An Epidemiological Data Mining Application Based on Census Databases

J. Pérez-Ortega, Alicia Martínez, E. Iturbide-Domínguez, M. Hidalgo-Reyes, A. Mexicano-Santoyo
Department of Computer Science
CENIDET
Cuernavaca, México
jpo_cenidet@yahoo.com.mx, {amartinez, emmanuel.iturbide10c, mh}@cenidet.edu.mx
amexicano@gmail.com

Crispin Zavala-Díaz
Faculty of Accounting, Administration and Informatics
UAEM
Cuernavaca, México
crispin_zavala@uaem.mx

***Abstract*— This paper shows the experience of a specialized data mining process that integrates mortality data collected by the official censuses of 2000 and 2010. The objective is the generation of patterns of interest based on the clustering of districts with high mortality rates for different causes of death. According to the specialized literature, few studies related to data mining applications using census databases have been reported, despite the potential census databases have. Contributions of this work are the implementation of a data preparation subsystem and the integration of a data warehouse that contains records of deaths, occurring in 2000 and 2010 for 2049 different causes of death. In order to validate the results, we analyzed four causes of death related to cancer C16 (stomach) and C34 (lung), and diabetes mellitus E11 (no insulin-dependent), and E14 (unspecified). Experimental results were satisfactory and they show some patterns of interest and an increase in the mortality rate in 2010 compared to 2000.**

Keywords - Data mining application; Mortality data; Census database.

I. INTRODUCTION

Censuses are important because the information collected is used to understand the reality that a country lives; in addition, this information is used by different organizations and institutions of the public, private and social sector [1]. In turn, it allows us to analyze and to understand the problems that occur in several areas of a country, e.g. economy, housing, health. In the public health area, this information is important for the organizations that manage public health services, allowing to identify vulnerable sectors in a population, managing and assigning the resources adequately, and implementing appropriate strategies to combat these diseases.

In Mexico, as in other countries, censuses collect information about different aspects of a given population, e.g. health, educational, economic and social features. Censuses help to support the research performed by academic and educational institutes in the socio-demographic field [2].

There are a few studies related to epidemiological data where data mining is used as a tool to explore them and cartographic visualization systems are used to represent the extracted knowledge. Data analysis performed in [3] shows a clear image of the concentration of cases of breast cancer in New Mexico. The results allow identifying the groups in populations using spatial data. This work is just focused on data related to a single state: New Mexico, USA. This represents an important drawback regarding the amount of available information.

The works described in [4] and [5] are part of a project that uses epidemiological and spatial data in order to identify, establish, and display, cartographically, possible relationships between patients with cancer and their proximity to factories and cell phone antennas. However, this work is closely linked to this hypothesis, and do not contemplate incidence or mortality rates.

The works [6] and [7] have used data mining in order to extract information from epidemiological data, however; the extracted knowledge and its representation follow a different approach to the one presented in this research.

This research takes part of a bigger data mining project in the epidemiological domain. Previous works [8], [9], and [10] were developed specifically for analyzing mortality causes related with cancer; these researches are an important antecedent and are closely related to our data mining objective. In [11] [12], a Data Warehouse was implemented and a data mining tool was developed in order to generate and to display patterns of interest represented as groups of districts with high rates of cancer mortality in maps of Mexico.

This paper shows the experience of a specialized data mining process whose primary objective is the generation of patterns of interest shown as groups of districts with high mortality rates. The process has been developed for several causes of death among which we have selected four causes related to cancer and diabetes mellitus as a practical case of study. As an additional contribution, a subsystem focused on the data preparation step for the epidemiological domain and a data warehouse, that integrates data from the official censuses, were implemented.

This document is organized as follows: Section II shows a brief description of the analysis, data preparation, and data integration performed on the mortality data. A data preparation subsystem is described in Section III. Section IV describes the cartographic visualization subsystem which displays maps of Mexico with groups of districts with high mortality rates for different causes. Section V shows the results obtained and the contributions made by this research. Conclusions and future work are shown in Section VI.

 ISBN: 978-1-61208-247-9

II. CENSUS DATABASES DESCRIPTION AND PREPROCESSING

We have collected databases from official censuses. In the epidemiological domain, data from census carried out by official organisms such as Statistics and Geography National Institute (INEGI, Instituto Nacional de Estadística y Geografía) in Mexico allow us to analyze and to understand the behavior of diseases and how they affect a population, showing information about what diseases have higher recurrence and the number of deaths occurred because of those diseases.

