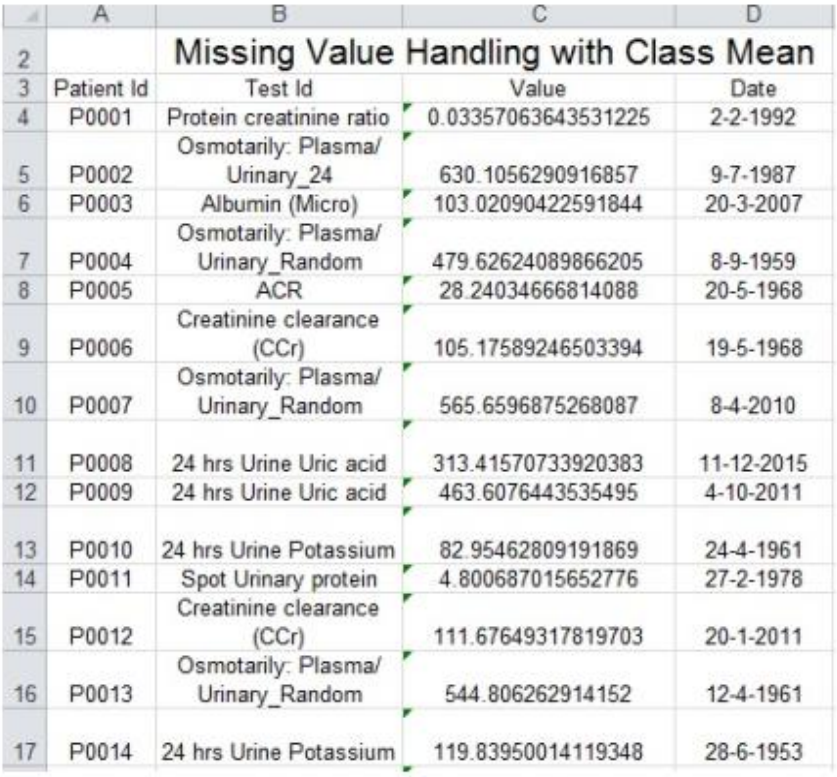
d. Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value.

e. Use the attribute mean or median for all samples belonging to the same class as the given tuple: For example, if classifying customers according to credit risk, we may replace the missing value with the mean income value for customers in the same credit risk category as that of the given tuple. If the data distribution for a given class is skewed, the median value is a better choice.

f. Use the most probable value to fill in the missing value: This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction.

Since all of our missing values are of numerical data type so we have preferred Class mean approach to handle the missing value. In our approach-

1. Each test item e.g. ACR has been considered as a class.
2. Here, each class two mean values namely female Mean and male Mean have calculated based on patient's gender.
3. If the patient whose test value is missing is male then the missing value will be replaced by male Mean of the test class. Otherwise the missing value will be replaced by female Mean of the test class. The data sample generated after replacing the missing values is shown in figure 3.2.3



	A	B	C	D
2		Missing Value Handling with Class Mean		
3	Patient Id	Test Id	Value	Date
4	P0001	Protein creatinine ratio	0.03357063643531225	2-2-1992
5	P0002	Osmotarily: Plasma/ Urinary_24	630.1056290916857	9-7-1987
6	P0003	Albumin (Micro)	103.02090422591844	20-3-2007
7	P0004	Osmotarily: Plasma/ Urinary_Random	479.62624089866205	8-9-1959
8	P0005	ACR	28.24034666814088	20-5-1968
9	P0006	Creatinine clearance (CCr)	105.17589246503394	19-5-1968
10	P0007	Osmotarily: Plasma/ Urinary_Random	565.6596875268087	8-4-2010
11	P0008	24 hrs Urine Uric acid	313.41570733920383	11-12-2015
12	P0009	24 hrs Urine Uric acid	463.6076443535495	4-10-2011
13	P0010	24 hrs Urine Potassium	82.95462809191869	24-4-1961
14	P0011	Spot Urinary protein	4.800687015652776	27-2-1978
15	P0012	Creatinine clearance (CCr)	111.67649317819703	20-1-2011
16	P0013	Osmotarily: Plasma/ Urinary_Random	544.806262914152	12-4-1961
17	P0014	24 hrs Urine Potassium	119.83950014119348	28-6-1953

