Business Intelligence and Nosocomial Infection decision making

Eva Silva
University of Minho, Computer Science and Technology Center (CCTC), Braga Portugal
Ana Alpuim
University of Minho, Computer Science and Technology Center (CCTC), Braga Portugal
Luciana Cardoso
University of Minho, Computer Science and Technology Center (CCTC), Braga Portugal
Fernando Marins
University of Minho, Computer Science and Technology Center (CCTC), Braga Portugal
Carlos Filipe Portela
University of Minho, ALGORITMI Research Center, Guimarães, Portugal
Manuel Filipe Santos
University of Minho, ALGORITMI Research Center, Guimarães, Portugal
José Machado
University of Minho, Computer Science and Technology Center (CCTC), Braga Portugal
António Abelha
University of Minho, Computer Science and Technology Center (CCTC), Braga Portugal

ABSTRACT

Nosocomial infection prevention in healthcare units it is very important to improve patient's well-being and safety. This prevention can be done by manipulating and analysing real data to identify critical processes and areas inside the healthcare unit, and monitoring indicators generated from data.
The main goal of this paper is to evaluate the applicability of the Business Intelligence tools and concepts to healthcare and their performance as a Clinical Decision Support System, analyzing the evolution of nosocomial infection in the *Centro Hospitalar do Porto*, by defining a set of indicators that can help the nosocomial infection management and inducing Data Mining models to predict the occurrence of nosocomial infections (sensitivity of 91%).
A Business Intelligence system composed by the presentation of a set of indicators and a Data Mining part capable of predict the occurrence of infection can provide important information to support healthcare professionals in their decisions.

Keywords: Business Intelligence, Clinical Decision Support System, Data Mining, Data Warehouse, Electronic Health Record, ETL, Health Information System, Nosocomial Infection.

INTRODUCTION

With technological advancement, health organizations have been increasingly adopting information systems. The Health Information Systems (HIS) handle data processing, information and knowledge in healthcare environments (Winter, Haux, Ammenwerth, Brigl, Hellrung, & Jahn, 2011). The Electronic Health Record (EHR) is a HIS which covers the various services and units of a healthcare organization. It consists in a set of standardized documents, compiled by health professionals, for the registration of the medical procedures provided to a particular patient. The EHR main goal is to improve the provision of healthcare to patients, which is the task of all of those who work in hospitals. The EHR brings the new information technologies in all its aspects, aiming to eliminate paper and consequently, to expedite and to improve the healthcare delivery to the patients (Duarte, Portela, Abelha, Machado, & Santos, 2011; Hasman, 1998; Salazar, Duarte, Pereira, Portela, Santos, Abelha, & Machado, 2013) .

Besides registration, consultation and research of all clinical information, resulting from the provision of healthcare to a particular patient, the EHR also allows the prescription of medicines and complementary means of diagnosis such as exams, called electronic medical prescription (Simões, Gomes, & Paiva, 2009).

The electronic medical prescription is a procedure performed through the use of information and communication technologies, in this case applications certified by the regulatory organization of healthcare services in Portugal, called _*Serviços Partilhados do Ministério da Saúde* (SPMS) (Ministério da Saúde, 2014).

The healthcare environment is characterized by a highly distributed and heterogeneous computational environment, where different systems and people need to be in contact and to communicate, exchanging data and knowledge that are indispensable for decision making (Cardoso, Marins, Portela, Santos, Abelha, & Machado, 2014; Machado, Alves, Abelha, & Neves, 2007). In addition, the healthcare environment and its processes are extremely dynamic, complex and multidisciplinary (Rebuge & Ferreira, 2012).

However, more and more healthcare organizations act constantly under financial pressure, forcing them to improve the efficiency of their activities and processes and applying their resources as efficiently as possible in order to improve the quality of services (Foshay & Kuziemsky, 2014; Rebuge & Ferreira, 2012).

In the healthcare sector it is very important to make fast and quality decisions because the decisions are frequently related to the human well-being. Furthermore, the healthcare decision making is always a very complex process and to be correct, it requires high quality information (Foshay & Kuziemsky, 2014; Lenz & Reichert, 2007).

Nowadays, with the implementation of Information Technologies (IT) in the healthcare organizations, the amounts of data collected have exponentially increased (Spruit, Vroon, & Batenburg, 2014). In the healthcare environment, decision support is mostly related to knowledge that can be extracted from the collected data, which makes the information management essential for these organizations (Lenz & Reichert, 2007). Thus, the extraction, analysis and presentation of the information in an useful and timely manner can reduce costs and improve the quality, safety and efficiency of healthcare delivery, once they allow a more rational decision. In consequence, the implementation of knowledge extraction techniques became a fundamental operation to support healthcare organizations, because they support the extraction of rich and quality information that can be applied in decision making. These decisions can be related to clinical and/or administrative issues (Mettler & Vimarlund, 2009; Spruit et al., 2014).

Thus, the decision process and the Clinical Decision Support Systems (CDSS) are fundamental for the actual healthcare organizations. The necessity of taking good and fast decisions in a very complex and competitive environment as the healthcare, concerning the quality of information and knowledge, can be achieved with the use of CDSS based on Business Intelligence (BI).

The main goal of this chapter is to evaluate the applicability of the BI concepts to healthcare data, through the utilization of the BI tool Jaspersoft Studio and through the application of Data Mining (DM) concepts for variables prediction. It is also aims to evaluate the performance of Jaspersoft Studio as CDSS. To validate the approach presented in this paper, it was conducted a case study in the *Centro Hospitalar do Porto* (CHP) a large hospital located in the north of Portugal.

This study analyses the evolution of nosocomial infection in the CHP, by defining a set of indicators that can help the nosocomial infection management, through the identification of important factors and parameters that characterize the nosocomial infection incidence. This study also evaluates the implementation of DM techniques to make clinical prediction related with the occurrence of a nosocomial infection when certain risk factors are present or absent in the patient.

Nosocomial infection or hospital-acquired infection is an infection that occurs in the period of 48 hours after the hospital admission, 3 days after the discharge or 30 days after an operation (Inweregbu, 2005).

This type of infection can be prevented with appropriate infection control measures and it is an important factor to evaluate the quality of the care delivered by a healthcare organization. Moreover, a patient with hospital-acquired infection spends a longer time in hospital, incurring additional costs.

The definition of indicators capable of capturing the rate of nosocomial infection and important factors that contribute to the presence of this infection in the organization, helps in the identification of processes, activities and departments where this rate is high and where should be applied measures to decrease the infection incidence.

