A Clinical Data Warehouse Architecture based on the Electronic Healthcare Record Infrastructure

Fabrizio Pecoraro[1], Daniela Luzi[1] and Fabrizio L. Ricci[2]
[1]Institute for Research on Population and Social Policies, National Research Council, Via Palestro 32, Rome, Italy
[2]Institute for Systems Analysis and Computer Science, National Research Council, Viale Manzoni 30, Rome, Italy

Keywords: Electronic Healthcare Record (EHR), Secondary Uses, Clinical Indicators, Data Warehouse, Dimensional Model.

Abstract: The development of clinical data warehouses is becoming increasingly important in the healthcare domain to support organizations in the improvement of decision-making, business processes as well as the communication between clinicians, patients and the administration. However, data and process integration is a big challenge considering the heterogeneous and distributed nature of healthcare information systems. This paper proposes a data warehouse architecture based on the Italian Electronic Health Record (EHR) technological infrastructure. It describes the main advantages in the application of EHR systems for secondary purposes and reports the data warehouse design framework outlining its architecture as well as a dimensional model based on a dashboard defined to manage the intervention of patients with diabetes. The adoption of EHR systems enhances interoperability given that these systems share standardized clinical data among multiple parties involved in different healthcare settings.

1 INTRODUCTION

The widespread diffusion of Information Technologies in different healthcare settings has led to a production of a massive amount of both clinical and administrative data. Despite all the efforts, these information are often stored in standalone heterogeneous information systems developed for specific specialties (radiology, admissions, general ledger, scheduling, pharmacy and patient records) that do not interchange data with each other (Wickramasinghe and Schaffer, 2006). Therefore, researchers doing data analysis still face interoperability and technical challenges in the support of administrative and clinical processes for purposes different than those they were gathered for (e.g. management information, quality assessment and research) (Kush et al., 2008, Taylor, 2008). In addition, data analysis can be a complex task considering that medical and sensitive information (Koh and Tan, 2011) are usually restricted in access due to ethical, legal and privacy issues. This makes it also necessary to collect and integrate these information before data analysis can be performed and to adopt specific techniques for data anonymisation and processing.

An important initial step toward the integration of data provided by heterogeneous multiple systems is the development of enterprise clinical data warehouses that are becoming increasingly important in the healthcare domain (Botsis et al., 2010, Kamal et al., 2005). This approach is adopted by different parties (e.g. hospital, GP, specialists) (Zhou et al., 2010, Sahama and Croll, 2007) and at different organizational level (i.e. local, regional and national authorities) (De Mul et at., 2012, Stow et al., 2006) in order to improve decision-making, business processes as well as communication between clinicians, patients and the administration. However, the implementation of data warehouses implies the solution of issues related to missing, corrupted, inconsistent or non-standardized data collected in different formats and data sources. In particular, the lack of a standard vocabularies is a serious barrier for the integration and analysis of data (Gillespie, 2000).

Electronic Health Record (EHR) systems represent a core source of information, managed and processed by multiple parties involved in healthcare settings. Their main purposes are to support physicians and other professionals in the delivery of care management services giving direct benefits to

Pecoraro F., Luzi D. and L. Ricci F..
A Clinical Data Warehouse Architecture based on the Electronic Healthcare Record Infrastructure.
DOI: 10.5220/0004764502870294
In *Proceedings of the International Conference on Health Informatics* (HEALTHINF-2014), pages 287-294
ISBN: 978-989-758-010-9

citizens in terms of speed and appropriateness of healthcare service delivery (primary uses). However, different recent experiences have recognized the successful use of EHRs for secondary purposes (Safran et al., 2007), such as clinical research (Hussain et al., 2012), epidemiological studies (Diomidous et al., 2009), ambulatory clinical care (Jensen et al., 2012), pharmacovigilance (Wang et al., 2009), comorbidity detection (Roque et al., 2011) or to alert providers of potential clinical risks (Lurio et al., 2010). However, these secondary uses of EHR are generally limited to a single institution (e.g. Hospital) or a single provider (e.g. General Practitioners) and/or on a specific target population (e.g. diabetics, investigational patients). Conversely, in our approach EHR is considered as a large-scale information infrastructure that integrates heterogeneous information systems managed by different organizations in a distributed environment. In our vision this EHR infrastructure can be the basis to develop a data warehouse as a business intelligence tool in a clinical governance framework. In this perspective EHR systems can be used to evaluate the effectiveness and appropriateness of healthcare services from the structural, organizational, financial and professional points of view. This improves the transparency of economic and clinical activities as well as the availability of real time information to decision makers (Mettler and Rohner, 2009). Moreover, the use of EHRs increases healthcare data quality and facilitates the interoperability between different systems and organizations. In particular, standardized data model as well as international widespread vocabularies and nomenclatures ensure a reliable data collection and a consistent data comparison even when collected from heterogeneous information systems. This approach can also provide additional values in different healthcare-related sectors including education, clinical research, public health, security and policy support (Committee on Data Standards for Patient Safety, 2003).