Figure 3.2: Data Sample after replacing missing values

3.4 Transformation analysis (e.g. Min-Max, Z-Score etc)

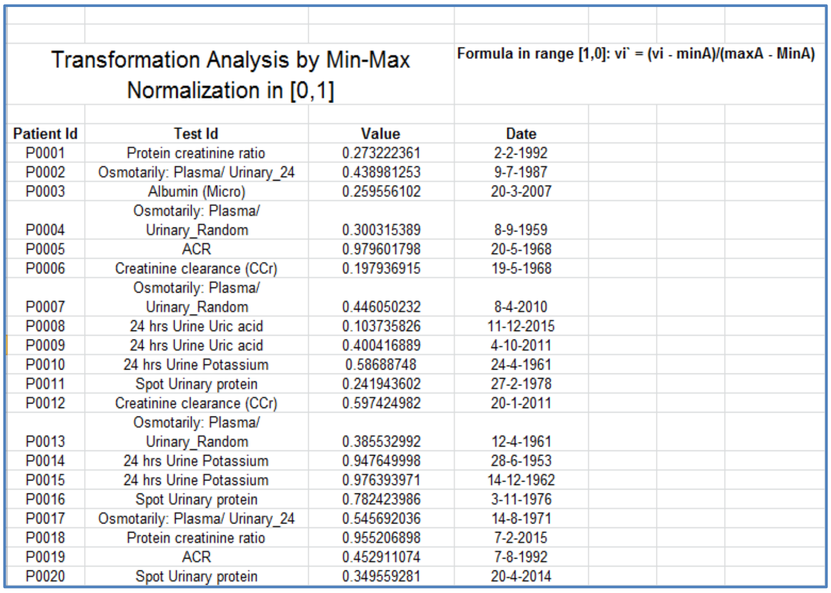
The task 4 is about transforming data to a suitable format for National Pathological Data Warehouse. Transformation of data is an essential task in building Data Warehouse. Since data are collected from multiple heterogeneous data sources to build DW so it is necessary to transform them into a common format that is suitable for data mining task. There are many ways of data transformation. In this task, we have used Min-Max and Z-score transformation to build our DW.

3.4.1 Data Transformation using Min-Max Normalization

Min-max normalization performs a linear transformation on the original data. Suppose that min A and max A are the minimum and maximum values of an attribute 'A'. Min-max normalization maps a value, vi, of A to vi 0 in the range [new min A; new max A] by computing$v_i^{'}$.

Here $v_i^{'} = \frac{v_i - min\ _A}{max\ _A - min\ _A}(new\_max_A - new\_min_A) + new\_min_A$

Now, we will apply this equation in our dataset and we will get the normalized dataset. Figure 3.3 shows the Min-max normalization of our sample dataset in the range [0, 1].[1]



Transformation Analysis by Min-Max Normalization in [0,1]

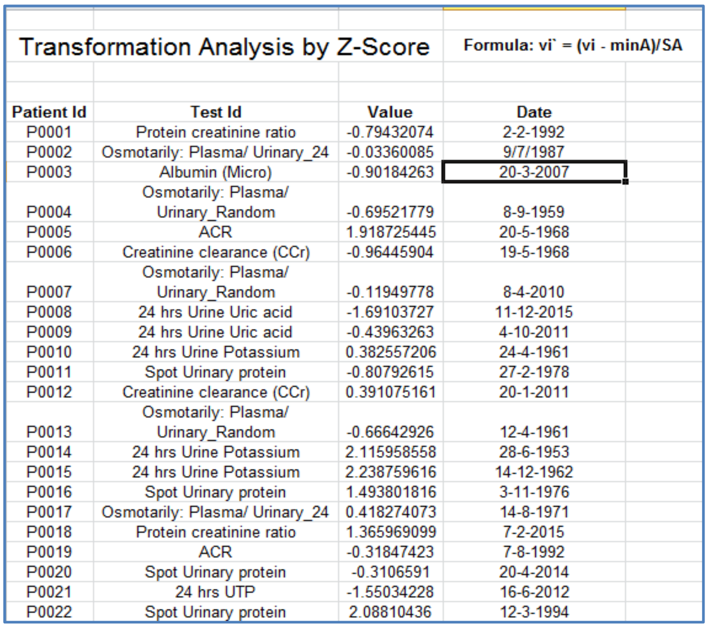
Formula in range [1,0]: vi` = (vi - minA)/(maxA - MinA)

