From Data to the Decision: A Software Architecture to Integrate Predictive Modelling in Clinical Settings

A. Martinez-Millana, C. Fernandez-Llatas *Member IEEE*, L. Sacchi, D. Segagni, S. Guillén *Member IEEE,* R. Bellazzi *Member IEEE* and V. Traver *Member IEEE*

***Abstract*— The application of statistics and mathematics over large amounts of data is providing healthcare systems with new tools for screening and managing multiple diseases. Nonetheless, these tools have many technical and clinical limitations as they are based on datasets with concrete characteristics. This proposition paper describes a novel architecture focused on providing a validation framework for discrimination and prediction models in the screening of Type 2 diabetes. For that, the architecture has been designed to gather different data sources under a common data structure and, furthermore, to be controlled by a centralized component (Orchestrator) in charge of directing the interaction flows among data sources, models and graphical user interfaces. This innovative approach aims to overcome the data-dependency of the models by providing a validation framework for the models as they are used within clinical settings.**

I. Introduction

Pre-diabetes is a pathological condition related to glucose metabolism deregulation and in which not all the diabetes criteria are met [1]. It is usually classified into Impaired Fasting Glycaemia (IFG) and Impaired Glucose Tolerance (IGT). IFG is characterized by elevated fasting blood glucose compared to the normal levels but is not high enough to be classified as Diabetes Mellitus (DM). Nonetheless, as the prevalence of DM is experimenting an outrage rising all around the world, and knowing that Type 2 Diabetes Mellitus (T2DM) accounts for more than the 95% [2] [3] researchers and epidemiologists are making efforts to produce classification algorithms and predictive models to understand why do the individuals develop this type of diabetes [4] [5]. Currently, there are studies assessing the improvement in the outcome of an early detection these pre-diabetic stages [6]. In this context, the use of modelling techniques has become highly developed in the UK with a wide range of industry-standard, research-based tools on the market [7]. Although in the European countries screening questionnaires continue to be extensively used to collect source data, in the NHS, where there is a consistent set of patient records in both secondary and primary care, predictive modelling has become largely the tool of choice. An application of the mentioned modelling techniques is the computation of risk scores. A risk score should accurately estimate the risk of a subject to develop a certain clinical condition. This scoring can be either based on discrimination, which classifies the individual on the basis of high, mid and low probability to go on that clinical condition, or prediction, if the individual is likely to develop that condition in the future. Discrimination and prediction algorithms are statistical models that combine information from several sources of data. Common types of models include logistic regression models, Bayesian networks, Support Vector Machines, Neural Networks, Cox proportional hazards models, and classification trees among others. Each type of model produces, for an individual, a predicted risk based on the information used to develop that model. However, various statistical and clinical factors may lead a model to perform poorly when applied to other individuals. It may happen that a model prediction is not reproducible because of deficiencies in the baseline data or modelling methods used in the study in which the model is derived, mostly due to over-fitting, differences between patient characteristics, measurement methods, healthcare systems particularities or data quality. The main ways to assess or validate the performance of a model on a new dataset are to compare observed and predicted event rates (calibration) and to quantify the model's ability to distinguish between patients who do or do not experience the event of interest (discrimination) [8]. There are two methodologies to carry out the validation of a model: internal validation and external validation [9].

A proper validation requires a full specification of the existing model (that is, both the input variables and their weights) to predict outcomes for the patients with the current dataset and then compare these predictions with the patients' prospective outcomes. Few predictive models are routinely used in clinical practice, probably because most have not been externally validated [10]. Moreover, the majority of the models published in the literature require the collection of data which is not available in the healthcare system, as it is obtained under the execution of a specific clinical trial [11]. To be considered useful, a risk score should be clinically credible, accurate (well calibrated with good discriminative ability), have generality (be externally validated), and, ideally, shown to be clinically effective—that is, provide useful additional information to professionals that improves decision making and thus patient outcome. It is crucial to quantify the

A. Martinez-Millana and S Guillen are with the company Tecnologias para la Salud y el Bienestar S.A. Ronda Auguste y Louis Lumiere 23, Paterna , Valencia 46980, Spain (corresponding author, phone: +34 96 387 7606; fax: +34 96 387 7279; e-maisl: {anmarmil; sguillen}@tsbtecnologias.es)

C. Fernández-Llatas and V. Traver are with SABIEN-ITACA Institute Universitat Politècnica de València, València 46022 Spain (e-mails: cfllatas@itaca.upv.es; vtraver@itaca.upv.es).