Besides the introduction, this article includes five more sections. The first is related to the background and provides an overview of the BI technology. In the second section it is discussed the utilization of BI in the healthcare sector. The third section presents the case study in terms of the indicators used in the analysis, the methodology followed, the main results and their discussion. The fourth section presents a study of the implementation of DM to predict nosocomial infections, the methodology used, the main results obtained and their discussion. The fifth section presents some future work suggestions and the last section presents the main conclusions of the chapter.

BACKGROUND

Business Intelligence

Business Intelligence (BI) is one area of the Decision Support Systems (DSS) discipline referring to a collection of decision support technologies capable of collecting, integrating, storing, analysing and presenting data about the organization activities and processes in order to support a more informed and, consequently, better decisions (Chaudhuri, Dayal, & Narasayya, 2011; Glaser & Stone, 2008; Prevedello, Andriole, Hanson, Kelly, & Khorasani, 2010).

The BI technology provides the means to transform the organizational data into relevant and strategic information that can be used to support the organization decision process, allowing knowledge workers such as executives, managers and analysts to make better and fast decisions (Bonney, 2013; Loshin, 2012). So, this technology is very important and its implementation is competitive advantage for the organization.

A BI system must be able to perform two fundamental tasks: integrate huge amounts of data coming from several different heterogeneous sources and provide the analytical tools for these data analysis (Popovič, Hackney, Coelho, & Jaklič, 2012). A BI system uses user-friendly tools that make this system available to the final user.

The principal benefits of BI technology are the saving of time in the access and in the analysis of data, the flexibility of this technology and the improvement in decision making by using information driven from real data. The real data that support decision making increase the probability of obtaining correct and reliable results with the BI system (Bonney, 2013). Also, BI improves the timeliness and the quality of inputs of the decision making process and that makes them capable of provide the correct information to the user on time, in order to assist in the decision process (Mettler & Vimarlund, 2009). In this way, the BI systems reveal to be a competitive advantage for the organization that implements them.

The BI technology includes several software features: Extraction, Transformation and Load (ETL), data warehousing, Online Analytical Processing (OLAP), DM, reporting, database querying and visualization (Bonney, 2013). A BI system integrates data coming from several heterogeneous sources, processes and converts these data into a unified format and load them in a data warehouse (DW) (Prevedello, Andriole, & Khorasani, 2008). This process can be very complex and time-consuming. After the implementation of this ETL process, the data stored in the DW can be used by analytical applications capable of aggregate the data using OLAP, perform DM, querying or reporting. In the last years, there has been a significant improvement in the BI tools development: the speed of data collection has been increasing, as well as the sophistication and interactivity of the data manipulation and querying and there has been an improvement of reporting tools and of the tools used to present the information (Prevedello et al., 2010).

Data Warehouse

The main component of a BI system is the DW, a repository of data coming from different sources that stores information about the activities that happen in an organization (Loshin, 2012). A Data Warehouse (DW) allows to store and consolidate them into a valid and consistent format for each organization. Also allowing its users to analyse and to explore the data using other tools (Santos & Ramos, 2006).

A DW aims to integrate data from different sources and formats, these data are subject oriented. This allows the analysis of the subject using different perspectives or dimensions. For these reasons the DW databases are considered multidimensional and are oriented to the subject, being the subject the set of information regarding a particular process in an organization, to facilitate the use of information. Besides the subject oriented, a DW has the ability to integrate data, being this one of its most important features. Thus, the process of data input is done in order to eliminate inconsistencies. To ensure consistency and integration among data, techniques of extraction, cleaning, transformation, loading and integration are used, also known as ETL techniques.

A DW is not volatile, in other words, the data update does not happen often because normally they are loaded in large amounts. Moreover, in DW, just two different operations are allowed, the insert and the consult of data. Finally, the variation in time is also a feature of DW, which allows the temporal analysis of the data. The data loading frequency varies depending on the needs of each organization (Park & Kim, 2013).

The storage of relevant data coming from different sources in that single location and format can improve the velocity and efficiency of the knowledge discovery process, allowing to make better, faster and more informed decisions (El-Sappagh, Hendawi, & El Bastawissy, 2011; Prevedello et al., 2010).

The data stored in the DW are consistent, integrated, historical and are available to be analysed by analytical and reporting BI tools, in order to extract information to apply in the decision making process. DWs differ from operational databases because they are integrated, organized by subject and they vary in time, which means that each entry in the DW corresponds to a specific moment in time and that allows the temporal analysis of the data. Besides that, DW are bigger in size, they support OLAP, they are not volatile because the data present in the DW are not delete or updated, and they are essentially used for decision support (El-Sappagh et al., 2011). Moreover, unlike the operational systems, DW systems are not worried about the fast and efficient processing of transactions. On the contrary, these systems are essentially worried about the fast access to information for analysis and creation of reports (Popovič et al., 2012).

Normally, a DW organizes its data according to the dimensional model that allows a more efficient representation of the data used by the analytic and decision support applications (Loshin, 2012; Thalhammer, Schre, & Mohania, 2001). This model is the most adequate technique to present analytical data because it presents the data in a very comprehensive way to the system users and, simultaneously, it has a high performance in queries processing (Kimball & Ross, 2002).

Most of the DW uses the star schema to represent data according to the dimensional model. In the star schema, there is a set of dimension tables connected to a fact table through their primary keys.

In a star schema, table facts are the main element of the model and they represent the events used to measure the performance and the outcomes of the processes. A fact is a measure of an event and all measurements in a fact table must have the same granularity or detail level. Any fact has a certain level of detail, different levels of detail should be represented in different facts tables, which are constituted by few attributes and many records, contrary to what happens with the dimension tables. A table fact contains a set of foreign keys that bind to the dimension tables, which contain the description of the measured facts and these attributes are often used to identify headers in the query results (Soler, Trujillo, Fernández-Medina, & Piattini, 2008).

The dimension tables contain the attributes that describe the facts stored in the fact table and these attributes make that facts unique and give them meaning (Chaudhuri et al., 2011; Thalhammer et al., 2001).

This star schema model is used in the development of DW systems to enable greater speed in knowledge discovery, especially in queries that involves multiple bonds, and its representation easily understood by the user (Pardillo, Mazón, & Trujillo, 2010).

To resume, the dimensional model star schema constituted by a fact table in the centre, which is connected to a set of dimensions that contain the description of the facts stored in the central table (Soler et al., 2008). The connection between dimensions and the fact table is done through a set of foreign keys present in the fact table, that bind to the dimension tables (Soler et al., 2008).

The dimensional model of a DW is intended to be a system for the publication of data and it aims to ensure easy interaction with the end user of the application, as well as a high performance query processing (Soler et al., 2008). As the main advantages of the utilization of the dimensional model one can point the simplicity of the data model, the benefits in the performance of the data access process and the efficient data representation in the DW (Kimball & Ross, 2002; Loshin, 2012).

