Methods in Molecular Biology 1246

Springer Protocols

Carlos Fernández-Llatas
Juan Miguel García-Gómez *Editors*

Data Mining in Clinical Medicine

Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

Data Mining in Clinical Medicine

Edited by

Carlos Fernández-Llatas

Instituto Itaca, Universitat Politècnica de València, València, Spain

Juan Miguel García-Gómez

Instituto Itaca, Universitat Politècnica de València, València, Spain

Humana Press

Editors
Carlos Fernández-Llatas
Instituto Itaca, Universitat
Politècnica de València
València, Spain

Juan Miguel García-Gómez
Instituto Itaca, Universitat
Politècnica de València
València, Spain

ISSN 1064-3745 ISSN 1940-6029 (electronic)
ISBN 978-1-4939-1984-0 ISBN 978-1-4939-1985-7 (eBook)
DOI 10.1007/978-1-4939-1985-7
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2014955054

Printed on acid-free paper

Humana Press is a brand of Springer
Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Data mining is one of the technologies called to improve the quality of service in clinical medicine through the intelligent analysis of biomedical information. From the enunciation of evidence-based medicine in early 1990s [1], the need for creating evidence that could be quickly transferred to physician daily practice is one of the most important challenges in medicine. The use of statistics to prove the validity of the treatment over discrete populations; the creation of predictive models for diagnosis, prognosis, and treatment; and the inference of clinical guidelines as decision trees or workflows from instances of healthcare protocols are examples of how data mining can help in the application of Evidence Based Medicine.

The great interest that emerges from the use of data mining techniques has caused that there was a large amount of data mining books and papers available in literature. The majority of techniques or methodologies that are available for use are published and can be studied by clinical scientist around the world. However, despite the great penetration of those techniques in literature, their application to real daily practice is far to be complete. For that, when we were planning this book, our vision was not just to compile a set of data mining techniques, but also to document the deployment of advance solutions based on data mining in real biomedical scenarios, new approaches, and trends.

We have divided the book into three different parts. The first part deals with innovative data mining techniques with direct application to biomedical data problems; in the second part we selected works talking about the use of the Internet in data mining as well as how to use distributed data for making better model inferences. In the last part of the book, we made a selection of new applications of data mining techniques.

In Chapter 1, Fuster-Garcia et al. describe the automatic actigraphy pattern analysis for outpatient monitoring that has been incorporated in the Help4Mood EU project for helping people with major depression recover in their own home. The system allows the reduction of inherent complexity of the acquired data, the extraction of the most informative features, and the interpretation of the patient state based on the monitoring. For this, their proposal covers the main steps needed to analyze outpatient daily actigraphy patterns for outpatient monitoring: data acquisition, data pre-processing and quantification, non-lineal registration, feature extraction, anomaly detection, and visualization of the information extracted. Moreover, their study proposes several modeling and simulation techniques useful for experimental research or for testing new algorithms in actigraphy pattern analysis. The evaluation with actigraphy signals from 16 participants including controls and patients that have recovered from major depression demonstrates the utility to visually analyze the activity of the individuals and study their behavioral trends.

Biomedical classification problems are usually represented by imbalanced datasets. The performance of the classification models is usually measured by means of the empirical error or misclassification rate. Nevertheless, neither those loss functions nor the empirical error are adequate for learning from imbalanced data. In Chapter 2, Garcia-Gomez and Tortajada define the loss function of LBER whose associated empirical risk is equal to the balanced

error rate (BER). In these problems, the empirical error is uninformative about the performance of the classifier and the loss functions usually produce models that are shifted to the majority class. The results obtained in simulated and real biomedical data show that classifiers based on the LBER loss function are optimal in terms of the BER evaluation metric. Furthermore, the boundaries of the classifiers were invariant to the imbalance ratio of the training dataset. The LBER-based models outperformed the 0–1-based models and other algorithms for imbalanced data in terms of BER, regardless of the prevalence of the positive class. Finally, the authors demonstrate the equivalence of the loss function to the method of inverted prior probabilities, and generalize the loss function to any combination of error rates by class. Big data analysis applied to biomedical problems may benefit from this development due to the imbalance nature of most of the interesting problems to solve, such as predictive of adverse events, diagnosis, and prognosis classification.

In Chapter 3, Vicente presents a novel online method to audit predictive models using a Bayesian perspective. This audit method is specially designed for the continuous evaluation of the performance of clinical decision support systems deployed in real clinical environments. The method calculates the posterior odds of a model through the composition of a prior odds, a static odds, and a dynamic odds. These three components constitute the relevant information about the behavior of the model to evaluate if it is working correctly. The prior odds incorporates the similarity of the cases of the real scenario and the samples used to train the predictive model. The static odds is the performance reported by the designers of the predictive model and the dynamic odds is the performance evaluated with the cases seen by the model after deployment. The author reports the efficacy of the method to audit classifiers of brain tumor diagnosis with magnetic resonance spectroscopy (MRS). This method may help on assuring the best performance of the predictive models during their continuous usage in clinical practice.

What to do when we obtain underperformed expectations of the predictive models during their real use of predictive models? Tortajada et al. in Chapter 4 propose an incremental learning algorithm for logistic regression based on the Bayesian inference approach that may allow to update predictive models incrementally when new data are collected or even to perform a new calibration of a model from different centers. The performance of their algorithm is demonstrated by employing different benchmark datasets and a real brain tumor dataset. Moreover, they compare its performance to a previous incremental algorithm and a non-incremental Bayesian model, showing that the algorithm is independent of the data model and iterative, and it has a good convergence. The combination of audit models, such as the proposal from Vicente, with incremental learning algorithms, such as that proposed by Tortajada et al., may help on the assurement of the performance of clinical decision support systems during their continuous usage in clinical practice.

New trends like interactive pattern recognition [2] aim at the creation of human understandable data mining models allowing them the correction of the models to make a direct use of data mining techniques as well as facilitate its continuous optimization. In Chapter 5 new possibilities about the use of process mining techniques in clinical medicine are presented. Process mining is a paradigm that comes from the process management research field and that provides a framework that allows to infer the care processes that are being executed in human understandable workflows. These technologies allow experts in the understanding of the care process, and the evaluation of how the process deployment affects the quality of service to the patient.

Chapter 6 analyzes the patient history from a temporal perspective. Usually data mining techniques are seen from a static perspective and represent the status of the patient in a specific moment. Using temporal data mining techniques presented in this chapter it is possible to represent the dynamic behavior of the patient status in an easy human understandable way.

One of the worst problems that affect data mining techniques for creating valid models is the lack of data. Issues as the difficulty for achieve specific cases and the data protection regulations are barriers for enabling a common sharing of data that can be used for inferring better models that can be used for a better understanding of the illnesses and for improving the cares to final patients. Chapter 7 presents a model to allow feed data mining system from different distributed databases allowing them in the creation of better models using more available data.

Nowadays, the greatest data source is the Internet. The omnipresence of the Internet in our lives has changed our communication channels and medicine is not an exception. New trends use the Internet to explore new kind of diagnoses and treatment models that are patient centered covering them in a holistic way. From the arrival of web 2.0 human cybercitizens use the net not only to get information, but also, Internet is continuously feeding about us. For that, there is a great amount of information available about single humans. Usually cyberhumans write in the Internet its sentiments and desires. Using data mining technologies with this information it will be possible to prevent psychological disorders providing new ways to diagnosis and treat this using the Net [5]. Chapter 8 presents new trends of using sentiment analysis technologies over the Internet.

As we have pointed previously, Internet is used for gathering information. But, not only patients use the Internet to gather information about their and their relatives' health status [4], but also junior doctors trust in the Internet for being continuously informed [3]. However, their universality makes Internet not always trustable. It is necessary to create mechanism to filter trustable information to avoid misunderstandings in patient information. Chapter 9 presents the concept of health recommender systems that use data mining techniques for support patients and doctors for finding trustable health data over the Internet.

However, Internet is not only for persons, but also for systems and applications. New trends, as Cloud Computing, see Internet as a universal platform to host smart applications and platforms for continuous monitoring on patients in a ubiquitous way. Chapter 10 presents an m-health context aware model based on Cloud Computing technologies.

Finally, we end the book with four chapters dealing with applications of data mining technologies: Chapter 11 presents an innovative use of classical speech recognition techniques to detect Alzheimer disease on elderly people; Chapter 12 shows how data mining techniques can be used for detecting cancer in early stages; Chapter 13 presents the use of data mining for inferring individualized metabolic models for controlling chronic diabetic patients; Chapter 14 shows a selection of innovative techniques for cardiac analysis in detecting arrhythmias. Chapter 15 presents a knowledge-based system for empower diabetic patients and Chapter 16 presents how serious games can help in the detection of specific elderly people.

We hope that the reader find our compilation work interesting. Enjoy it!

Valencia, Spain

Carlos Fernandez-Llatas
Juan Miguel García-Gómez

References

1. Davidoff F, Haynes B, Sackett D, Smith R (1995) Evidence based medicine. BMJ 310(6987): 10851086. doi:10.1136/bmj.310.6987.1085. http://www.bmj.com/content/310/6987/1085.short
2. Fernndez-Llatas C, Meneu T, Traver V, Benedi JM (2013) Applying evidence-based medicine in telehealth: an interactive pattern recognition approximation. Int J Environ Res Public Health 10(11):5671–5682. doi:10.3390/ijerph10115671. http://www.mdpi.com/1660-4601/10/11/5671
3. Hughes B, Joshi I, Lemonde H, Wareham J (2009) Junior physician's use of web 2.0 for information seeking and medical education: a qualitative study. Int J Med Inform 78(10):645–655. doi:10.1016/j.ijmedinf.2009.04.008. PMID: 19501017
4. Khoo K, Bolt P, Babl FE, Jury S, Goldman RD (2008) Health information seeking by parents in the internet age. J Paediatr Child Health 44(7–8):419–423. doi:10.1111/j.1440-1754.2008.01322.x. PMID: 18564080
5. van Uden-Kraan CF, Drossaert CHC, Taal E, Seydel ER, van de Laar, MAFJ (2009) Participation in online patient support groups endorses patients' empowerment. Patient Educ Couns 74(1):61–69. doi:10.1016/j.pec.2008.07.044. PMID: 18778909

Contents

Preface . . . . . *v*
Contributors . . . . . *xi*

PART I INNOVATIVE DATA MINING TECHNIQUES FOR CLINICAL MEDICINE

1 Actigraphy Pattern Analysis for Outpatient Monitoring 3
Elies Fuster-Garcia, Adrián Bresó, Juan Martínez Miranda, and Juan Miguel Garcia-Gómez

2 Definition of Loss Functions for Learning from Imbalanced Data to Minimize Evaluation Metrics 19
Juan Miguel Garcia-Gómez and Salvador Tortajada

3 Audit Method Suited for DSS in Clinical Environment 39
Javier Vicente

4 Incremental Logistic Regression for Customizing Automatic Diagnostic Models 57
Salvador Tortajada, Montserrat Robles, and Juan Miguel Garcia-Gómez

5 Using Process Mining for Automatic Support of Clinical Pathways Design 79
Carlos Fernandez-Llatas, Bernardo Valdivieso, Vicente Traver, and Jose Miguel Benedi

6 Analyzing Complex Patients' Temporal Histories: New Frontiers in Temporal Data Mining 89
Lucia Sacchi, Arianna Dagliati, and Riccardo Bellazzi

PART II MINING MEDICAL DATA OVER INTERNET

7 The Snow System: A Decentralized Medical Data Processing System 109
Johan Gustav Bellika, Torje Starbo Henriksen, and Kassaye Yitbarek Yigzaw

8 Data Mining for Pulsing the Emotion on the Web 123
Jose Enrique Borras-Morell

9 Introduction on Health Recommender Systems 131
C.L. Sanchez-Bocanegra, F. Sanchez-Laguna, and J.L. Sevillano

10 Cloud Computing for Context-Aware Enhanced m-Health Services 147
Carlos Fernandez-Llatas, Salvatore F. Pileggi, Gema Ibañez, Zoe Valero, and Pilar Sala

Part III New Applications of Data Mining in Clinical Medicine Problems

11 Analysis of Speech-Based Measures for Detecting and Monitoring Alzheimer's Disease 159
A. Khodabakhsh and C. Demiroglu

12 Applying Data Mining for the Analysis of Breast Cancer Data 175
Der-Ming Liou and Wei-Pin Chang

13 Mining Data When Technology Is Applied to Support Patients and Professional on the Control of Chronic Diseases: The Experience of the METABO Platform for Diabetes Management 191
Giuseppe Fico, Maria Teresa Arredondo, Vasilios Protopappas, Eleni Georgia, and Dimitrios Fotiadis

14 Data Analysis in Cardiac Arrhythmias 217
Miguel Rodrigo, Jorge Pedrón-Torecilla, Ismael Hernández, Alejandro Liberos, Andreu M. Climent, and María S. Guillem

15 Knowledge-Based Personal Health System to Empower Outpatients of Diabetes Mellitus by Means of P4 Medicine 237
Adrián Bresó, Carlos Sáez, Javier Vicente, Félix Larrinaga, Montserrat Robles, and Juan Miguel García-Gómez

16 Serious Games for Elderly Continuous Monitoring 259
Lenin-G. Lemus-Zúñiga, Esperanza Navarro-Pardo, Carmen Moret-Tatay, and Ricardo Pocinho

Index *269*

Contributors

Maria Teresa Arredondo • *Life Supporting Technologies, Universidad Politécnica de Madrid, Madrid, Spain*

Riccardo Bellazzi • *Dipartimento di Ingegneria Industriale e dell'Informazione, Università degli Studi di Pavia, Pavia, Italy*

Johan Gustav Bellika • *Norwegian Centre for Integrated Care and Telemedicine (NST), Tromsø, Norway*

Jose Miguel Benedi • *PHRLT, Universitat Politècnica de València, València, Spain*