Patient Id	Test Id	Value	Date
P0001	Protein creatinine ratio	0.273222361	2-2-1992
P0002	Osmotarily: Plasma/ Urinary_24	0.438981253	9-7-1987
P0003	Albumin (Micro)	0.259556102	20-3-2007
P0004	Osmotarily: Plasma/ Urinary_Random	0.300315389	8-9-1959
P0005	ACR	0.979601798	20-5-1968
P0006	Creatinine clearance (CCr)	0.197936915	19-5-1968
P0007	Osmotarily: Plasma/ Urinary_Random	0.446050232	8-4-2010
P0008	24 hrs Urine Uric acid	0.103735826	11-12-2015
P0009	24 hrs Urine Uric acid	0.400416889	4-10-2011
P0010	24 hrs Urine Potassium	0.58688748	24-4-1961
P0011	Spot Urinary protein	0.241943602	27-2-1978
P0012	Creatinine clearance (CCr)	0.597424982	20-1-2011
P0013	Osmotarily: Plasma/ Urinary_Random	0.385532992	12-4-1961
P0014	24 hrs Urine Potassium	0.947649998	28-6-1953
P0015	24 hrs Urine Potassium	0.976393971	14-12-1962
P0016	Spot Urinary protein	0.782423986	3-11-1976
P0017	Osmotarily: Plasma/ Urinary_24	0.545692036	14-8-1971
P0018	Protein creatinine ratio	0.955206898	7-2-2015
P0019	ACR	0.452911074	7-8-1992
P0020	Spot Urinary protein	0.349559281	20-4-2014

Figure 3.3: Data Transformation using Min-Max Normalization in [0,1] range

3.4.2 Data Transformation using Z-Score Normalization

In z-score normalization (or zero-mean normalization), the values for an attribute 'A', are normalized based on the mean (i.e., average) and standard deviation of A. Z-Score normalization is calculated by the equation $v_i' = \frac{v_i - \bar{A}}{\sigma_A}$. Where $\bar{A}$ and σ_A are the mean and standard deviation, respectively, of attribute A.[1]



Transformation Analysis by Z-Score

Formula: vi` = (vi - minA)/SA

Patient Id	Test Id	Value	Date
P0001	Protein creatinine ratio	-0.79432074	2-2-1992
P0002	Osmotarily: Plasma/ Urinary_24	-0.03360085	9/7/1987
P0003	Albumin (Micro)	-0.90184263	20-3-2007
P0004	Osmotarily: Plasma/ Urinary_Random	-0.69521779	8-9-1959
P0005	ACR	1.918725445	20-5-1968
P0006	Creatinine clearance (CCr)	-0.96445904	19-5-1968
P0007	Osmotarily: Plasma/ Urinary_Random	-0.11949778	8-4-2010
P0008	24 hrs Urine Uric acid	-1.69103727	11-12-2015
P0009	24 hrs Urine Uric acid	-0.43963263	4-10-2011
P0010	24 hrs Urine Potassium	0.382557206	24-4-1961
P0011	Spot Urinary protein	-0.80792615	27-2-1978
P0012	Creatinine clearance (CCr)	0.391075161	20-1-2011
P0013	Osmotarily: Plasma/ Urinary_Random	-0.66642926	12-4-1961
P0014	24 hrs Urine Potassium	2.115958558	28-6-1953
P0015	24 hrs Urine Potassium	2.238759616	14-12-1962
P0016	Spot Urinary protein	1.493801816	3-11-1976
P0017	Osmotarily: Plasma/ Urinary_24	0.418274073	14-8-1971
P0018	Protein creatinine ratio	1.365969099	7-2-2015
P0019	ACR	-0.31847423	7-8-1992
P0020	Spot Urinary protein	-0.3106591	20-4-2014
P0021	24 hrs UTP	-1.55034228	16-6-2012
P0022	Spot Urinary protein	2.08810436	12-3-1994