In the data warehousing field, there are two different approaches according to the used methodology, the Bill Inmon’s paradigm and the Ralph Kimball’s paradigm (1Keydata, 2014).

In the Bill Inmon’s paradigm, the BI is divided in several parts being one of them the data warehouse. Each enterprise has one data warehouse, and data marts to source their information from data warehouse. In this paradigm, the data warehouse information is stored in 3rd normal form. In the Ralph Kimball's paradigm, the data warehouse is the conglomerate of all data marts within the enterprise and the information is always stored in the dimensional model.

Knowledge Discovery in Database

The term Knowledge Discovery in Database (KDD) defines a process of a set of ongoing activities that produce new knowledge from databases. This set consist of five stages like selection, Pre-processing, Transformation, DM and finally Interpretation/Evaluation.

In the first stage, occurred the selection of the important data to perform the DM, and then the pre-processing which includes cleaning and processing of data in order to make them consistent.

According to the target, the data are worked out, this is the transformation stage.

Finally, in the DM stage, the objectives and the type of result wanted to achieve were defined. According to the type of problem and the type of task being performed they are defined and identified the techniques to be used. Subsequently will be selected DM scenarios in order to design prediction models and obtain patterns. The interpretation or evaluation is the last stage and consists in the interpretation and evaluation, as the name implies, of the patterns obtained. The validity of the results obtained is verified by applying the patterns found at new datasets (Azevedo, 2011).

Data Mining

DM is an interdisciplinary subfield of computer science which has evolved rapidly due the introduction of new methods, methodologies and findings in various applications related to many areas, such as, medicine, computer science, bioinformatics and stock market prediction, weather forecast, text, audio and video processing (Heinrichs & Lim, 2003; PhridviRaj & GuruRao, 2014).

DM is a process of analyzing relationships of information and discovering patterns from the existing data based on open-ended user queries, involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the DM process is to extract information from a data set and transform it into an understandable structure. In DM the first tasks are concentrated on cleansing the data to make it feasible for further processing. The process of cleansing the data is also called as noise elimination, noise reduction or feature elimination. The process of cleansing data can be either made by using tools such as ETL, tools available in the market or may be done by using various suitable techniques available (Heinrichs & Lim, 2003; Hema & Malik, 2010; PhridviRaj & GuruRao, 2014). Then a set of scenarios are explored using DM techniques in order to create new knowledge.

Data Mining Techniques

The DM stage has many different techniques that could be used depending on the problem, but those who better solved this problem were the Decision Tree, the Naive Bayes and the Support Vector Machine. The Decision Tree technique automatically generates rules, which are conditional statements that reveal the logic used to build the tree. The Naïve Bayes uses Bayes' Theorem, a formula that by counting the frequency of values and combinations of values in the historical data, calculates probability. The Support Vector Machine is a powerful algorithm based on linear and nonlinear regression (Paramasivam, Yee, Dhillon, & Sidhu et al., 2014).

Statistical Measures

Using these DM techniques and through the use of Confusion Matrix it is possible to obtain four types of results. The first situation could be a true positive (TP) result that corresponds to the number of positive examples correctly classified. The second situation could be a false positive (FP) result that corresponds to the number of positive examples classified as negative. We could also get a true negative (TN) result, that corresponds to the number of negative examples actually classified as negative and, finally, the false negative (FN), that corresponds to the number of negative examples classified as positive.

From this type of values resulting models, there are statistical metrics for assessing data quality that could be estimated, in particular, the sensitivity, specificity and acuity.

Sensitivity is the ability to correctly detect the occurrence of the procedure. It is the result of the ratio of true positive (TP) values on all the values corresponding to positive (TP + FN). On the other hand, specificity it is the ability to correctly identify in a model the non-occurrence of a procedure. It is measured by the ratio of correctly identified as negative values (TN) and all values corresponding to negative (TN + FP). Finally, acuity is the total percentage of agreement between the values detected correctly and the actual values. It is measured by the proportion of all the results measured correctly (TP + TN) from the models of all cases liable to be obtained (TP + TN + FP + FN) (Amaral, 2007).Table 1 displays the expressions that characterize each of the metrics described above.

Table 1. Expressions that define the sensitivity, specificity and acuity

	Positive Result	Negative Result	Sensitivity	Specificity	Acuity
Positive Value	TP	FP	$\frac{TP}{TP+FN}$	$\frac{TN}{TN+FP}$	$\frac{TP+TN}{TP+FP+FN+TN}$
Negative Value	FN	TN			

BUSINESS INTELLIGENCE IN HEALTHCARE ENVIRONMENT

The implementation of BI tools in healthcare organizations help the managers and the healthcare professionals in their decision making process because of the analysis of data that provide relevant information about the activities and processes that happen inside the organization and considering the complex and ever changing environment experienced in healthcare. Thus, these tools can improve the quality and safety of the health care delivery and the efficiency and financially performance of the organization (Bonney, 2013; Prevedello et al., 2008).

Many HIS, like the EHR, contain huge amounts of clinical information of high relevance to clinical decision making (Bonney, 2013).

Many studies suggest that the application of BI tools to the content of EHR is the only way to ensure that the clinical data are efficiently explored because these tools allow the extraction of relevant and quality information from the data. The extracted data can be used by healthcare professionals to support decision making in real-time, contributing, consequently, to improved results for the healthcare delivery (Bonney, 2013).

Moreover, the data analysis requests are today more frequent, sophisticated and based on different data that must be integrated from the different information systems implemented by the organization (Glaser & Stone, 2008). In most of the healthcare organizations, the data are stored in different systems and sometimes it is necessary to correlate data coming from them. Usually, these systems are poorly integrated, which makes the extraction of the data a difficult operation. This complexity implies that people without a solid knowledge in databases are not capable of accessing data present in these systems (Prevedello et al., 2010). For this reason, it is important and there is interest to develop applications that can facilitate the access to HIS data and the extraction of information from that data.

BI tools are capable of working with healthcare data in an efficient manner, to generate real-time information and knowledge and this is the reason why they are very attractive to the healthcare sector and that is why in the last years the interest in applying BI tools in the healthcare sector has grown and different solutions have been reported (Bonney, 2013).

The implementation of a BI solution in the healthcare environment improves the decision making process by allowing fact-based decision making and changes the culture of the organization contributing to a higher level of transparency in all the processes (Nagy, Warnock, Daly, Christopher, Christopher, & Reuben, 2009).