A Martinez-Millana, C. Fernandez Llatas and V Traverare also within Unidad Mixta de Reingeniería de Procesos Sociosanitarios (eRPSS), Instituto de Investigación Sanitaria del Hospital Universitario y Politécnico La Fe, Bulevar Sur S/N, Valencia 46026, Spain

L Sacchi and R Bellazzi are with Dipartimento di Ingenieria Industriale e de'llInformazione in Università degli Studi di Pavia, via Ferrata 1 in Pavia 27100, Italy (e-mails: lucia.sacchi@unipv.it; riccardo.bellazzi@unipv.it)

D. Segagni is with Laboratorio di Informatica e Sistemistica in Fondazione Salvatore Maugeri IRCCS, via Roncaccio, 16 Tradate 21049 Italy (e-mail: daniele.segagni@fsm.it)

978-1-4244-9270-1/15/$31.00 ©2015 IEEE

performance and importance of a predictive model on a new series of patients before applying the model in daily practice to guide patient care [9]. There are several criteria for assessing the selection of a decision support tool, but it should include the wide-known indicators as effectiveness (Sensitivity and Predictive Value), the predictive power and application to all risk categories [12]. Moreover, it should be also considered its accessibility to the clinical staff, the possibility for time-line evaluations (provide a baseline to evaluate the intervention over time or costs), ease of use and positioning to support wider considerations [13]. In this context, a proper use of Information and Communication Technologies (ICT) can be a solution to overcome the problem of meeting some of the needs reported above. The specification of a model is usually approached by mathematicians and biomedical engineers, then the model is wrapped into software pieces by designers and computer engineers and finally used by clinicians in a web or desktop application. The interaction of these stakeholders during the design, development and release of the decision support tool for the pre-diabetic screening is a process that has to be coordinated and well-documented. Workflow technologies [14] allow the specification and complex-process definition by non-programming experts. However, although workflows can describe process flows, they are not able to execute the interaction among different components. Workflows are designed to be orchestrated as the orchestration assumes that the flow is managed by a central component which decides the next step in the process [15]. In this paper a novel architecture to overcome the main limitations of the validation of discrimination and prediction models is presented. We herein define a principal aim: to provide a platform capable of performing both internal and external validation. This principal aim surrounds two secondary aims: to use a common data repository structure integrating several data sources (SA1); to be based on the controlled execution of independent components (SA2). The scope of the manuscript is to provide high-level description of a framework in which researchers, engineers and clinicians can develop, validate and use tools for the screening of T2DM.

II. MATERIALS AND METHOD

The design of a comprehensive and dynamic architecture has been approached from a set of initial baseline requirements. These requirements have been classified into three conceptual modules: Data Storage, Models and User Interfaces. Each of them must provide a set of functionalities among them with the peculiarity that the interaction is not intended to be 1:1 and fixed over time. The relationships may be 1:n and evolve during the system performance.

In order to create a dynamic distributed architecture accessible to modify the interaction flow and perform a hot-update of the models, the concept of Service Oriented Architecture [16] was identified as the best fitting solution. The use of SOA paradigm assumes the distribution of functionalities all across the system wrapped into services with a common defined language to exchange messages. There are two strategies to define this common language: syntactical and semantic [17]. The use of syntactical messages ensures the correct understanding of the data being exchanged. However, it limits the ability of services to understand the message content and thus, can affect their performance. The semantic approach is an alternative that provides advantages for the previous limitations, and hence, the services themselves are capable to understand the message content and underpin the requested actions in a most efficient way. The use of ontologies to semantically describe the services in the system gives the chance to the SOA to coordinate the functionalities from three main modules of the proposed architecture, detailed in the following sections.

A. Data Storage Module

Five different data sources were available: Local Healthcare Agency of Pavia (ASL), IRCSS Fondazione Salvatore Maugeri from Pavia (FSM), Health Department Hospital La Fe (LF), and the Hippokration Hospital in Athens (HP) and finally data collected during Botnia-PPP clinical study [18]. Data from hospital settings included parameters regarding patients consumption of hospital services (planned/unplanned visits), laboratory tests, drug prescriptions, date and type of diagnose among others. After a descriptive analysis of all the parameters, variables and structure of the retrieved datasets, the objective was to build a common-mapping with the following goals: First, to identify common parameters and homogenize the representation of the variables (measurement units, timestamp, coding system and representation) and second, to define a common data structure and transform it into a common ontology to gather the mentioned datasets – even they are from different countries and purposes. The main goal is to provide a unique conceptual module to logically store the data sources using a flexible common ontology. The Integrating Biology and the Bedside (I2B2) technology [19] was used for this purpose. One of the goals of I2B2 is to provide clinical investigators with the software tools necessary to collect and manage project-related clinical research data in the big data as a cohesive entity, a software suite to construct and manage the modern clinical research chart. The I2B2 splits the data storage from the data management. On the one hand, data is stored in Data-Warehouses, a dedicated virtual machine that acts as a server, capable to run different engines (MySQL, Oracle and SQLServer). Beneath it, he "hive" acts as the Data Access Layer. The HIVE is a set of software modules called "cells" that have a common messaging protocol that allow them to interact using web services and Extended-Markup Languages (XML) messages. The data gathering strategy and organization has been already published by coauthors **[20]**.

