Towards a universal architecture for disease data models sharing and evaluation

Ali-Eldin, Amr M.T.; Hafez, Eman Abdelrazek

DOI
10.1109/ISNCC.2017.8071995

Publication date
2017

Document Version
Final published version

Published in
Proceedings of International Symposium on Networks, Computers and Communications, ISNCC 2017

Towards A Universal Architecture for Disease Data Models Sharing and Evaluation

Amr M. T. Ali-Eldin [1,2,3] and Eman Abdelrazek Hafez [4]

[1] Leiden Institute of Advanced Computer Science, Leiden University, Leiden, the Netherlands.

[2] Faculty of Technology, Policy and Management, Delft University of Technology, Delft, the Netherlands.

[3] Computer and Control Systems Dept., Faculty of Engineering, Mansoura University, Mansoura, Egypt.

[4] Rheumatology and Rehabilitation Department, Faculty of Medicine, Mansoura University, Mansoura, Egypt.

Abstract— Usually, medical researchers find it cumbersome to find disease data profiles that suit their research experiments requirements. In this paper, we propose a functional architecture where medical researchers can share disease data profiles after removing patients' sensitive information. In addition, the proposed architecture is equipped with some features that facilitate collaborative discussions among researchers. Besides, some machine learning techniques are adopted for analysis and modelling of disease datasets. This way, it is expected that medical researchers can better collaborate together and perform their researches on larger patient samples obtaining more accurate and representative results. The main functionalities of the architecture are introduced. One component of the architecture, which is the evaluation engine, was implemented using Matlab showing its advantages as a tool for researchers. The case of analyzing a model for a chronic disease known as Juvenile idiopathic arthritis has been studied. Obtained results show the applicability and effectiveness of the proposed approach.

Keywords— ***rare and chronic diseases; shared architectures; privacy; access control; Disease Data Model; Juvenile Idiopathic Arthritis (JIA); Adaptive Network Fuzzy Inference Systems (ANFIS).***

I. INTRODUCTION

Rare diseases are difficult to observe especially when there is little collaboration or integration between clinics and hospitals. This is because of the lack of a common patient record infrastructure or an integrated medical system across the country. Therefore, physicians find difficulties in collecting needed disease data for their medical research. Most of the time medical researchers complain about the lack of enough information about certain rare and chronic diseases.

In this work, we propose and define the architectural components of a framework for collecting rare disease data and share it among medical researchers. The architecture provides basic functionality for sharing disease data such as search; download; publish; and evaluate diseases. Further, we show how to effectively model diseases by studying the case of Juvenile Idiopathic Arthritis (JIA) patients. An assessment model for health status of JIA patients is built to automate the process of disease evaluation. This model is retained in the architecture repository after removing all patients' sensitive and identifying information. The dataset, collected at Mansoura University Hospital (MUH), is used to analyze and evaluate the disease model. Further a survey is conducted with a number of physicians to evaluate the architecture usefulness and applicability.

It is expected that other medical studies outcomes will also be retained in this shared architecture and disease assessment models will be created. The use of such architecture will help in fine tuning these models and making sure they provide accurate results. In addition, it will make research profiles available for other medical research purposes. This architecture will eventually become a big data repository of all disease profiles that can be shared among medical hospitals for research purposes.

This paper is organized as follows: in section II, the proposed functionalities are presented. Section III provides an overview of the studied disease. Section VI introduces the collected data and how it was prepared. Section V provides the experimental work while section VI provides results of the evaluation survey. Section VII evaluates obtained results and discusses related work. Section VIII concludes the paper with recommendations for future work.

II. PROPOSED ARCHITECTURE FUNCTIONALITIES

In this section, the needed functionalities for such an architecture are discussed followed by the possible technical implementations.

A. Disease Data Models

In order to store disease information in the architecture, a master data model is needed for diseases. The cause effect model is used to model a disease as shown in Fig. 1 where factors cause (influence) the disease output that is to be measured. We can have till *n* factors. At the same time, there can be multiple outputs (*m*) to be measured for a disease.

