# Decision Support through Data Integration: Strategies to Meet the Big Data Challenge

**Enea Parimbelli**[1,2], **Lucia Sacchi**[1,2], **Riccardo Bellazzi**[1,2,3]

[1] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy

[2] Interdepartmental Centre for Health Technologies, University of Pavia, Italy

[3] IRCCS Foundation "S. Maugeri", Pavia, Italy

## Abstract

**Objectives:** Presentation of an overview of the reasons why data integration initiatives should be seen as enablers for effective decision support in data-intensive healthcare settings.
**Methods:** Typical challenges rising from the information requirements of clinical decision support systems are highlighted. We then propose a methodological solution where several heterogeneous data sources are integrated by the means of a common data model on top of which the DSS is built.
**Results:** We report on two successful case studies based on the DSSs developed in the context of the MobiGuide and Mosaic projects, funded by the European Union in the Seventh Framework Program. The MobiGuide patient guidance system has been successfully validated during a recent pilot study involving 30 patients (10 with atrial fibrillation and 20 with gestational diabetes), while Mosaic is currently undergoing a validation phase involving 1000 type 2 Diabetes patients.
**Conclusions:** In the era of big data, effective data integration strategies are an essential need for medical informatics solutions and even more for those intended to support decision processes. Building generic DSSs based on a stable (but easily extensible) data model, specifically designed to meet the information requirements of DSSs and analytics, has proven to be a successful solution in the two presented use cases.

## Keywords

Decision support; data integration; big data

**Correspondence to:**

**Riccardo Bellazzi**
Department of Electrical, Computer and Biomedical Engineering, University of Pavia
Address: Via Ferrata 5, 27100, Pavia, Italy
E-mail: riccardo.bellazzi@unipv.it

**EJBI 2016; 12(1):en10-en14**
received: April 19, 2016
accepted: April 25, 2016
published: May 20, 2016

## 1 Introduction

Almost all the stakeholders involved in healthcare processes have to face complex decisions on a regular basis. Regardless of them being patients, clinicians or health policy makers they have the common need of balancing a wide range of objectives, often competing among themselves: make difficult diagnoses, avoid errors, ensure highest quality, choose between alternative treatments, maximize efficacy and save money all at the same time. For these reasons, decision support functionalities are among the most sought after capabilities of medical informatics systems. Indeed the need for effective support to decision-making is even more urgent today than in the days of early adoption of these systems [1]. The advances of biomedical discovery including genomics (as well as other "omics" like proteomics or exposomics), the improved understanding of diseases, availability of new technologies for mobile and self-monitoring devices, the exponential increase in the use and penetration of the internet are only some of the factors contributing to the growth of the two main components needed for effective decision making: information (i.e. data) and knowledge on how to use these information. As a consequence, to thoroughly support decision processes, it is desirable for a decision support system (DSS) to consider the widest possible set of available information. In most advanced systems these might include patient history coming form EHRs [2], several clinical parameters collected by patients using self-monitoring devices [3, 4] (e.g. blood pressure or blood glucose measurements), information available from local health agencies [5] (e.g. purchases and refills of medications), genetic data [6], environmental data (e.g. pollution), patients preferences and lifestyle habits [7]. However there are at least three main challenges to face in this scenario: (i) the information that is relevant for a decision task is typically scattered across several data sources; (ii) these sources can have different data representation formats; (iii) the

 

Table 1: Fact sheet of the MobiGuide and Mosaic projects compared.

| | **MobiGuide** | **Mosaic** |
|---|---|---|
| Project duration | 4 years (Nov 2011-Nov 2015) | 40 months (Jan 2013-Apr 2016) |
| Clinical domain | Potentially supports any domain. Pilot implementation on atrial fibrillation and gestational diabetes | Type II diabetes |
| Users of DSS | Clinicians and patients | Clinicians and health care managers |
| Type of DS | Guideline-based | Based on predictive models and visual analytics |
| Type of output | Clinical recommendations | Risk scores, temporal patterns and alerts |
| Data sources | Hospital EHR, patients' body area network, patients personal preferences | Hospital EHR, administrative data, environmental data |
| Data Integration main component | Data Integrator | Orchestrator |
| Data model | HL7 vMR + openEHR archetypes | i2b2 star schema |
| Data Integration technologies | BaseX noSql XML-based storage + REST web services for querying | Oracle DBMS + i2b2 hive ecosystem + REST web services for querying |

volume and heterogeneity of the data may often raise "big data" challenges, which require proper technological and architectural solutions.