Figure 3.4: Data Transformation using Z-Score Normalization

IV. ARCHITECTURE OF PATHOLOGICAL DATA WAREHOUSE

The architecture of NPDW is shown in fig.4.1The development of pathological data warehouse is a huge undertaking, and requires a considerable commitment of time and effort. Firstly historical data will be collected from multiple operational systems, Actually Heterogeneous pathological data will be collected from different governmental and private sources of Bangladesh. By Using Extraction, Transform and load (ETL) these data will be integrated into a temporary repository. After that Determination of Data Types, Synthetic random data generation in MS Excel format, missing value handling for each test, Transformation analysis (e.g. Min-Max, Z-Score etc.) are done. Then data are loaded into the ODS. From ODS data are finally stored in DW. Afterward OLAP Analysis, Reporting and Data mining task are performed from NPDW.

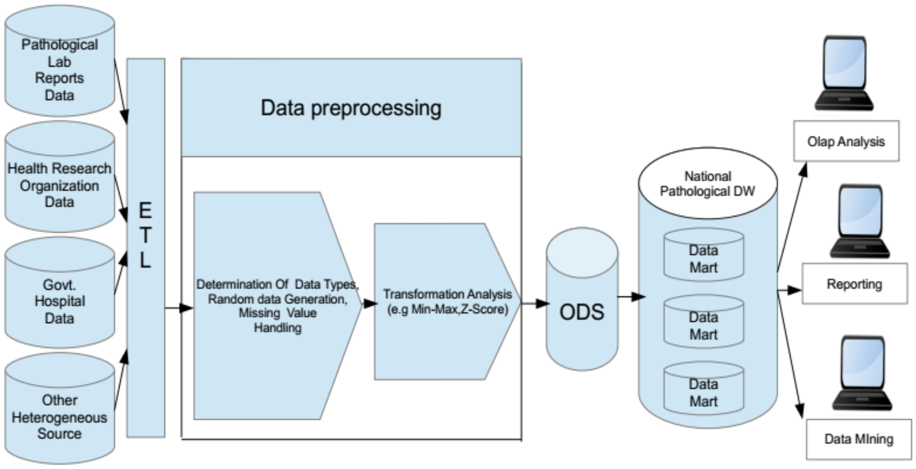

Pathological Lab Reports Data
Health Research Organization Data
Govt. Hospital Data
Other Heterogeneous Source
E T L
Data preprocessing
Determination Of Data Types, Random data Generation, Missing Value Handling
Transformation Analysis (e.g Min-Max,Z-Score)
ODS
National Pathological DW
Data Mart
Data Mart
Data Mart
Olap Analysis
Reporting
Data MIning

Fig 4.1 Architecture of pathological data warehouse

Now, we will focus on developing the logical understanding of the National Pathological Data Warehouse (NPDW). This section also consisted of four tasks as follow:
Task 1: Identification of all facts and dimensions for NPDW
Task 2: Designing star schemas for all facts and dimensions tables
Task 3: Identification of all uses of NPDW
Task 4: Showing some results from the synthetic data generated in section 3 that can be loaded to the warehouse to serve the uses as per task 3.
The details of these tasks are discussed in the subsequent sections.

4.1 Identification of all facts and dimensions for NPDW

We have identified two facts tables and five dimensions tables for the NPDW.
The facts and dimensions are-

Fact and Dimensions Table	
Fact 1	Test Result [patient key, lab key, test key, disease id, time key, test result, test cost, no of tests done, no of normal result]
Fact 2	Disease info [patient key, test key, disease key, no of related Disease, no of affected Patients, disease likely hood]
Dimensions	1. Patient [patient key, age, gender, district, division] 2. Diagnostic Lab [lab key, lab name] 3. Test [test key, test name, test type] 4. Time [time key, day, month, year] 5. Disease [disease id, disease name]

Table 4.1: Fact and Dimensions of star schema

Here, the ‘disease likelihood measure’ gives the statistical probability of how a patient likely affected by a disease or how a test likely related to a disease. For example likelihood (test key, disease key) > 0.5 means the test is related to the disease.