978-1-5090-4260-9/17/$31.00 ©2017 IEEE

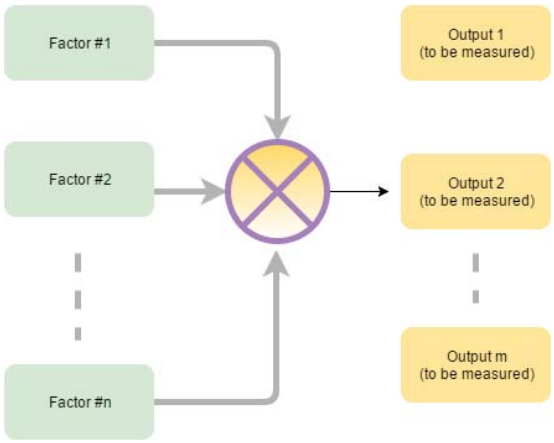

Fig. 1 Disease Data Model

B. *The Functional Architecture*

As mentioned earlier, an architecture is needed in order to share the different disease models among medical researchers remotely. Privacy techniques are needed in order to protect patients' privacy. For that, access control mechanisms and anonymity approaches are applied on the disease datasets. The disease data architecture will have the following functionalities (see Fig. 2):

- Search Diseases: the search functionality will be done by searching on some disease properties such as symptoms, diagnosis, disease management and statistics etc. Expected outcome is the disease data models showing the disease outputs and influencing factors.
- Publish Disease Data: a medical researcher can choose to publish disease data to the platform to be used by other researchers as well. Before publishing, a number of measurements need to be taken by the privacy filtration functionality first to guarantee patients privacy without impacting the quality of the data.
- Discuss and Evaluate Diseases: through this functionality medical researchers can discuss together in the form of an internal social site and share their opinions, on certain disease properties such as diagnosis or treatment, with other experts who have had similar experiences. Further this component implements some data mining and artificial intelligence techniques to evaluate disease properties and provide some reports.
- Access Management and Privacy Filtration: access control is needed in order to restrict access to this environment for authorized persons only. These may include: registered physicians, radiology technicians and decision makers etc. Further, sensitive data and privacy identifying information are removed before publishing by the privacy filtration component.
- Evaluation Engine: this is responsible for providing an autonomous evaluation of a disease based on input taken from medical staff and based on comparison to similar cases available in the repository. In this paper, the evaluation is done based on a clustered neuro-fuzzy technique to learn from collected data and create a generic data model for each disease. The fuzzy inference system is configured in run time to meet the newly collected dataset. More machine learning features and capabilities will be added to this component in future research such as deep learning, stochastic analysis, genetics etc.
- Request for Download: via this functionality, system users can have access to the data they want for a certain disease after taking the necessary approvals.
- User Roles: three types of users are distinguished; management, authorized users and operators. The management is responsible for granting access to users. Further, they approve diseases data being uploaded by the operators to be published in the environment. Physicians registered to the system are known as authorized users.

III. JUVENILE IDIOPATHIC ARTHRITIS *(JIA)*

In this work, the case of Juvenile Idiopathic Arthritis (JIA) is studied. JIA, a disabling inflammatory rheumatic disease affecting children, is accompanied by chronic arthritis and deformity [1]. By affecting joints this can have effect on the child normal growth. Functional disability is a common problem in JIA patients. A child with JIA suffers from pain, fatigue and decreased functional activity. In severe cases of JIA, patients can suffer from deformities in the upper and lower limb joints and as a result, lower quality of life and high absence percentage in schools [2, 3].

Since a complete cure is not possible in JIA patients, the primary aim is to manage this disease to ensure the best possible health status of patients with JIA [4]. Evaluation of patients' health status and prediction of its determinants is critical to enhance the achievement of such satisfactory health status [5]. Statistical analysis revealed a number of factors that significantly impact patients' health status. These factors will be presented later in the paper.