Adrián Bresó • *IBIME-ITACA, Universitat Politècnica de València, València, Spain*

Wei-Pin Chang • *Yang-Ming University, Taipei, Taiwan*

Andreu M. Climent • *Fundación para la Investigación del Hospital Gregorio Marañón, Madrid, Spain*

Arianna Dagliati • *Dipartimento di Ingegneria Industriale e dell'Informazione, Università degli Studi di Pavia, Pavia, Italy*

C. Demiroglu • *Faculty of Engineering, Ozyegin University, İstanbul, Turkey*

Dimitrios Fotiadis • *Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, University of Ioannina, Ioannina, Greece*

Carlos Fernandez-Llatas • *SABIEN-ITACA, Universitat Politècnica de València, València, Spain*

Giuseppe Fico • *Life Supporting Technologies, Universidad Politécnica de Madrid, Madrid, Spain*

Elies Fuster-Garcia • *Veratech for Health, S.L., Valencia, Spain*

Juan Miguel García-Gómez • *IBIME-ITACA, Universitat Politècnica de València, València, Spain*

Eleni Georgia • *Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, University of Ioannina, Ipiros, Greece*

María S. Guillem • *BIO-ITACA, Universitat Politècnica de València, València, Spain*

Torje Starbo Henriksen • *Norwegian Centre for Integrated Care and Telemedicine (NST), Tromsø, Norway*

Ismael Hernández • *BIO-ITACA, Universitat Politècnica de València, València, Spain*

Gema Ibañez • *SABIEN-ITACA, Universitat Politècnica de València, València, Spain*

A. Khodabakhsh • *Ozyegin University, Istanbul, Turkey*

Félix Larrinaga • *Elektronika eta Informatika saila, Mondragon Goi Eskola Politeknikoa, España, Spain*

Lenin-G. Lemus-Zúñiga • *Instituto ITACA, Universitat Politècnica de València, València, Spain*

Alejandro Liberos • *BIO-ITACA, Universitat Politècnica de València, València, Spain*

Der-Ming Liou • *Yang-Ming University, Taipei, Taiwan*

Juan Martínez Miranda • *IBIME-ITACA, Universitat Politècnica de València, València, Spain*

Jose Enrique Borras-Morell • *University of Tromso, Tromsø, Norway*

Carmen Moret-Tatay • *Departamento de Neuropsicología, Metodología y Psicología Social, Universidad Católica de Valencia San Vicente Mártir, València, Spain*
Esperanza Navarro-Pardo • *Departamento de Psicología educativa y de la educación, Facultad de Psicología, Universitat de València, Valencia, Spain*
Jorge Pedrón-Torecilla • *BIO-ITACA, Universitat Politècnica de València, València, Spain*
Salvatore F. Pileggi • *Department of Computer Science, The University of Auckland, Auckland, New Zealand*
Ricardo Pocinho • *Instituto de Psicologia Cognitiva, Desenvolvimento Vocacional e Social da, Universidade de Coimbra, Coimbra, Portugal*
Vasilios Protopappas • *Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, University of Ioannina, Ipiros, Greece*
Montserrat Robles • *IBIME-ITACA, Universitat Politècnica de València, València, Spain*
Miguel Rodrigo • *BIO-ITACA, Universitat Politècnica de València, València, Spain*
Lucia Sacchi • *Dipartimento di Ingegneria Industriale e dell'Informazione, Università degli Studi di Pavia, Pavia, Italy*
Carlos Saez • *IBIME-ITACA, Universitat Politècnica de València, València, Spain*
Pilar Sala • *SABIEN-ITACA, Universitat Politècnica de València, València, Spain*
C.L. Sanchez-Bocanegra • *NORUT (Northern Research Institute), Tromsø, Norway*
F. Sanchez-Laguna • *Virgen del Rocío University Hospital, Seville, Spain*
J.L. Sevillano • *Robotic and Technology of Computers Lab, Universidad de Sevilla, Seville, Spain*
Salvador Tortajada • *Veratech for Health, S.L., Valencia, Spain*
Vicente Traver • *SABIEN-ITACA, Universitat Politècnica de València, València, Spain*
Bernardo Valdivieso • *Departamento de Calidad, Hospital - La Fe de Valencia, Valencia, Spain*
Zoe Valero • *SABIEN-ITACA, Universitat Politècnica de València, València, Spain*
Javier Vicente • *IBIME-ITACA, Universitat Politècnica de València, València, Spain*
Kassaye Yitbarek Yigzaw • *Norwegian Centre for Integrated Care and Telemedicine (NST), Tromsø, Norway*

Part I

Innovative Data Mining Techniques for Clinical Medicine

Chapter 1

Actigraphy Pattern Analysis for Outpatient Monitoring

Elies Fuster-Garcia, Adrián Bresó, Juan Martínez Miranda, and Juan Miguel García-Gómez

Abstract

The actigraphy is a cost-effective method for assessing specific sleep disorders such as diagnosing insomnia, circadian rhythm disorders, or excessive sleepiness. Due to recent advances in wireless connectivity and motion activity sensors, the new actigraphy devices allow the non-intrusive and non-stigmatizing monitoring of outpatients for weeks or even months facilitating treatment outcome measure in daily life activities. This possibility has propitiated new studies suggesting the utility of actigraphy to monitor outpatients with mood disorders such as major depression, or patients with dementia. However, the full exploitation of data acquired during the monitoring period requires the use of automatic systems and techniques that allow the reduction of inherent complexity of the data, the extraction of most informative features, and the interpretability and decision-making. In this study we purpose a set of techniques for actigraphy patterns analysis for outpatient monitoring. These techniques include actigraphy signal pre-processing, quantification, nonlinear registration, feature extraction, detection of anomalies, and pattern visualization. In addition, techniques for daily actigraphy signals modelling and simulation are included to facilitate the development and test of new analysis techniques in controlled scenarios.

Key words Actigraphy, Outpatient monitoring, Functional data analysis, Feature extraction, Kernel density estimation, Simulation

1 Introduction

The activity-based monitoring, also known as actigraphy, is a valuable tool for analysing patients' daily sleep-wake cycles and routines and it is considered as a cost-effective method for assessing specific sleep disorders such as diagnosing insomnia, circadian rhythm disorders, or excessive sleepiness [1]. A growing number of studies have been published analysing the validity of actigraphy, their utility to analyze sleep disorders, their utility to study circadian rhythms, and their use as treatment outcome measure [2–5].

In the last years a high number of commercial devices for research, clinical use, and even for sport and personal well-being have been developed. The last developments in actigraphy sensors allow the monitoring of motion activity for several weeks and

Carlos Fernández-Llatas and Juan Miguel García-Gómez (eds.), *Data Mining in Clinical Medicine*, Methods in Molecular Biology, vol. 1246, DOI 10.1007/978-1-4939-1985-7_1,

also to embed the sensors in discrete and small devices (e.g. watches, smartphones, key rings, or belts). Moreover, most of these actigraphy devices are able to establish wireless communication with the analysis infrastructure such as preconfigured personal computers [6]. These advances have allowed a non-intrusive and non-stigmatizing monitoring of outpatients facilitating treatment outcome measure in daily life activities as extension of face-to-face patient care.

The main studies of activity monitoring have been done in the context of sleep and circadian rhythm disorders. However in the last years, the non-intrusive and non-stigmatizing monitoring of motion activity has been found especially interesting in the case of patients with mood disorders. Different studies suggest that actigraphy-based information can be used to monitor patients with mood disorders such as major depression [7–10], or patients with dementia [11, 12]. In those patients it is highly desirable to facilitate the execution of normal life routines, but minimizing the risks associated with the disease by designing efficient outpatient follow-up strategies. These goals are currently being addressed in international projects such as Help4Mood [13] and Optimi [14].

An efficient outpatient follow-up system must include three main tasks: (1) acquisition of information (through physiological and/or environmental sensors), (2) processing and analysis of information acquired, and (3) support of clinical decision. In this sense a follow-up system based on actigraphy information needs to automatically extract valuable and reliable information from signals acquired during monitoring period. Moreover, the extracted information should be presented in a way that helps clinical decision-making by the use of high dimensionality reduction techniques and visualization strategies. These needs are even greater when considering long-term studies, where evaluating changes in daily activity patterns and detection of anomaly patterns are desirable.

To contribute to this goal, in this study we cover the main steps needed to analyse outpatient daily actigraphy patterns for continuous monitoring such as: data acquisition, data pre-processing and quantification, non-lineal registration, feature extraction, anomaly detection, and visualization of the information extracted. Finally, in addition, to these main steps, modeling and simulation techniques are included in this study. These techniques allow modeling the actigraphy patterns of a patient or a group of patients for the analysis of their similarities or dissimilarities. Moreover, these models allow the simulation of new actigraphy signals for experimental research or for testing new algorithms in this field.

To illustrate the use of this methodology in a real application, we have considered the use of data acquired in the Help4Mood EU project [13]. The main aim of this project is to develop a system that will help people with major depression recover in their own home. Help4mood includes a personal monitoring system

mainly based on actigraphy data to follow up patient behaviour characteristics such as sleep or activity levels. The actigraphy signals obtained are used by the system to feed a decision support system that tailor each session with the patient to the individual needs, and to support clinicians in the outpatient monitoring. Specifically the data used in this work consist in actigraphy signals from participants including controls and patients that have recovered from major depression, and acquired in the framework of the project.

2 Data Acquisition, Pre-processing, and Quantification

In this section we introduce the basic processes and techniques to acquire actigraphy signals, pre-process them to detect and replace missed data, and finally quantify a set of valuable parameters for the monitoring of daily patient activity. A schema of the dataflow in this first stage of actigraphy patterns analysis can be seen in Fig. 1.

2.1 Data Acquisition

In recent years it has been a wide inclusion of accelerometer sensors on non-intrusive and non-stigmatizing devices. These devices include a wide diversity of wearable objects such as smartphones, wristwatches, key rings, and belts, and even other devices that can be installed at outpatient home such as under-mattress devices. The use of these technologies allows performing long-term monitoring studies of outpatients without modifying their normal activity. At this point it is important to consider three main characteristics that an actigraphy device for long-term outpatients monitoring needs to have. Firstly, it has to be non-intrusive, non-obstructive and non-stigmatizing. Secondly, the device has to minimize the user's responsibility in the operation of the system; and finally the device and the synchronization system have to be able to avoid failure situations that can alter the patient and their behaviour.

Following these requirements, in this work we have used the Texas Instruments ez430 Chronos wristwatch device to obtain free

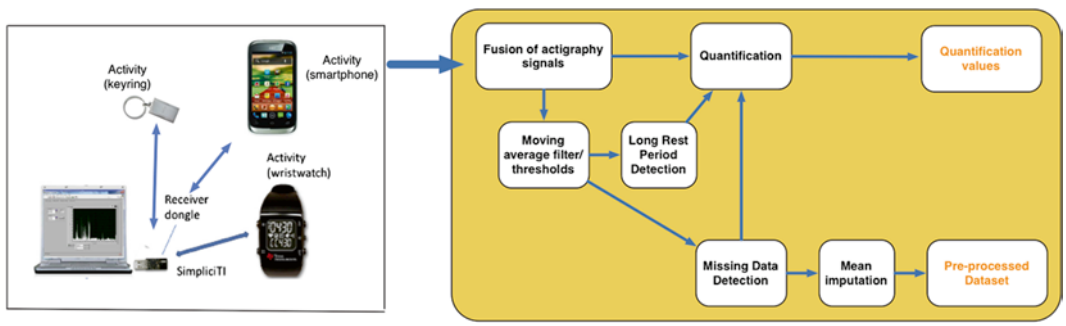

Fig. 1 Schema of the main steps of actigraphy pre-processing and quantification

living-patient activity signals. These signals will be used along the study to present the methodology for actigraphy patterns analysis for outpatient monitoring. The main characteristics of this device are RF wireless connectivity (RF link at 868 MHz), 5 days of memory without downloading (recording one sample per minute), more than 30-day battery life, and fully programmable. For additional technical information of this device *see* ref. 15. The ez430 Chronos wristwatches used in this study were programmed to acquire the information from the three axis with a sampling frequency of 20 Hz, and to apply a high pass second order Butterworth filter at 1.5 Hz on each axis signal. Afterwards, the activity for each axis was computed by using the value of Time Above a Threshold (TAT) of 0.04 g in epochs of 1 min. Finally the resulting actigraphy value was selected as the maximum TAT value of the three axes.

The real dataset used in this work were acquired during the Help4Mood EU project and comprises the activity signals of 16 participants monitored 24 h a day. Half of these participants correspond to patients previously diagnosed with major depression but in the recovered stage at the moment of the study. The other half of the participants is composed by controls, which were aimed to follow their normal life. As a result a total of 69 daily activity signals were compiled.

2.2 Actigraphy Data Pre-processing and Quantification

Before analysing daily actigraphy signals they need to be pre-processed. The pre-processing includes at least three basic steps: (a) fusion of actigraphy signals provided by the different used sensors, (b) detection of periods containing missed data, (c) detection of long resting periods (including sleep), and (d) missed data imputation.

The fusion of actigraphy signals could be done using different strategies. If the different devices use a similar accelerometer sensor and the device in which they are embedded are used by the patient in a similar way (e.g. wristwatch and smartphone), we can assume that the response to the same activity will be similar, and therefore we can use a simple mean between both signals in the periods where they do not contain missed data. However in the case of major differences such as in the case of the wristwatch and under-mattress sensor [16] a more complex fusion strategy is required [17]. In this case the strategy will take into account non-linear relations between both signals and differences in sensitivities between both sensors.