This paper proposes a clinical data warehouse architecture based on the EHR infrastructure developed in Italy. To demonstrate the feasibility of the data warehouse architecture a set of indicators related to diabetes pathology is proposed in particular to implement a chronic disease management intervention. Afterwards, starting from these indicators a business process is presented highlighting the design methodology of a dimensional model.

2 DATA WAREHOUSE DESIGN

2.1 EHR Infrastructure and Conceptual Model

In Italy the development of a local EHR system is delegated to each regional administration. To ensure the interoperability between the different local solutions the Department for the digitization of Public Administration and Innovation Technology in collaboration with the Department of Information and Communication Technologies of the National Research Council (CNR) have carried out the InFSE project (EHR technological infrastructure) (Ciampi et al., 2012) that defines a set of technological requirements with the aim of developing an interoperable EHR national infrastructure. InFSE provides a set of infrastructural components that notify clinical events to the involved local EHR systems through the adoption of a publish-subscribe pattern. Moreover, it archives clinical documents as generated by authorized users during a clinical event guaranteeing their persistency, security and reliability. To ensure semantic interoperability among local systems clinical documents stored in the relevant repository are structured using HL7 CDA (Clinical Document Architecture) Release 2 standard (Dolin et al., 2006). This standard allows to structure the content of both header and body of a document using the XML standard based on HL7 Reference Information Model (RIM) (Schadow et al., 2006), coupled with terminology. Moreover, to facilitate documents retrieval and localization InFSE components manage a set of descriptive metadata, such as document type, patient ID, document author, organization responsible for the document security, date of creation and update of the document, location of the document (URI), etc. However, to simplify the relationship between documents produced in different clinical events as well as to facilitate their sharing between different actors it is necessary to introduce a higher level of document aggregation and classification schema, providing a set of concepts that represents both content and context of healthcare services. These concepts were defined in the CONTsys standard (EN 13940, 2007) to describe different aspects of clinical and organizational processes such as health issue, contact and episode of care. They enable the information management of the healthcare delivery process to an individual subject of care as well as its continuity, taking into consideration data handling, decision processes, quality control and resource management. These concepts and their relationships have been the basis of the EHR conceptual model as shown in Figure 1 mapped in the HL7 RIM.

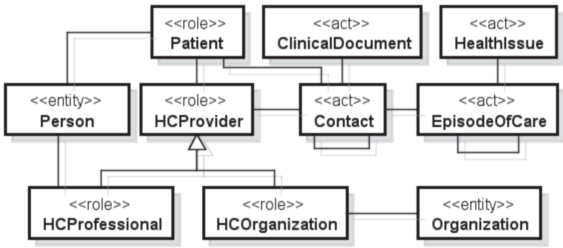

Figure 1: Portion of the EHR domain model based on the CONTsys concepts.

The RIM is not just the starting point to represent a clinical document backbone but it is also used as a standard for defining the structure of a message to be exchanged between heterogeneous information systems to achieve semantic interoperability (Oemig and Blobel, 2005). The main class of this model is the *Contact* that describes an encounter between the *Patient* and a *Healthcare Provider* (e.g. GP). Each *Contact* is associated with one or more *Episode Of Care* (e.g. weight gain), each one related to one or more *Health Issue* (e.g. diabetes) suffered by the *Patient*. During a *Contact* a set of *Clinical Documents* are produced by the *Provider*. An extended model of the proposed diagram has been used to define the conceptual model of LuMiR, a region-wide EHR system (Contenti et al., 2008). Based on the conceptual model shown in Figure 1 it is possible to define relevant indicators that can be used for secondary purposes such as the comparison of different phenomena over a period of time and between different areas of the same region.