## 2 Methods

A methodology based on a data integration strategy is hereby proposed as a solution to the challenges highlighted in the previous section. The proposed methodological approach has been successfully applied in the context of two EC-funded projects, namely MobiGuide [8] and Mosaic [9], whose architecture and results will be presented in detail in the following section.

The need for a DSS to access different data, stored in different formats and originating from different sources is inevitable to achieve optimal decision support capabilities. On the other hand, tightly coupling the specific DSS implementation to each of the several data formats very likely results in increased complexity, high change requests frequency, and ultimately poor maintainability of the produced system.

An alternative approach consists in adding a data integration layer to the architecture and to represent all the needed information using a common information model that serves as a single data provider for the DSS. Several approaches to the design and implementation of such common data models have been proposed in the medical informatics literature, some of which have been specifically developed to support clinical DSS [10, 11, 12, 13]. Among these, some approaches only provide the specifications for a logical data model that defines entities and their relationships at a high level of abstraction (e.g. HL7 vMR [14]) while others also define a technological structure (e.g. i2b2 [15], which comprises a database management system, services for querying the data, etc.) able to support the implementation of the defined abstract model. In both cases one of the most important steps to perform is the reconciliation process, often referred to as mapping, of different information items to fit the entities available in the chosen target data model. This is often a resource-intensive phase of the development process both at a knowledge engineering level and in terms of required implementation efforts [16]. In fact the core of many data integration solutions consists of a set of specifically developed extraction-transformation-loading (ETL) procedures that allow collecting the information from the different sources in a single integrated repository. Given their importance, mapping and ETL processes have recently gained attention in medical informatics research literature where several advanced methodologies including semantic [17, 18, 19] and real-time approaches [20] have been described. It is also important to stress the fact that sharing a common data model also enables to create DSSs that rely on the data provided by multiple centers. This could be accomplished using two different strategies: (i) physically aggregating the data in a single corporate-level warehouse populated through periodical ETL procedures or (ii) building a federation of local repositories which share the same logical model allowing them to be collectively queried [21].

## 3 Results

The described methodology for building a clinical DSS on top of a data integration solution has been applied in the context of two different European research projects (Table 1) carried out between November 2011 and April 2016. The projects activities involved the University of Pavia and the IRCCS Foundation "S. Maugeri" Hospital, in collaboration with international academic and industrial partners.

MobiGuide aims at developing an intelligent guideline-based decision-support system for patients with chronic illnesses [8]. The system accompanies the patients wherever they go and helps them and their care providers in managing their condition, whether they are at home, at

 