4.2 Designing star schemas for all facts and dimensions tables

The star schemas of the facts and dimensions identified above are shown in figure 4.2. There are two star schemas for the DW, one for Test Result fact table and another for Disease info fact table.

4.3 Identification of all uses of NPDW

There are so many usages of National Pathological Data Warehouse for Bangladesh. Some of the usages are given below.

1. Supports strategic planning and quality management for medical sectors of Bangladesh.
2. Fosters improved outcomes for patients, population and the provider organization.
3. Enables public health initiatives at the state and national level.
4. Permits healthcare providers to influence national level policy.
5. Monitors National level disease trends and helps in taking initiatives.
6. Defines national level reference value for a particular test.
7. Detects redundant costly tests prescribed by doctor.

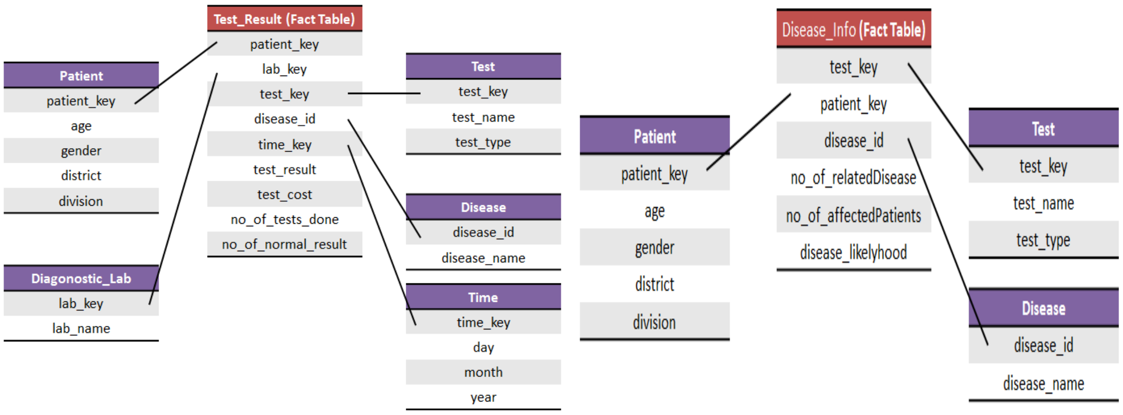


Fig 4.2: Star Schemas for NPDW of Bangladesh

4.4Showing some results from the synthetic data

This section shows some example usages that can be performed by our proposed star schemas. Suppose, the government of Bangladesh is taking a strategic decision about buying some machines for diagnostic purpose The NPDW can help in taking such decision by computing the measure ’no of tests done’ from Test Result fact table using the dimensions Time, Patient, Test and Disease. Different combinations of dimensions e.g <Patient, Disease>can be used to calculate the total number of tests done by the patient for that particular disease. Redundant costly tests also can be detected using disease likelihood measure along different dimensions.

V. CONCLUSION

Data warehouses save time and energy and also guarantee the accuracy of the data. Pathological data warehouse represents an exciting area of current development and offer the potential to shape the way we utilize and manage Pathological data. In this paper we have developed pathological data warehouse architecture and showing the development stages of it. We have also discussed logical design approaches star schema for large DW. Some preprocessing tasks are done by using suitable transformation techniques. If the clinical systems become more habituated to use data warehouse then the efficiency of the health care providers increased day by day. If the proper authority takes initiative to setup data warehouse for pathological sector then it will benefit people in many ways. Patients can access their pathological information whenever they want. Many health researchers can make knowledge discovery by using pathological data warehouse. Clinic and pathological center who wants to get more advantages from technology they should use data warehouse by creating it.

REFERENCES

[1]. Roddick JF, Fule P, Graco WJ (2003) *Exploratory medical knowledge discovery: experiences and issues. SIGKDD Explor. Newsletter, 5(1)*: 9499

[2]. Cios K (2002) *Uniqueness of medical data mining. Artificial intelligence in medicine*. 26:1-24

[3]. Sahama TR, Croll PR (2007) *A Data Warehouse Architecture for Clinical Data Warehousing. Australasian Workshop on Health Knowledge Management and Discovery (HKMD 2007)*

[4]. Lyman JA, Scully K, Harrison JH (2008) *The development of healthcare data warehouses to support data mining. Clin Lab Med. 28(1)*:55-71

[5]. T. Manjunath, S. Ravindra, and G. Ravikumar, *“Analysis of data quality aspects in data warehouse systems,” International Journal of Computer Science and Information Technologies, Vol. 2, No. 1*, 2010, pp. 477- 485.

