A Data Gathering Framework to Collect Type 2 Diabetes Patients Data

Arianna Dagliati, Lucia Sacchi, Mauro Bucalo, Daniele Segagni, Konstantia Zarkogianni, Antonio Martinez Millana, Jorge Cancela, Francesco Sambo, Giuseppe Fico, Maria Teresa Meneu Barreira, Carlo Cerra, Konstantina Nikita, Claudio Cobelli, Luca Chiovato, Maria Teresa Arredondo, and Riccardo Bellazzi

Abstract— **In this work, we present a framework implemented within the EU project MOSAIC, funded under the FP7 framework, to gather Type 2 Diabetes (T2D) patients' data coming from three European hospitals and a local health care agency. A subset of the MOSAIC activities is centered on the development of Temporal Data Mining models to identify relevant clinical pathways in patients' histories and will in particular benefit from the data coming from the medical centers involved in the project. To best exploit this repository, the need for creating a common and sharable data model becomes immediately apparent. This model is the main subject of this paper. The proposed approach relies on the Informatics for Integrating Biology and the Bedside (i2b2) and the Shared Health Research Information Network (SHRINE) open source software tools. It provides an integrated research setting to merge clinical and environmental data that will enable obtaining a broader vision of individual patients' histories, which will be then mined with multivariate models to identify relevant clinical pathways.**

I. Introduction

Clinical data coming from hospital information systems and collected during routine practice, together with data recorded for administrative purposes, has been intensively used in clinical and epidemiological studies in the last years [1, 2, 3]. The integration of these different streams of data is fundamental to get the best knowledge out of them and build complete and rich patients' histories. Patients' and clinicians' actions can be monitored for several years and involve acute events, follow-ups, chronic disease management, medications administration, etc. The activities devoted to the development of models to identify relevant clinical pathways in such complex patients' histories can be conveniently supported by the availability of an integrated data model, able to store data coming from heterogeneous sources in a homogeneous way [4,5,6,7].

The MOSAIC project is devoted to the development of models and methodologies to enhance the currently available tools for the diagnosis and management of diabetic patients. In particular, the focus is put on the identification of a set of risk profiles, characterized by clinical and environmental factors and by specific patterns of care, able to stratify the population with the objective of delivering a more targeted and personalized care. Within the project consortium, a number of data sets has been made available, including data coming from three hospitals and one local healthcare agency. To best exploit this repository, the need for creating a common and sharable data model becomes immediately apparent. The solution we propose is to build a data warehouse (DW) to integrate, visualize and query data in an informative way, considering both their temporal nature and complexity.

In this paper, we will describe how this strategy has been implemented within the MOSAIC project.In Section II, we will introduce the hospital databases available in MOSAIC and the mapping process that has been carried out to reach a common parameters representation. In Section III, we will present the technological solutions selected to provide all the medical centers with a common substrate to store their data. Furthermore, we propose a software solution that gives the possibility of aggregating the database instances under a common framework, to perform integrated queries while maintaining the data inside the facilities of each hospital. In Section IV we present the shared data model, known as the Core Ontology, which has been developed as a collaboration of the project partners.

II. Hospitals Data Mapping

The medical centers participating in the project are: the Local Healthcare Agency of Pavia in Italy (ASL), the IRCCS Fondazione Salvatore Maugeri Hospital of Pavia in Italy (FSM), the Health Department Hospital La Fe in Valencia, Spain (La Fe) and the Hippokration General Hospital in Athens, Greece (Hippokration).

As a matter of fact, each hospital database has a different structure and different variables might potentially be collected by different centers. On the other hand, though, to query the data in a consistent way, it is important to share a common data model with a homogeneous representation of the collected parameters. To this end, the data available in

A.D., L.S., M.B. and R.B. Authors are with the Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy (corresponding author: A.D. phone: +39 0382 592044; e-mail: arianna.dagliati@unipv.it Other Authors' email:, lucia.sacchi@unipv.it, riccardo.bellazzi@unipv.it).