In the radiology field, several authors report the use of the BI technology to generate Key Performance Indicators (KPIs) capable of evaluating the performance of the radiology department (Nagy et al., 2009; Prevedello et al., 2010). These indicators can present relevant information useful to evaluate the efficiency and financial performance of the department, as well as monitoring the quality and safety of the care delivery and they allow a deep insight of the factors involved in a process or procedure. Decision makers apply the knowledge obtained with the analysis of these indicators in their decisions, and that knowledge can be used to adjust or modify the organization current behaviour and help it to achieve its goals, allowing a continuously performance improvement (Prevedello et al., 2010)

Open source tools were used to build a DW to analyse KPIs for a radiology department that allows the integration and visualization of data coming from all the department information systems in an only visualization (Prevedello et al., 2010). The method proposed in this paper is particularly useful when applied to the situations where data are continuously generated and the reports need to be created regularly and based on updated data.

It was also presented an automated system for the extraction, processing and visualization of KPIs used to identify problems and improvement opportunities related to the radiology department performance (Nagy et al., 2009). The data analysis is performed over the DW that contains data extracted from the different information systems and the indicators are generated using a Web based dashboard, which facilitates the analysis of results. The results obtained during 24 after the solution implementation suggest that it allowed obtain significant data to improve the efficiency of the department management (Nagy et al., 2009).

These concepts can be extended to other healthcare sectors to deal, for example, with data of nosocomial infection and to obtain indicators capable of helping with the nosocomial infection management in a healthcare organization.

Business Intelligence and Nosocomial Infections

Nosocomial infections have a large impact in patient’s mortality and morbidity, especially in intensive care units where the nosocomial infections rate is higher as the result from the invasive procedures executed in these units and the compromised immune system of the patients hospitalized in these units (Inweregbu, 2005; Rigor, Machado, Abelha, Neves, & Alberto, 2008). Moreover, a patient with a nosocomial infection remains longer times in the hospital and sometimes it has to be readmitted, resulting in additional costs for the healthcare institution (Inweregbu, 2005; Rigor et al., 2008). So, the nosocomial infection management is essential for healthcare organizations.

The nosocomial infection management can be done by preventing and controlling the infection with the application of specific infection control measures. Thus, by defining a set of parameters that can help summarizing important facts and causes, associated with the presence of infection and present in data, one can identify risk processes and activities inside the healthcare unit. With that information, specific prevention and control measures can be created and implemented in order to diminish the nosocomial infection rate. BI technology concepts and tools can be used to generate and present these parameters, by treating and analyzing data in an efficient way. So, the information presented with the BI tool can be used to help in the decision making and in the identification of problems present in the healthcare environment that can be harmful to the patient's well-being.

Through the application of BI technology concepts to nosocomial infection data, the work presented in this chapter pretends to validate and evaluate the applicability of BI technology to healthcare environments and, at the same time, evaluate and analyse the utilization of BI tools as DSS for the healthcare environment.

CASE STUDY: NOSOCOMIAL INFECTION INCIDENCE INDICATORS IN CHP

The evaluation of the applicability of BI tools in healthcare environments was performed through the creation of a DW, lately explored and analysed by BI tools.

In this study, an analysis of the nosocomial infections incidence in 2013 in CHP was performed.

From the 391 nosocomial infection reports considered in the study, only 33 were associated with the presence of a nosocomial infection, which means that the nosocomial infection percentage in that period is circa 8,49%.

To begin this study, the parameters with interest to the analysis were identified and then considered in the development of the DW that is used for analytic purposes.

The following indicators were used as the parameters to study:

- The total capacity (number of beds available in the specialty), the average number of hospitalization days, the number of discharges, the number of nosocomial infections reports registered and the percentage of registries made by the physicians, per specialty in 2013. All of these parameters help in the characterization of the population in study.
- Total amount records of invasive devices and amount of records of invasive devices in the presence of nosocomial infection, per specialty in 2013 and per invasive device. This set of indicators help in the characterization of the relationship between the usage of invasive devices, such as catheters and intubation, and the presence of nosocomial infections.
- Total amount records of intrinsic risk factors and amount of records of intrinsic risk factors in the presence of nosocomial infection, per specialty in 2013 and per intrinsic risk factor. This set of indicators help in the characterization of the relationship between the presence of intrinsic risk factors, such as diabetes, malnutrition and usage of corticoids, and the occurrence of nosocomial infections.

The CHP data referring to nosocomial infection registries were extracted from the hospital database and manipulated and cleaned in order to build the DW with the relevant data for the study.

In this case study, the approach used to build de DW was the Ralph Kimball's paradigm.

So, the DW is constituted by two data marts, one related with the characterization of the population and the other related to the relationship between risk factors and the presence of nosocomial infection.

To create the data mart for the analysis of the indicators related with the population in study a dimensional model following the star schema (Figure 1) was developed. This dimensional model contains all the information necessary to perform the analysis of the indicators.

Figure 1. Dimensional model of the data mart developed to represent the population in study.

This schema is the dimensional model developed for the population data mart and it is composed by one fact table (“Population”) and three dimensional tables (“Date”,” Specialty”, “Patient”).

In the dimensional model creation it is essential to choose the right dimensions and facts that will represent the desired analysis indicators. These facts and dimensions are stored in the DW, which is used to support querying.

The fact table contains the measures that characterize the indicators and the keys of the dimensional tables. Each dimensional table contains the attributes that can be used to explore the indicators and the primary key referenced in the fact table.

The dimension “Patient” contains information about the patients, the dimension “Specialty” contains information about the different specialties in study and the dimension “Date” contains information about the date. The data used in the case study developed in this chapter refers to the nosocomial infection reports registered in CHP in the year of 2013. Nevertheless, the system can be expanded to deal with data from other years and the “Date” dimension allows the analysis of data per year, even when the DW is composed of data respecting different years.

To create the data mart for the analysis of the relationship between the occurrence of risk factors and the presence of nosocomial infections a dimensional model following the snowflake schema (Figure 2) was developed. This dimensional model contains all the information necessary to perform the analysis of the indicators.

Figure 2. Dimensional model of the data mart developed to represent the risk factors present and their relationship with the presence of nosocomial infection.

This schema is the dimensional model developed for the risk factors data mart and it is composed by one fact table (“Facts”) and five dimensional tables (“Date”, “Specialty”, “Patient”, “Conditions” and “Groups”). The dimension tables “Date”, “Specialty”, “Patient” are the same used by the data mart of the population in study. The “Conditions” table contains all the possible conditions present in data and it includes for example the pairs “Presence of nosocomial infection – Yes”, “Presence of nosocomial infection – No”, “Presence of urinary catheter – Yes” and “Presence of urinary catheter – No”. The dimension table “Groups” contains the description of the type of the condition and thus allows to relate all the conditions with their subject, for example it relates all the conditions related with nosocomial infection with the description “Nosocomial infection” and all the conditions related with urinary catheterization with the description “Urinary catheterization”. The fact table “Facts” contains all the keys related with the dimension tables present in the data mart.

