Using Data Warehousing and OLAP in Public Health Care

Dimitar Hristovski[1], M.Sc., Mitja Rogac[2], M.Sc., Mladen Markota[2], M.D., M.Sc.
[1]Institute of Biomedical Informatics, Medical Faculty, University of Ljubljana
Vrazov trg 2/2, 1105 Ljubljana, Slovenia
[2]Public Health Institute of the Republic of Slovenia
Trubarjeva 2, 1000 Ljubljana, Slovenia
e-mail: dimitar.hristovski@mf.uni-lj.si, mitja.rogac@ivz-rs.si, mladen.markota@ivz-rs.si

ABSTRACT

The paper describes the possibilities of using data warehousing and OLAP technologies in public health care in general and then our own experience with these technologies gained during the implementation of a data warehouse of outpatient data at the national level. Such a data warehouse serves as a basis for advanced decision support systems based on statistical, OLAP or data mining methods. We used OLAP to enable interactive exploration and analysis of the data. We found out that data warehousing and OLAP are suitable for the domain of public health and that they enable new analytical possibilities in addition to the traditional statistical approaches.

INTRODUCTION AND BACKGROUND

Health statistics has a long-standing tradition in Slovenia. It is almost fully centralised and the Public Health Institute of the Republic of Slovenia (PHIRS) is responsible for collecting, analysing and processing of medical data from all HC levels. PHIRS analyse statistical data to asses the morbidity of Slovenian population and to find out the trends in different time periods.

Every year PHIRS prepares the Health statistical annual. Data concerning organisation of health care, tasks performed and findings from all health services domains in the Republic of Slovenia are published in the statistical annual. Demographic and vital statistical data are included. The field of outpatient health care statistics contains for example:

- Number of curative and preventive attendances in outpatient health care services;
- Causes for attendances in outpatient health care according to the ICD-10 chapters and by age groups for the entire population;
- Causes for attendances in outpatient health care according to the ICD-10 chapters and by health region;
- First five most frequent causes for attendances of the unemployed in outpatient health care services by ICD-10 chapters etc;

These materials provide a basis for analysis of activities and definition of health services tasks in communities and health regions of Slovenia. PHIRS provides data for government (Health plan till 2004) and other institutions like WHO, UNICEF, World Bank etc.

For the processing the Health statistical annual PHIRS uses the Statistical Package for Social Sciences (SPSS tool).

In Slovenia we still use predominantly diskette media to transfer data from regions to the PHIRS and vice versa. We have developed and implemented electronic data interchange network in communication between these institutions.

Nevertheless our data structure is not yet completely unified with the EDIFACT standard. Increasing electronic communication with medical data raises some questions. Current information technologies offer some closed networks like HC network, Health Insurance network, Government Centre for Informatics network etc. All of them are still in process of independent and mutually uncoordinated development. Nevertheless the needs for data transmission exceed the capacity of the networks mentioned above.

1067—5027/00/$5.00 © 2000 AMIA, Inc.

MOTIVATION

In addition to the numerous reports produced at PHIRS using the traditional statistical tools, we wanted to give the health care data analysts the possibility to perform interactive exploration, ad-hock analysis and discover trends in a user-friendly way. Because of that, we decided to use data warehousing and OLAP (On-line Analytical Processing) technology for exploring and analysing the data. At the moment we have built a data warehouse of outpatient data, but later we plan to include the inpatient data as well.

DECISION SUPPORT AND DATA WAREHOUSING

The systems that run the every day operations of an organisation are usually called on-line transaction systems and the mode of operation is usually referred to as operational processing. For example, in outpatient clinics the details about the patient visit are recorded in the operational system. These systems are usually continually updated through the day.

Analytical systems are systems that provide information used for analysing a domain or situation. Analytical processing is primarily done through comparisons, or by analysing patterns and trends. For example an analytical system might show the main reasons for hospital visits for different regions of a country. By comparing the values for several consecutive years some trends may be discovered.

The data used for analytical processing is usually organised in a data warehouse. According to [1] a data warehouse is a subject oriented, integrated, non-volatile, and time variant collection of data in support of management's decisions. In other words a data warehouse is used as a foundation of a decision support system.

WHAT IS OLAP?

The phrase On-Line Analytical Processing (OLAP) was coined by Codd in [2] to characterise the requirements for summarising, consolidating, viewing, applying formulae to, and synthesising data according to multiple dimensions. OLAP systems provide an information structure that allows an analyst to have very flexible access to data, to slice and dice data any number of ways and to dynamically explore the relationship between summary and detail data.