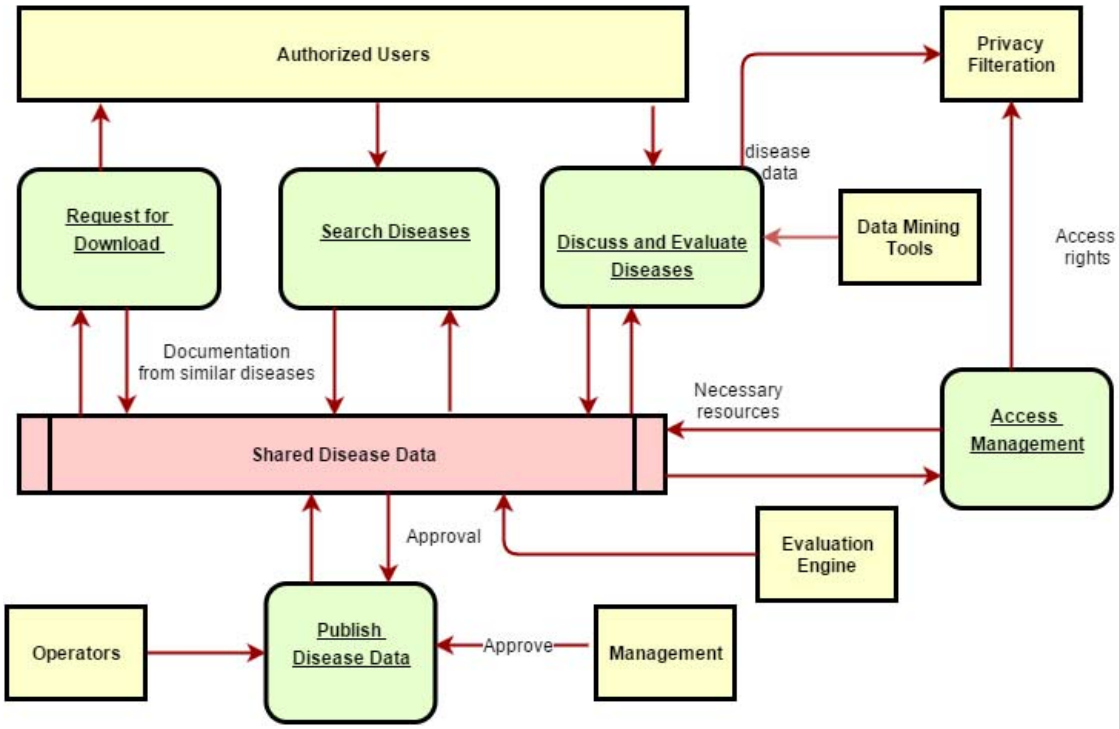

Fig. 2 The Proposed Architecture Functional Components

IV. Data Collection & Preparation

In this study, thirty two JIA patients were gathered from Mansoura University Hospital. The disease collected data was used for implementing the evaluation engine after removing patients' sensitive and identifying information such as names, age, gender and addresses etc. The collected data was divided into two sets; a training set and a testing set. Table 1 shows a selection of variables of the testing dataset while Table 2 shows the same selection in the training dataset.

Table 1 Testing Dataset

Input					**Output**
patient	***VAS* (0-3)**	***Joints* (0-3)**	***CHAQ* (0-3)**	***Activity* (0-3)**	***GE* (0-3)**
#1	0.1	1	2	2	0.9
#2	1.5	3	3	0	2.1
#3	0.3	3	2	1	0.6
#4	0.1	0	0.25	1	1.5
#5	1.6	3	2.6	3	2.1
#6	0.2	1	0.25	3	1.2
#7	1.5	3	1.395	2	1.8
#8	0.9	1	0.275	1	0.9
#9	0.2	3	0.75	1	0.9
#10	1.9	3	0.5	2	1.2
#11	2.0	3	1	3	1.8
#12	0.1	0	0.4	2	0.6
#13	0.2	1	0.3	2	0.6
#14	0.0001	0.1	0.0001	0	0