However, it is important to completely exploit this information in order to obtain more significant knowledge that permits health organisms to make decisions to prevent these diseases or to improve the living conditions of those who suffer them. Data mining is a tool that facilitates the task of extracting knowledge from data, which may be useful to understand a phenomenon.

The knowledge extracted represents information of interest to health organisms not only in Mexico, but also in other parts of the world where the censuses are carried out. The above mentioned represents a possibility to apply this project to other countries. In the following sections, the data mining process followed in this research is briefly described:

A. Database description

The data were extracted from different official information sources from Mexico. The information sources and the data description are shown below:

- Mortality database: records of deaths occurring in 2000 and 2010 for different causes of death [13], extracted from: National Health Information System (SINAIS, Sistema Nacional de Información en Salud).
- Geographic database: records of the geographical position of the districts of Mexico [14], extracted from: Database District System (SIMBAD, Sistema Municipal de Bases de Datos).
- Population database: records of the total population by districts in Mexico, for 2000 and 2010 [15], extracted from: INEGI.
- International catalogue of diseases (CIE-10) [16]: code classification and names of diseases. It includes 2049 causes of death (Updated to 2009), extracted from: Collaborating Center for the Family of International Classifiers (CEMECE, Centro Colaborador para la Familia de Clasificadores Internacionales de la OMS en México).

SINAIS and SIMBAD systems have data provided by INEGI. Table I shows the number of attributes and records of the databases for 2000 and 2010.

TABLE I. NUMBER OF ATTRIBUTES AND RECORDS.

Database	Attributes		Records	
	2000	***2010***	***2000***	***2010***
Mortality	38	40	437,667	592,018
Geographic	7		2475	
Population	3		2475	
CIE-10	24		2049	

B. Data analysis and pre-processing

Databases were analyzed in order to understand the database schema, to select the attributes of interest, and to understand the meaning of these attributes. This analysis was performed using a data description file, which is provided by the information sources from where the databases were extracted.

Additionally, the data analysis allowed us to identify attributes with errors which should be corrected or deleted in order to prevent the knowledge extracted by the data mining process to be wrong. In a first review, attributes with null values or empty attributes were deleted for all tuples; dependent attributes were identified and deleted using a correlation analysis. Then, the attributes of interest, needed to achieve the data mining goal, were selected by an expert domain.

Records that do not represent information of interest to the data mining goal, e.g., records related to districts with population less than 100,000 inhabitants, were eliminated.

Finally, the data construction tasks were performed. In this case, two new attributes, needed to reach the data mining objective, were identified and incorporated. The specific operations to calculate them are: calculation of the mortality Incidence and the calculation of the Mortality Rate; these are relevant indicators for the epidemiological domain experts.

The mortality *Incidence* is the number of cases observed for a particular disease in a specific district, in a particular year. The calculation of the *Mortality Rate* is performed for each district with population over 100,000 inhabitants, by convention in the health area, using the Expression (1):

$$\text{Rate} = \frac{\textit{Incidence}}{\textit{Population}} * 100{,}000 \qquad (1)$$

where *Population* is the number of inhabitants in a district for a specific year. This value is extracted from the population data provided by INEGI.

C. Data integration

A data warehouse is a database that integrates data from one or more information systems in an organization and it is oriented to assist in the decision making process [17]. The data were integrated in a data warehouse with a similar structure to that proposed in [18]. A data warehouse stores historical and non-volatile information related with a fact, in this case, the information is related to all deaths occurred in 2000 and 2010, in districts with a population over 100,000

 ISBN: 978-1-61208-247-9

inhabitants. Additionally, the data warehouse structure facilitates the integration of data for different years.

The data warehouse includes three dimensions. The *fact* table stores the results of the operations to calculate the mortality Incidence and the Mortality Rate for a particular disease among 2049 causes contained in the CIE-10.

Figure 1 shows the multidimensional model of the data warehouse. It is represented as a bucket with three views or dimensions that include CAUSE of death related to SPACE (districts) and TIME (year) of occurrence.

III. DATA PREPARATION SUBSYSTEM

In order to automate the data preparation process for the epidemiological domain, a data preparation subsystem was developed. The subsystem was implemented on Java language using Java Database Connectivity (JDBC) and SQL (Structure Query Language).