D.S, L.C. Author is with Fondazione Salvatore Maugeri of Pavia (e-mail: daniele.segagni@fsm.it, luca.chiovato@fsm.it)

C.C. Author is with Department of Purchasing and Control, ASL Pavia, Italy (e-mail: carlo_cerra@asl.pavia.it)

K.Z, K.N. Authors are with the Faculty of Electrical and Computer Engineering, National Technical University of Athens, Zografou, Athens 15780, Greece (e-mail: kzarkog@biosim.ntua.gr, knikita@ece.ntua.gr)

F.S, C.C Authors are with the Department of Information Engineering University of Padova, Italy (e-mail: francesco.sambo@dei.unipd.it, cobelli@dei.unipd.it)

J.C., G.F., M.T.A Authors are with the Universidad Politecnica de Madrid, Spain (e-mail jcancela@lst.tfo.upm.es, gfico@lst.tfo.upm.es)

A.M.M., M.T.M.B., Authors are with Universidad. Politécnica de Valencia, Spain (e-mail anmarmil@itaca.upv.es, tmeneu@upvnet.upv.es)

978-1-4799-2131-7/14/$31.00 ©2014 IEEE

each of the involved medical centers were analyzed, with the aim of defining the most efficient strategy to gather the data using a common framework. Hospital clinical partners have actively participated to these activities, providing detailed descriptions of their clinical databases and the necessary medical and scientific knowledge to set up the processes of data mapping and parameters selection.

The strategy we have adopted to map the data coming from the medical centers had several objectives:

- Identification of the parameters that are in common between the different centers;
- For such parameters, share a common representation of the variables (same units of measurement, same coding system, same type of representation);
- Define a common data structure to: build a data warehouse for each hospital, facilitate data sharing for analysis purposes, allow DWs aggregation.

The data mapping was performed through a multi-step process: first, each hospital shared a list of the available variables and parameters. An initial matching was performed and, after iterative refinement phases, a final agreed version of the mapping was delivered.

Analyzing the data coming from the medical centers, we identified a set of macro sections under which we were able to group the available variables. This is presented in Table I.

TABLE I.

	Macro-categories of variables
1	Demographics
2	Lifestyle-related variables
3	Variables related to the physical examination performed at each visit
4	Laboratory Test Results
5	Complications/Comorbidities
6	Therapy Prescriptions
7	Administrative Information: - Pharmaceutical Data, related to the purchase of a drug by the patient; - Hospitalizations details; - Consumed Outpatient Services.

As shown in Table I, data can come from different sources: on the one hand there are clinical data, collected by the hospitals during their routine medical activities, while on the other there are administrative data, which are mainly process data collected for billing purposes. A peculiar example in this sense is represented by the databases coming from the Pavia area: the ASL database stores administrative information, while the FSM database collects clinical data. Being able to provide an integrated version of these data sources can deliver a complete view on the clinical histories of diabetes patients, ranging from their clinical data to the different accesses they performed to national healthcare services.

While the data collected at the hospital level provide detailed information about clinical parameters, they have the intrinsic limit of not supplying complete information about the history of the patient in terms of time and space. The reason is that the patient, although assigned to a specific center for the diabetes pathology, may undergo visits or laboratory tests elsewhere through the national health care service. The administrative data allow a broader view of the patient's history, supplying information about his/her treatment pathway, contacts with the regional health services, drug prescriptions, etc. The integration of these data sources seems particularly suitable to manage chronic patients' histories. This wider view of the diabetic population will represent an added value to the project, and the data analysis techniques can greatly benefit from it.

III. Methods

The data gathering strategy that has been designed for collecting hospital data in the MOSAIC project is synthesized in Figure 1. The approach is based on the transformation of the datasets coming from the different centers, and currently stored in a variety of formats, into data warehouses implemented using the same technology and with an underlying homogeneous data model. As shown in Figure 1, FSM and ASL data will be joined to form a single Data Warehouse, while each of the other hospitals will have its own Data Warehouse. Dedicated Extraction, Transformation and Loading (ETL) steps have been implemented to collect data from the original data sources.