2.2 Data Warehouse Architecture

From the architectural point of view, the data warehouse has been designed as a three-tier architecture considering the information collected and the relevant infrastructure of the EHR described in the previous paragraphs. As highlighted in Figure 2, data are extracted from the legacy operational systems (EHR data sources) and subsequently cleaned-up and integrated in a data staging area represented by an Operational Data Store (ODS). Moreover, ETL (Extract, Transform, Load) tools feed the data warehouse (and data marts) with already integrated data based on a shared message model that do not require transformation. This makes it easier to extract the information from the data source layer and load them in the data warehouse for statistical purposes.

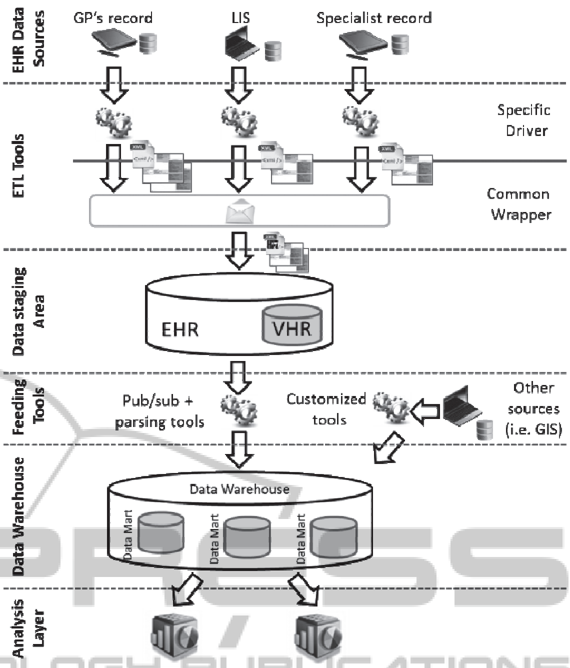

Figure 2: Data warehouse architecture proposed.

The source layer is represented by the set of legacy systems and repositories (e.g. GP's electronic healthcare record, laboratory information systems, radiology information systems) that manage healthcare and administrative information related to citizens. Identification of data sources represents a critical process to be carried out for the success of a data warehouse project. It is important to obtain information about selected data sources identifying the role played by each information system in the development of the data warehouse and data marts. Given that data collected in source information systems are usually stored in different formats and on a variety of platforms, it is necessary to define a common data model that integrates data handled by multiple sources into a single database. To achieve this aim some authors have proposed to use a module called wrapper (Roth and Schwarz, 1997), that is responsible for data gathering from different sources, data cleansing, format conversions as well as data integration. Data managed by each wrapper can therefore be loaded in the data warehouse and thus in the relevant data mart. However, a change in a data warehouse schema makes the revision of each wrapper not straightforward. Therefore, it is necessary to include in the architecture an intermediate stage between the data sources and the data warehouse tiers. This middleware system, called ODS can also be used as a database for operational processing, as proposed by Inmon