JDBC allows to establish the connection between the java language and the data warehouse; SQL standard allows to write queries to access data stored in it.

Figure 2 shows the conceptual model of the data preparation subsystem. The subsystem contains two modules that automate some tasks of data construction and data integration. A brief description of these modules is provided below:

A. Module of data construction tasks

In this module, the subsystem access to the data stored in the data warehouse and selects all data related to a specific cause of death for a particular year, and then the operations to calculate the mortality Incidence and Mortality Rate are executed. For each operation the subsystem executes:

- Calculation of mortality Incidence. A query sentence that counts all records related to deaths occurred for a specific cause of death in a particular year.
- Calculation of Mortality Rate. The results obtained for the mortality incidence are used to execute arithmetic operations to calculate the Mortality Rate using Expression 1 (showed in Section II.B).

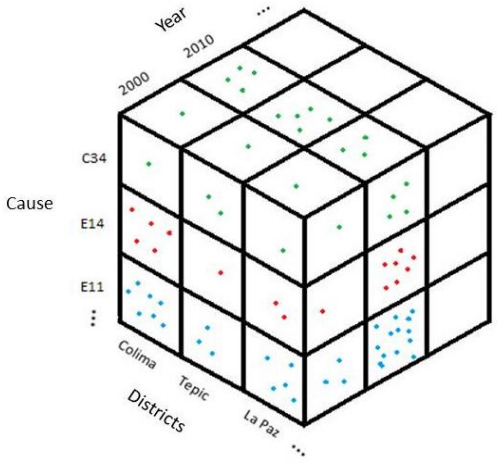


Figure 1. Conceptual view of the data warehouse.

The subsystem executes the previous tasks and the results are stored in the *fact* table described in Section II.C.

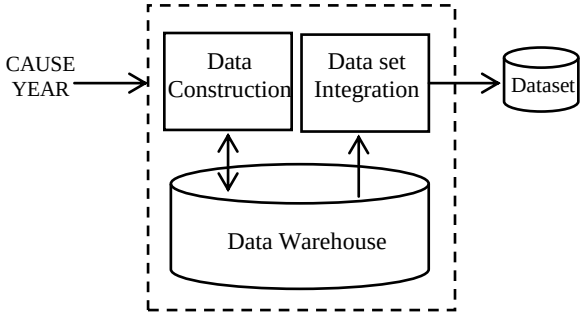


Figure 2. Data preparation subsystem.

B. Module of final dataset integration

Additionally, the subsystem automates the creation of a dataset whose structure is described in Table II. The data preparation subsystem takes, from the *fact* table, all the records related to CAUSE and YEAR of death introduced as input.

The dataset structure has four attributes; the first one contains the value for the cause of death to be analyzed, second and third columns contain the values of latitude and longitude of the district used to find its position on the map. The last column corresponds to the calculation of the mortality rate.

The subsystem has the ability to perform these calculations and to build the dataset, receiving as input only the values of CAUSE and YEAR of death; the tests were performed changing these values. For CAUSE, the causes tested were C16, C34, E11, and E14; for YEAR 2000 and 2010.

IV. CARTOGRAPHIC VISUALIZATION SUBSYSTEM

The dataset generated by the data preparation subsystem is used as input by the cartographic visualization subsystem described in [4]; this subsystem contains two modules: a pattern generator and a cartographic display.

The pattern generator takes the dataset (described in Section 4), converts it into an .arff file and, using the K-means clustering algorithm implemented on Weka, classifies the *n*-items into a given number of *k*-groups of districts with similar "Latitude", "Longitude", and "Mortality rate" for a specific cause of death. The results of the clustering algorithm are lists of element groups including the centroid of each group.

TABLE II. DATA SET STRUCTURE.

Cause	Latitude	Longitude	Mor_Rate
Cause of death	Latitude of the district. Position on the map.	Longtude of the district.Position on the map.	Mortality rate.