OLAP can also be put in the context of data mining and knowledge discovery in databases. Data mining aims at discovering interesting, non-trivial and new patterns in a large amount of detail data. The systems used for data mining differ significantly according to the way of automation they offer the user. OLAP tools fall on the lower end of the spectrum and are used for manual data exploration.

The data in an OLAP system in organised in a multidimensional data structure, usually called a multidimensional data cube. Dimensions usually have associated with them hierarchies that specify aggregation levels and hence granularity of viewing data. Thus, day –> month –> year is a hierarchy on the TIME dimension that specifies various aggregation levels. Other dimensions that appear in outpatient data: patient age group, sex, diagnosis,.... In the intersection of the dimensions are the measures (or facts). A typical measure for health care is the number of attendances.

CONTENTS OF THE REPORTING DATA STRUCTURE

Outpatient health care activities are performed in health care centres, specialist outpatient offices in hospitals, and private practices. There are more than 1000 such health care providers in Slovenia. During their work with patients, health care personnel enters the required data into computers.

At mid-year and at year's end, health care providers are obliged to prepare the required data structure [3] on the basis of their daily entered records, according to the instructions of PHIRS, and send them to their regional Health Care Institute (HCI) by the prescribed deadline. There are nine regional HCIs in Slovenia. Regional HCIs collect data from the health care providers in their region, check them using the ZUBSTAT program and, if errors are found, send them back to the originating provider with a list of errors.

This new method of data reporting was introduced in the following basic health care activities [4]:

- General practise
- General practise - prevention
- Occupational medicine
- School health care - curative
- School health care - prevention (regular check-ups)
- Paediatric health care - curative
- Paediatric health care - prevention (consultancies)
- Developmental (paediatric) outpatient facilities
- Women's health care (Gynaecology and obstetrics)
- Women's health care - (consultancies for pregnant women)

- Women's health care - (consultancies for contraceptive practices)
- Specialist outpatient health care: by medical specialty

Data reporting using the new computerised method according to the instructions of the Ministry of Health, was introduced on 1 January 1997.

PHIRS has decided on a new, uniform record structure for reporting data on outpatient health care statistics at the national level [5], on the assumption that the set of data entered by health care personnel on a daily basis remains unchanged.

The objectives of this new reporting structure are as follows:

- Uniforming of methods for data reporting on attendances and referrals, as well as those on diseases and conditions from basic and specialist health care.
- Data reporting using the new three-digit municipal codes, in effect since 1 January 1997.
- Monitoring of patient structure by municipality of residence for each health care centre, i.e. patient gravitation towards health care centres (which is enabled by the new reporting structure of data on attendances and referrals).
- Data reporting by more detailed age groups and by sex.
- Collection of data on diagnoses by the tenth revision of the International Classification of Diseases (ICD-10), which has been used in the field of health care statistics since 1 January 1997.
- More detailed definitions and standards for data collected.
- Checking of data with regard to valid code schemes and logical associations between data.
- Higher quality of output data given an unchanged amount of input data.

In connection with attendances and referrals in outpatient health care activities at the national level, PHIRS collects and monitors the following data [6]:

- Designation for the reporting period
- Health-care provider's municipal code
- Health care provider's code
- Designation of the provider's basic activity
- Code of the provider's health care service
- Code of the provider's location
- Designation of the type of attendance
- Municipal code of the patient's residence
- Number of patients by age group and sex
- Number of patients by health insurance category
- Number of pregnant women grouped by duration of pregnancy
- Number of patients referred to the hospital
- Number of patients referred to a specialist

The key attribute here is "Designation of the Type of Attendance", to which all other attributes are associated. In this case as well, each attribute has its own definition of content and reference to the corresponding code table, which has been harmonised at the national level.

There are 14 basic patient age groups which are grouped into three higher level groups: pre-school children, school children and adults.

BUILDING THE OUTPATIENT DATA WAREHOUSE

We performed the following steps when building the data warehouse of outpatient data. First the data sent from the regional Health Care provides was checked for consistency and in case of errors it was returned to the providers. We have to stress here that the reporting data does not contain any personal information.

After that, we did some data transformation of the reporting files. The reporting files are actually preaggregated for space saving reasons. They were transformed in a more atomic form suitable to be loaded into the data warehouse. This was accomplished with scripts written in the AWK programming language.