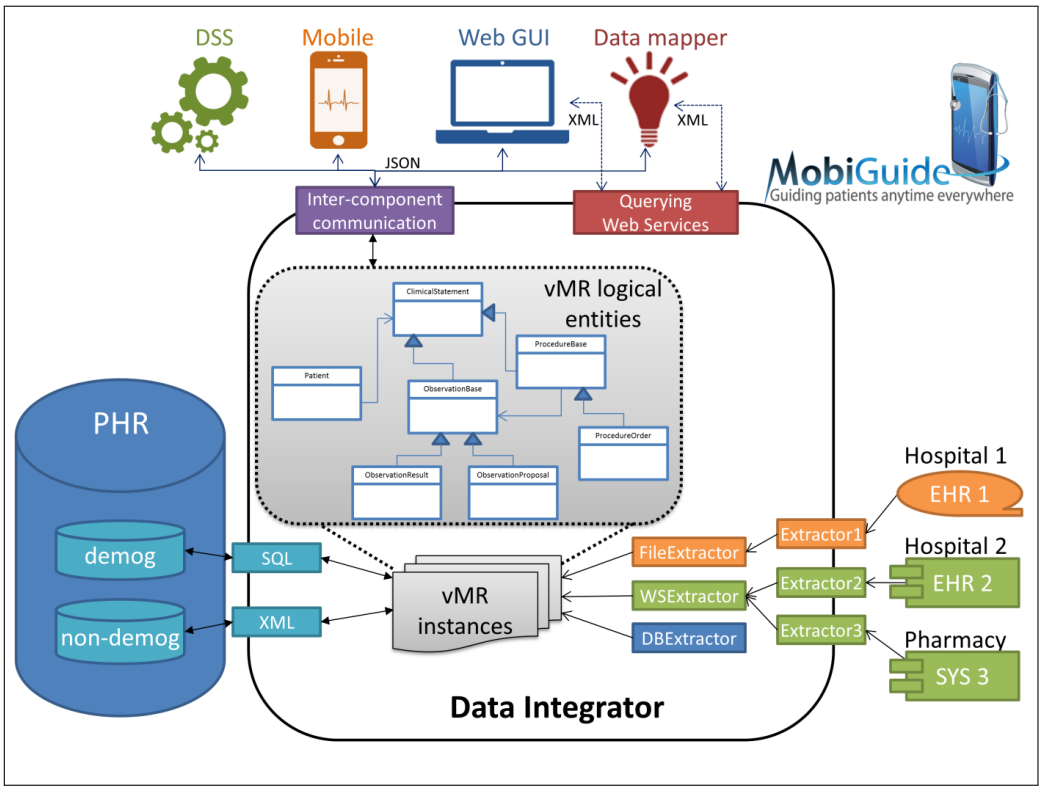


Figure 1: The MobiGuide architecture is centered around its Data Integrator component.

work, out and about or travelling abroad on holiday or for business. One peculiar characteristic of the MobiGuide DSS is that it provides guidance both to clinicians and patients directly. Patients' smartphones act as the centralized connection hub for the MobiGuide Body Area Network (BAN), which comprises a set of sensors able to provide data to the DSS while patients continue with their everyday life with the benefit of being constantly monitored by the system. Moreover, a dedicated smartphone app is used as the primary interface for the patients to interact with the system and receive feedbacks and guidance from the DSS. The DSS engine is also distributed in its nature: it features a full-fledged back-end DSS and a lightweight mobile DSS, which runs directly on the patient smartphone and is able to provide essential guidance even if the connection to the back-end system is unavailable. The MobiGuide DSS applies the knowledge contained in computer-interpretable clinical practice guidelines [22] to a continuously monitored set of patient data, and outputs personalized clinical recommendations about disease management. These recommendations include suggestions on clinical actions for the physicians to take (e.g. prescribe a certain drug or perform a specific diagnostic procedure) and advices directly delivered to patients (e.g. reminders about taking drugs on schedule or taking an additional measurement to check potentially harmful, unforeseen situations) [23]. To accomplish this, the guideline execution engine needs bio-signals originating from patient smartphones, clinical findings collected in the hospital HER and patient preferences collected with dedicated interfaces. All of these sources of information are aggregated in a centralized personal health record (PHR) [13]. A dedicated software component, namely the data integrator (DI) [24], manages all the insertions and retrievals of data from the PHR and periodically updates it through ETL procedures. The data model on which the DI and PHR are based has been derived from the HL7 vMR standard. The same logical model is also used for inter-component communications inside the MobiGuide environment [25]. The vMR standard alone does not provide a specific technical solution for its implementation. This, while complying to the same logical data model across the entire system, allowed to choose an XML representation to implement data persistence while using a lighter JSON format inter-component communication (Figure 1).