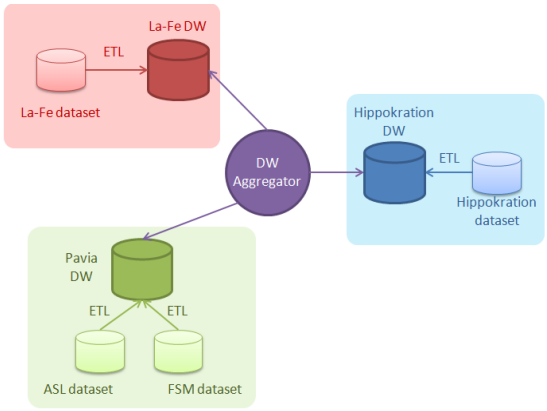

Figure 1. Architecture of the proposed data gathering strategy.

In addition to this transformation step, an infrastructure able to "aggregate" the obtained Data Warehouses has been set up. The technological solutions that have been selected to build the Data Warehouses and the Data Warehouse aggregator that will be implemented on top of the described architecture are respectively the Informatics for Integrating Biology and the Bedside (i2b2) [8, 9] and the Shared Health Research Information Network (SHRINE) open source software tools [4,5]. The SHRINE infrastructure allows integrating several local i2b2 instances under a shared framework so that locally recorded diseases, environmental details and outcomes will be queried in parallel and exported in a format suitable for further analyses.

A. MOSAIC i2b2

The IT software solution that has been selected to build a Data Warehouse in the MOSAIC project is i2b2 (Informatics for Integrating Biology and the Bed Side [8]). i2b2 is one of the seven centers funded by the NIH Roadmap for Biomedical Computing. The mission of i2b2 is to provide clinical investigators with a software infrastructure able to integrate clinical records and research data [9]. The key feature of this tool is that it "fits together" medical record data and clinical trial data at a person-level so that diseases,

genes, and outcomes can be related to each other. i2b2 allows storing multidimensional data with a common representation in a star relational database, where facts are hierarchically structured as ontologies. It provides a query tool interface to extract sets of interesting patients. For this reason, it is perfectly suitable to the needs of a project such as MOSAIC. Within MOSAIC data gathering activities, the i2b2 platform is exploited on top of different models for fast data exploration and interrogation, as well as for retrieving data.

There are several reasons why i2b2 has been selected among the variety of available Data Warehouse development tools [10,11], a crucial one is that i2b2 is the only software that is patient-specific and supports the use of ontologies for querying the Data Warehouse. For this reason, there's no need to use Data Warehouse dedicated languages to perform a query.

The i2b2 Data Warehouse meets the requirements to reach the goals of an efficient data gathering strategy as it enables to:

- Collect multidimensional data.
- Integrate different sources of information.
- Aggregate data and export them in a format suitable for temporal dimension analysis.

Moreover, security and privacy are always important concerns within the healthcare ecosystem, but this becomes even more important working in a Consortium environment. i2b2 allows the data sharing with a controlled and restricted data transfer making this process safer, quicker and easier from the administrative point of view.

Within the MOSAIC project, we implemented three different i2b2 instances, one for each of the clinical centres involved. As shown in Figure 1, each centre established its own necessary ETL procedures in order to fill in its own i2b2 local data warehouse with clinical and administrative data. While the ontologies of each i2b2 instance have a common core, each local i2b2 framework is independent of the other instances. In this way each centre has the possibility of creating and including specific concepts, thus extending the original core ontology.

B. MOSAIC SHRINE

The data gathering strategy that has been designed for collecting hospital data includes the setup of an infrastructure able to aggregate data stored in the different DWs. The main objective of this aggregator is to make the results of integrated queries performed on the three Data Warehouses available to the researchers. The technological solution selected to perform this task is SHRINE (Shared Health Research Information Network), a project developed by the Harvard Medical School in Boston and strictly connected to the i2b2 project.