Actigraphy signals from outpatients usually contain missing data. In most of the cases this missed data is related to not wearing the actigraph; however it can be also related with empty batteries, empty memory, or even communication errors. Detecting this missed data is mandatory for a robust analysis of activity patterns in recorded data. To detect this missed data, a two steps threshold-based strategy was used. The first step consists in applying a moving average filter to the actigraphy signal. In this study a

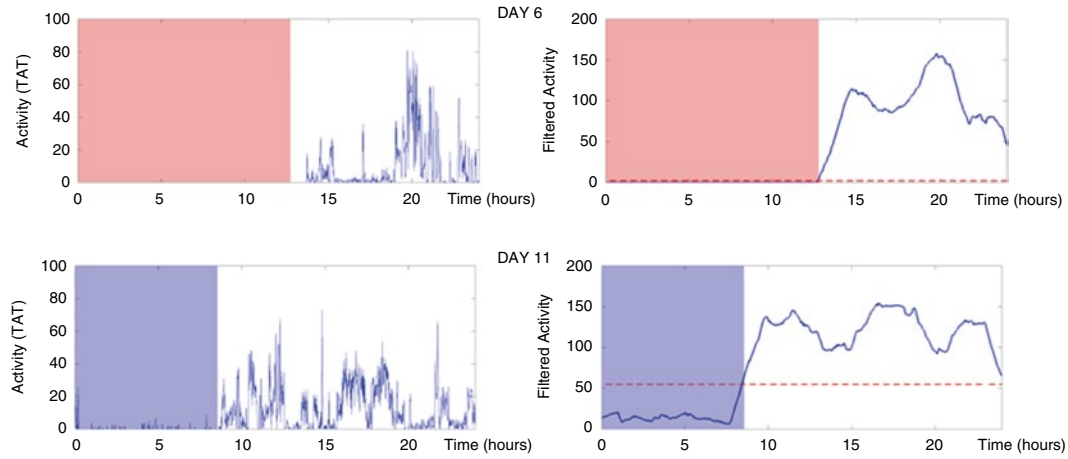

Fig. 2 Example of two daily actigraphy signals s (*left*), and their corresponding filtered signal fs (*right*). The *red-shaded* region shows missed data, and the *blue-shaded* region shows sleep periods. The thresholds values are presented as *dashed lines*

window size of 120 min was used. As a result we obtained a smoothed signal that represents in each point the mean value of activity in a region centred on it. The second step consists in applying a threshold to detect periods with actigraphy values equal to zero or near zero. In this study the threshold value th_{md} was equal to 2. An example of the result of the missing data detection algorithm is presented in Fig. 2 (top). Posteriorly, a missed data imputation method (e.g., mean imputation or knn imputation [18]) could be applied to the daily actigraphy signal to fill the missed data periods when they are not so long.

The analysis of sleep/awake cycles and circadian rhythms represents valuable information for the clinicians when monitoring outpatients, and mostly when they are patients with mental disorders such as (major depression or anxiety). Different algorithms have been presented in the literature to identify sleep-wake periods. These are based on linear combination methods (e.g., Sadeh's algorithm [19] or the Sazonov's algorithm [20]), or based on pattern recognition methods such as artificial neurons or decision trees [21]. However the parameters of these algorithms need to be computed for each different type of actigraphy device using annotated datasets. In our case we have used a simple linear model to segment actigraphy signals into two main types of activity: long resting periods (including sleep) and active awake periods. To do so we followed the same strategy used for missing data detection but using a higher threshold value. This value depends on the actigraphy device and on the algorithm used to quantify the activity. In our study a threshold value th_{sd} of 50 was used. An example of the result of the segmentation algorithm is presented in Fig. 2 (bottom).

Finally, after the detection of missed data and the segmentation processes we calculated relevant parameters for outpatients monitoring such as:

- Mean daily activity, to represent the average of the actigraphy signal values along the whole awake period.
- Standard deviation of daily activity, to represent the standard deviation of the actigraphy signal values over the whole awake period.
- Maximum sustained activity, to characterize the maximum sustained activity over a period of 30 min during the whole day. This value is defined as the maximum value of the daily actigraphy signal filtered using a moving average filter with a span of 30 min.
- Total hours of sleep, to represent the total sleep time detected in a day.
- Sleep fragmentation, to measure the number of periods of uninterrupted sleep during a day.
- Mean activity during sleep, to measure the mean value of the activity signal in the detected sleep periods.
- Total time of missing data, to represent the total missed data time detected in a day.

3 Analysis of Daily Actigraphy Patterns

Once the acquired signals have been pre-processed and quantified, we can proceed with the analysis of daily actigraphy patterns. The main objective of this section is to describe each daily actigraphy signal by using a minimum number of variables, analyse its similarity with the rest of daily signals to detect anomalies, and finally display the results optimally to facilitate the patient follow up and clinical decision making. To do so, in this section we introduce basic steps to perform this task such as the nonlinear registration of the daily actigraphy signals, the extraction of features, the detection of anomaly patterns, and finally a way to present and visualize the information extracted (*see* Fig. 3).

3.1 Nonlinear Registration of Daily Acigraphy Signals

The actigraphy signals contain a strong daily pattern due to sleep-wake cycles, work schedules, and mealtimes executed by the subject. Although these patterns are present on the signals, they do not need to coincide exactly in time every day. This variability increases the complexity of the automatic analysis of the signals, and makes the comparison between daily activity patterns difficult. To reduce this variability we can apply a non-lineal registration algorithm capable of aligning the different daily activity signals that

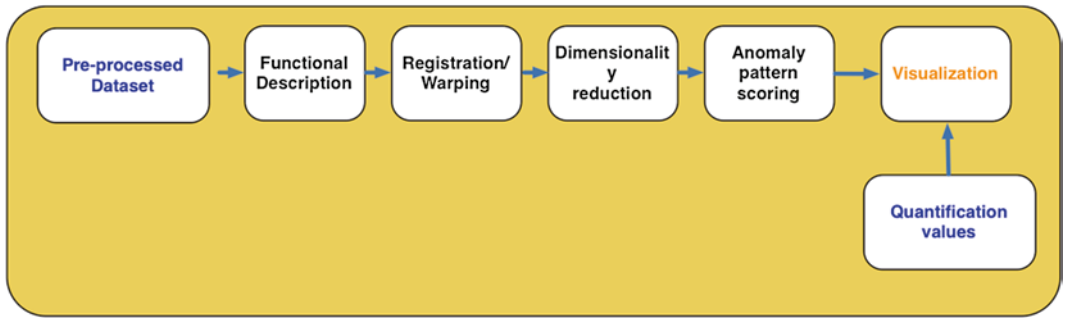

Fig. 3 Schema of the main steps in the analysis of actigraphy patterns

are slightly phase-shifted. In this study we propose the use of the time warping algorithm based on functional analysis and described by Ramsay in ref. 22, and implemented in the FDA MATLAB toolbox [23].

In this algorithm, the daily actigraphy signal is represented in terms of a B-spline basis [24] with uniformly distributed knots. The B-spline basis is defined by two main parameters: the number of knots and the level n. The number of knots defines the number of partitions in the signal in which it will be approximated by a polynomial spline of $n-1$ degree.

To represent our actigraphy signals using the B-spline basis the smoothing algorithm described by Ramsay et al. in ref. 22 is used. The goal of this algorithm is to estimate a curve x from observations $s_i = x(t_i) + \epsilon_i$. To avoid over-fitting, it introduces a roughness penalty to the least-square criterion used for fitting the observations, resulting in a penalized least squares criterion (PENSSE):

$$\mathrm{PENSEE}_{\lambda}(x) = \sum_{i=1}^{\mathrm{length}(s)} \left(s_i - x(t_i)\right)^2 + \lambda J(x) \qquad (1)$$

where $J(x)$ is a roughness parameter, and λ is a coefficient that controls the amount of penalty introduced due to roughness of x. When λ values are higher the generated model is smoother. In order to automatically select an optimum λ value for a specific dataset, the Generalized Cross-Validation measure developed by Craven and Wahba [25] has been used.

The roughness parameter $J(x)$ is based in the concept of curvature or squared second derivative $(D^2x(t))^2$ of a function,

$$J(x) = \int \left(D^2 x(t)\right)^2 \mathrm{d}t \qquad (2)$$

Once the functional description of each daily activity signal is obtained, we are able to register the signals. That is, to construct a transformation h_i for each daily actigraphy curve such that the registered curves with values,

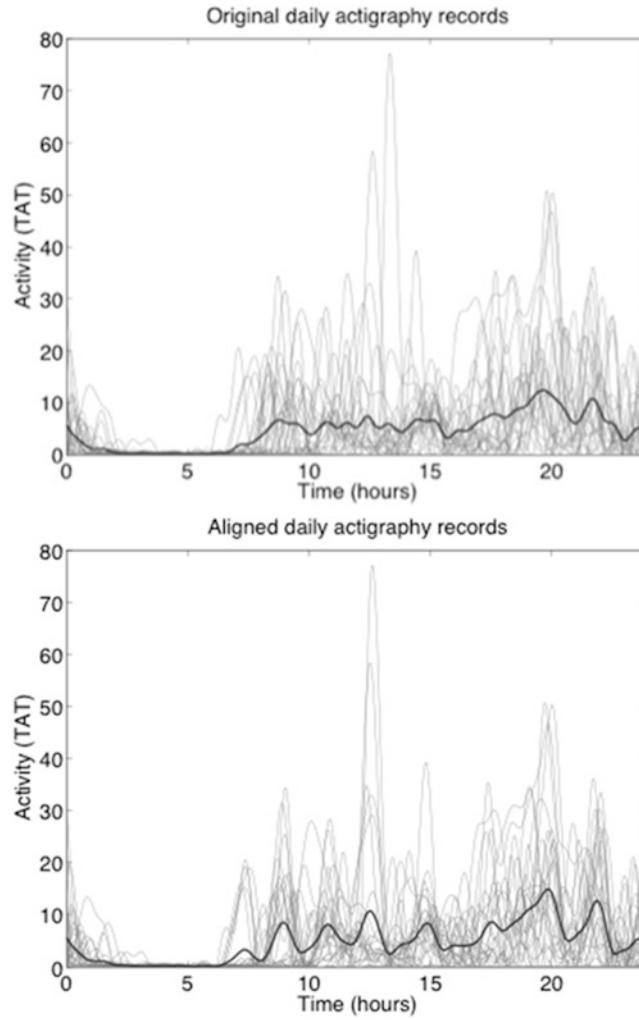

Fig. 4 Daily activity signals included in the study and their associated mean for both non-registered signals (*top*) and for registered signals (*bottom*)

$$x^*(t) = x_i\left[h_i(t)\right] \tag{3}$$

have more or less identical argument values for any given landmark (i.e., local maxima/minima, zero crossings of curves). This requires the computation of function h_i for each curve, called a time-warping function as described in ref. 22.

An illustrative example of the benefits of the registration of daily acigraphy signals is the improvement of the mean actigraphy pattern (*see* Fig. 4). On this example it is easy to see how the registering processing allows the visualization of hidden activity patterns in the mean daily actigraphy related to daily activity routines.

3.2 Feature Extraction

Once the daily actigraphy signals are pre-processed and registered we need to extract the features allowing us to explain the most relevant information included in the signals, but using only a few number of descriptors. The quantification parameters described in Subheading 2.2 can be seen as a features extracted based on prior knowledge. However, these descriptors do not explain global features such as the signal shape or the activity patterns observed in the daily signals, and do not allow the comparison between different daily activity behaviours. To do so, feature extraction methods based on machine learning algorithms could be used, and specifically feature extraction methods based on global features such as independent component analysis, principal component analysis (PCA) [26], or even newer techniques such as nonnegative matrix factorization [27], or feature extraction based on manifold learning [28].

In this study a standard feature extraction method based on PCA was used. PCA uses orthogonal transformation to convert the initial variables, such that the first transformed variables describe the main variability of the signal. When using PCA to reduce the number of variables, we need to choose a criterion to decide the number of principal components is enough to describe our data. The most widely used criterion to select the number of principal components is the % of variability explained. In this case we have fixed the % of variability explained above 75 %, resulting in the first 15 principal components.

3.3 Anomaly Detection

The detection of anomaly activity patterns could be very useful for the analysis of outpatient's actigraphy patterns by detecting non-usual behavior in the monitored patients or even creating alerts for the clinicians. In the case we have an annotated dataset with activity signals tagged as normal for each patient we can use classification-based anomaly detection techniques such as neural-networks, Bayesian networks, support vector machines or even rule systems. However in most of cases this information is not available. In these cases a useful approach to the computation of an anomaly measure for a daily activity signal is based on the nearest neighbour analysis. The anomaly score for a specific signal (represented in the 15th dimensional space of PCA components) is based on the distance to its kth nearest neighbors in a given data set. The hypothesis of this method is that normal activity signals occur in dense neighborhoods, while anomalies occur far from their closest neighbors. To avoid that activity patterns that recur even once a week can be considered as anomalous, we purpose the use of a k value equal to the number of weeks included in the study. A detailed introduction to different anomaly detection methods can be found on ref. 29.

3.4 Visualization

To ensure an effective monitoring of patient activity, the design of effective visualizations is mandatory. These visualizations will include valuable information for the patient state monitoring such as total daily

activity recorded, the amount of lost data, the number of hours slept, anomaly score, or the notion of similarity among patterns in the plot.

In the case of outpatient monitoring, we propose an enriched comparative visualization of the signals consisting in the daily actigraphy-monitoring plot described in ref. 30. This plot is based in the two first components extracted by the feature extraction technique selected (e.g., PCA). Once we have reduced the information of the daily actigraphy signals to only two dimensions, we are able to display them as circles in 2D scatter plot. In this way the distance between circles in the plot is proportional to the similarity between activity patterns. Moreover, we need to include other useful and complementary information for patient monitoring. To do so we propose to add (1) the level of total daily activity, by varying the radius of the circle, (2) the amount of data lost, by varying alpha value (transparency) of the circle colour, and finally (3) the anomaly score for each daily actigraphy signal, by changing the circle colour according to a colour map as can be seen in Fig. 5.