Table 2 Training Data

Input					**Output**
patient	***VAS* (0-3)**	***Joints* (0-3)**	***CHAQ* (0-3)**	***Activity* (0-3)**	***GE* (0-3)**
#16	0.6	1	0.4	1	0.9
#17	1	1	0.56	1	0.6
#18	2.5	3	1	3	2.4
#19	1	1	0.56	1	0.9
#20	0.7	1	0.56	1	0.4
#21	0.8	1	0.4	1	0.6
#22	0.5	1	1	1	0.6
#23	1.5	1	3	1	2.1
#24	0.3	3	2	3	0.6
#25	0.1	3	0.25	3	1.5
#26	1.6	0	2.6	0	2.1
#27	0.2	3	0.25	3	1.2
#28	1.5	1	1.395	1	1.8
#29	0.9	3	0.275	3	0.9
#30	0.2	1	0.75	1	0.9
#31	1.9	3	0.5	3	1.2
#32	0.1	0	0.12	0	1.8

V. Experimental Work

For simulation purposes, Matlab r2012 on a UNIX server was used offline to build the evaluation rule engine component. The proposed architecture is to be built using PHP and MySQL and will be hosted online using an Apache web server. It will be accessed from Internet using web interfaces through Representational State (REST) protocol services known as restful web services. Authorized users via their web browser can navigate to the PHP application website launching a user interface (web forms). Disease data transfer and system commands will be executed via

hypertext transfer protocol (HTTP) through Extensible Markup Language (XML) or JavaScript Object Notation (JSON). There are many frameworks that can be used to build PHP Restful web services where mostly the common ones are Zend [6] and Tonic [7].

A. Evaluation Engine

An evaluation model using a clustering fuzzy technique is created in the Evaluation Engine for each disease collected by the system. This model captures and visualizes the relationships between the different factors impacting the disease output to be measured. From the statistical analysis, health status was shown to be significantly correlated to factors such as pain (VAS), number of joints deformed, CHAQ (child health assessment questionnaire output), disease functional activity and body mass index (BMI). A fuzzy inference system of the type ANFIS is used to model these factors as fuzzy sets. Fuzzy logic is useful in creating generic models for better representation of the relationships between variables impacting certain outputs [8]. Data clustering is used to identify natural categories in the dataset which is helpful for better understanding of relationships embedded in the data. The clustering approach implemented in this work using Matlab is the subtractive clustering approach [9].

B. ANFIS Modeling

The ANFIS approach calculates the target value as follows:

Step #1: also known as fuzzification, the fuzzy system creates the fuzzy values from the corresponding input ones. Assuming number of input factors and membership functions are n and m respectively, then the k^{th} rule is applied by the fuzzy system as follows (k =1,m):

$$if\ Y_1 = Y_{1k}\ and\ Y_2 = Y_{2k} \dots\dots and \dots Y_n = Y_{nk}\ Then\ f_k = \sum_{j=1}^{n} q_{\ jk} * Y_j + r_j \quad (1)$$

Step #2: based on rules firing rates w_k, target weights are calculated per rule by aggregating them as follows:

$$w_k = \prod_{j=1}^{n} w_j \quad (2)$$

Step #3: from target weights, normalized weights $\overline{w_k}$ are calculated:

$$\overline{w_k} = \frac{w_k}{\sum_{k=1}^{m} w_k} \quad (3)$$

Step #4: at this step, the normalized output $\bar{f}_k$ is calculated:

$$\bar{f}_k = \overline{w_k} * f_k \quad (4)$$

Step #5: also known as the defuzzification phase, the final target is calculated:

$$f = \sum_{i=1}^{m} \bar{f}_k \quad (5)$$

Besides the previous steps, FIS adapts the parameters q_{kj}, r_j using the hybrid algorithm introduced in [10]. The hybrid algorithm uses the forward pass and the backward pass methods. The forward pass method applies least square to calculate the parameters in step #4 while the backward pass method propagates the errors backward and updates the parameters using the gradient descent approach.