Then we defined the necessary dimensions and variables in the multidimensional OLAP server. Special attention was paid to the definition of hierarchies for some of the dimensions. E.g. health care providers are organised in a hierarchy with the following levels: *Country* → *Region* → *Municipality* → *Health Care Provider*. The OLAP server we used was Oracle Express 6.2.

Afterwards, the previously prepared data was loaded into the OLAP server and some additional variables were also defined. Then, the variables were aggregated along the hierarchical dimensions. This turned out to be a very time and space consuming process because of the high dimensionality of the resulting data cube.

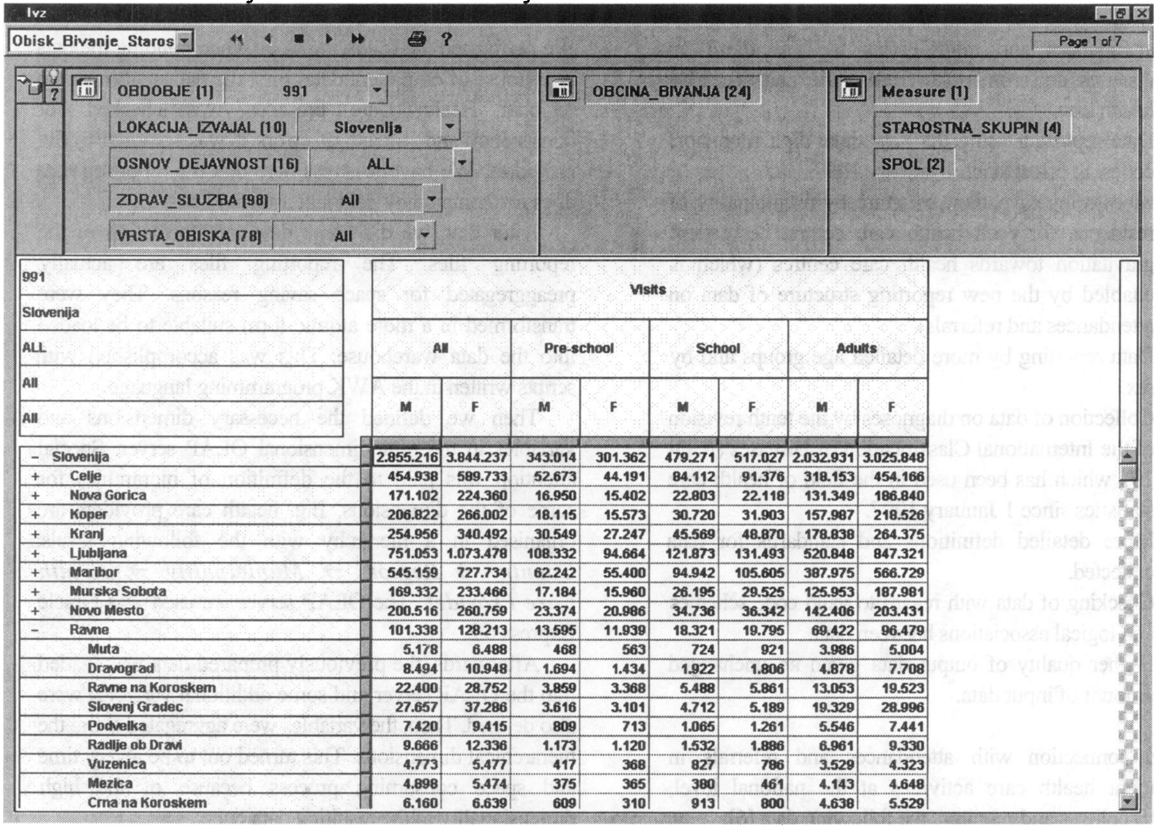
At the end, we developed the end-user applications for data exploration and analysis using the Oracle Express Analyzer tool.