(Inmon, 1999). It contains detailed and integrated data with specific constraints including referential integrity that ensures data accessibility by relevant units. In the proposed architecture the ODS is represented by two inter-related systems: a) the EHR repository that contains individual's structured clinical documents together with the EHR registry that manages metadata for indexing clinical documents stored in relevant repositories; and b) the Virtual Healthcare Record (VHR) that manages data extracted from the documents contained in EHR repositories, parsing the HL7 CDA document. The integration of these systems is an appropriate approach to define the ODS considering both conceptual and operational point of views. Referring to the first one, a common data model has already been defined to ensure semantic interoperability between heterogeneous legacy systems and the EHR/VHR system based on the conceptual model shown in Figure 1. This model comprises message information models to exchange data about events, health issues, episode of care, etc. as well as to structure and share clinical documents produced in each event. Considering the operational point of view message exchange between legacy software and EHR/VHR is based on a standard protocol already implemented in some regional EHR infrastructure. In particular, each source system has installed a specific driver that extracts relevant information from the proprietary database describing the event happened (e.g. GPs visit) along with data and documents provided already in HL7 CDA format (e.g. drug prescription). The driver creates an XML standard file that is then provided to a generic wrapper that transforms it in a standardized HL7 message subsequently loaded in the EHR/VHR system. In this way the specific driver installed in each legacy system and the generic wrapper can be seen as an ETL tool in the data warehouse architecture. This makes it easier to integrate data provided by different systems in the EHR/VHR and thus in the data analytical process. The data warehouse collects information stored in the EHR/VHR in an OLAP (On-Line Analytical Processing) approach that facilitates the integrated analysis to develop specific dashboards based on business processes and clinical indicators. Given that VHR is composed by data extracted from clinical documents, there is a need to develop a specific tool that parses relevant information from EHR/VHR and store them in the data warehouse. At this stage of the design process an important role is played by the Hierarchical Event Manager the InFSE module responsible for a real-time notification of clinical data and documents generated during each notified event (Ciampi et al., 2012). This component can be used not only to integrate different systems in the EHR infrastructure, but also to feed the data warehouse as well as the data marts. In particular, this manager notifies users with events that occurred and they are interested in, through a federation of brokers operating on a publish-subscribe paradigm. Of course other techniques are allowed to extract data from the EHR and load them in the data warehouse, in particular in case of frequent monitoring of services. It is important to emphasize that confidential information exchanged between the ODS and the data warehouse must be anonymized to preserve patient's privacy. At this architectural level clinical information managed by the EHR/VHR can be also integrated with other data such as social, demographic, economic, etc.

Finally, the analysis layer concerns tools and techniques for data analysis, such as data mining, reporting and OLAP tools. For instance, they can be used to define a set of clinical indicators, as highlighted in the following.

2.3 Dashboard

In order to identify a set of reliable indicators, in this study the attention has been focused on the diabetes mellitus type 2 pathology. This disease leads to several extremely debilitating complications that can impair the function of vital organs such as heart (e.g. myocardial infarction), kidney (e.g. renal failure), blood vessels (e.g. cardiovascular diseases, stroke) and eyes (e.g. retinopathies). These health issues are monitored in complex processes with frequent contacts between patient and healthcare facilities and professionals. This makes it necessary to timely collect data and to foresee the verification of the healthcare assistance processes in a regional and national benchmarking framework. In Italy, a comprehensive strategy for implementing a chronic disease management intervention for people with diabetes is defined in the IGEA project (Maggini, 2009). Within this project a set of clinical indicators has been defined to measure the process, structure and outcomes of patient care in order to assess a particular clinical situation and to indicate whether the healthcare delivered is appropriate. They are used by different bodies at different levels to identify areas of concerns, which might require further review or development. Clinical indicators are usually included in a benchmarked framework comparing the outcome of each local agency over different period of time with other agencies or with the literature results published, for instance, for other populations or pathologies. In this paper the attention is focused on this type of indicators that are further divided into two categories: 1) process: to assess whether a program is properly implemented by the provider; 2) outcome: to measure the program's level of success in improving service accessibility, utilization or quality. They capture the effect of care processes on the health and wellbeing of patients and populations. Outcome indicators are intermediate if they reflect changes in biological status or in life-style or final indicators if they assess the incidence of new episodes, events or comorbidity within a specific period of time. Examples of clinical indicators are reported in Table 1 highlighting: 1) the type of indicator, 2) the source class (based on the Figure 1 conceptual model), and 3) an example of the indicator based on the IGEA project dashboard (Maggini, 2009). Moreover, for each example the attribute to be taken into account to extract the relevant information is reported. Process indicators are assessed considering the proportion of patients that have been treated in a specified number of visits/controls/measurements over the total number of diabetic patients involved in the program. From the EHR/VHR point of view these indicators can be measured considering on the one hand the contact type (e.g. GP-patient encounter to assess, for instance, the proportion of patients who carried out at least two GP visits in one year) and on the other hand a specific therapy used to treat a patient's health problem (e.g. proportion of patients treated with antihypertensive drug). Intermediate outcome indicators are assessed considering the proportion of patients that have values for a clinical parameter (e.g. micro albuminuria, glycated hemoglobin, central arterial pressure, weight) within a relevant threshold over the total number of diabetic patients involved in the program with at least one measure of that parameter. This type of indicators is measured similarly to the process indicators thus considering a specific healthcare service (e.g. laboratory test) provided to measure a metabolic parameter (e.g. glycated hemoglobin). Another type of intermediate indicators is related to the patient change in life style (e.g. number of patients that has reduced the number of cigarettes smoked during a day) and is assessed considering the patient summary document established and updated by the GP to snapshot, at a specific time point, the health status of a patient with a given pathology. Final outcome indicators are measured considering the proportion of patients that have suffered an event during a specific period of time over the total number of patients who did not suffer from the same event at the beginning of the period.