After creating the dimensional model for the DW, and considering the indicators to study, ETL was executed in order to clean the data and to load it to the DW. ETL is a crucial step for the efficient loading of huge volumes of data and to ensure quality data and, consequently, good results in the indicators presented by the BI tools (Chaudhuri et al., 2011). In this work the ETL process was performed with the application of PL/SQL procedures. The Oracle database 11g was used as the repository for the input data used by the ETL process and for the data transformed by the ETL techniques.

The DW tables were created in the database and populated with PL/SQL procedures that extract data from the sources already treated with the ETL process. After that, a BI tool was used to perform data analysis and present the results.

The resulting database follows the dimensional models previously defined and it contains the fields and the measures that define the relevant indicators to analyse. Thus, this database can be used by other applications, for example, Jaspersoft Studio, in order to create reports, to analyse data and to aggregate data using, for example, OLAP tools.

For the data analysis and the indicators presentation and after evaluate a set of tools it was chosen the Jaspersoft Studio Community version. Jaspersoft Studio was used to query the DW and show the results of the queries in the form of graphics and tables.

Jaspersoft Studio is free and has an open source report designer for JasperReports and JasperReports Server that allow the creation of sophisticated layouts containing, for example, charts, tables, and access data from any data source. The reports can be published in a wide range of formats. It also allows the publication of results in a web platform when connected to JasperReports Server (Jaspersoft Corporation, 2014).

Jaspersoft was the Business Intelligence chosen tool because, Jaspersoft allows to make faster decisions by bringing them timely, actionable data inside their apps and business processes through an embeddable reporting and analytics platform.

Jaspersoft enables us to easily self-serve and get the answers that we need inside our preferred app or on our favourite devices. Unlike desktop visualization tools and scales architecturally, his platform, economically to reach everyone, it is an open source tool, well documented and simple to use (Jaspersoft Corporation, 2014). Moreover, it allows the connection to Oracle databases and the publication of results in several formats, including pdf, the format chosen in this study.

Results and Discussion

Exploring the features of Jaspersoft Studio, the results presented in the Figures 3, 4 and 5 were obtained.

Figure 3. Example of results obtained with Jaspersoft Studio: characterization of the population in study per specialty in 2013.

The table in the Figure 3 is related to the set of indicators that characterize the population in study.

Figure 4. Example of results obtained with Jaspersoft Studio: nosocomial infections relationship with extrinsic risk factors per specialty in 2013.

The table of the Figure 4 respects to the indicators that explore the relationship between the presence of nosocomial infections and the usage of invasive devices per specialty in 2013 and per type of invasive device.

Figure 5. Example of results obtained with Jaspersoft Studio: nosocomial infections relationship with intrinsic risk factors per specialty in 2013.

The table of the Figure 5 respects to the indicators that explore the relationship between the occurrence of nosocomial infections and the presence of intrinsic risk factors, per specialty in 2013 and per type of intrinsic risk factor.

The results obtained with this approach allow to identify important characteristics of nosocomial infection incidence in CHP and of the population in study. For example, by analyzing the results of the Figure 3, “Medicine (Type B)” is the specialty with the highest average number of hospitalization days and “Medicine (Type A)” is the specialty with the highest amount of nosocomial infection filled reports and also the specialty with the higher capacity. Figure 3 also shows that all clinical specialties under analysis have a percentage of filled reports of 100%, which means that physicians use the system to record the nosocomial infection reports.

According to Figure 4, “Medicine (Type A) is the specialty with the highest number of nosocomial infections in the presence of catheters or intubation. However, it is important to consider that “Medicine (Type A)” is the specialty with more records of catheterization and intubation usage and that the difference from the other specialties is relevant, which means that naturally it has a higher possibility of having nosocomial infection occurrences.

Figure 5 shows the number of registered occurrences of several intrinsic risk factors in the absence and presence of nosocomial infections per specialty in 2013. For example, the specialty “Medicine (Type A)” has the highest amount of records of diabetes, but the number of patients with diabetes and nosocomial infection at the same time for this specialty is much lower. The same situation occurs with “Medicine (Type C)”, although the number of records of patients with diabetes is much higher for “Medicine (Type A)”.

The analysis of these results helps healthcare professionals in the identification of critical processes and clinical specialties where specific infection control measures are essential for the patient's safety and well-being, helping them in their decisions, defining primary areas where infection control measures should be implemented faster and planning specific measures to diminish the rate of nosocomial infection incidence.

The methodology described in this paper can be helpful to investigate the application of BI concepts in healthcare. BI tools help dealing with the complex, heterogeneous and distributed environment experienced in healthcare by identifying critical activities where measures can be applied and helping the healthcare professionals in their decision making process.

The utilization of ETL techniques allows the integration and the transformation of data in order to create DWs for specific analytic purposes. Using the process indicated in this paper, the healthcare data can be processed in order to obtain the desired indicators.

The dimensional model used allows drill-down on the data because the "Date" dimension has attributes such as day, month and year that can be used to aggregate data in different ways and according different hierarchic levels (e.g. day, month, year, etc.). The "Specialty" dimension allows the data arrangement by specialty code or by specialty description. Moreover, the dimension "Patient" allows seeing the data by patient code and it also allows to see information related to the patient. The utilization of a dimensional model facilitates the querying process and allows drill-down on specific attributes. Furthermore, graphs and reports can be created presenting data aggregated by the desired attributes.

Through the application of BI tools, the database is not only used for storage purposes, but also to be used for knowledge extraction. The data processing is a crucial step to obtain good results and the presentation of these in a graphical way makes easier the results interpretation. So BI tools can be very useful to manipulate and analyse healthcare data and with the results obtained, the BI tool can be used as a DSS for healthcare professionals that depend on these data to perform their job.

DATA MINING FOR NOSOCOMIAL INFECTION PREDICTION

DM models can be applied to make predictions in a real environment using real healthcare data. In this work it was analyzed the influence that some risk factors have in the occurrence of a nosocomial infection. These clinical predictions are important for the healthcare professionals to understand better the incidence of nosocomial infection. In this case the DM module was included in the BI Platform.

Business Understanding

The main DM aim of this work was to study the influence of certain factors in the occurrence of a nosocomial infection in CHP. To solve this problem, a model which predicts if a patient will have a nosocomial infection when certain risk factors are presented was built considering data from past nosocomial infection records. In order to apply DM techniques to build this model, the problem to solve was formulated as "*How likely are the extrinsic or the intrinsic risk factors, such as catheters and intubation, responsible for the occurrence of a nosocomial infection in CHP?"*

The data that are likely to contain a relationship between the variables and the occurrence of nosocomial was chosen, analyzed and pre-processed before inducing the model.