USING OLAP FOR INTERACTIVE EXPLORATION AND ANALYSIS

Figure 1 shows a typical OLAP view of the data cube. In the upper part of the screen the dimensions are listed. The table under the dimension list shows the number of outpatient visits variable dimensioned by patient's residence and age group and sex. This table can be used for interactive data exploration. The patient's residence dimension shown in the first column is a hierarchical one as well as the age groups. The age groups are shown at the first two hierarchical levels. The multidimensional view allows hierarchies associated with each dimension also to be viewed in a logical manner. The numbers in the first row show the aggregated (totalled) number of visits. We can explore the next level of detail by clicking the plus sign to the left of a particular dimensional value, which is called drill-down in the OLAP terminology. Drill-down is essential because often users want to see only aggregated data first and selectively see more detailed data. It is possible to view the data dimensioned by some other dimension by simply exchanging the dimensions positions. E.g. we can view the data by the location of the health care provider by dragging that dimension over the patient's residence dimension.

An analyst might want to see only a subset of the data and thus might view some attributes and within each selected attribute might restrict the values of interest. In OLAP terminology [7], these operations are called pivoting (rotate the hyper-cube to show a particular face) and slicing-dicing (select some subset of the cube).

The multidimensional data can be also shown in a graphical form. Figure 2 shows the gravitation of patients towards health care centres. Both dimensions are selected at the regional level.



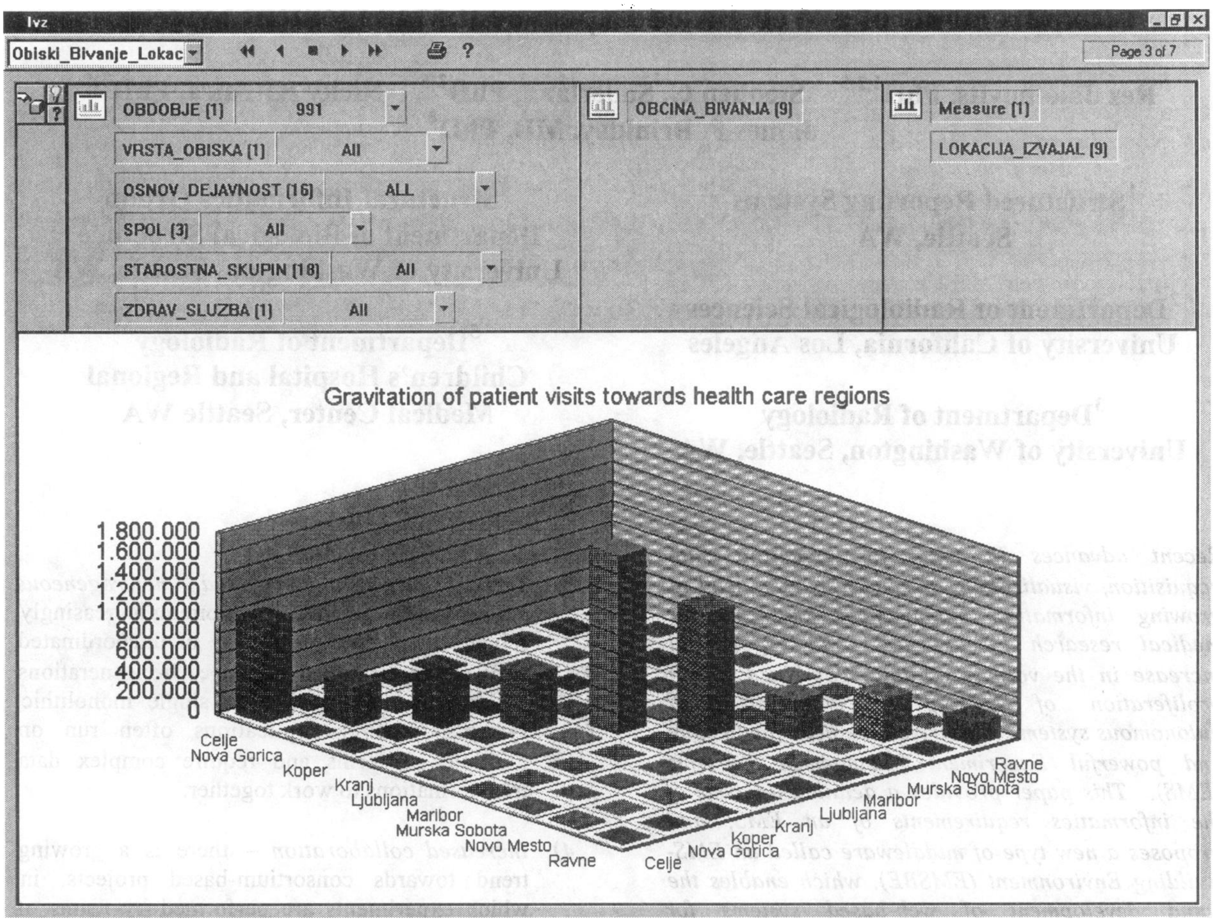
OBDOBJE [1]: 991
LOKACIJA_IZVAJAL [10]: Slovenija
OSNOV_DEJAVNOST [16]: ALL
ZDRAV_SLUZBA [98]: All
VRSTA_OBISKA [78]: All