 ISBN: 978-1-61208-247-9

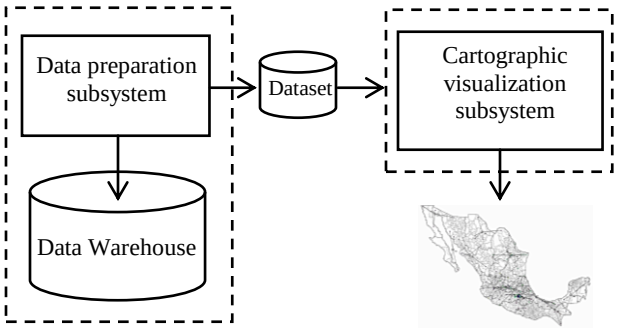


Figure 3. Interaction between the data preparation subsystem and the cartographic visualization subsystem.

The cartographic display permits to select and draw a map of Mexico with all the groups generated after analyzing the data with k-means clustering algorithm. Figure 3 shows the interaction between the data preparation subsystem and the cartographic visualization subsystem.

A set of experiments were conducted with the cartographic visualization subsystem for the same causes tested by the data preparation subsystem. We have used the k-means algorithm and established the value of *k* parameter in 15, 20 and 25 for each cause corresponding to data in 2000. For data in 2010, the *k* parameter was established in 20, 25, and 30 because in 2010, the number of districts was greater than in 2000. In both cases, the best result was obtained with k =20.

Figures 4 and 5 show the maps displayed by the cartographic visualization subsystem for the mortality cause E11 related to diabetes mellitus no insulin-dependent. We have highlighted three patterns that represent the groups of districts with the higher mortality rates for this cause in 2000 and 2010.

Figures 6 and 7 show the maps displayed by the cartographic visualization subsystem for the mortality cause E14 related to unspecified diabetes mellitus. We have highlighted three patterns that represent the groups of districts with the higher mortality rates for this cause in 2000 and 2010.

V. RESULTS AND CONTRIBUTIONS

We have implemented a data warehouse that contains information about all deaths occurred in districts of Mexico with populations over 100,000 inhabitants in 2000 and 2010. In addition, we have implemented a data preparation subsystem which has the capacity to generate prepared datasets for 2049 different causes of death registered in the CIE-10.

In order to validate the results obtained by the automated data preparation subsystem, we have performed tests for the causes of death C16 (stomach cancer) and C34 (lung cancer), and we have compared these results with the results obtained by previous works [5], [6], and [7] where the data preparation process were manually performed. The values generated for the mortality *Incidence* and the *Mortality Rate* by the data preparation subsystem correspond exactly to the values obtained in previous works. Table III shows the results obtained for the first group identified in previous works as a group of interest for the cause of death C16 in 2000:

TABLE III. GROUP OF INTEREST FOR THE CAUSE C16 IN 2000.

District	Incidence	Mortality rate
Rio Bravo	14	13,43
Matamoros	54	12,91
Torreon	65	12,27
Monterrey	113	11,97
Piedras Negras	15	11,7
San Nicolas de los G.	53	10,67
Reynosa	42	9,98
Gomez Palacio	27	9,88
Santa Catarina	21	9,25

Table IV shows the results obtained for the first group identified for the cause of death C34 in 2000 in previous works:

TABLE IV. GROUP OF INTEREST FOR THE CAUSE C34 IN 2000.

District	Incidence	Mortality rate
Guaymas	15	11.52
Hermosillo	48	7.87
La Paz	14	7.11
Los Cabos	7	6.64

Another important contribution is the automation of the data preparation tasks that calculate the mortality *Incidence* and the *Mortality Rate*, which represented the largest effort in the data preparation process. We carried out tests for four causes of death (C16, C34, E11 and E14) for 2000 and 2010, Table V shows a comparison between the time required to perform manually and automatically these tasks:

TABLE V. TIME COMPARISON.

Task	Manual average (min)	Automatic average (min)	% Time reduction
Calculation of Incidence	53.255’	.101’	99.81
Calculation of mortality rate	5.5’	.544’	90.11

Additionally, other experiments were performed; datasets with information for the causes of death E11 and E14 for the years 2000 and 2010 and their corresponding maps (Figures 4, 5, 6 and 7) were generated. Group three (Figure 6c) is the one with the largest average mortality rate, but this group is not meaningful because its districts are more dispersed than in other groups.

 ISBN: 978-1-61208-247-9

VI. CONCLUSIONS AND FUTURE WORK

Data collected by official censuses in a country permits to analyze interesting aspects of a population. In the health domain, census databases allow us to have a perspective of how diseases affect some sectors of the population and the evolution of these diseases through time and space.