Data Understanding

The data necessary to the analysis were extracted from the CHP database, and the quality of the possible variables to be used in the DM process was analyzed. The data in study respects to 391 nosocomial infections registries recorded in CHP in 2013.

Not all the attributes registered in the nosocomial infection record were relevant for the analysis, thus a selection of the most significant attributes for the analysis was performed.

The variables considered in the models construction were:

- *Nosocomial Infection (NI):* target variable that dictates the outcome of the diagnosis process. The result of this variable can be 'Yes' or 'No';
- *Age (A), Sex (S), Clinical Specialty (CS), Days of Hospitalization (DH):* specific attributes that characterize the patient registry in the database;
- *Risk Factors (RF):* variable that represents the presence or the absence of intrinsic risk factors.
- *Intubation (I):* variable that represents the presence or the absence of invasive devices for intubation during the patient's hospitalization;
- *Catheters (C):* variable that represents the presence or the absence of invasive devices for catheterization during the patient's hospitalization;

All of these variables chosen is modelling the problem in study allowing the prediction of the target variable.

Data Preparation

After selecting the data and the variables to use, a pre-processing task was performed. The pre-processing task eliminated all null values from data, leaving only 283 records at the end. During this stage, classes to aggregate data were formed in order to model the problem in a more accurately way. The following classes were created:

- *Age Class (AC)*: aggregation of the age in intervals of ages;
- *Intubation (I)*: aggregation of all the invasive devices related with intubation in a single class;
- *Catheters (C)*: aggregation of all the invasive devices related with catheterization in a single class.

This pre-processing step allowed the construction of a dataset with all the cases of interest, in which DM techniques were applied. Oversampling was also applied to data in order to replicate the data with nosocomial infection to get a number of records with nosocomial infection approximate to the number of records without nosocomial infection.

Modelling

Considering the variables previously presented and their possible combinations, several scenarios were taken in account for achieving the desired models for DM. To model the data, classification models were used, and three different classification techniques were applied to the models: Support Vector Machine, Decision Tree and Naive Bayes.The selection of these DM techniques was based in the interpretability of the models and engine efficiency.

Thus, four scenarios were modeled by the different classification DM techniques, one target variable and three different combinations of variables were used, which means that, at the end, 36 models were obtained. The DM techniques were applied to the dataset of interest in order to obtain the best model to model the problem to solve.

The scenarios considered to build the models were:

- *Without Risk Factors (Scenario 1)*: all data, except the variable *RF*, were used in the model;
- *Without Intubation (Scenario 2)*: all data, except the variable *I*, were used in the model;
- *Without Catheters (Scenario 3)*: all data, except the variable *C*, were used in the model;
- *All data (Scenario 4)*: all data present in the data repository were used in the model.

These scenarios were modeled for three different approaches: models with the original data (*Approach A*), models with a dataset with replicated data (*Approach B*) and models with a dataset with replicated data and the age aggregated in classes (*Approach C*).

The developed models can be represented by the following expression, which means that the model (M_n) belongs to the technique (A) classification and is composed by a scenario (S) and a DM technique (TDM):

$$M_n = A_f + S_i + TDM_y$$

where for *Approach A*

$$A_f = \{Classification\}$$
$$TDM_y = \{A, S, CS, DH, RF, I, C\}$$
$$S_i = \{Scenario\ A1 \dots Scenario\ A4\},$$

for *Approach B*

$$A_f = \{Classification\}$$
$$TDM_y = \{A, S, CS, DH, RF, I, C\}$$
$$S_i = \{Scenario\ B1 \dots Scenario\ B4\}$$

and for *Approach C*

$$A_f = \{Classification\}$$
$$TDM_y = \{AC, S, CS, DH, RF, I, C\}$$
$$S_i = \{Scenario\ C1 \dots Scenario\ C4\}.$$

Evaluation

All of these models were evaluated with statistical metrics that assess the results achieved by the DM models. The metrics used for the evaluation were the specificity, the sensitivity and the acuity. Some of these models allowed achieving the best four overall results for each of the DM techniques used (Table 2).

Table 2. Top 4 models for each DM technique

	Support Vector Machine		
	Specificity	**Sensitivity**	**Acuity**
Scenario B1	0.763	0.919	0.838
Scenario B2	0.741	0.942	0.832
Scenario C1	0.731	0.793	0.766
Scenario C3	0.675	0.845	0.754
	Decision Tree		
	Specificity	**Sensitivity**	**Acuity**
Scenario C1	0.855	0.798	0.818
Scenario C4	0.673	0.982	0.786
Scenario B4	0.673	0.982	0.786
Scenario B2	0.67	1	0.786
	Naive Bayes		
	Specificity	**Sensitivity**	**Acuity**
Scenario B1	0.733	0.941	0.825
Scenario B2	0.733	0.941	0.825
Scenario B4	0.733	0.941	0.825
Scenario C4	0.733	0.941	0.825

Deployment

After the model evaluation, the knowledge obtained can be used by healthcare professionals to predict when a nosocomial infection will occur in the presence of certain risk factors. The models and results obtained with DM were added to the BI platform.

Discussion

Using the DM techniques chosen it was possible to obtain acceptable results for each model, because the ideal behavior for classification models would be that the predictions yielded sensitivity values in general upper than 90% and sensitivity values upper 91.90% were obtained. It is important to note that the models with a high percentage of sensitivity are capable of correctly detect the occurrence of a nosocomial infection.

The best models were selected based on the value of the sensitivity, since it is important to know all the cases of nosocomial infection even though some false positives may occur. Thus, it is preferable to think that a nosocomial infection may occur when that is not true, than not considering the possibility of nosocomial infection and put the patient's safety and well-being in risk.

The best value of sensitivity was 100% and it was obtained for *Scenario B2* (*Scenario 2* of *Approach B)* with the Decision Tree technique, which means that the age data where aggregated in classes. In this case, a value of 78.60% of acuity were achieved.Thus, it can be concluded that for this case the model can predict the presence of infection with 100% of certainty. Thus, it can be concluded that of all combinations of variables and techniques, this is the best to predict the *NI* variable.

The best value of specificity was 85.50% and it was obtained for *Scenario C1* (*Scenario 1* of *Approach C)* with the Decision Tree technique, which means that the age data where aggregated in classes. In this case, a value of 81.80% of acuity was achieved.

Table 3 shows the results (accuracy) of the *Scenario C1*, identifying the number of cases in that the model hit and the number of instances where the model failed for each of the algorithms used.