A. Model Host Module

The Models Host Module intends to be the container of the Discrimination and Prediction models. According to the mathematical method and the purpose of the model, two model families have been defined: First, the Bayesian Network model, which provides the functionalities of value imputation and individuals' discrimination among three classes: IFG, IGT and type 2 pre-diabetic [21]. This model is wrapped into an executable file built in R language, an open source statistical language developed by the Carnegie Mellon University [22]. Second, the Support Vector Machine model for ranking a set of patients in terms of risk of T2DM onset in the future. It is meant to work on a population to stratify all patients into higher/lower risk of developing the disease. The model has been developed using the same language as the Bayesian Network.

C. *Interface Module*

The interface module contains the Graphical User Interfaces (GUIs), which are software components that provide control and functionalities to the end users: the clinicians. These interfaces should be able to deal with the different data sources, the execution of the models and display results in a user-friendly fashion. Depending on the client application of each clinical setting the plugin should be implemented in Java or under .NET framework. To be as much as compliant with the criteria of tools acceptability [13], the target of the plugin wrapper will be to embed the discrimination or prediction tool functionality into the already existing hospital software application.

III. RESULTS

The proposed architecture is based on three main modules which exchange information through the distributed instantiation of services. This architecture is associated to the traditional three tier model (Data-Business-User Interface) but adapted to tackle the PA and subsequently SA1 and SA1. Figure 1 shows a schema of this architecture. Data Storage Module (DSM) gathers data warehouses form the mentioned data sources and provides access to them to the whole system. The Models Host Module (MHM) groups the algorithms and engines which contain discrimination and predictive models. This module also contains the Orchestrator Component (Figure 1), which is the functional director that controls all the service interactions across the system. The Interface Module (IFM) is actually a virtual module (physically hosted in the Models Host module but deployed on client side such a web browser or disease management application) that provides the functionalities to the end users. The core of the architecture is the message dispatcher engine (the Orchestrator) and a data base that contains the services that are registered and semantically described within it. The services may be connected to the core locally, when the services are allocated in the same server of the MHM (e.g.: Bayesian Network), or remotely by using a TCP protocol service wrapper (e.g.: Data Access Services in the DSM). An ontology reasoner is connected to the Orchestrator and is able to infer knowledge from registered services where semantic information is available. Using this approach, the end users and researchers can create their own interaction flows by defining execution flows and loading them into the Orchestrator. The Orchestrator component dispatches messages among the modules using a specific XML message protocol called XMSG [23]. This protocol is based on the combination of FIPA [24] and SOAP [25] protocols. The classic FIPA protocol, defined for Multi Agent Systems communication allows sharing knowledge using several protocols. XMSG is based on FIPA headers to route and characterize the messages. At the same time, he content of XMSG is based in the SOAP protocol. SOAP is a well-known and widely used protocol to perform service calls. The XMSG protocol allows broad and multicast as well as Peer to Peer (P2P) message calls using custom symbols in the destination address.

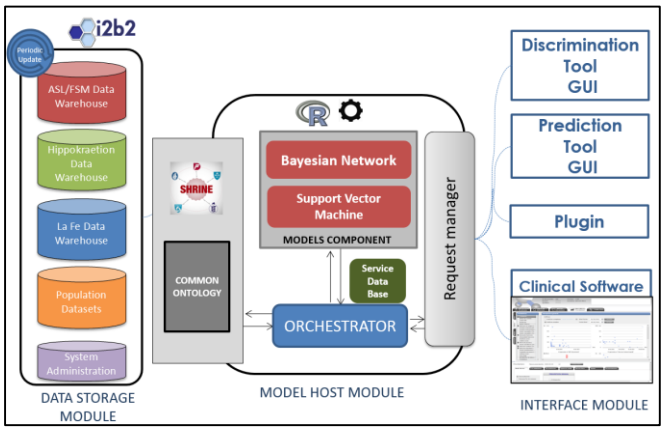

Figure 1. Service Oriented Architecture based on process orchestration. The architecture is based on the controlled interaction of three modules: Data storage, Model Host and Interface. The centralized control is performed by the Orchestrator, which is able to run the models over different datasets simultaneously.