We have used official census databases from 2000 and 2010 and applied a specialized data mining process in order to analyze different causes of death related to cancer and diabetes.

In the causes of death selected for cancer (C16 and C34) and diabetes (E11 and E14), we observed an increase in the mortality rates in 2000 compared to 2010. For the cause of death E11, we have observed in the maps an increase in the number of districts with high mortality rate and the concentration of these districts in the central region of the country (see Figure 6b).

Other patterns of interest were identified for the cause of death E14. In 2000 a group of districts in the state of Chihuahua were identified (see Figure 6a). Figure 7a shows another group of interest in the northwest region which has the second largest mortality rate; its districts are located between the states of Sonora and Baja California.

As a future work, we propose to make a trend analysis of deaths occurred between 2000 and 2010 for different causes of death. Additionally, we propose the integration of the data preparation subsystem and the cartographic visualization subsystem.

REFERENCES

[1] Consumer's Federal Attorney, PROFECO. Available: http://www.profeco.gob.mx/encuesta/brujula/bruj_2005/b05_censos.asp. Last visited: December, 2012.

[2] Censuses and population counts. Available: http://www.inegi.org.mx/est/contenidos/proyectos/ccpv/presentacion.aspx. Last visited: December, 2012.

[3] B. Zhan "Monitoring geographic concentration of female breast cancer using cluster analysis: the case of New Mexico", Papers and Proceedings of Applied Geography Conference (AGC 02). Vol 25, New York, Oct. 2002, pp 1-8.

[4] V. Bogorny, P. Engel, and L. Alvares "Spatial data preparation for knowledge discovery", IFIP Academy on the state of software theory and practice – PhD Colloquium. Porto Alegre, Brazil, 2005.

[5] V. Bogorny, P. Engel, and L. Alvares "A reuse-based spatial data preparation framework for data mining", International Conference on Sotware Engineering and Knowledge Engineering (SEKE 05), Taiwan, July 2005, pp. 649 - 652.

[6] N. Labib and M. Malek "Data mining for cancer management in Egypt case study: childhood acute lymphoblastic leukemia", Transaction on Engineering Computing Technology, Vol 8,Oct. 2005, pp. 309-314.

[7] M. Izadi, D. Buckeridge, and K. Charland, "Mining epidemiological data sources in H1N1 pandemic using probabilistic graphical models", International Conference on Advances in Information Mining and Management (IMMM 11), Spain, Oct. 2011, pp. 1-6.

[8] J. Salinas, Adaptation of a data mining methodology for its application to a real population-based database of cancer records, 2007. Master thesis, Computer science. Cenidet, Cuernavaca, Mexico.

[9] A. Mexicano, Development of a methodology for feature selection and indicator generation for the application of data mining to a real population-based cancer database, 2007. Master thesis, Computer science. Cenidet, Cuernavaca, Mexico.

[10] M. Barron, Development of a prototype for the application of data mining techniques on a real population-based cancer database, 2008. Master thesis, Computer science. Cenidet, Cuernavaca, Mexico.

[11] J. Pérez et al., "Spatial Data Mining of a Population-Based Data Warehouse of Cancer in Mexico", International Journal of Combinatorial Optimization Problems and Informatics (IJCOPI 10). Vol 1, May 2010, pp. 61 – 67.

[12] J. Pérez, O. Fragoso, R. Santaolaya, A. Mexicano, and F. Henriques, "A data mining system for the generation of geographical C16 cancer patterns" International Conference on Software Engineering Advances (ICSEA 10), France, Aug. 2010, pp. 417-421.

[13] National Health Information System, SINAIS. Available: http://www.sinais.salud.gob.mx/basesdedatos/estandar.html. Last visited: December, 2012

[14] Database District System, SIMBAD. Available: http://sc.inegi.org.mx/sistemas/cobdem/contenido-arbol.jsp. Last visited: December, 2012

[15] Statistics and Geography National Institute, INEGI. Available: http://www.inegi.org.mx/. Last visited: December, 2012

[16] Collaborating Center for the Family of International Classifiers, CEMECE. Available: http://www.cemece.salud.gob.mx/fic/cie/index.html. Last visited: December, 2012

[17] J. Moral, "Introduction to data Warehousing", Annals of Mechanical and Electrical. Vol. LXXXVI, Spain, Mar. 2009, pp. 42-47.