The main goal of SHRINE is the development of an IT infrastructure that allows communication between different i2b2 instances, installed in different hospitals, thus allowing data sharing among different centres. As i2b2, SHRINE is freely available and open source [12]. There are many commercial products for the creation of distributed database systems (e.g. Oracle Streams Database) or that combine data from multiple sources (e.g. Microsoft SQL Server Integration Services). However, a tool to run combined queries on clinical data must meet the technical restrictions imposed by the supervisory bodies of the hospitals, for example in the area of data privacy, and should address the specific challenges related to medical data that are often inaccurate and incomplete. As for i2b2, SHRINE is able to efficiently handle administrative issues and privacy request. Since it is difficult due to legal barriers to transfer healthcare data from one hospital database to another, SHRINE allows to run queries in multiple databases at the same time, preserving the privacy, security and legal requirements and allowing researchers to identify potential datasets for research purposes. SHRINE exploits digital certificates to protect network communication and to identify accredited institutions at different levels:

To create a homogeneous environment that allows performing queries over multiple databases, one of the most important components of the SHRINE architecture is the Core Ontology. The Core Ontology is the instrument through which all the medical concepts are represented; it is organized in a hierarchical fashion, to facilitate navigation and selection of query-specific concepts. Following suitable procedures (detailed in [13]), the single data repositories of the hospitals participating to SHRINE can be mapped to the Core Ontology. Each medical centre, despite the common mapping of data, has the possibility of using different methods for encoding parameters and specific ontologies for recording patients' observations. Furthermore, the three different instances of i2b2 will be installed independently. For these reasons, the first step is mapping each local database ontology with the SHRINE Core Ontology. In this way, each centre will be able to work both autonomously on the local i2b2 and make data available to the consortium in a common and aggregated format. The Core Ontology of SHRINE has been designed to provide maximum breadth of concepts related to the project and commonly available in electronic health record systems and administrative data warehouses.

IV. RESULTS

A. Definition of the common data model

Once the aggregation and data storing procedures have been defined, the following task has been dedicated to the creation of the integrated ontology representing common concepts. One of the most important components of the whole infrastructure is the core ontology. To build the Ontology, we started from the results of the mapping procedures described in Section II. As a general guideline, we have chosen to include in the core ontology those concepts that are represented in at least two out of the three participating medical centres.

According to the i2b2 structure, the main idea is to represent each clinical event happening to the patient (follow-up visit, hospitalization, lab test, drug prescription) as a specific instance in the Data Warehouse, thus providing it with a start and end time and connecting it to the specific concepts related to that particular event. This is reflected in the resulting ontology structure (Figure 2).

Relying on these criteria, the ontology that we have defined contains 7 high level concepts, which are:

- Patient data: collects all the concepts related to a patient and that are not depending on the follow-up ("static" data). This information is collected once, usually during the first encounter
- ICD9 codes: this node contains all the concepts related to the ICD9 coding system. ICD9 codes can be related to several types of events, such as hospitalizations, outpatient services, follow-up visits, comorbidities, complications, etc. To represent information related to ICD9-CM codes in a standardized way, we will rely on the NCBO BioPortal web services [14, 15]. The strategy we plan to adopt is to pull data from NCBO via dedicated rest web services and then reorganize the results into the format used by the i2b2 ontology cell.
- Complications and comorbidities: these are specific concepts related to diabetes that the medical centers always collect. Complications are directly related to the disease while comorbidities have an external cause.
- Contact details: collects all concepts that can be measured during an encounter (follow-up visit, hospitalization and outpatient visit).
- Laboratory exams: collects all the concepts related to the main laboratory tests related to the disease.
- Drugs data: collects all the data related to the pharmacological therapy prescribed to the patient within health centers (Therapy prescriptions) and retrieved form administrative flows (Pharmaceutical Data).

B. *Current State of the System*

Up to now, the data mapping phase is concluded and all the instances of the i2b2 DW have been implemented, relying on the core ontology. ETL procedures have been defined and tested for the four medical centers. The SHRINE architecture for aggregating the DWs is currently being set up. Future work will regard finalizing the data uploading process and the aggregator.

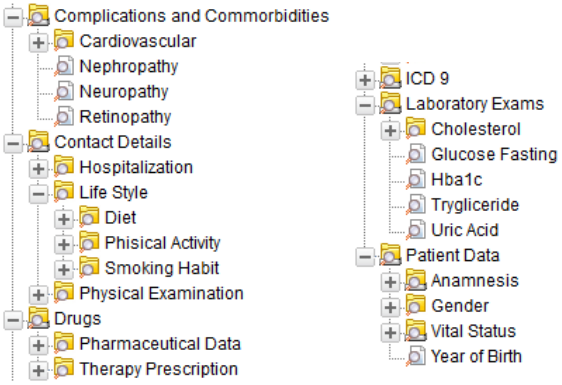

Figure 2 – The Ontology structure as represented in the i2b2 system.