As an example: a message is sent from a service, say PredictiveService, whose logical address is Models.R.BayesianNetwork, to the PredictionGUIService, whose logical address is HTML5.PredictionGUIService. Both sender and receiver information and the type of message sent (request, inform, event…) are defined in the message header. Following it, in the content part of the message, the call to the specific method of the service is defined. In this example, the method invoked is Execute Prediction on a previously selected population, which needs clinical data from a specific patient. Figure 1 shows the main services and components involved in the proposed architecture. From left to right, the schema shows the data storage module, based on the I2B2 technology and development framework. This module contains several single data entities, each of them gathering data from different sources: Hospital Data Warehouse (from FSM, Hyppokration and La Fe) and Population Datasets from the previous Botnia-PPP study. From a logical point of view, the Data Storage Module is a unique conceptual part from the overall architecture and relies on the defined Common Ontology (CO), however, from a physical point of view, each warehouse and data set is an isolated virtual machine located elsewhere and reachable though the internet network. The connection of the DSM and the MHM is performed by the "shrine" service layer (grey shaded). It is a set of interoperability services that allow performing federated queries to the whole data storage warehouses, regardless its physical location and data structure, thanks to the use of the CO. This configuration permits researchers and clinicians to choose which is the target population of the queries, and furthermore, enables them to perform internal and external validation of the models with different data sources.

The services are gathered within the Orchestrator Data Base (deep green).This component is in charge of executing the predefined workflows for each tool and model. This kind of complex process execution is solved by using process orchestration which assumes that the processes are able to exchange data to execute processes in a distributed way. This means that the orchestrated processes are independent and can communicate with each other, in what we know as the "defined execution flow" (workflow). An example of

workflow could be: Retrieve Data a data warehouse- Execute Bayesian Network Algorithm- Display Results. Using this architecture, it is possible to connect and disconnect components and modules dynamically if an external validation is needed (e.g.: change the data warehouse). Components can offer their functionalities and services can consume them without the necessity of knowing the concrete architecture of the deployed service. This facilitates the aggregation of new data sets, the update of new models or the validation of them against different data sources.

IV. CONCLUSION

This manuscript proposes a semantically tagged service oriented architecture based on process orchestration for the exploitation of discrimination and prediction models in the screening of T2DM. The reasons why clinicians and researchers are not prone to use predictive modelling are identified as a lack of reliability and inadequacy of the models validation, as in most cases it is done just as internal validation. The presented architecture proposes a centralized coordination of services by merging the main three components needed to overcome the mentioned barriers. By defining a common exchange messaging format and a semantic definition of the services, the proposed architecture is capable of modifying the interaction flow to improve the outcome. Using I2B2 technology, a set of five different data sources have been integrated from a conceptual approach, by defining a common ontology that embraces all the different parameters across them into the DSM, meeting the defined SA1. Next to this module, the Model Host Module gathers the internally validated discrimination (Bayesian Network) and predictive (Support Vector Machine) models to be assessed with the data sources in the DSM. These modules use the model script code to generate automatically the executable model to be used on the Discrimination and Prediction of T2DM independendently (meeting SA2), enabling to make improvements in the model performance without the need of re-debugging the entire module. In this terms, a model can be externally validated within the same system infrastructure, and thus, be provided to the end users through the integration of the discrimination or predictive tool in the current software management system used in the clinical setting (meeting the defined PA). The proposed architecture is being deployed currently in two hospitals: IRCSS Fondazione Salvatore Maugeri from Pavia in Italy and Health Department Hospital La Fe from Valencia in Spain. Future work looks at assessing the performance of the architecture in a real setting with real end users. Security issues and traceability of the architecture will be also tackled.

ACKNOWLEDGMENT

The authors wish to acknowledge the consortium of the MOSAIC project (funded by the European Commission, Grant nr. FP7-ICT 600914) for their commitment during concept development.