[18] R. Boone, Identification of regions with high rate of cancer incidence by integrating and using data mining techniques, data warehouse, clustering and geographic information systems, 2011. Master thesis, Computer science. Cenidet, Cuernavaca, Mexico.

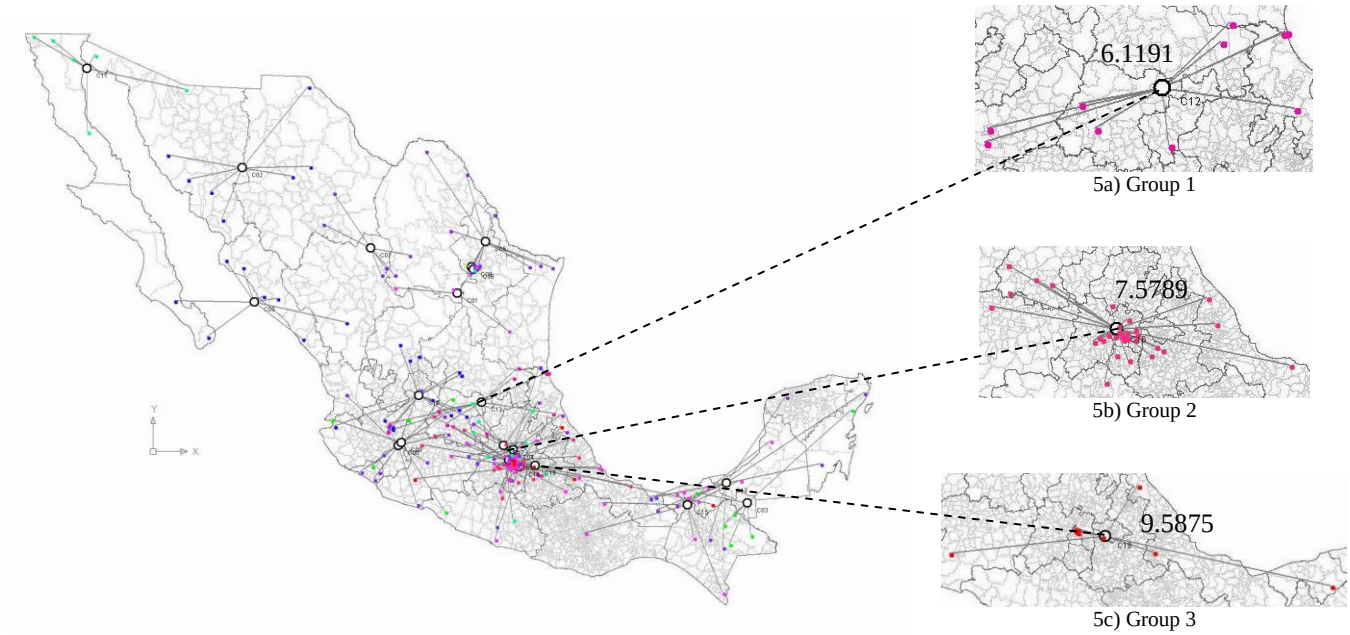
 ISBN: 978-1-61208-247-9

4a) Group 1

4b) Group 2

4c) Group 3

Figure 4. Identification of groups of interest for the cause of death E11 (diabetes mellitus no insulin-dependent) in 2000. The groups 1, 2, and 3 have the highest mortality rates. Figures 4a, 4b, and 4c show the average value for each group.



5a) Group 1

5b) Group 2

5c) Group 3

Figure 5. Identification of groups of interest for the cause of death E11 (diabetes mellitus no insulin-dependent) in 2010. The groups 1, 2, and 3 have the highest mortality rates. Figures 5a, 5b, and 5c show the average value for each group.

 ISBN: 978-1-61208-247-9

6a) Group 1

6b) Group 2

6c) Group 3

Figure 6. Identification of groups of interest for the cause of death E14 (unspecified diabetes mellitus) in 2000. The groups 1, 2, and 3 have the highest mortality rates. Figures 6a, 6b, and 6c show the average value for each group.

7a) Group 1

7b) Group 2

7c) Group 3

Figure 7. Identification of groups of interest for the cause of death E14 (unspecified diabetes mellitus) in 2010. The groups 1, 2 , and 3 have the highest mortality rates. Figures 7a, 7b, and 7c show the average value for each group.

 ISBN: 978-1-61208-247-9