The MOSAIC project is aimed at developing novel analytics methods and tools for managing type 2 diabetes (T2DM) and its complications. Differently from MobiGuide, the type of decision support delivered by the MOSAIC system is based on a set of models, developed within the project [26], able to estimate the risk of a patient to develop T2DM or its complications, and to guide users in the management of the temporal evolution of the disease. The MOSAIC system offers two perspectives to its users: on the one hand, it allows managing patients during visits through a single-patient view. On the other hand, it allows analyzing sets of patients thanks to a population view. Being focused on these two use cases, the MOSAIC system addresses mainly physicians and health care managers, but it also proved to be a useful instrument to be shown to the patients during visits.

 

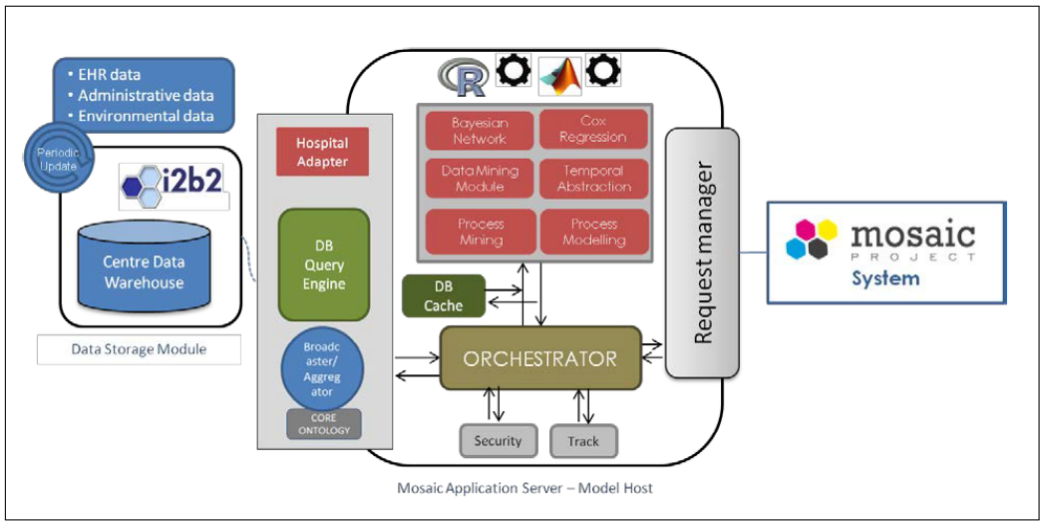


Figure 2: The MOSAIC system architecture.

The MOSAIC system has been designed to be potentially used in any context dealing with T2DM patients, including the GP practice, where diagnosis is performed and first treatments are delivered, the hospitals, where more complex cases are managed, and public health premises, where strategic and organizational decisions are taken on the basis of the analysis of patients' populations. To be able to apply the portfolio of analysis models developed in the project to the data of the different participating centers, a common data model was defined and implemented using the i2b2 technology [15]. In the MOSAIC system, the i2b2 data warehouse allows integrating clinical information coming from hospital EHRs, administrative data from the local health care agencies, and environmental data collected from satellites [27]. The data model is defined using the i2b2 core ontology, and data are loaded to the repository thanks to ETL procedures properly defined at each participating center.

The MOSAIC system has been designed as a Service Oriented Architecture (SOA), where different components from different modules access the whole functionality of the system through a set of Web Services (Figure 2). These components are linked together in an asynchronous way through a message oriented architecture and interact with each other over internet using SOAP messages, conveyed using HTTPS and other Web standards. All of them work in a collaborative way thanks to the orchestrator, a deployment engine that allows the distributed coordination among different modules and services based on choreography principles. The choreographer is a system that allows the intercommunication among services providing tools for registering, multicast and broadcast communication as well as message filtering. This approach allows an easy deployment, efficiency, and independence from the programming language thanks to the intercommunication among services developed in different platforms and technologies.