Table 1: Examples of clinical indicators.

Indicator	Concept / Class	Example
Process indicators		
Services accessibility	HCService	Proportion of patients with at least two GP visits in a year • HCService.Type: GP visit
Treatment of medical condition	HCService	Proportion of patients treated with antihypertensive drug • HCService.Type: Pharmacotherapy • Pharmacotherapy.Type: Antihypertensive drug
Intermediate outcome indicators		
Services outcome monitoring	HCService Laboratory test	Proportion of patients whit hemoglobin < 7% • HCService.Type: LaboratoryTest • LaboratoryTest.Measure: Hemoglobin • Measure.Value: < 7%
Lifestyle behavior changes	Document Patient summary	Proportion of patients who reduced the number of cigarettes smoked • Document.Type: PatientSummary • PatientSummary.Entry: CigarettesSmoked • Entry.Value (to be compared over time)
Final outcome indicators		
Services accessibility	Episode of Care HCService Hospitalization	Proportion of patients hospitalized by episode category • HCService.Type: Hospitalization • EpisodeOfCare.Type
Episode of care observation	Episode of Care	Incidence of proliferative diabetic retinopathy • EpisodeOfCare.Type: Diabetic retinopathy

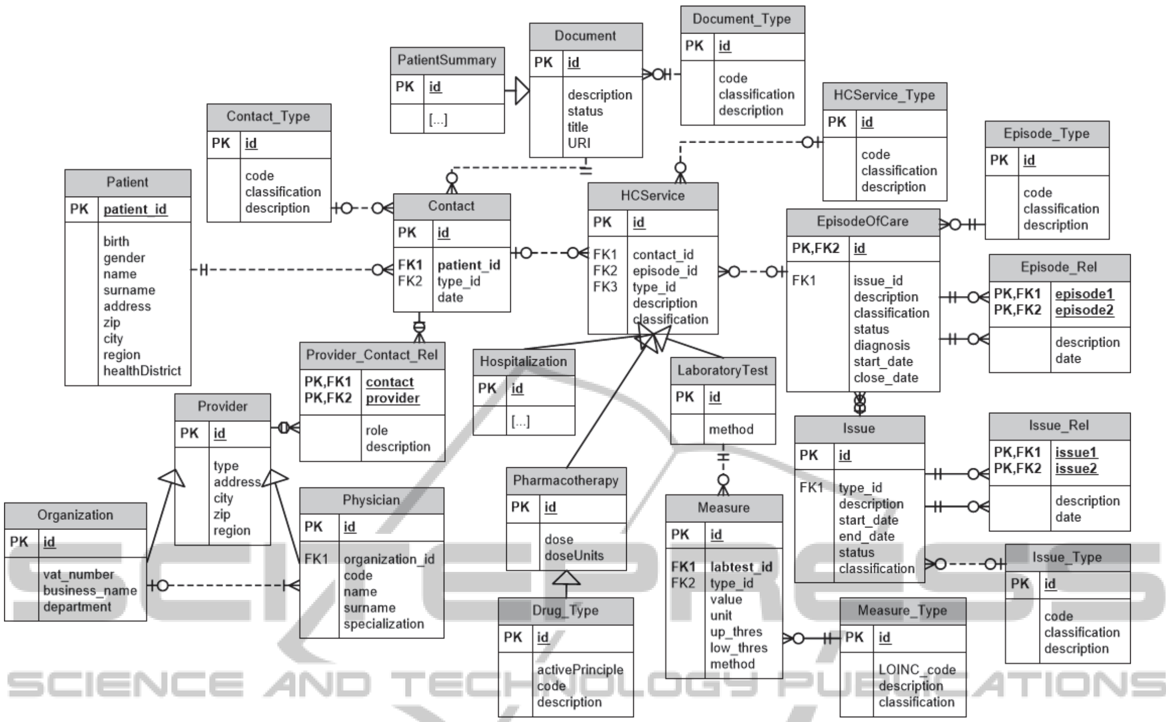

Figure 3: EHR high-level Entity Relationship diagram.