C. Properties of the Fuzzy Inference System

Matlab generated fifteen rules using fifteen membership functions and a subtractive clustering approach. In addition, membership function types and parameters are selected that suit the dataset. The Gaussian membership function was selected such that:

$$f(Y, \sigma, d) = e^{\frac{-(Y-d)^2}{2\sigma^2}} \quad (6)$$

Where, Y represents the crisp value of the input variable, σ as the standard deviation, and d as the mean. Each input variable $\bar{Y}$ is represented using membership function values as follows:

$$\bar{Y} = \sum_{k=1}^{m} \mu_k * e^{\frac{-(Y-d_k)^2}{2{\sigma_k}^2}} \quad (7)$$

Where m represents the number of membership functions and μ as the membership value. Fig. 3 shows example of VAS input membership functions.

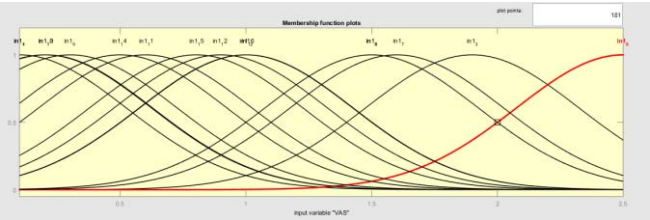

Fig. 3 VAS input membership functions

D. Results

In this section, we highlight how the evaluation engine could further assist physicians in studying the relationships between the different factors in the disease model from the collected patients' data. The evaluation engine provides some visual illustrations on the analysis of the data from which some conclusions can be made. For example, pain severity increase will have severe impact on GE in patients with more than two affected joints. By using the training functionality in the ANFIS model, the fuzzy output would be fine-tuned to meet the dataset pattern. The more data fed to the system, the more accurate becomes the disease data model. Fig. 4 shows the mean absolute error obtained. The evaluation engine can provide also visualization into the relationship among the different input factors and the disease output (see Fig. 5 & Fig. 6 - vertical axis is GE).

VI. Validation Results

There has been a number of methodologies for validating information systems frameworks [11]. Most mythologies focused on validating behavioral intention of use, perceived usefulness and ease of use. Other factors were also validated such as voluntaries, visibility and career opportunities [11]. The proposed framework was evaluated using a questionnaire with a set of six physicians. The conducted survey focused on behavioral intention of use, perceived usefulness and ease of use. The physicians were asked to evaluate the statements in Appendix A with values

ranging from 1 (strongly disagree) to 5 (strongly agree). The collected assessment results are shown in Table 3. Based on the collected answers from the survey, we can see that the proposed framework is useful and applicable. Further, we have seen that the framework can be implemented easily using REST web services which are known for their fast performance. The major challenge that needs still to be solved is convincing governmental authorities to support the sharing architecture initiative and to collect existing paper-based registries of rare diseases at the ministry of health.

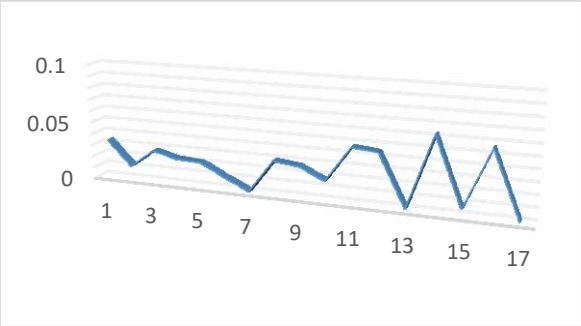

Fig. 4 mean absolute error between actual and predicted.