[6]. S. I. Khan and A. S. M. L. Hoque, "Development of national health data warehouse Bangladesh: Privacy issues and a practical solution," 2015 *18th International Conference on Computer and Information Technology (ICCIT), Dhaka*, 2015, pp. 373-378.

[7]. Prather JC, Lobach DF, Goodwin LK, et al. *Medical data mining: knowledge discovery ina clinical data warehouse. Proc AMIA Symp 1997;101–5.*

[8]. Cao H, Markatou M, Melton GB, et al. *Mining a clinical data warehouse to discover diseasefinding associations using co-occurrence statistics. Proc AMIA Symp 2005*;106–10.

[9]. Wang, H, Wang S (2008) Medical knowledge acquisition through data mining,. *IEEE International Symposium ITME.*

[10]. Dewitt JG, Hampton PM. *Development of a data warehouse at an academic health system:knowing a place for the first time. Acad Med 2005;80*:1019–25.

[11]. Lee IN, Liao SC, Embrechts M (2000) *Data mining techniques applied to medical information.Medical Informatics & the Internet in Medicine 25(2)*: 81-102.

[12]. Obenshain MK, *Application of Data Mining Techniques to Healthcare Data,Infection Control and Hospital Epidemiology, vol.25, no 8*, pp. 690-695, 2004.

[13]. Sellick JA Jr. *The use of statistical process control charts in hospital epidemiology. Infect Control Hosp Epidemiol 1993;14*:649-656.

[14]. Marleen de Mul , Peter Alons , Peter van der Velde , Ilse Konings , Jan Bakker , Jan Hazelzet (2010) *“Development of a clinical data warehouse from an intensive care clinical information system”* Computer Methods and Programs in Biomedicine.

[15]. Kumari S, Singh A (2013) A data mining approach for the diagnosis of diabetes mellitus. *IEEE 7th International Conference on Intelligent Systems and Control.*

[16]. Fayyad UM, Shapiro GP, Smyth P (1996) *From Data Mining to Knowledge Discovery: An Overview.Advances in Knowledge Discovery and Data Mining* 1–36.

[17]. Khosla R, Dillon T (1997) *Knowledge Discovery, Data Mining and Hybrid Systems. Engineering Intelligent Hybrid Multi-Agent Systems, Kluwer Academic Publishers* 143–177.

[18]. Stolba N, Banek M, and Tjoa AM (2006) The Security Issue of Federated Data Warehouses in the Area of EvidenceBased Medicine. *First International Conference on Availability, Reliabilityand Security (ARES'06, IEEE).*

[19]. Zhu X, Khoshgoftaar T, Davidson I,Zhang S (2007) *Special issue on mining low-quality data, Knowledge and Information Systems, 11*:131-136.

[20]. Brown ML, Kros JF (2003) *Data mining and the impact of missing data. Industrial Management & Data Systems 103: 611-621.*

[21]. *Lavrač N (1999) Selected techniques for data mining in medicine.Artificial intelligence in medicine 16(1)*: 3-23.

[22]. World Health Organization (2008).*Worldwide prevalence of anaemia 1993–2005. Geneva: World Health Organization*. ISBN 978-92-4-159665-7.

[23]. DeckerMD. *Continuous qualityimprovement. Infect Control HospEpidemiol 1992;13*:165-1.

[24]. *Committee on Quality of Health Care in America, Crossing the Quality Chasm: A New Health System for the 21st Century*, National Academy Press, Washington, 2001.

[25]. R.R. Kimball, M. Ross, *The Data Warehouse Lifecycle Toolkit*,2nd edition, Wiley, 2002.

[26]. M.A. Geisler, D. Will, *Implementing enterprise wide databases: a challenge that can be overcome, Topics in Health Information Management* 19 (1998) 11–18.

[27]. https://www.merckmanuals.com/professional/appendixes/normal-laboratory-values/urine-tests-normal-values