From the EHR/VHR point of view final outcome indicators are mainly assessed considering two EHR concepts: 1) the episode of care (e.g. incidence of proliferative diabetic retinopathy) and 2) the healthcare service (e.g. proportion of patients who accessed the emergency room) that is usually calculated considering the related disease (e.g. episode of care classified using the ICD classification).

2.4 Business Process Modelling

According to the dimensional modelling framework the first step of the data warehouse design was to identify the business processes to be modelled. In this paper the attention is focused on the treatment of medical condition to define process indicators, such as the percentage of patients treated with antihypertensive drug. This business process has been used as a starting point to identify the main tables and attributes to be used to design the high level Entity Relationship (ER) diagram reported in Figure 3. This diagram is based on the conceptual model proposed in Figure 1 where essential concepts are further identified, such as *Patient* who interacts with the healthcare *Provider* that can be either an *Organization* (e.g. the diabetologic centre) or a *Physician* (e.g. the GP). Each *Physician* can also belong to an *Organization* (e.g. the radiologist who works in an hospital). Moreover, the *HCService* describes the different activities performed during each *Contact*. A *HCService* can be for instance a *Laboratory test* or a *Pharmacotherapy*. Given a portion of the ER schema needed to properly describe the investigated business process the attribute tree has been defined and afterward pruned and grafted in order to remove the unnecessary levels of detail such as, irrelevant attributes or relationships (Figure 4). Starting from the attribute tree a dimensional fact model was defined (Figure 5) with the following characteristics: a) Fact table (*TreatmentIndex*) is based on the atomic data extracted from the structured clinical Document and stored in the VHR that describe drugs delivered to a patient to treat a specific episode of care; b) four dimensions have been detected: *Drug* (type of pharmacological product delivered by the pharmacists to the patient, e.g. antihypertensive), *Patient*, *Date* (when the drug has been delivered) and *District* (where the drug has been provided to the patient, i.e. the pharmacy district). To determine costs attributable to a particular drug or treatment and related to a specific episode of care the model depicted in Figure 5 includes an additional dimension (*EpisodeOfCare*) and two measures: the *quantity* of product sold and the *price* for each unit of product. These information are provided by the Italian Medicines Agency (AIFA) that is the national authority responsible for drugs regulation in Italy.

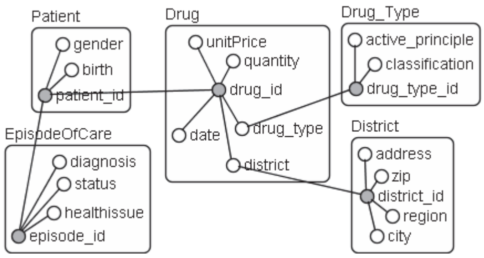

Figure 4: Attribute tree after pruning and grafting.

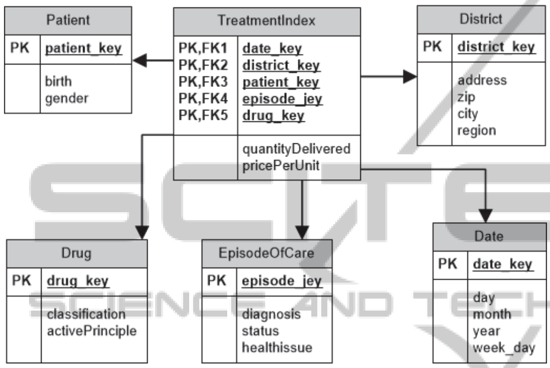

Figure 5: Dimensional model based on service accessibility and outcome monitoring business processes.