Table 3. Number of correct and incorrect cases for Scenario C1, for each of the algorithms used

	Incorrect	Correct	% of Correct
Support Vector Machine	36	118	76.62
Decision Tree	28	126	81.82
Naive Bayes	30	124	80.52

According to these results, it appears that the most efficient algorithm applied to the *Scenario C1* was the Decision Tree algorithm, since it is the algorithm that has highest percentage of correct answers (81.82%).

In general the specificity values are acceptable, being the lowest value 67.00%, which means that the models are specific and, thus, they are capable of detect correctly the non-occurrence of a nosocomial infection.

The smallest acuity value was 75.40% and the highest was 83.80%, which means that there is s good percentage of agreement between the values correctly detected and the real value.

The best combination of scenario and approach for multiple DM techniques was *Scenario B2* (*Scenario 2* and *Approach B)* because this combination has the higher value of sensitivity for all the DM techniques used. These models have also high values of acuity.

FUTURE RESEARCH DIRECTIONS

As future work, it would be interesting to test Jaspersoft tool for other indicators to perform a deeper exploration of the BI tool and to test them too. It would also be interesting to test other BI tools and compare them with Jaspersoft.

To explore this case study, it would also be interesting to test the tool used in this paper with another DW with a different dimensional model.

As future work, it would also be interesting to consider different DM techniques, incorporate other variables in the predictive models and repeating the DM experiments with other data.

CONCLUSION

The work presented in this chapter allows studying the applicability of BI concepts in healthcare and the evaluation of BI tools as DSS for healthcare. With the application of BI technology, the database is not only used for storage purposes, but also for decision support. Through the extraction and analysis of the data collected from the different information systems present in the organization and supporting the decision making process with the knowledge extracted, BI tools help overcoming the problems associated with the complexity, heterogeneity and the distribution present in the healthcare environment.

Healthcare organizations act under constant financial pressure as well as competitive pressure and healthcare professionals need to improve productivity, efficiency and, at the same time, the quality of the care. BI concepts can be extended to the healthcare environment and using them it is possible to capture indicators used to characterize and evaluate the healthcare unit and use them to monitor activities and processes within the healthcare environment.

Moreover, nowadays, the implementation of BI tools in healthcare is essential for healthcare organizations to treat and to analyse their data. The utilization of BI tools like Jaspersoft Studio enables improvements in healthcare efficiency, quality, safety and financial performance because they can be used to present relevant indicators that help in the identification of important parameters and in the analysis of data about the activities and processes performed inside the healthcare organization. The knowledge obtained with the analysis of these indicators can then be applied by healthcare professionals in their decision making, helping them perform their jobs.

The construction of a DW for analytical purposes must be carefully projected in order to properly represent the information needed to generate the indicators to analyse. The ETL process is also very important to this process because it ensures the quality and consistency of the data stored in the DW.

Although the implementation of BI tools is not an easy process, it is essential for healthcare institutions because it allows the deep analysis of healthcare data, resulting in relevant and strategic results about the data to study.

Open source tools, like Jaspersoft Studio, can be a good help to explore data because they allow the generation of new knowledge in real-time and without additional costs for the organization.

Finally, the methodology proposed developed to generate indicators for nosocomial infection data analysis can be extended to other healthcare data and to obtain other indicators for nosocomial infection, even with data from other healthcare institution.

The DM study demonstrated that it is possible to obtain classification DM models to predict if a patient with nosocomial infection risk factors will have or not a nosocomial infection, by using real healthcare data from previous nosocomial infection records.

Three distinct techniques were used to perform classification tasks: Decision Tree, Naïve Bayes and Support Vector Machine. It can be concluded that using the classification techniques and the past nosocomial infection data of patients it is possible to predict the future the nosocomial infection incidence.

A BI system composed by the presentation of a set of indicators and a DM part capable of predict the occurrence of infection can provide very useful information for healthcare professionals.

The work presented in this chapter is of good worth for the society because the system presented is capable of helping in the prevention of nosocomial infections in healthcare units diminishing, thus, the risk of complications for the patients and improving their well-being and safety.

ACKNOWLEDGEMENTS

This work is funded by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project PEst-OE/EEI/UI0752/2014.

REFERENCES

1Keydata. (2014). Bill Inmon vs. Ralph Kimball. *1keydata.com*. Retrieved from http://www.1keydata.com/datawarehousing/inmon-kimball.html

Amaral, J. J. F. (2007). Avaliação de Artigos Científicos. *Bases da Epidemiologia Clínica.* Faculdade de Medicina da Universidade Federal do Ceará, Fortaleza, Brasil.

Azevedo, A. (2011). *Data mining languages for business intelligence* (Doctoral dissertation). Universidade do Minho, Portugal.

Bonney, W. (2013). Applicability of Business Intelligence in Electronic Health Record. *Procedia - Social and Behavioral Sciences*, *73*, 257–262.

Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. *Communications of the ACM*, *54*(8), 88.

Cardoso, L., Marins, F., Portela, F., Santos, M., Abelha, A. & Machado, J. (2014). The Next Generation of Interoperability Agents in Healthcare. *Int. J. Environ. Res. Public Health 11*(5), 5349-5371.

Duarte, J., Portela, F., Abelha, A., Machado, J. & Santos, M. (2011). Electronic Health Record in Dermatology Service. *ENTERprise Information Systems. Communications in Computer and Information Science* (pp. 156–164). Springer Berlin Heidelberg.

El-Sappagh, S. H. A., Hendawi, A. M. A., & El Bastawissy, A. H. (2011). A proposed model for data warehouse ETL processes. *Journal of King Saud University - Computer and Information Sciences*, *23*(2), 91–104.

Foshay, N., & Kuziemsky, C. (2014). Towards an implementation framework for business intelligence in healthcare. *International Journal of Information Management*, *34*(1), 20–27.

Glaser, J., & Stone, J. (2008). Effective use of business intelligence. *Healthcare Financial Management: Journal of the Healthcare Financial Management Association*, *62*(2), 68–72.

Hasman, A. (1998). Education an health informatics. *International Journal of Medical Informatics*, *52*, 209–216.

Heinrichs, J. H., & Lim, J.-S. (2003). Integrating web-based data mining tools with business models for knowledge management. *Decision Support Systems*, *35*(1), 103–112.

Hema, R., & Malik, N. (2010). Data Mining and Business Intelligence. In *Proceedings of the 4th National Conference; INDIACom-2010*. New Delhi, India.

Inweregbu, K. (2005). Nosocomial infections. *Continuing Education in Anaesthesia, Critical Care & Pain*, *5*(1), 14–17.

Jaspersoft Corporation. (2014). *Jaspersoft Studio | Jaspersoft Community*. Retrieved from http://community.jaspersoft.com/project/jaspersoft-studio

Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling.* (2nd ed.). New York, NY: John Wiley & Sons Inc.