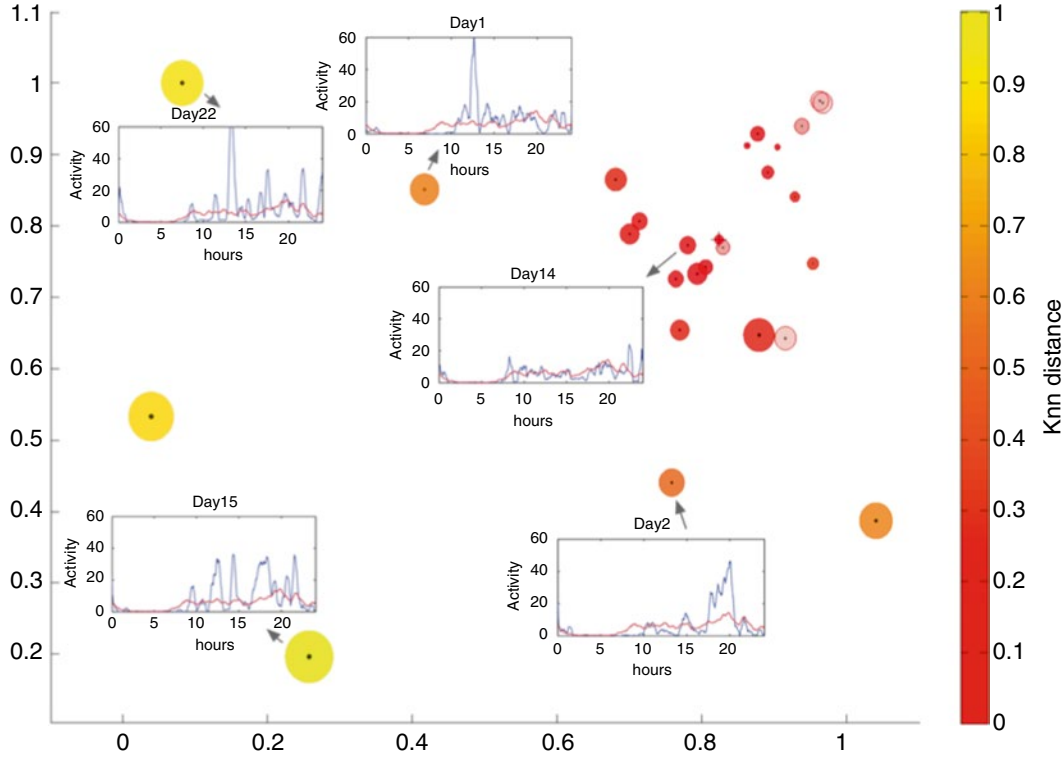

Fig. 5 Daily actigraphy patterns visualization including 14-day samples from a single participant represented as *circles*. The radius of the circle represents the total daily activity, and their transparency represents the amount of missing data. Moreover, the daily actigraphy signals (*blue lines*) are presented for some of the most representative days, including the mean actigraphy signal (*red lines*) for comparison purposes. The anomaly score for each daily actigraphy signal was included in the plot by changing the circle color according to the color map. The median is indicated as + symbol

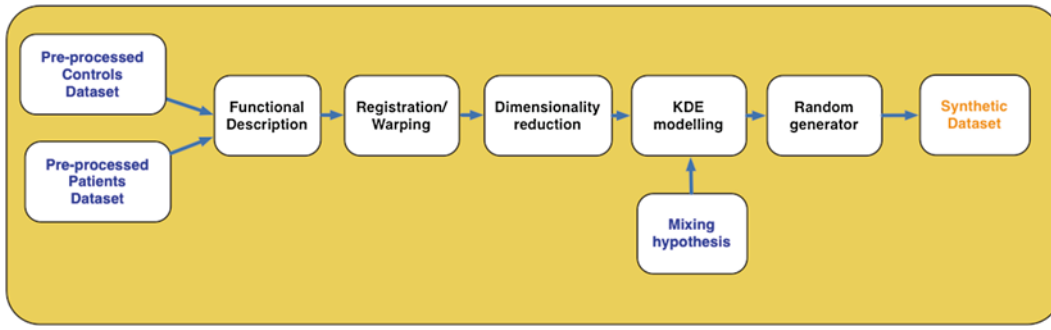

Fig. 6 Schema of the main steps for modelling and simulating actigraphy data

This daily actigraphy-monitoring will be useful for clinicians to visually detect days containing anomaly activity patterns, and to identify relevant events. Moreover, this plot organize the daily activity signals acording to their shape helping to visualize periods of stable behaviour or periods where the patient do not follow daily routines.

4 Actigraphy Data Modeling and Synthetic Datasets Generation

Finally, in this section, a methodology for actigraphy data modelling and synthetic datasets generation is presented. This methodology is feed by the pre-processed data, and uses some of the techniques explained above such as registration and dimensionality reduction as can be seen in Fig. 6. In order to avoid repetition, in this section we will consider that the actigraphy data is registered and that the relevant features are already extracted. This methodology could be used to model the behaviour of a specific set of participants (e.g., patients-like or control-like), and identify daily activity patterns related to a specific disease. Moreover, it allows us to generate synthetic datasets based on specific set of real data to test our new algorithms and techniques in controlled scenarios.

4.1 Modeling and Mixture

In order to simulate new daily actigraphy samples, we need to build a generative model. To do so, a strategy based on Multivariate Kernel Density Estimation (MKDE) [31] is proposed. MKDE is a nonparametric technique for density estimation that allows us to obtain the probability density function of the features extracted from actigraphy signals. Let $s_1, s_2, \ldots, s_n$ be a set of n actigraphy signals represented as a vectors of extracted features (e.g., principal components). Then the kernel density estimate is defined as

$$f_H(s) = \frac{1}{n}\sum_{i=1}^{n} K_H(s - s_i) \tag{4}$$

where f_H is the estimated probability density function, H is the bandwidth (or smoothing) matrix which is symmetric and positive definite and K is the kernel function which is a symmetric multivariate density. For the generative model presented in this work a MKDE based on a 15-D Gaussian kernel was used based on the 15 principal components used to describe daily actigraphy signals. In MKDE algorithm, the choice of the bandwidth matrix H is the most important factor size that defines the amount of smoothing induced in the density function estimation. Automatic bandwidth selection algorithms could be used to do so, such as the 1D search using max leave-one-out likelihood criterion, the mean integrated squared error criterion, or the asymptotic approximation of the mean integrated squared error criterion.

The MKDE model allows weighting the relevance of each of the input samples for the computation of the probability density function. This property allows us to obtain models based on a subset of the available samples, or even a controlled mixture them.

This property is useful to obtain control-like models, patient-like models or models that represent the disease evolution of a person from healthy to ill or vice versa (*see* Fig. 7).

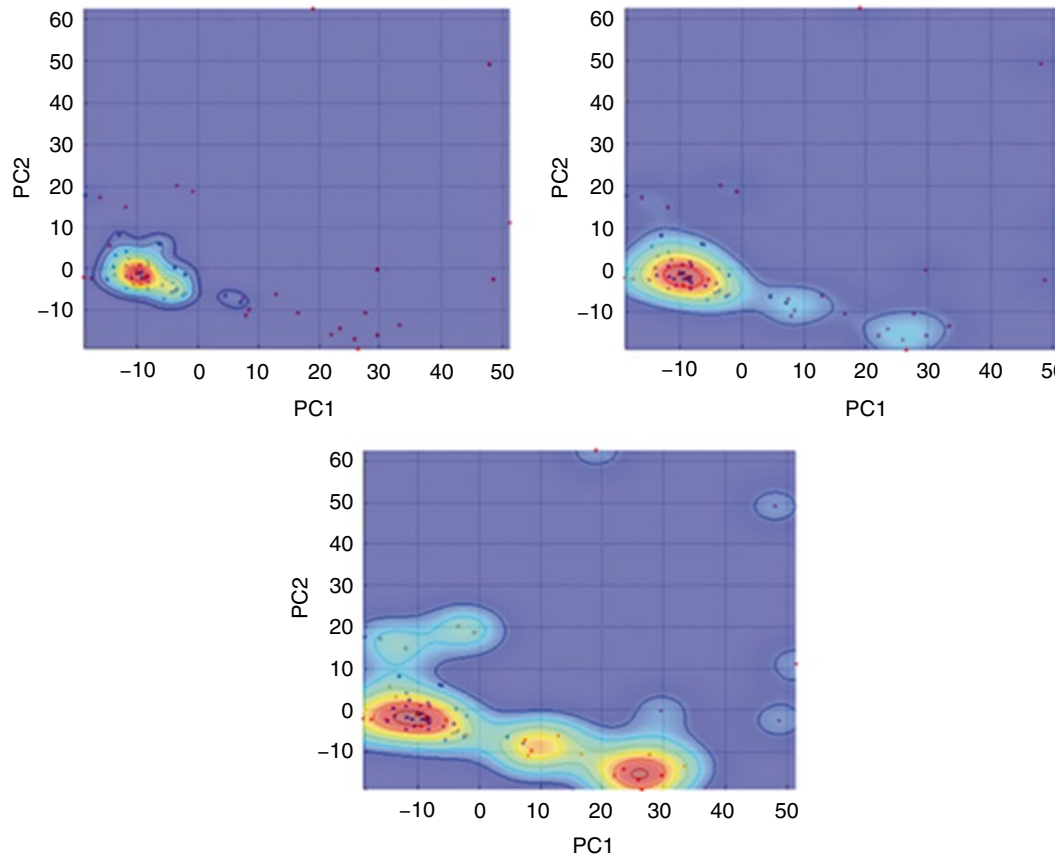

Fig. 7 Different KDE models: Control-like model (*left*), mixed model (*centre*), depressed patient-like model (*right*)

4.2 Random Sample Generation

Based on the probability density function obtained by the KDE model, we can generate random points in the 15th dimensional space of the PCs. Each of these randomly generated points represents a daily actigraphy signal. Form these points, the signals can be reconstructed using the coefficients of the PCA. An example of the random daily actigraphy samples obtained can be seen in Fig. 8 (bottom). Moreover a plot of real data can be seen on Fig. 8 (top) to allow the comparison of the shape of the daily actigraphy samples generated with the real data used to generate the model. For this study we have simulated 20 daily activity patterns using 1 min of temporal resolution (1,440 data points per pattern). Half of this patterns (10) are patient-like activity patterns and the other half (10) are control-like activity patterns.

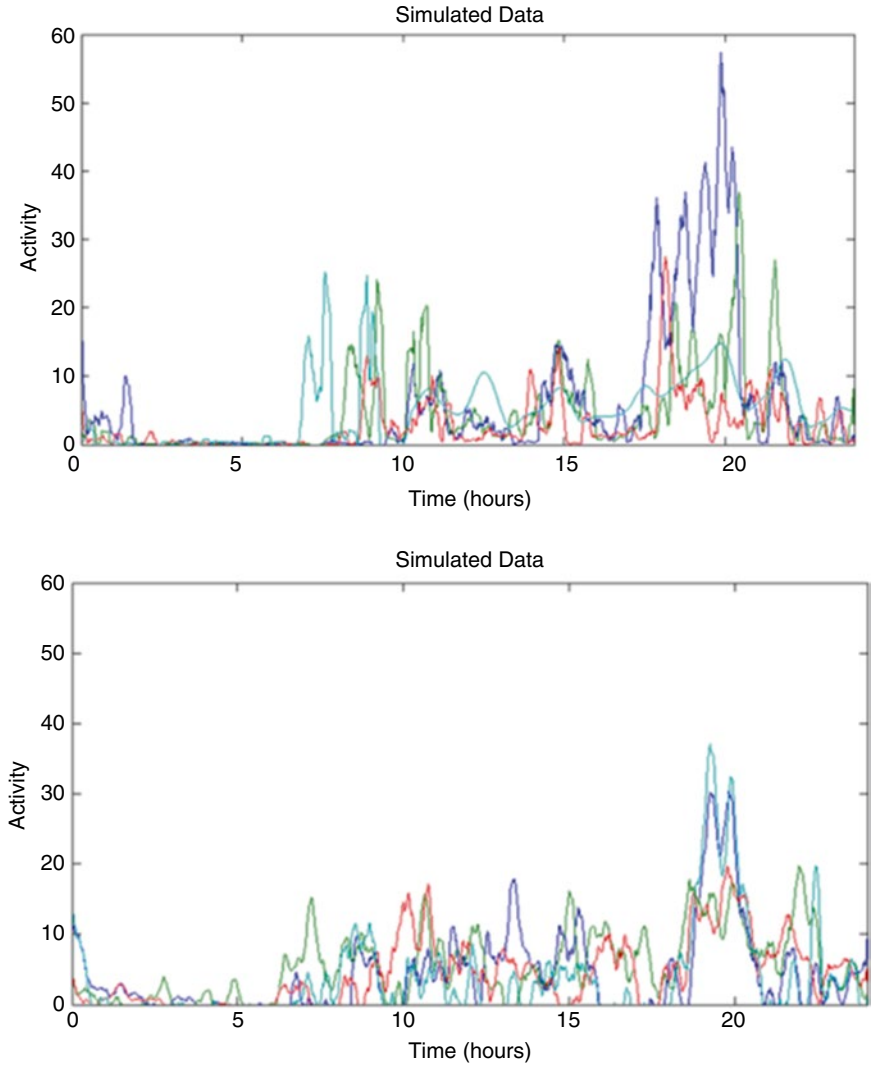

Fig. 8 Real daily actigraphy signals (*top*) vs. simulated daily actigraphy signals (*right*)

5 Conclusions

The main objective of this chapter was to introduce the basic steps for the analysis of actigraphy patterns in the context of outpatient monitoring. These steps include: (1) data acquisition, (2) preprocessing, (3) quantification, (4) analysis, and (5) visualization. For each of these tasks specific solutions were proposed, and real examples of their applications were included. Additionally, a new modelling and simulation method for actigraphy signals was presented. This method allows the modelling of physical activity behaviour of the subjects as well as the simulation of synthetic datasets based on real data. The aim of the simulation method is to facilitate the development of new techniques of actigraphy analysis in controlled scenarios. Summarizing, the methodologies purposed in this chapter are intended to provide a robust processing of actigraphy data, and to improve the interpretability actigraphy signals for outpatients monitoring.