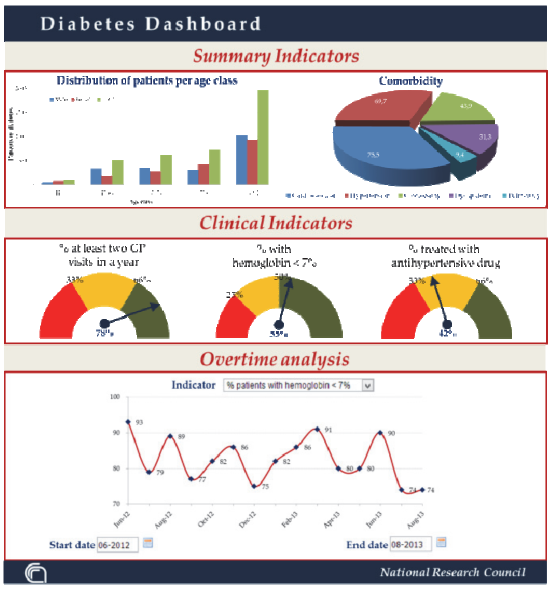

Figure 6: Diabetes dashboard displaying clinical indicators based on the IGEA project (see paragraph 2.2).

A diagram picturing a hypothetical dashboard for diabetes patients is reported in Figure 6. It provides a snapshot of a set of clinical indicators to monitor the effectiveness of healthcare service delivery in a benchmarking framework.

3 CONCLUSIONS

In this paper we demonstrated the feasibility of secondary uses of EHR to develop an enterprise data warehouse architecture in a clinical governance framework. To our knowledge this is a novel approach that intends to exploit the entire EHR infrastructure to develop a Business Intelligence tool that supports the evaluation of healthcare activities under different points of view. As shown in the case study reported in this paper this approach can be used for instance to benchmark local/national healthcare services as well as to monitor the effectiveness of applied guidelines. The EHR represents the core of the proposed architecture as it entails different advantages. First of all, it ensures data quality using standardized data and documents already integrated in a health infrastructure. In particular, HL7 CDA used to exchange structured clinical documents makes it easier to retrieve and process atomic data that are usually coded using international nomenclatures and dictionaries such as ICPC (International Classification of Primary Care), SNOMED (Systematized Nomenclature of Medicine Clinical Terms) and ICD (International Classification of Diseases). This also supports semantic interoperability between different source systems. Data are provided using a publish-subscribe paradigm that guarantees health data to be promptly exchanged at the moment an event is published, with an automatic detection of relevant information directly from source systems. This ensures a timely and continuously updated information flow as well as data integrity and consistency based on a sample size that covers the entire target population. Moreover, this approach makes it possible to gradually integrate other applications within the data warehouse infrastructure such as geographical or socio-demographic information systems. EHR represents the operational data store that guarantees the separation between the transactional and the analytical processing. The feasibility study described in this paper will be implemented in a regional EHR environment.

REFERENCES

Botsis T., Hartvigsen G., Chen F., Weng C., 2010. Secondary use of EHR: data quality issues and informatics opportunities. In *AMIA Summits Translational Science Proceedings* pp. 1–5.

Ciampi M., De Pietro G., Esposito C., Sicuranza M., Donzelli P., 2012. On Federating Health Information

Systems. In *International Conference on Green and Ubiquitous Technology (GUT)*, pp. 139-143.

Committee on Data Standards for Patient Safety, 2003. Key Capabilities of an Electronic Health Record System. *Institute of Medicine Report* 5.

Contenti M., Mercurio G., Ricci F.L., Serbanati L.D., 2008. LuMiR: A Region-wide Virtual Healthcare Record. In *Proceeding of the 9th International HL7 Interoperability Conference.* Crete, Greece. pp. 80-3.

De Mul, M., Alons, P., Van der Velde, P., Konings, I., Bakker, J., Hazelzet, J., 2012. Development of a clinical data warehouse from an intensive care clinical information system. *Computer methods and programs in biomedicine* 105 (1) pp. 22-30.

Diomidous M., Zimeras S., Mantas J., 2009. Spatial Electronic Health Record for the epidemiological clinical data. *Travel Health Informatics and Telehealth* 1, pp. 66-72.

Dolin R. H., Alschuler L., Boyer S., Beebe C., Behlen F. M., Biron P. V., Shvo A. S., 2006. HL7 clinical document architecture, release 2. *Journal of the American Medical Informatics Association* 13 (1) pp. 30-9.

EN 13940-1:2007 Health informatics - System of concepts to support continuity of care - Part 1: Basic concepts. CEN 2007. British Standards Institution.