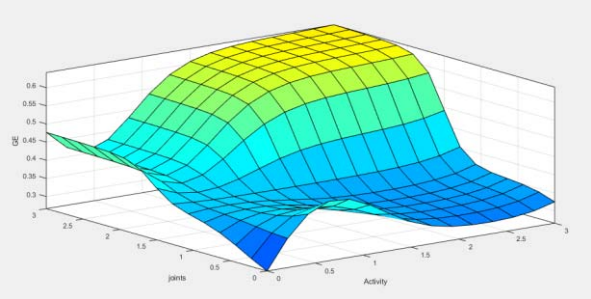

Fig. 5 Activity and number of affected joints effect on GE

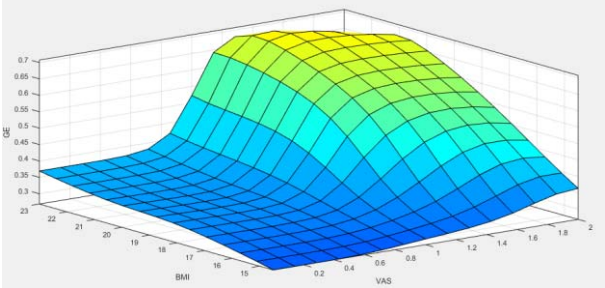

Fig. 6 Relation between pain (VAS), BMI and GE

VII. Results Analysis and Discussion

The idea of using a shared architecture for services provisioning could not be accomplished without the availability of applicable technical solutions. Such technical solutions have been there since more than two decades with the emergence of web services technology by Microsoft [12]. The concept of shared architectures was introduced for integrating large enterprise systems such as governmental agencies [13]. It has been applied in many cases where different companies merged together and their enterprise infrastructure needed to be integrated. The early implementations of this approach depended on the high flexibility offered by web services technology. A couple of years later, and with the development of service-oriented architecture paradigm, enterprises invested in big IT transformation projects towards the services approach leading to more adoption of web services and hence more developments in this paradigm. This lead to introduction of web2.0 technology resources such as Representational State Transfer (REST) protocol and RESTFUL web services. In this work, RESTFUL web services are recommended because REST provides better performance than SOAP web services. In addition, its implementation is easier than that of SOAP [14].

Table 3 Survey Outcome

Evaluation Criteria	Average Evaluation (max. is 5)	Comments
Behavioral Intention	4	Most physicians liked the approach and supported the idea
Perceived Usefulness	3.2	Most physicians believe in the usefulness of such approach. Some physicians opted in for the need for specifying more data types such as x-ray images
Ease of Use	3.5	Most physicians found no difficulty with using such platform

With the development of Health 2.0, attention has been given to web 2.0 technology integration in health platforms leading to the adoption of sharing medical records concept on a larger scale. However most of efforts in this area have focused on patients' records exchange in the form of commercial efforts connecting pharmacies, insurance companies and hospitals. Some examples are the electronic health records by Telus [15], and the shared medical platform for insurance underwriting [16]. However few efforts have worked on developing a shared architecture for rare diseases all over the world. RareConnect [17] provides a forum for connecting patients with rare diseases globally. One of the earliest efforts, Frost, Massagli [18] shows how a study on the effect of lithium on Amyotrophic Lateral Sclerosis (ALS) was triggered through an online community. The EPIRARE project [19] called up for building a platform to register rare disease in the EU and to address any related regulatory, ethical and technical issues. Taruscio et al. [20] argue that despite that most existing registries in Europe are based on rather unstructured efforts and are focused on one or group of related diseases, they are mature enough to be integrated in a uniform architecture.

Our approach, however, does not aim at providing registries but rather a platform for disease data exchange for research purposes. Therefore, there are less challenges than those applicable in the case of EPIRARE. We further showed how the use of intelligent techniques can help in evaluating disease data models that can be helpful for disease understanding and analysis.

VIII. CONCLUSIONS

Scarcity of diseases data represents a challenge for medical researchers. Medical researchers have to wait very long time till they are able to collect enough patients' records so that they can perform adequate experimental research. In this paper, this problem is tackled by proposing a shared architecture to collect and share data about rare and chronic diseases. Such platform functionalities were presented together with detailing the evaluation engine used to model disease data that need to be measured. Juvenile Idiopathic Arthritis (JIA) is used as a case study. Results show the applicability and effectiveness of the proposed approach. Further studies will be conducted on harvesting current rare disease registries and see how they can be connected to such a platform. Besides, we will be studying legal and ethical requirements for such a platform to succeed. Last and not least, more components of the proposed architecture will be implemented.