Acknowledgements

This work was partially funded by the European Commission: Help4Mood (contract no. FP7-ICT-2009-4: 248765). E. Fuster Garcia acknowledges to Programa Torres Quevedo from Ministerio de Educación y Ciencia, co-founded by the European Social Fund (PTQ-12-05693).

References

1. Morgenthaler T, Alessi C, Friedman L et al (2007) Practice parameters for the use of actigraphy in the assessment of sleep and sleep disorders: an update for 2007. Sleep 30(4): 519–529
2. Sadeh A, Acebo C (2002) The role of actigraphy in sleep medicine. Sleep Med Rev 6:113–124
3. Ancoli-Israel S et al (2003) The role of actigraphy in the study of sleep and circadian rhythms. Sleep 26:342–392
4. De Souza L et al (2003) Further validation of actigraphy for sleep studies. Sleep 26:81–85
5. Sadeh A (2011) The role and validity of actigraphy in sleep medicine: an update. Sleep Med Rev 15:259–267
6. Perez-Diaz de Cerio D et al (2013) The help-4mood wearable sensor network for inconspicuous activity measurement. IEEE Wireless Commun 20:50–56
7. Burton C et al (2013) Activity monitoring in patients with depression: a systematic review. J Affect Disord 145:21–28
8. Todder D, Caliskan S, Baune BT (2009) Longitudinal changes of day-time and night-time gross motor activity in clinical responders and non-responders of major depression. World J Biol Psychiatry 10:276–284
9. Finazzi ME et al (2010) Motor activity and depression severity in adolescent outpatients. Neuropsychobiology 61:33–40
10. Lemke MR, Broderick A, Zeitelberger M, Hartmann W (1997) Motor activity and daily variation of symptom intensity in depressed patients. Neuropsychobiology 36:57–61
11. Hatfield CF, Herbert J, van Someren EJW, Hodges JR, Hastings MH (2004) Disrupted daily activity/rest cycles in relation to daily cortisol rhythms of home-dwelling patients with early Alzheimer's dementia. Brain 127:1061–1074
12. Paavilainen P, Korhonen I, Ltjnen J (2005) Circadian activity rhythm in demented and non-demented nursing-home residents measured by telemetric actigraphy. J Sleep Res 14: 61–68

13. Help4mood project website. http://help-4mood.info
14. Optimi project website. http://www.opti-miproject.eu
15. Texas Instrumetns (TI) Chronos watch information [Online] Available: http://processors.wiki.ti.com/index.php/EZ430-Chronos
16. Mahdavi H et al. (2012) A wireless under-mattress sensor system for sleep monitoring in people with major depression. A: IASTED International Conference Biomedical Engineering (BioMed). Proceedings of the ninth IASTED International Conference Biomedical Engineering (BioMed 2012). Innsbruck, 2012. p 1–5
17. Fuster-Garcia E et al. (2014) Fusing actigraphy signals for outpatient monitoring. Informat Fusion (in press). http://www.sciencedirect.com/science/article/pii/S156625351400089X
18. Little RJ, Rubin DB (2002) Statistical analysis with missing data. Wiley, New York
19. Sadeh A, Sharkey KM, Carskadon MA (1994) Activity-based sleep-wake identification: an empirical test of methodological issues. Sleep 17: 201–207
20. Sazonov E, Sazonova N, Schuckers S, Neuman M, CHIME Study Group (2004) Activity-based sleep-wake identification in infants. Physiol Meas 25:1291–1304
21. Tilmanne J, Urbain J, Kothare MV, Wouwer AV, Kothare SV (2009) Algorithms for sleep–wake identification using actigraphy: a comparative study and new results. J Sleep Res 18:85–98
22. Ramsay JO, Silverman BW (2005) Functional data analysis, Springer series in statistics. Springer, New York
23. Software available at http://www.psych.mcgill.ca/misc/fda/
24. Deboor C (1978) A practical guide to splines. Springer, Berlin
25. Craven WG (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the methods of generalized cross-validation. Numer Math 31: 377–403
26. Jolliffe IT (2002) Principal component analysis. Springer, New York
27. Lee DD, Seung HS (2000) Algorithms for non-negative matrix factorization. In: NIPS. MIT Press, Cambridge, pp 556–562
28. Lee JA, Verleysen M (2007) Nonlinear dimensionality reduction. Springer, New York
29. Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv 41:3–61
30. Fuster-Garcia E, Juan-Albarracin J, Breso A, Garcia-Gomez JM (2013) Monitoring changes in daily actigraphy patterns of free-living patients. International work-conference on bioinformatics and biomedical engineering (IWBBIO) proceedings, pp 685–693
31. Simonoff JS (1996) Smoothing methods in statistics. Springer, New York

Chapter 2

Definition of Loss Functions for Learning from Imbalanced Data to Minimize Evaluation Metrics

Juan Miguel García-Gómez and Salvador Tortajada

Abstract

Most learning algorithms for classification use objective functions based on regularized and/or continuous versions of the 0-1 loss function. Moreover, the performance of the classification models is usually measured by means of the empirical error or misclassification rate. Nevertheless, neither those loss functions nor the empirical error is adequate for learning from imbalanced data. In these problems, the empirical error is uninformative about the performance of the classifier and the loss functions usually produce models that are shifted to the majority class. This study defines the loss function L_{BER} whose associated empirical risk is equal to the BER. Our results show that classifiers based on our L_{BER} loss function are optimal in terms of the BER evaluation metric. Furthermore, the boundaries of the classifiers were invariant to the imbalance ratio of the training dataset. The L_{BER}-based models outperformed the 0-1-based models and other algorithms for imbalanced data in terms of BER, regardless of the prevalence of the positive class. Finally, we demonstrate the equivalence of the loss function to the method of inverted prior probabilities, and we define the family of loss functions L_{WER} that is associated with any WER evaluation metric by the generalization of L_{BER}.

Key words Cost-sensitive learning, Imbalanced datasets, Machine learning, Loss function

Abbreviations

ACC	Accuracy
BER	Balanced Error Rate
CSL	Cost-Sensitive Learning
ERR	Error
ERR1	Error of the positive class (1-sensitivity)
FN	False Negative
FP	False Positive
TN	True Negative
TP	True Positive
WER	Weighted Error Rate

Carlos Fernández-Llatas and Juan Miguel García-Gómez (eds.), *Data Mining in Clinical Medicine*, Methods in Molecular Biology, vol. 1246, DOI 10.1007/978-1-4939-1985-7_2,

1 Introduction

Cost-Sensitive Learning (CSL) studies the problem of optimal learning with different types of loss [1]. It is based on the Bayesian decision theory that provides the procedure to perform optimal decision given a set of alternatives. CSL has been studied to solve learning from imbalanced datasets. Learning from imbalanced datasets is a difficult problem that is often found in real datasets and limits the performance and utility of predictive models when combined to other factors such as overlapping between classes. The current digitalization of massive data is uncovering this problem in multiple applications from different scopes, such as social media, biomedical data, massive sensorization, and quantum analytics. Moreover, incremental learning has to deal with changing prevalences of imbalanced datasets from which multi-center predictive analyses are required [2].

Chawla in [3] classified cost-sensitive learning within those solutions for learning from imbalanced data at algorithmic level. He compiled some advantages of using CSL for learning from imbalanced datasets. First, CSL is not encumbered by large sets of duplicated examples; second, CSL usually outperforms random resampling methods; and third, it is not always possible to apply other approaches such as smart sampling in data level algorithms. On the contrary, a general drawback of CSL is that it needs a cost-matrix to be known for different types of errors—or even examples—but this cost-matrix is usually unknown and some assumptions have to be made at design-time. Another characteristic of CSL is that it does not modify the class distribution of data the way re-sampling does, which can be considered an advantage or a drawback depending on the author or the application [3].

Breiman et al. [4] studied the connection among the distribution of training samples by class, the costs of mistakes on each class, and the placement of the decision threshold. Afterwards, Maloof [5] reviewed the connection of learning from imbalanced datasets and cost-sensitive learning. Specifically, he observed the same ROC curve when moving the decision threshold and adjusting the cost-matrix. Visa and Ralescu in [6] studied the concept learning in the presence of overlapping and imbalance in the training set and developed solutions based on fuzzy classifiers.

The conclusions of the AAAI-2000 workshop and ICML-2003 pointed out the relevance of designing classifiers which performs well across a wide range of costs and priors. The insensitiveness to class imbalance of learning algorithms may lead to better control of their behavior in critical applications and streaming data scenarios. Furthermore, He and García in [7] supported the proposition addressed by Provost in [8] to concentrate the research on the theoretical and empirical studies of how machine learning algorithms can deal most effectively with whatever data they are given.

Weiss in [9, 10] supported the idea that the use of error and accuracy lead to poor minority-class performance. Moreover, Weiss determined the utility of the area under the ROC curve such as a measure to assess overall classification performance, but useless to obtain pertinent information for the minority class. He suggested appropriate evaluation metrics that take rarity into account, such as the geometric mean and the F-measure. In conclusion, he pointed out the value of using appropriate evaluation metrics and cost-sensitive learning to address the evaluation of results and to guide the learning process, respectively.

Although recent literature focuses its attention on the characterization and use of evaluation metrics which are sensitive to the effect of imbalanced data, the evaluation metrics used in class imbalance problems have not been studied in terms of the loss function under the empirical risk that defines them. To our concern, this is the first time a loss function is defined to equal its associated empirical risk to an evaluation metric different from the empirical error. Furthermore, it is our objective to observe its optimal behavior in terms of the selected evaluation metric, illustrate its stability, and compare its performance to other approaches for learning from imbalanced datasets.

2 Theoretical Framework

A predictive model (or classifier in classification problems), $\hat{y} = f(\mathbf{x}, \alpha)$, is a function with parameters α that gives a decision from the discrete domain $\hat{y} \in \hat{y}$ defined by the supervisor, given the observation of a sample represented by $\mathbf{x} \in \hat{y}$.

In Bayesian decision theory, the loss (or cost) function $L(y, \hat{y})$ measures the consequence of deciding $\hat{y}$ given the sample $\mathbf{x}$ that actually belongs to class y. When a predictive model decides $\hat{y}$ after observing $\mathbf{x}$, it assumes a *conditional risk*,

$$R(\hat{y} \mid \mathbf{x}) = E_{y|\mathbf{x}}[L(y, \hat{y})]. \tag{1}$$

In the case of classification problems, (1) is the sum of the weighted loss over the space of possible classes,

$$R(\hat{y} \mid \mathbf{x}) = \sum_{y \in \hat{y}} L(y, \hat{y}) p(y \mid \mathbf{x}). \tag{2}$$

As a consequence, the prediction model assumes a *functional risk* that is equal to the expected conditional risk over the possible values of $\mathbf{x}$,

$$R(\alpha) = E_{\mathbf{x}}[R(\hat{y} \mid \mathbf{x})] \tag{3}$$

$$= E_{\mathbf{x}}[E_{y|\mathbf{x}}[L(y,\hat{y})]] \tag{4}$$

$$= \int_{\mathbf{x}} E_{y|\mathbf{x}}[L(y,\hat{y})]p(\mathbf{x})d\mathbf{x} \tag{5}$$

$$= \int_{\mathbf{x}} \sum_{y\in\hat{y}} L(y,\hat{y})p(y \mid \mathbf{x})p(\mathbf{x})d\mathbf{x}. \tag{6}$$

Equation 6 requires knowing the joint distribution $p(\mathrm{x},y) = p(y \mid \mathrm{x})p(\mathrm{x})$, which is not always possible. Hence, it is common to estimate an *empirical risk* by means of an observed sample $\hat{y} = \{(\mathbf{x}_i, y_i)\}, i = 1,\ldots,N, \mathbf{x} \in \hat{y}, y_i \in \hat{y}$,

$$R_{\hat{y}}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} E_{y|\mathbf{x}}[L(y,\hat{y})] \tag{7}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\sum_{y\in\hat{y}} L(y,\hat{y})p(y \mid \mathbf{x}), \tag{8}$$

where $p(y_i \mid \mathrm{x}_i) = 1$ and $p(y_{j\neq i} \mid \mathrm{x}_i) = 0$ are assumed to be observed in supervised learning. Thus, the empirical risk can be calculated as

$$R_{\hat{y}}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} L[y_i, f(\mathbf{x}_i,\alpha)]. \tag{9}$$

Furthermore, the evaluation metric that has historically been used in classification is the ERR, or its positive equivalent, accuracy. Nevertheless, when dealing with class imbalance problems, it is necessary to evaluate the performance using metrics that take into account the prevalence of the datasets. In this paper, we focus our attention on the Balanced Error Rate (BER) and on the Weighted Error Rate (WER) family to define their associated loss functions.

The evaluation of a predictive model implies the estimation of its performance in future samples by means of an evaluation metric. Ideally, this evaluation metric is the estimation of the empirical risk given an independent and representative set of test cases. For instance, when the loss function used for the evaluation is the 0-1 loss function, then the *functional risk* is the *generalization error* and the *empirical risk* is the *test error* (or their equivalents in terms of accuracy). Similarly, the L_{BER} loss function defined in Subheading 3 and the family of loss functions defined in Subheading 5 ensure the equality of their respective empirical risks with the BER and WER evaluation metrics, respectively.