Gillespie G., 2000. There's gold in them thar' databases. *Health Data Management* 8 (11) pp. 40-52.

Hussain S., Ouagne D., Sadou E., Dart T., Jaulent M. C., De Vloed B., Daniel C., 2012. EHR4CR: A semantic web based interoperability approach for reusing electronic healthcare records in protocol feasibility studies. In *Proceedings of the 5th International Workshop on Semantic Web Applications and Tools for Life Sciences.* Paris, France.

Inmon W., 1999. *Building the Operational Data Store*. John Wiley & Sons. New York. 2nd edition.

Jensen P. B., Jensen L. J., Brunak S., 2012. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics* 13 (6) pp. 395-405.

Kamal J., Pasuparthi K., Rogers P., Buskirk J., Mekhjian H., 2005. Using an information warehouse to screen patients for clinical trials: a prototype. In *Proceeding of American Medical Informatics Association Symposium*. Washington DC, USA. pp. 1004.

Koh H. C., Tan G., 2011. Data mining applications in healthcare. *Journal of Healthcare Information Management* 19 (2) p. 65.

Kush R. D., Helton E., Rockhold F. W., Hardison C. D., 2008. Electronic health records, medical research, and the Tower of Babel. *New England Journal of Medicine* 358 pp. 1738–40.

Lurio J., Morrison F. P., Pichardo M., Berg R., Buck M. D., Wu W., Calman N., 2010. Using electronic health record alerts to provide public health situational awareness to clinicians. *Journal of the American Medical Informatics Association* 17 (2) pp. 217-9.

Maggini M., 2009. IGEA-A chronic disease management project for people with diabetes. *Annali Istituto Superiore di sanità* 45 pp. 349-352.

Mettler T., Rohner P., 2009. Supplier Relationship Management: A Case Study in the Context of Health Care. *Journal of Theoretical and Applied Electronic Commerce Research* 4 pp. 58-71.

Oemig F., and Blobel B., 2005. Does HL7 Go towards an architecture standard? *Studies in Health Technology and Informatics* 116 pp. 761-6.

Roque F. S., Jensen P. B., Schmock H., Dalgaard M., Andreatta M., Hansen T., Søeby K., Bredkjær S., Juul A., Werge T., Jensen L. J., Brunak S., 2011. Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. *PLoS Computational Biology* 7 (8), pp. 1-10.

Roth M. T., Schwarz P., 1997. Don't scrap it, wrap it! a wrapper architecture for legacy data sources. In *Proceeding of the Conference on Very Large Data Bases (VLDB)*. Athens, Greece. pp. 266–275.

Safran C., Bloomrosen M., Hammond W. E., Labkoff S., Markel-Fox S., Tang P. C., Detmer D. E., 2007. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. *Journal of the American Medical Informatics Association* 14 (1) pp. 1-9.

Sahama T. R., Croll P. R., 2007. A data warehouse architecture for clinical data warehousing. In *Proceedings of the fifth Australasian symposium on ACSW frontiers* 68. Ballarat, Victoria, Australia. pp. 227-32.

Schadow G., Mead C. N., Walker D. M., 2006. The HL7 reference information model under scrutiny. *Studies Health Technology Informatics* 124 pp. 151–6.

Stow P. J., Hart G. K., Higlett T., George C., Herkes R., McWilliam D., Bellomo R., 2006. Development and implementation of a high-quality clinical database: the Australian and New Zealand Intensive Care Society Adult Patient Database. *Journal of Critical Care* 21 (2) pp. 133-41.

Taylor P., 2008. When consent gets in the way. *Nature* 456 pp. 32–3.

Wang X., Hripcsak G., Markatou M., Friedman C., 2009. Active Computerized Pharmacovigilance using Natural Language Processing, Statistics, and Electronic Health Records: a Feasibility Study. *Journal of the American Medical Informatics Association* 16 (3) pp. 328-37.

Wickramasinghe N., Schaffer J.L., 2006. Creating knowledge-driven healthcare processes with the intelligence continuum. *International Journal of Electronic Healthcare* 2 (2) pp. 164-74.

Zhou X., Chen S., Liu B., Zhang R., Wang Y., Li P., Guo Y., Zhang H., Gao Z., Yan X., 2010. Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. *Artificial Intelligence in Medicine* 48 (2) pp. 139-52.