Both the MobiGuide and Mosaic systems have been validated in pilot trials involving Foundation "S. Maugeri" Hospital to prove the feasibility of the approach described in this paper. In particular the MobiGuide system has been successfully validated in a 3-months-long study involving 10 patients form the atrial fibrillation domain and 20 patients with gestational diabetes (diabetic patients were enrolled at the Parc Tauli' Hospital in Sabadell [28]). The MOSAIC system is currently being evaluated at the Endocrinology Unit of the Fondazione Salvatore Maugeri Hospital. Physicians working at the Diabetology outpatient service have started a pre-post evaluation phase, where they have been seeing patients without the system for two months and they are performing visits using the system for three months. During this period, a set of interesting variables are being monitored, and will be compared at the end of the study to evaluate the benefits of introducing the system in the clinical practice. Of the MOSAIC cohort, which consists of 1000 patients, up to now, 500 patients have been analyzed during the phase without the system and 440 patients have been visited using MOSAIC. In parallel, a set of meetings have been organized among physicians working at the hospitals and healthcare managers working at the local healthcare agency of Pavia, to evaluate the tool working on patients' populations.

## 4 Conclusions

Modern DSSs need to interact with multiple, distributed and heterogeneous data sources. For these reasons, effective data integration strategies are an essential need for medical informatics solutions and even more for those intended to support decision processes. Methodologies for creating and maintaining centralized data repositories allow building DSSs on top of single data providers whose data model has been specifically designed to meet the requirements of integrated decision and analytical processes. Moreover, distributing the responsibility of adapting to the repository to the ETL procedures at the single centers improves system scalability, maintainability and provides the possibility of integrating additional informa-

 

tion sources even at late stages of the development or after deployment. Two different implementations of the described approach have proven to be successful in the two presented research projects.

### Acknowledgements

The work carried out in the MobiGuide and Mosaic projects has been supported by the European Union's Seventh Framework Program for research, technological development and demonstration under the grant numbers 287811 and 600914.

## References

[1] Greenes RA, editor. Clinical Decision Support: the road to a broad adoption. Oxford: Academic Press; 2014.

[2] Roshanov PS, Fernandes N, Wilczynski JM, Hemens BJ, You JJ, Handler SM, Nieuwlaat R, Souza NM, Beyene J, Van Spall HGC, Garg AX, Haynes RB. Features of effective computerised clinical decision support systems: meta-regression of 162 randomised trials. BMJ. 2013;346:f657.

[3] Bailey KJ, Little JP, Jung ME. Self-Monitoring Using Continuous Glucose Monitors with Real-Time Feedback Improves Exercise Adherence in Individuals with Impaired Blood Glucose: A Pilot Study. Diabetes Technol Ther. 2016 Feb 17;

[4] Meng K, Musekamp G, Schuler M, Seekatz B, Glatz J, Karger G, Kiwus U, Knoglinger E, Schubmann R, Westphal R, Faller H. The impact of a self-management patient education program for patients with chronic heart failure undergoing inpatient cardiac rehabilitation. Patient Educ Couns. 2016 Feb 16;

[5] Dagliati A, Marinoni A, Cerra C, Decata P, Chiovato L, Gamba P, Bellazzi R. Integration of Administrative, Clinical, and Environmental Data to Support the Management of Type 2 Diabetes Mellitus: From Satellites to Clinical Care. J Diabetes Sci Technol. 2015;10(1):19–26.

[6] Welch BM, Kawamoto K. Clinical decision support for genetically guided personalized medicine: a systematic review. Journal of the American Medical Informatics Association. 2013 Mar 1;20(2):388–400.

[7] Quaglini S, Sacchi L, Lanzola G, Viani N. Personalization and Patient Involvement in Decision Support Systems: Current Trends. Yearb Med Inform. 2015 Aug 13;10(1):106–18.

[8] MobiGuide - Home [Internet]. [cited 2014 May 4]. Available from: `http://www.mobiguide-project.eu/`

[9] Mosaic [Internet]. [cited 2016 Mar 2]. Available from: `http://www.mosaicproject.eu/`

[10] Kawamoto K, Del Fiol G, Lobach DF, Jenders RA. Standards for scalable clinical decision support: need, current and emerging standards, gaps, and proposal for progress. Open Med Inform J. 2010;4:235–44.