Without loss of generality, we define the evaluation metrics for a two-class discrimination problem $\hat{y} = \{y_1, y_2\}$, with y_1 as the positive class and y_2 as the negative class. The problems with imbalanced data are usually defined such that the positive class is under represented (minority class) compared to the negative class (majority

class) [11]. Let the test sample be a sample of N cases, where n_1 cases are from class y_1 and n_2 cases are from class y_2. The *confusion matrix* of a predictive model takes the form:

	$\hat{y}_1$	$\hat{y}_2$	
y_1	n_{11} (TP)	n_{12} (FN)	n_1
y_2	n_{21} (FP)	n_{22} (TN)	n_2
	$\hat{n}_1$	$\hat{n}_2$	N

where n_{11} is the number of positive cases that are correctly classified (True Positive (TP)), and n_{21} is the number of negative cases that are misclassified (False Positive (FP), or type I errors). Similarly, n_{22} is the number of negative cases that are correctly classified (True Negative (TN)), and n_{12} is the number of positive cases that are misclassified (False Negative (FN), or type II errors). The evaluation metrics for a model with parameters can be defined in terms of the values from the confusion matrix:

1. Err3or (ERR)

$$\mathrm{ERR}(\alpha) = \frac{n_{12} + n_{21}}{N} \tag{10}$$

2. Error of the positive class (1-sensitivity) (ERR1)[1]

$$\mathrm{ERR}_1(\alpha) = \frac{n_{12}}{n_1} \tag{11}$$

3. Balanced Error Rate (BER)

$$\mathrm{BER}(\alpha) = \frac{1}{2}\left(\frac{n_{12}}{n_1} + \frac{n_{21}}{n_2}\right) \tag{12}$$

4. Weighted Error Rate (WER)

$$\mathrm{WER}(\alpha) = w\frac{n_{12}}{n_1} + (1-w)\frac{n_{21}}{n_2}, 0 \le w \le 1 \tag{13}$$

Observe that WER is a convex combination with parameter w that defines the family and the BER is the specific case of this family when $w = 1\hat{y}2$. Figure 1 shows the BER loss function with respect to the errors by class. It is worth noting that this function takes the value 0 when both errors are 0, 1 when both errors are 1 and the value 0.5 when one of them is 0 and the other is 1. We have used colored lines to highlight the results obtained by different imbalances.

[1] Similarly, the Error of the negative class is defined by $\mathrm{ERR}_2(\alpha) = \frac{n_{21}}{n_2}$.

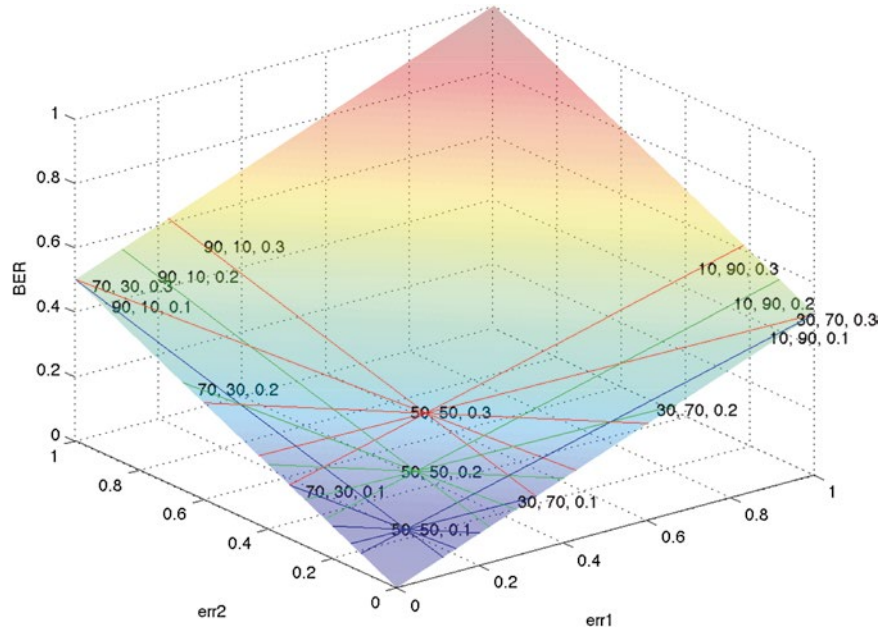

Fig. 1 BER loss function in relation to the error by class ($\mathrm{ERR}_1 = \frac{n_{12}}{n_1}$ and $\mathrm{ERR}_2 = \frac{n_{21}}{n_2}$). The *colored lines* correspond to the evaluation of class imbalance problems with the ratios [10,90], [30,70], [50,50], [70,30], and [90,10]. The labels are composed by three values: the first two show the prevalence of each class and the third shows the result of the evaluation in terms of ERR. The third label of each line is the ERR of the classifier

3 Definition of the L_{BER} Loss Function

Let L_{BER} be the loss function that defines the empirical risk that is equivalent to the BER evaluation metric:

$$L_{\mathrm{BER}}(y,\hat{y}) = \frac{N}{n_y \mid \hat{y} \mid}(1-\delta(y,\hat{y})), \tag{14}$$

where N is the number of cases, from which n_y is the number of cases of class y and

$$\delta(y,\hat{y}) = \begin{cases} 1, & \text{if} \quad y = \hat{y}, \\ 0, & \text{if} \quad y \neq \hat{y}. \end{cases} \tag{15}$$

For reasons of clarity, let us focus our demonstration on a classification problem with $\mid \hat{y} \mid = 2$, such that (14) can be specified as a 2×2 loss-matrix:

	$\hat{y}_1$	$\hat{y}_2$
y_1	0	$\frac{N}{2n_1}$
y_2	$\frac{N}{2n_2}$	0

By substitution of (14) in (9), we can demonstrate that its empirical risk is equal to the evaluation metric BER(α) (12):

$$
\begin{aligned}
R_{\hat{y}}(\alpha) &= \frac{1}{N}\sum_{N}^{i=1} L_{\mathrm{BER}}[y_i, f(\mathbf{x}_i,\alpha)] \\
&= \frac{1}{N}\sum_{N}^{i=1} \frac{N}{2n_{y_i}}(1-\delta(y_i, f(\mathbf{x}_i,\alpha))) \\
&= \frac{N}{2N}\sum_{N}^{i=1} \frac{1}{n_{y_i}}(1-\delta(y_i, f(\mathbf{x}_i,\alpha))) \\
&= \frac{1}{2}\left(\sum_{i:y_i=y_1} \frac{1}{n_1}(1-\delta(y_1, f(\mathbf{x}_i,\alpha))) \right. \\
&\quad \left. + \sum_{i:y_i=y_2} \frac{1}{n_2}(1-\delta(y_2, f(\mathbf{x}_i,\alpha))) \right) \\
&= \frac{1}{2}\left(\frac{1}{n_1}\sum_{i:y_i=y_1} (1-\delta(y_1, f(\mathbf{x}_i,\alpha))) \right. \\
&\quad \left. + \frac{1}{n_2}\sum_{i:y_i=y_2} (1-\delta(y_2, f(\mathbf{x}_i,\alpha))) \right) \\
&= \frac{1}{2}\left(\frac{1}{n_1} n_{12} + \frac{1}{n_2} n_{21} \right) = \mathrm{BER}(\alpha).
\end{aligned}
$$

4 Experiments

The following experiments are designed to (1) observe the behavior of the classifiers based on the L_{BER} from imbalanced datasets when varying the overlapping of the classes, (2) observe the sensitiveness or stability of the boundaries obtained by L_{BER} for different class imbalances, and (3) compare the performance of the *LBER-based* classifiers with SMOTE [12], which is a reference method for learning from imbalanced datasets using oversampling. Finally, we report the performance of predictive models based on the L_{BER} loss function in several real discrimination problems.

For our experiments with synthetic data, we studied two-class classification problems based on a $\mathbb{R}^2$ input space. Hence, we generated datasets following prior distributions and bidimensional Gaussian distributions that were parameterized for each experiment.

We compared classifiers based on the 0-1 loss function (c01) with those defined by the L_{BER} loss function (cLBER) which minimizes the conditional risk given the observation of **x**,

$$\hat{y}^* \leftarrow \arg\min_{\hat{y} \in \hat{y}} R(\hat{y} \mid \mathbf{x}). \tag{16}$$

Specifically, we compare our cost-sensitive learning classifiers based on L_{BER} with classical Gaussian classifiers based on generative models with free covariate matrices.

4.1 Behavior of L_{BER} When Varying the Overlapping Between Classes

The first experiment compared L_{BER}-based classifiers (cLBER) with the Gaussian classifier (c01) after training with imbalanced datasets in terms of ERR (10), BER (12), and ERR_1 (11). As pointed out by [3], the effect of the overlapping between classes is amplified when dealing with imbalanced problems. Hence, we studied the performance of the classifiers with respect to the overlapping ratio between classes.

We introduced a parameter D to control the overlapping between classes. Varying D from 1 to 5, we randomly generated a training sample composed by $N = 10{,}000$ cases with 10% of cases from class y_1 and the rest from class y_2 using the following distributions,

- Class 1, $p(y_1) = 0.1$, $p(x \mid y_1) = \hat{y}(\mu_1, \Sigma_1)$, $\mu_1 = (0,0)^T$, $\Sigma_1 = I$,
- Class 2, $p(y_2) = 0.9$, $p(x \mid y_2) = \hat{y}(\mu_2, \Sigma_2)$, $\mu_2 = (0, D)^T$, $\Sigma_2 = I$.

As can be observed, the parameter D controls the overlapping between the classes. Additionally, we randomly generated test samples composed by other $N = 10{,}000$ cases from the same distribution.

Figure 2 shows the results of the experiment. In general, cLBER always outperforms c01 in terms of BER, whereas c01 outperforms cLBER in terms of ERR. This shows the optimization that the L_{BER} loss function produces in terms of the evaluation metric BER. It is worth noting that for the cLBER classifiers the ERR and BER lines are equal.

This is due to the equilibrium that L_{BER} produces in the number of false negative and false positive cases. Meanwhile, c01 classifiers tend to classify most of the new cases as the majority class, obtaining different ERR and BER lines as a result. As expected, the evaluations of the two approaches and the two metrics converge when the overlapping decreases. Nevertheless, it is more interesting to observe that the discrepancy of the ERR and the BER for the c01

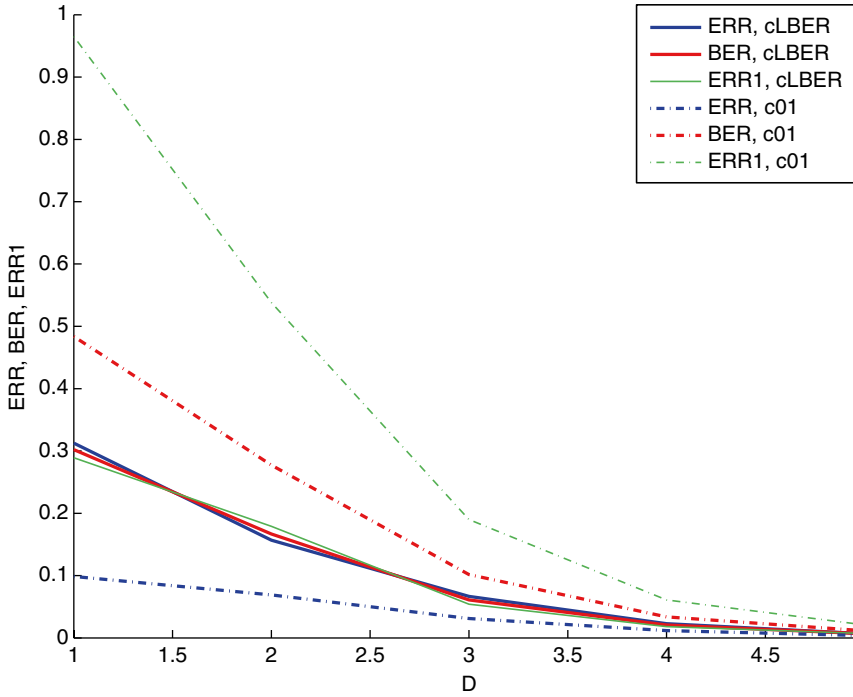

Fig. 2 ERR, BER, and ERR_1 of the cLBER and c01 classifiers with respect to D (grade of overlapping). cLBER classifiers obtain optimal results in terms of BER and stability between ERR and BER. The good behavior of cLBER classifiers is due to the control of ERR1 obtained with the L_{BER} loss function

classifiers increases with the overlapping. However, the ERR and the BER of the cLBER classifiers stay the same. This behavior is mainly explained by ERR_1, whereas the behavior is balanced for the L_{BER}-based classifiers, it is extremely high for the c01 classifiers.

4.2 Stability of the Boundaries When Varying the Class Imbalance

In the second experiment, we studied the behavior of the decision boundary of the L_{LBER}-based classifiers when varying the imbalance ratio. We generated 10^6 samples from the following distributions,

- Class 1, $p(y_1)=p$, $p(x \mid y_1)=\hat{y}(\mu_1,\Sigma_1)$, $\mu_1=(0,0)^T$, $\Sigma_1=I$
- Class 2, $p(y_2)=1-p$, $p(x \mid y_1)=\hat{y}(\mu_1,\Sigma_1)$, $\mu_2=(0,2)^T$, $\Sigma_2=I$,

where the prior probability of the positive class (y_1) took the values $[0.01, 0.1, 0.3, 0.5]$. We compared our results with those obtained when using c01 classifiers.