OBCINA_BIVANJA [24]

Measure [1]
STAROSTNA_SKUPIN [4]
SPOL [2]

991 Slovenija ALL All All	Visits							
	All		Pre-school		School		Adults	
	M	F	M	F	M	F	M	F
− Slovenija	2.855.216	3.844.237	343.014	301.362	479.271	517.027	2.032.931	3.025.848
+ Celje	454.938	589.733	52.673	44.191	84.112	91.376	318.153	454.166
+ Nova Gorica	171.102	224.360	16.950	15.402	22.803	22.118	131.349	186.840
+ Koper	206.822	266.002	18.115	15.573	30.720	31.903	157.987	218.526
+ Kranj	254.956	340.492	30.549	27.247	45.569	48.870	178.838	264.375
+ Ljubljana	751.053	1.073.478	108.332	94.664	121.873	131.493	520.848	847.321
+ Maribor	545.159	727.734	62.242	55.400	94.942	105.605	387.975	566.729
+ Murska Sobota	169.332	233.466	17.184	15.960	26.195	29.525	125.953	187.981
+ Novo Mesto	200.516	260.759	23.374	20.986	34.736	36.342	142.406	203.431
− Ravne	101.338	128.213	13.595	11.939	18.321	19.795	69.422	96.479
Muta	5.178	6.488	468	563	724	921	3.986	5.004
Dravograd	8.494	10.948	1.694	1.434	1.922	1.806	4.878	7.708
Ravne na Koroskem	22.400	28.752	3.859	3.368	5.488	5.861	13.053	19.523
Slovenj Gradec	27.657	37.286	3.616	3.101	4.712	5.189	19.329	28.996
Podvelka	7.420	9.415	809	713	1.065	1.261	5.546	7.441
Radlje ob Dravi	9.666	12.336	1.173	1.120	1.532	1.886	6.961	9.330
Vuzenica	4.773	5.477	417	368	827	786	3.529	4.323
Mezica	4.898	5.474	375	365	380	461	4.143	4.648
Crna na Koroskem	6.160	6.639	609	310	913	800	4.638	5.529

Figure 1. Patient visits dimensioned by patient's residence and age group and sex.



Ivz

Obiski_Bivanje_Lokac

Page 3 of 7

OBDOBJE [1] 991

VRSTA_OBISKA [1] All

OSNOV_DEJAVNOST [16] ALL

SPOL [3] All

STAROSTNA_SKUPIN [18] All

ZDRAV_SLUZBA [1] All

OBCINA_BIVANJA [9]

Measure [1]

LOKACIJA_IZVAJAL [9]

Figure 2. A graphical OLAP view showing the gravitaion of patient visits towards health care regions.

CONCLUSIONS AND FURTHER WORK

Our experience with building an outpatient data warehouse at the national level showed that data warehousing and OLAP are suitable technologies for building decision support systems in the domain of public health care. In the future we plan to build a data warehouse of hospital visits data and apply data mining methods for advanced data analysis.

References

[1] Inmon WH. Building the Data Warehouse. Second Edition. John Wiley & Sons. 1996.

[2] Codd EF, Codd SB, and Salley CT. Beyond decision support. Computerworld, 27(30), July 1993.

[3] Ministrstvo za zdravstvo, Konceptualni okviri nadaljnjega razvoja zdravstvenega informacijskega sistema v Republiki Sloveniji. Ljubljana, 1992.

[4] WHO, Elements of the Role of Informatics in the Health for All Policy of the World Health Organisation. Copenhagen, 1994.

[5] Inštitut za varovanje zdravja RS, Projekt Elementi enotnosti zdravstvenega informacijskega sistema v Republiki Sloveniji. II. fazno poroèilo, Ljubljana, 1994.

[6] Inštitut za varovanje zdravja RS, Zunajbolnišnièna zdravstvena statistika. Metodološko gradivo, Ljubljana, 1999.

[7] Agraval R. Modeling Multidimensional Databases. Research Report. IBM Almaden Research Center. 1995.