V. CONCLUSION

The introduced system is a concrete example of an integrated ICT architecture able to collect large and heterogeneous data sets, in order to better understand the mechanisms underlying the evolution of diabetes through the analysis of temporal events and behavioural factors. These data will be useful for investigators interested in generating new research hypotheses, planning research requiring large sample sizes not easily available at any single institution and conducting research in the areas of population health and health services.

The MOSAIC framework will manage data of more than 5.000 T2D patients, in three different countries across Europe in a seamlessly way.. Such data have been collected for about 10 year for clinical and administrative purposes and will now be "reused" for research investigation on the basis of a multisource data structure.

VI. ACKNOWLEDGEMENT

The MOSAIC project is funded by the European Commission under the 7thFramework Program, Theme ICT--2011.5.2 Virtual Physiological Human (600914).

REFERENCES

[1] S Dieren, JW Beulens, AP Kengne, et al.: Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. Heart. 2012; 98: 360-369

[2] GS Collins, S Mallett, O Omar, et al. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Medicine 2011; 9:1-14

[3] M. Skevofilakas, K. Zarkogianni, B. G. Karamanos, K. S. Nikita, "A hybrid Decision Support System for the Risk Assessment of retinopathy development as a long term complication of Type 1 Diabetes Mellitus", 32nd International Conference of the IEEE Engineering in Medicine and Biology Society, August 31 - September 4, 2010, Buenos Aires, Argentina.

[4] L. Sacchi; D. Segagni; A. Dagliati; A. Zambelli; R. Bellazzi. Mining Careflow Patterns in data warehouses of breast cancer patients. AMIA 2013 Annual Symposium, podium presentation.

[5] Segagni D, Tibollo V, Dagliati A, Malovini A, Zambelli A, Napolitano C, Priori SG, Bellazzi R. Clinical and research data integration: the i2b2-FSM experience. AMIA Summits Transl Sci Proc. 2013 Mar 18;2013:239-40. PubMed PMID: 24303274; PubMed Central PMCID: PMC3845786.

[6] Segagni D, Tibollo V, Dagliati A, Napolitano C, G Priori S, Bellazzi R. CARDIO-i2b2: integrating arrhythmogenic disease data in i2b2. Stud Health Technol Inform. 2012;180:1126-8. PubMed PMID: 22874375.

[7] Segagni D, Tibollo V, Dagliati A, Zambelli A, Priori SG, Bellazzi R. An ICT infrastructure to integrate clinical and molecular data in oncology research. BMC Bioinformatics. 2012 Mar 28;13 Suppl 4:S5. doi: 10.1186/1471-2105-13-S4-S5. PubMed PMID: 22536972; PubMed Central PMCID: PMC3303735.

[8] http://www.i2b2.org/software (last visited January 9th, 2013)

[9] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S. et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2) Journal of the American Medical Informatics Association. 2010;17:124–130.

[10] Business Intelligence Software. http://www54.sap.com/pc/analytics/business-intelligence.html. Last accessed 10/10/2013

[11] IBM Cognos Software. http://www-01.ibm.com/software/analytics/cognos/. Last accessed 10/10/2013

[12] McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, Bickel J, Wattanasin N, Gilbert C, Trevvett P, Churchill S, Kohane IS. SHRINE: enabling nationally scalable multi-site disease studies. PLoS One. 2013;8(3):e55811.

[13] https://open.med.harvard.edu/display/SHRINE/Welcome+to+SHRINE;jsessionid=B2FA703B7ED135ACED245CB43884F38B (last visited January 9th, 2013)

[14] Mark A. Musen, Natasha F. Noy, Nigam H. Shah, Christopher G. Chute, Margaret-Anne Storey, Barry Smith, and the NCBO team. The National Center for Biomedical Ontology. J Am Med Inform Assoc. 2012 Mar-Apr;19(2):190-5.

[15] http://www.bioontology.org/ (last visited March 20th, 2014)