Figure 3 shows the bidimensional space with the 1,000 cases following the previous distribution when $p=0.1$. The boundaries obtained by the cLBER classifier are represented by thick lines, whereas the boundaries obtained by the c01 classifier are represented by thin lines. This figure clearly shows the stability of the cLBER

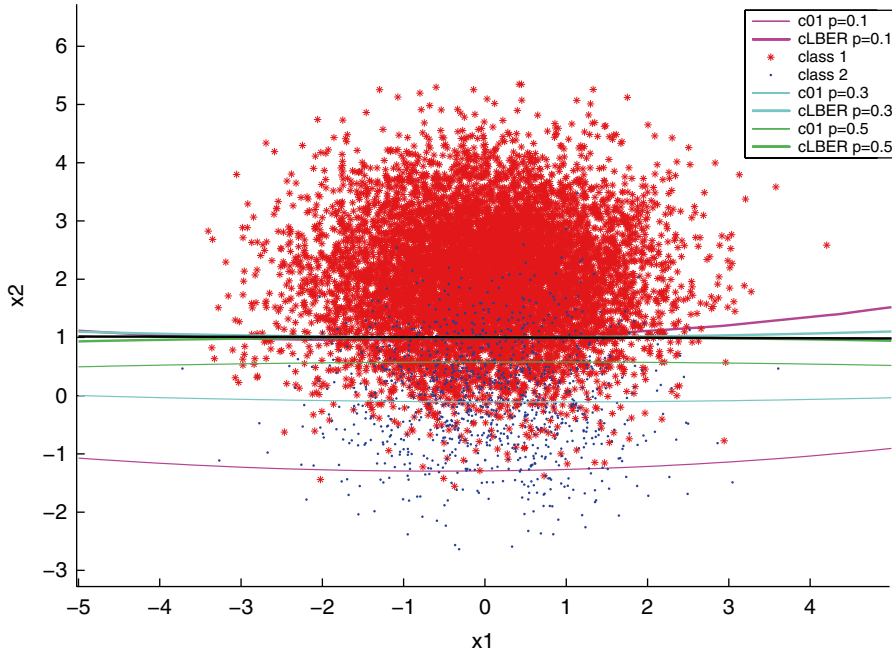

Fig. 3 Decision boundaries obtained by the cLBER classifiers (*thick lines*) and by the c01 classifiers (*thin lines*). The stability of the boundaries shows that our approach is invariant to the imbalance of the training sample and that the location of the boundary can be controlled by the loss function

with respect to the variation of the number of samples used for training. Moreover, the boundaries obtained by the cLBER classifier correspond to the boundary of the c01 classifier that is trained with a balanced dataset ($p=0.5$). This result shows that our approach is invariant to the imbalance of the training sample and that the location of the boundary can be controlled by the loss function.

4.3 Performance of L_{BER} Classifiers Compared to SMOTE

After characterizing our approach in terms of optimality and stability, we are interested in comparing its behavior with other approaches for learning from imbalanced datasets. SMOTE is a well-known algorithm that deals with imbalanced datasets by applying a synthetic minority oversampling. Specifically, we have used the implementation of SMOTE by Manohar at MathWorks based on [12] with the default parameters, which also performs a random subsampling of the majority class.

A characteristic effect of SMOTE is the local directionality of the samples in the oversampling distribution.

In this experiment, we compared the performance of our L_{BER} approach with SMOTE and c01 classifiers in terms of ERR, BER, and ERR_1. The learning process after applying SMOTE was the estimation of the Gaussian distributions, which is similar to the process for c01 classifiers. We were interested in seeing the stability of the performance when varying the imbalance ratios, i.e.,

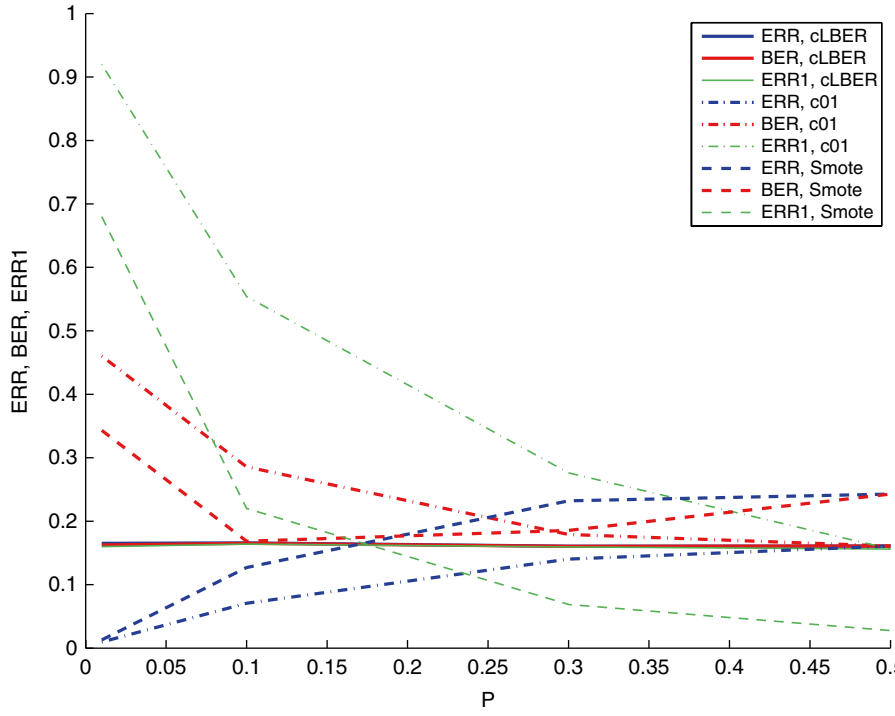

Fig. 4 ERR, BER, and ERR_1 of the cLBER, c01, and SMOTE classifiers with respect to the imbalance ratio (P). The L_{BER} is stable and invariant to the imbalance ratio in terms of ERR, BER, and ERR_1 (*solid lines*) in contrast to SMOTE, which shows a low performance for extreme and low imbalance datasets

$p=[0.01, 0.1, 0.3, 0.5]$. In order to complement our previous results, we used the same distribution as the one in Subheading 4.1 but fixing $D=2$. Figure 4 shows the results of the experiment. Both cBER and SMOTE overperformed the c01 classifiers in terms of BER and ERR_1 for moderate ($p=[0.1, 0.3[$) and extreme ($p=[0.01, 0.1[$) imbalance ratios.

The most important result of this experiment is the stability of the cLBER classifiers to changes in the imbalance ratio. In fact, the L_{BER} approach obtained constant values in the three evaluation metrics, ERR, BER, and ERR_1 (solid lines) that are directly relative to the overlapping between the distributions. This good result contrasts with the behavior obtained by SMOTE. In terms of ERR_1 (the green dashed line), the SMOTE algorithm is able to compensate moderate imbalances ($p=[0.1, 0.3[$), but it fails for extreme imbalances ($p=[0.01, 0.1[$) and low imbalances ($p=[0.3, 0.5]$). This results in a BER function (the red dashed line) with a minimum at $p=0.1$ but with worse behavior for extreme and low imbalances. Moreover, when approaching extreme imbalances, the slope of the BER function is high. As in the first experiment, we consider ERR (the blue lines) not to be

Table 1
Computational time of the c01, SMOTE, and cLBER approaches. SMOTE is computationally costly in comparison with cLBER approach. This is due to its re-sampling strategy, whereas our CSL approach is based on the modification of the conditional risk, which minimizes the learning algorithm

Training	Computational time (s)
Gaussian 0-1 (c01)	0.0013
Smote + Gaussian 0-1 (Smote)	19.0955
Gaussian + L_{BER} (cLBER)	0.0016

significant of the study the performance of the classifiers, especially for extreme imbalances.

We also studied the computational time required by each approach to perform the learning process of a classifier. Specifically, we measured the time of the learning processes of the c01, SMOTE, and cLBER classifiers when the imbalance is $p = 0.1$ on a 1.8 GHz dual-core Intel Core i5 with Mac OS X v10.7.5 and Matlab R2011a 64 bit. The results presented in Table 1 clearly show a significant difference between the SMOTE algorithm and the L_{BER} approach. This is due to the different strategy of each approach. Whereas SMOTE is a re-sampling method that involves generating a new dataset, L_{BER} implies the estimation of fix number of parameters to be used in the decision process or during a minimization process.

4.4 Performance of L_{BER} Classifiers in Real Datasets

We trained predictive models based on the L_{BER} loss function for three real datasets. One of them is the reference Contraceptive dataset from UCI [13], whereas the other two are biomedical datasets previously studied by machine-learning techniques: Brain Tumor [14, 15, 16, 17] and Postpartum depression [18]. All of them are two-class datasets. The Contraceptive dataset contains 1,473 cases ($p_1 = 0.427$) and 9 variables. The Brain Tumor dataset includes 571 cases ($p_1 = 0.257$) and 15 variables, and the Postpartum dataset is composed by 1,008 cases ($p_1 = 0.128$) represented by 19 variables. The predictive models for Contraceptive and Brain were trained with different subsets to study the response of the L_{BER} for different prevalences. Finally, we report the evaluation of the predictive models for the Postpartum depression dataset. The evaluation results include the evaluation metrics ERR, BER, and ERR_1 which were estimated by a bootstrap strategy with 200 repetitions. For the Contraceptive and the Brain Tumor datasets, we repeated the bootstrap estimation for each prevalence ten times, in order to avoid spurious results from specific subsets.

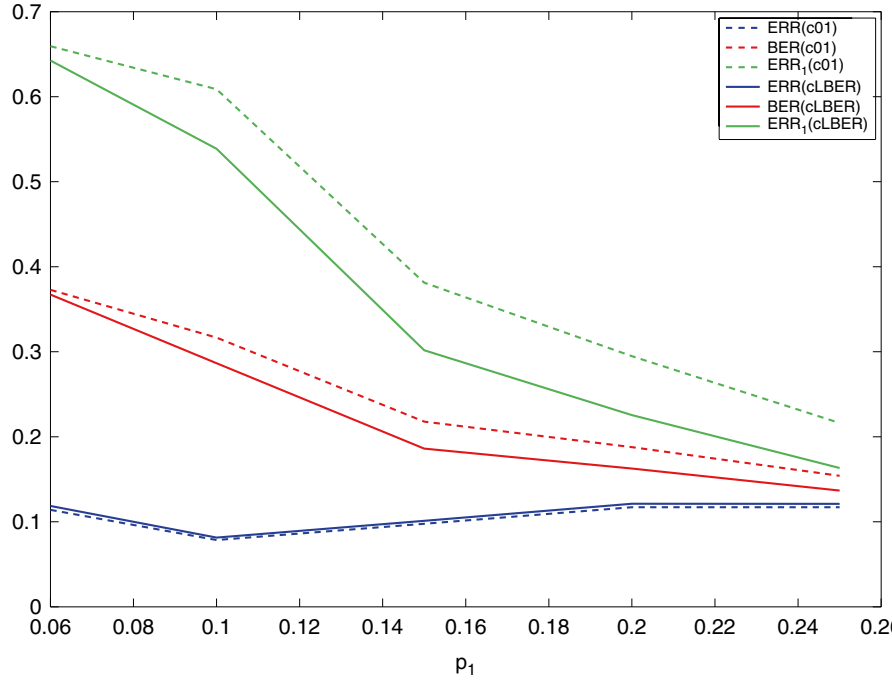

Fig. 5 BER, BER, and ERR_1 of the Brain Tumor problem obtained by the cLBER and c01 predictive models trained with datasets with different $p(y_1)$. The cLBER models outperform the c01 models in terms of BER and ERR1 independently of the prevalence

Figure 5 shows the BER, BER, and ERR_1 of the Brain Tumor problem obtained by the cLBER and c01 predictive models trained with datasets with different prevalences of the positive class ($[0.06, 0.10, 0.15, 0.20, 0.25]$). Table 2 shows the relative improvement in the Brain Tumor problem obtained by the predictive models based on the LBER loss function with respect to the 0-1 loss function in terms of BER and ERR_1.

Figure 6 shows the *BER*, *LBER*, and ERR_1 of the Contraceptive problem obtained by the cLBER and c01 predictive models trained with datasets with different prevalences of the positive class ($[0.05, 0.10, 0.15, 0.2, 0.3, 0.4]$). Table 3 shows the relative improvement in the Contraceptive problem obtained by the predictive models based on the LBER loss function with respect to the 0-1 loss function in terms of BER and ERR_1.

For the Postpartum Depression dataset, we prepared a predictive model with the full dataset $p(y_1) = 0.0128$ based on LBER, and we compared its performance with 0-1-based models. BER improved 5.6 % and ERR_1 improved 25.8 % using cLBER models with respect to the c01 models.