REFERENCES

[1] Arkela-Kautiainen M, H.J., Kautiainen H, Vilkkumaa I, Malkia E, Leirisalo-Repo M., Favourable social functioning and health related quality of life of patients with JIA in early adulthood. Ann Rheum Dis, 2005: p.; 64:875–80.

[2] Abdul-Sattar A, M.S., Negm MG., Associates of school impairment in Egyptian patients with juvenile idiopathic arthritis: Sharkia Governorate. Rheumatol Int, 2014: p. 34(1):35-42.

[3] Klotsche, J., Minden, K., Thon, A., Ganser, G., Urban, A. and Horneff, G., Improvement in Health-Related Quality of Life for Children With Juvenile Idiopathic Arthritis After Start of Treatment With Etanercept. Arthritis Care Res, 2014: p. 66: 253–262.

[4] Flato B, L.G., Smerdel A, Vinje O, Dale K, Johnston V, et al.,, Prognostic factors in juvenile rheumatoid arthritis: a case-control study revealing early predictors and outcome after 14.9 years. J Rheumatol, 2003: p.;30:386–93.

[5] Selvaag AM, F.B., Lien G, Sørskaar D, Vinje O, Førre Ø, Measuring health status in early juvenile idiopathic arthritis: determinants and responsiveness of the child health questionnaire. J Rheumatol., 2003: p.;30(7):1602-10.

[6] Zend Framework available at https://framework.zend.com - last visited on 11/9/2016.

[7] Tonic available at http://tonic.sourceforge.net/ - last visited on 11/9/2016.

[8] Zadeh, L., Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems 1(1), 1978: p. 3-28.

[9] Chiu, S.L., Fuzzy Model Identification Based on Cluster Estimation. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology 2(3), 1994: p. 267-278.

[10] Jang, J.S.R., "ANFIS: adaptive-network-based fuzzy inference system," IEEE Transactions on Systems, Man, and Cybernetics, , 1993: p. vol. 23, no. 3, pp., 665-685.

[11] Riemenschneider, C.K. and B.C. Hardgrave, Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Theoretical Models. IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, 2002. 28(12): p. 1135-1145.

[12] Josuttis, N.M., SOA in Practice: The Art of Distributed System Design. 2007: O'Reilly.

[13] Ali-Eldin, A.M.T., Towards A Shared Public Electronic Services Framework. International Journal of Computer Applications, 2014. 93(14).

[14] Abeysinghe, S., Restful PHP Web Services. 2008: PACKT.

[15] Telus Framework available at https://www.telushealth.com - last visited 10/9/2016.

[16] Kemp, D.R. and S. Jensen, Shared Medical Data Platform for Insurance Underwriting 2014: USA.

[17] RareConnect available at https://www.rareconnect.org/en - last visited 10/9/2016.

[18] Frost, J.H., et al. How the Social Web Supports patient experimentation with a new therapy: The demand for patient-controlled and patient-centered informatics. in AMIA Annual Symposium Proceedings 2008.

[19] Taruscio, D., et al., EPIRARE survey on activities and needs of rare disease registries in the European Union. Orphanet Journal of Rare Diseases, 2012.

[20] Taruscio, D., et al., The Current Situation and Needs of Rare Diseases Registries in Europe. Public Health Genomics (16), 2013: p. 288–298.

Appendix A Validation Survey

Evaluation Criteria	Statement	Strongly disagree (1)	Somewhat disagree (2)	Neither agree nor disagree (3)	Somewhat agree (4)	Strongly agree (5)
Behavioral intention of use	1. I intend to use the proposed architecture					
	2. Getting the opportunity, I will use the proposed architecture					
	3. I will recommend the proposed architecture to my colleagues					
Perceived usefulness	4. Using the proposed architecture would improve my job performance					
	5. Using the proposed architecture would improve my productivity					
	6. Using the proposed architecture would increase the quality of my work					
	7. The advantages of using the proposed architecture outweigh the disadvantages					
Ease of use	8. I think the proposed architecture is clear and understandable					
	9. I find the proposed architecture easy to use					