[11] González-Ferrer A, Peleg M. Understanding requirements of clinical data standards for developing interoperable knowledge-based DSS: A case study. Computer Standards & Interfaces. 2015 Nov;42:125–36.

[12] Beale T. Archetypes: Constraint-based domain models for future-proof information systems. OOPSLA 2002 workshop on behavioural semantics. 2002.

[13] Detmer D, Bloomrosen M, Raymond B, Tang P. Integrated Personal Health Records: Transformative Tools for Consumer-Centric Care. BMC Med Inform Decis Mak. 2008 Oct 6;8:45.

[14] HL7 Standards Product Brief - HL7 Version 3 Standard: Clinical Decision Support; Virtual Medical Record (vMR) Templates, Release 1 [Internet]. [cited 2015 Jul 16]. Available from: `https://www.hl7.org/implement/standards/product_brief.cfm?product_id=339`

[15] i2b2: Informatics for Integrating Biology & the Bedside [Internet]. [cited 2016 Feb 26]. Available from: `https://i2b2.org/index.html`

[16] Johnson PD, Tu SW, Musen MA, Purves I. A virtual medical record for guideline-based decision support. Proceedings / AMIA . Annual Symposium AMIA Symposium. 2001;294–8.

[17] Peleg M, Keren S, Denekamp Y. Mapping computerized clinical guidelines to electronic medical records: Knowledge-data ontological mapper (KDOM). Journal of Biomedical Informatics. 2008 Feb;41(1):180–201.

[18] Post AR, Krc T, Rathod H, Agravat S, Mansour M, Torian W, Saltz JH. Semantic ETL into i2b2 with Eureka! AMIA Jt Summits Transl Sci Proc. 2013;2013:203–7.

[19] Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. J Am Med Inform Assoc. 2016 Feb 5;

[20] Majeed RW, Röhrig R. Automated realtime data import for the i2b2 clinical data warehouse: introducing the HL7 ETL cell. Stud Health Technol Inform. 2012;180:270–4.

[21] Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, Kohane IS. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009 Oct;16(5):624–30.

[22] Peleg M. Computer-interpretable clinical guidelines: A methodological review. JBI. 2013 Aug;46(4):744–63.

[23] Sacchi L, Fux A, Napolitano C, Panzarasa S, Peleg M, Quaglini S, Shalom E, Soffer P, Tormene P. Patient-tailored workflow patterns from clinical practice guidelines recommendations. Stud Health Technol Inform. 2013;192:392–6.

[24] Marcos C, González-Ferrer A, Peleg M, Cavero C. Solving the interoperability challenge of a distributed complex patient guidance system: A data integrator based on HL7's Virtual Medical Record standard. JAMIA. 2015 Apr 15;22(3):587–99.

[25] Arturo González-Ferrer, Mor Peleg, Enea Parimbelli, Erez Shalom, Carlos Marcos Lagunar, Guy Klebanov, Iñaki Martínez-Sarriegui, Nick Lik San Fung, Tom Broens. Use of the Virtual Medical Record Data Model for Communication among Components of a Distributed Decision-Support System. Proceedings of 2nd IEEE Biomedical and Health Informatics International Conference. Valencia; 2014.

[26] Martinez-Millana A, Fernandez-Llatas C, Sacchi L, Segagni D, Guillen S, Bellazzi R, Traver V. From data to the decision: A software architecture to integrate predictive modelling in clinical settings. Conf Proc IEEE Eng Med Biol Soc. 2015 Aug;2015:8161–4.

[27] Bellazzi R, Dagliati A, Sacchi L, Segagni D. Big Data Technologies: New Opportunities for Diabetes Management. J Diabetes Sci Technol. 2015 Sep;9(5):1119–25.

[28] García-Sáez G, Rigla M, Martínez-Sarriegui I, Shalom E, Peleg M, Broens T, Pons B, Caballero-Ruíz E, Gómez EJ, Hernando ME. Patient-oriented Computerized Clinical Guidelines for Mobile Decision Support in Gestational Diabetes. J Diabetes Sci Technol. 2014 Mar 6;1932296814526492.

 