Lenz, R., & Reichert, M. (2007). IT support for healthcare processes – premises, challenges, perspectives. *Data & Knowledge Engineering, 61*(1), 39–58.

Loshin, D. (2012). *Business Intelligence: The Savvy Manager's Guide*. (2nd ed.). San Francisco, CA: Morgan Kaufman Publishers Inc.

Machado, J., Alves, V., Abelha, A. & Neves, J., (2007). Ambient Intelligence via Multiagent Systems. *Medical arena; International Journal of Engineering Intelligent Systems*, *15*(3), 167-173.

Mettler, T., & Vimarlund, V. (2009). Understanding Business Intelligence in the Context of Health Care. *Health Informatics Journal*, *15*(3), 254–264.

Ministério da Saúde. (2014). SPMS. Retrieved from http://spms.min-saude.pt/

Nagy, P. G., Warnock, M. J., Daly, M., Christopher, B. S., Christopher, T., & Reuben, D. M. (2009). *Informatics in Radiology Automated Web-based Graphical Dashboard for Radiology Operational*. *RadioGraphics*, *29*(7), 1897–1907.

Paramasivam, V., Yee, T. S., Dhillon, S. K., & Sidhu, A. S. (2014). A methodological review of data mining techniques in predictive medicine: An application in hemodynamic prediction for abdominal aortic aneurysm disease. *Biocybernetics and Biomedical Engineering,* 1–7.

Pardillo, J., Mazón, J.-N., & Trujillo, J. (2010). Extending OCL for OLAP querying on conceptual multidimensional models of data warehouses. *Information Sciences*, *180*(5), 584–601.

Park, T., & Kim, H. (2013). A data warehouse-based decision support system for sewer infrastructure management. *Automation in Construction*, *30*, 37–49.

PhridviRaj, M. S. B., & GuruRao, C. V. (2014). Data Mining – Past, Present and Future – A Typical Survey on Data Streams. *Procedia Technology, 12*, 255–263.

Popovič, A., Hackney, R., Coelho, P. S., & Jaklič, J. (2012). Towards business intelligence systems success: Effects of maturity and culture on analytical decision making. *Decision Support Systems, 54*(1), 729–739.

Prevedello, L. M., Andriole, K. P., Hanson, R., Kelly, P., & Khorasani, R. (2010). Business intelligence tools for radiology: creating a prototype model using open-source tools. *Journal of Digital Imaging, 23*(2), 133–41.

Prevedello, L. M., Andriole, K. P., & Khorasani, R. (2008). Business intelligence tools and performance improvement in your practice. *Journal of the American College of Radiology*, *5*(12), 1210–1211.

Rebuge, Á., & Ferreira, D. R. (2012). Business process analysis in healthcare environments: A methodology based on process mining. *Information Systems*, *37*(2), 99–116.

Rigor H., Machado J., Abelha A., Neves J., & Alberto C. (2008). A Web-Based System to Reduce the Nosocomial Infection Impact in Healthcare Units. In *Proceedings of the WEBIST 2008- International Conference on Web Information Systems*. Madeira, Portugal.

Santos, M. Y., & Ramos, I. (2006). *Business Intelligence : tecnologias da informação na gestão de conhecimento.* Lisboa, Portugal: FCA - Editora de Informática.

Salazar, M., Duarte, J., Pereira, R., Portela, F., Santos, M., Abelha, A., & Machado, J., (2013). Step towards Paper Free Hospital through Electronic Health Record, in *Advances in Information Systems and Technologies, Advances in Intelligent Systems and Computing* (pp. 685-694). Springer Berlin Heidelberg.

Simões, N., Gomes, B., & Paiva, P. (2009). *Análise da Viabilidade Económica das Aplicações SAM e SAPE*. Universidade Nova de Lisboa, Lisboa, Portugal.

Soler, E., Trujillo, J., Fernández-Medina, E., & Piattini, M. (2008). Building a secure star schema in data warehouses by an extension of the relational package from CWM. *Computer Standards & Interfaces*, *30*(6), 341–350.

Spruit, M., Vroon, R., & Batenburg, R. (2014). Towards healthcare business intelligence in long-term care. *Computers in Human Behavior, 30,* 698–707.

Thalhammer, T., Schre, M., & Mohania, M. (2001). Active data warehouses : complementing OLAP with analysis rules. *Data & Knowledge Engineering*, *39*(3), 241–269.

Winter, A., Haux, R., Ammenwerth, E., Brigl, B., Hellrung, N., & Jahn, F. (2011). Strategic Information Management in Hospitals. *An Introduction to Hospital Information Systems*. London, UK: Springer London.

ADDITIONAL READING SECTION

Chaudhuri, S, Dayal, U. (1997). An Overview of Data Warehousing and OLAP Technology. *SIGMOD Rec.*, *26*(1), 65–74.

Coiera, E., Westbrook, J., & Wyatt, J. (2006). The safety and quality of decision support systems. *Yearbook of Medical Informatics*, 20–5.

Fayyad, U., Piatetsky-shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases, *17*(3), 37–54.

Ferreira, J., Miranda, M., Abelha, A., & Machado, J. (2010). O Processo ETL em Sistemas Data Warehouse. *INForum 2010 - II Simpósio de Informática*, 757–765.

Inmon, W. H. (2002). *Building the Data Warehouse* (3rd ed.). New York, NY: John Wiley & Sons, Inc.

Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. (1998). *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses* (1st ed.). New York, NY: John Wiley & Sons, Inc.

Koh, H. C., & Tan, G. (2005). Data mining applications in healthcare. *Journal of Healthcare Information Management : JHIM*, *19*(2), 64–72.

KEY TERMS & DEFINITIONS

Business Intelligence: Set of technologies capable of treating and analyzing data in order to present relevant and strategic information, helpful in the decision making process of an organization.

Clinical Decision Support System: Software system designed to assist healthcare professionals in their decision making process.

Data Mining: Process of discovering interesting and relevant patterns in large datasets.

Data Warehouse: Component of a BI system that stores and consolidates data into a valid and consistent format and allows the analysis and exploration of that data using other tools.

Electronic Health Record: HIS that covers the different services and units of a healthcare institution and consists of a set of standardized documents for the registration of the medical procedures provided to a particular patient.

ETL: Process that transforms and converts data coming from several sources to a unified format, fitted to the DW schema where these data will be stored.

Health Information System: System responsible for handling data processing, information and knowledge in healthcare environments.

Nosocomial Infection: Infection acquired in the hospital environment in the period of 48 hours after hospital admission, 3 days after discharge or 30 days after an operation.