REFERENCES

[1] Nichols GA, Hillier TA, Brown JB. "Progression From Newly Acquired Impaired Fasting Glusose to Type 2 Diabetes" *Diabetes Care* vol. 30, no. 2, pp. 228–233, Feb. 2007

[2] Wild S, Roglic G, Green A, Sicree R, King H. "Global Prevalence of Diabetes. Estimates for the year 2000 and projections for 2030" *Diabetes Care* vol. 27, no. 5, pp. 1047-1053, May 2004

[3] Shaw JE, Sicree RA, Zimmet PZ. "Global estimates of the prevalence of diabetes for 2010 and 2030" *Diabetes Res Clin Pract* vol. 87 no. 1, pp. 4-14, Jan 2010

[4] Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P. "Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QRisk2" *BMJ* vol 336, pp. 1475-1482, May. 2008

[5] Meigs JB, Schrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, et al. "Genotype score in addition to common risk factors for prediction of type 2 diabetes" *N Engl J Med*. vol. 359, pp. 2208-2219, Nov. 2008

[6] Gillies CL, Lambert PC, Abrams KR, Sutton AJ, Cooper NJ, Hsu RT, et al. "Different strategies for screening and prevention of type 2 diabetes in adults: cost effectiveness analysis" *BMJ* vol. 336, pp. 1180-1185, May 2008

[7] Tremblay M. Predictive health: policy for predictive modelling and long-term health conditions. United Kingdom: Department of Health, 2005

[8] Harrell FE Jr, Lee KL, Mark DB. "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors". *Stat Med*. vol. 15, pp. 361-387, Feb. 1996

[9] Altman D, Vergowe Y, Royston P, Moons K. "Prognosis and prognostic research: validating a prognostic model. Research Methods & Reporting" *BMJ* 338, pp. 1432-1435, May 2009

[10] Reilly BM, Evans AT. "Translating clinical research into clinical practice: 25 impact of using prediction rules to make decisions" *Ann Intern Med* vol. 144, pp. 201-209, Feb. 2006

[11] Noble D. Mathur R, Dent T, Meads C, Greenhalgh T. "Risk models and scores for type 2 diabetes: systematic review". *BMJ* vol 343, pp.1-31, Nov. 2011

[12] Steyberg E, Vickers A, Cook N et altres. "Assessing the performance of prediction models: a framework for some traditional and novel measures" *Epidemiology* vol 21 num. 1, pp. 128-138, Jan. 2010

[13] Cochrane Tools Assessment. United Kingdom:Tribal, 2008

[14] Workflow Management Coalition. *Workflow Management Coalition Terminology Glossary*. 1999.

[15] Fernández-Llatas C, Mocholí JB, Sala P and Naranjo JC. "Process choreography for human interaction computer-aided simulation" *Proceedings of the 14th international conference on Human-computer interaction: design and development approaches* Berlin, 2011, pp. 214-220

[16] Erl T. *Service-Oriented Architecture: Concepts, Technology and Design*. New Jersey: Prentice Hall, 2005

[17] Cena F, Furnani R. "A SOA-based Framework to Support User Model Interoperability" *Adaptive Hypermedia and Adaptive Web-Based Systems. Lecture Notes in Computer Science* vol 5149, pp. 284-287, Jul. 2008

[18] Pyykkönen AJ, Räikkönen K, Tuomi T, Eriksson JG, Groop L, Isomaa B. "Stressful life events and the metabolic syndrome: the Prevalence, Prediction and Prevention of Diabetes (PPP)-Botnia Study". *Diabetes Care* vol. 33, pp. 378–384, Feb. 2010

[19] http://www.i2b2.org. Last access April 2015.

[20] Dagliati A, Sacchi L, Bucalo M et altres. "Data Gathering Framework to Collect Type 2 Diabetes Patients Data" *EMBS Biomedical and Health Informatics Conference* (BHI) Spain, 2014 pp. 244-247

[21] Sambo F. Facchinetti A et altres. "A Bayesian network for probabilistic reasoning and imputation of missing risk factors in type 2 diabetes". *AIME* Italy 2015 unpublished

[22] http://www.r-project.org/ last access February 2015

[23] Fernandez-Llatas C, Mocholi JB, Moyano A, Meneu T. "Semantic Process Chorepgraphy for Distributed Sensor Management" *International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management – SSW* Spain 2010, pp. 31-37

[24] P. D. O'Brien and R. C. Nicol. "Fipa towards a standard for software agents". *BT Technology Journal*, vol. 3, pp. 51-59, Jul. 1998.

[25] Henrik F. Nielsen, Noah Mendelsohn, Jean J. Moreau, Martin Gudgin, and Marc Hadley. "SOAP version 1.2 part 1: Messaging framework" W3C recommendation, Jun. 2003.