Table 2
Relative improvement of the *cBER* models in the Brain Tumor problem. The greatest improvement in terms of BER is obtained for a prevalence of $p(y_1) = 0.15$

$p(y_1)$	0.06	0.10	0.15	0.20	0.25
(BER(c01)-BER(cLBER))/BER(c01)	0.015	0.095	0.146	0.134	0.112
(ERR1(c01)-ERR1(cLBER))/ERR1(c01)	0.0256	0.115	0.208	0.235	0.244

ε

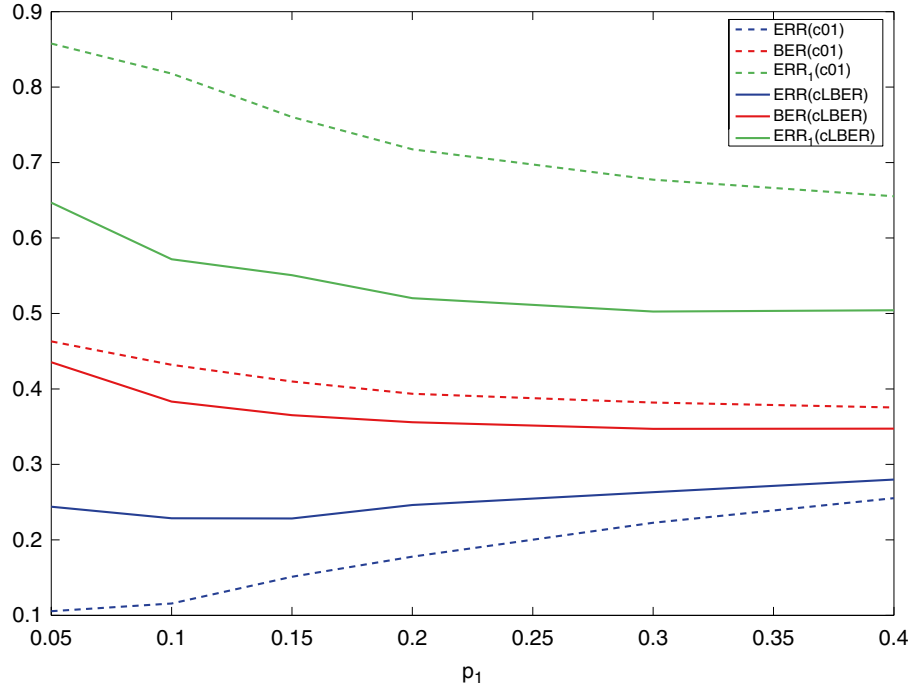

Fig. 6 BER, LBER, and ERR_1 of the Contraceptive problem obtained by the cLBER and c01 predictive models trained with datasets with different $p(y_1)$. cLBER models overperform c01 models in terms of BER and ERR1 independently of the prevalence

5 Discussion

Our experiments have shown that L_{BER} is a loss function that minimizes an empirical risk that is equal to the BER evaluation metric. As a result, the cost-sensitive classifiers based on the L_{BER} loss-function obtain the best performance in terms of their associated evaluation metric, which is BER. Furthermore, learning with L_{BER} has the theoretical property

Table 3
Relative improvement of the cBER models in the Contraceptive problem. The highest improvement in terms of BER is obtained for a prevalence of $p(y_1) = 0.10$

$p(y_1)$	0.05	0.1	0.15	0.2	0.3	0.4
(BER(c01)-BER(cLBER))/BER(c01)	0.0595	0.1128	0.1090	0.0960	0.0913	0.0746
(ERR1(c01)-ERR1(cLBER))/ERR1(c01)	0.2459	0.3008	0.2756	0.2751	0.2581	0.2307

of being insensitive to the imbalance of the datasets in terms of the challenges summarized by Chawla et al. [3]. Specifically, for the Gaussian models used in our experiments, the different values of the loss function cause a shift in the decision boundary.

As the results of the first experiment in Subheading 4.1 show, the classifiers based on the L_{BER} outperformed those based on the 0-1 loss functions in terms of BER. This result is compatible with the fact that the L_{BER}-based classifiers minimize a risk equivalent that is to the BER evaluation metric, whereas the 0-1 classifiers minimize the empirical error. The extension of this result to the L_{WER} family may allow designers to adapt their classifiers to the desired behavior in terms of their final results for class imbalance problems.

The bias (or independent term) of the boundaries of the *LBER-based* classifiers are independent of the prevalence of the classes in the training dataset. This fact can be easily observed in the results of our second experiment, in contrast to the behavior of the Gaussian classifier. To demonstrate this behavior, we can consider each term of the empirical risk of a classifier based on a generative model, where the product of $\frac{N}{n_y \mid \hat{y} \mid}$ and the posterior probability $p(y|\mathbf{x})$ (using the Bayes theorem), obtains the result $\frac{p(\mathbf{x} \mid y)}{\mid \hat{y} \mid p(\mathbf{x})}$, independent of the prior probability. Hence, an L_{BER} classifier that is trained with an imbalanced dataset is equivalent to a 0-1 classifier with equal prior probabilities.

The results of the third experiment in Subheading 4.3 demonstrate the good behavior of the approach in comparison with SMOTE. Moreover, our CSL approach has three advantages over SMOTE: (1) the computational time is significantly shorter; (2) the performance of the cLBER classifiers is insensitive to the imbalance of the classes; and (3) the estimation of the distributions (e.g., for generative models) are not disturbed by the assumptions

of the over-sampling or under-sampling procedures that modify the training sample.

In real datasets, as expected, the performance of the classifiers is dependent on the dataset; nevertheless, several features can be observed in all of the experiments. BER(cLBER) is equal to or less than BER(c01) in every dataset and every prevalence. ERR_1(cLBER) is lower than ERR_1(c01) in all the experiments. In return, ERR(cLBER) is worse than BER(c01); this is consistent with the trade-off effect produced by the BER metric.

The mean relative improvement in our experiments in terms of BER is about 10%. The best improvement in terms of LBER is obtained when the prevalence p_1 is in the interval $[0.10 \ldots 0.15]$. When the prevalence p_1 of the positive class is very small ($[0.05 \ldots 0.06]$), the improvement in terms of BER is small. Moreover, the improvement in terms of ERR_1 for Brain Tumor is also small. The cause is twofold. First, the improvement in ERR_1 is due to a great decrease in ERR_2; hence, the mean of both errors obtains a small improvement. Second, the y_1 class is not correctly represented by the small number of cases; hence, it is worse represented.

In our results with real datasets, the stability of LBER decreases with respect to the results obtained with synthetic data. This can be due to the limitation of the Gaussian models when applied to real problems. Nevertheless, our results demonstrate the improvement of the LBER approach in terms of BER for low to moderate class imbalance problems.

5.1 The L_{WER} Loss Function Family

We can generalize the L_{BER} loss function to define the L_{WER} loss function family that defines the empirical risk equivalent to WER for a given vector of weights **w**, such that $\sum_{y \in \hat{y}} w_y = 1$:

$$L_{WER}(y, \hat{y}) = w_y \frac{N}{n_y}(1 - \delta(y, \hat{y})), \tag{17}$$

where $\delta(y, \hat{y})$ is defined in (15). When $|\hat{y}| = 2$, we can establish $w_{y_1} = w$ y $w_{y_2} = 1 - w$, so the previous expression can be written as

	$\hat{y}_1$	$\hat{y}_2$
y_1	0	$w\frac{N}{n_1}$
y_2	$(1-w)\frac{N}{n_2}$	0

One more time, by substitution of (17) in (9):

$$R_{\hat{y}}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} L_{WER}[y_i, f(\mathbf{x}_i, \alpha)]$$
$$= w\frac{n_{12}}{n_1} + (1-w)\frac{n_{21}}{n_2} = WER(\alpha),$$

so we demonstrate that the empirical risk given by the L_{WER} loss function (17) is equal to WER(α) (13).

5.2 Relation to Previous Studies

We have observed some connections of our methodology with previous studies about learning from imbalanced datasets by means of cost-sensitive learning. The Library for Support Vector Machines implemented by Chang and Lin in [19] implements a similar effect than our approach for training SVMs by assigning different soft-margin constants for the positive C_1 and negative C_2 cases [20, 21]. Specifically, similar effect to our proposal can be obtained choosing the ratio $\frac{C_1}{C_2} = \frac{n_1}{n_2}$ between constants. Nevertheless, compared to the Chang and Lin implementation, our approach is directly applicable to any approach to solve decision problems and to multi-class problems.

Raskutti and Kowalczyk in [11] investigated two methods of extreme imbalance compensation for SVM, one of which was based on cost-sensitive learning. They proposed a weight balancing through the regularization constants for the minority and majority class data. Their function is equivalent to our L_{BER} loss function when $B = 0$ and $C = N$. Nevertheless, they do not explain which risk function is minimized by the learning process, or the effect of their approach in terms of the risk function. Instead, they explain the effect in terms of the ROC curve.

One of the interesting conclusions obtained in [11] is the suitability of positive one-class classification for extremely imbalanced datasets. For this result, they observed that the performance of the classifiers was slightly better at higher values of C, which is connected to large L_{BER} loss for false negative cases when the training size N is very large.

Thai-Nghe et al. in [22] proposed two empirical methods to learn SVM from imbalanced datasets, one of them by optimizing the cost ratio. Their method is able to minimize user-selected evaluation metrics by means of a grid search where the cost ratio between positive and negative classes is a hyperparameter that must be adjusted. On the one hand, the advantage of their approach is that it can be used for any classification method. On the other hand, the disadvantages of the method are the need for a tuning dataset, the computational cost, and the sensitivity to the imbalance ratio that was demonstrated in their results.

6 Conclusions

We have defined the L_{BER} loss function to make the empirical risk of a classifier equal to the BER evaluation metric. To our concern, this is the first time a loss function is defined analytically to equal its associated empirical risk to an evaluation metric. The training of classifiers based on this loss function with imbalanced datasets is equivalent to the training of those based on the 0-1 function when they are trained with balanced data. The same concept has also been generalized to the L_{WER} loss function family. Our results, which are based on synthetic data, show that our approach obtains the optimal performance of the classifiers in terms of the evaluation metric associated with the loss metrics. Moreover, we have observed a trend to the stability of the classifiers with respect to the imbalance of the dataset. The approach is computationally efficient and allows the use of the available data. The classifiers based on the L_{BER} loss function outperformed the classifiers based on the 0-1 loss function in all our experiments with real data. Finally, we have also discussed some interesting properties of the loss function derived by its definition and its use in the calculation of conditional risks.

In further work, we plan to extend the approach to other evaluation metrics and introduce the loss function into other families of classification techniques, and study the stability of the classifiers with respect to the imbalance ratios. Our final objective is to incorporate the methodology in incremental learning frameworks for biomedical streaming data and multi-center repositories.

7 Acknowledgements

We thank Carlos Sáez and the rest of the IBIME group for their interesting discussions about biomedical problems and the need for clinical decision support systems that justify the application of this work to real environments. The work presented in this paper is funded by the Spanish grant *Modelo semántico y algoritmos de Data Mining aplicados al tratamiento del Cáncer de Mama en centros de Atención Especializada* (IPT-2011-1126-900000), Ministerio de Ciencia e Innovación (INNPACTO 2011), the Spanish and EU grant *Servicio remoto de atención sanitaria basado en la prevención, autonomía y autocontrol de los pacientes* (IPT-2011-1087-900000), Ministerio de Ciencia e Innovación (INNPACTO 2011) and FEDER (Fondo Europeo de Desarrollo Regional), and the EU grant *Help4Mood: A Computational Distributed System to Support the Treatment of Patients with Major Depression* (FP7-ICT-2009-4; 248765), European Commission (Seventh Framework Program).

References

1. Elkan C (2001) In: Proceedings of the seventeenth international joint conference on artificial intelligence, pp 973–978
2. Quiñonero-Candela J, Sugiyama M, Schwaighofer A (2009) Dataset shift in machine learning. MIT Press, Cambridge
3. Chawla NV, Japkowicz N, Kotcz A (2004) SIGKDD Explor Newslett 6(1):1. doi:10.1145/1007730.1007733. http://doi.acm.org/10.1145/1007730.1007733
4. Breiman L, Stone CJ, Friedman JH, Olshen RA (1984) Classification and regression trees. Chapman & Hall, New York
5. Maloof MA (2003) In: ICML-2003 workshop on learning from imbalanced data sets II
6. Visa S, Ralescu A (2003) Learning imbalanced and overlapping classes using fuzzy sets. University of Ottawa, Washington
7. He H, García E (2009) IEEE Trans Knowl Data Eng 21(9):1263. http://dx.doi.org/10.1109/TKDE.2008.239
8. Provost F (2000) In: Proceedings of the learning from imbalanced datasets: papers from the American Association for Artificial Intelligence workshop
9. Weiss GM, Provost F (2003) J Artif Intell Res 19:315
10. Weiss GM (2004) SIGKDD Explor Newslett 6(1):7. doi:10.1145/1007730.1007734. http://doi.acm.org/10.1145/1007730.1007734
11. Raskutti B, Kowalczyk A (2004) ACM Sigkdd Explor Newslett 6(1):60. http://dl.acm.org/citation.cfm?id=1007739
12. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) J Artif Intell Res 16(1):321. http://dl.acm.org/citation.cfm?id=1622407.1622416
13. Lim T-S, Loh W-Y, Shih Y-S (2000) A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Mach Learn 40(3):203–228. doi:10.1023/A:1007608224229
14. García-Gómez JM, Tortajada S, Vidal C, Julià-Sape M, Luts J, Moreno-Torres À, Van Huffel S, Arus C, Robles M (2008) NMR Biomed 21(10):1112. doi:10.1002/nbm.1288. http://onlinelibrary.wiley.com/doi/10.1002/nbm.1288/abstract
15. García-Gómez JM, Luts J, Julià-Sape M, Krooshof P, Tortajada S, Robledo JV, Melssen W, Fuster-García E, Olier I, Postma G, Monleon D, Moreno-Torres À, Pujol J, Candiota AP, Martinez-Bisbal MC, Suykens J, Buydens L, Celda B, Van Huffel S, Arus C, Robles M (2009) Magma (New York, NY) 22(1):5. doi:10.1007/s10334-008-0146-y. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2797843/. PMID: 18989714 PMCID: PMC2797843
16. Fuster-García E, Navarro C, Vicente J, Tortajada S, García-Gómez JM, Saez C, Calvar J, Griffiths J, Julià-Sape M, Howe FA, Pujol J, Peet AC, Heerschap A, Moreno-Torres À, Martinez-Bisbal MC, Martinez-Granados B, Wesseling P, Semmler W, Capellades J, Majos C, Alberich-Bayarri À, Capdevila A, Monleon D, Marti-Bonmati L, Arus C, Celda B, Robles M (2011) Magn Reson Mater Phys Biol Med 24(1):35. doi:10.1007/s10334-010-0241-8. http://link.springer.com/article/10.1007/s10334-010-0241-8
17. Fuster-García E, Tortajada S, Vicente J, Robles M, García-Gómez JM (2012) NMR Biomed. doi:10.1002/nbm.2895. http://onlinelibrary.wiley.com/doi/10.1002/nbm.2895/abstract
18. Tortajada S, García-Gómez JM, Vicente J, Sanjuán J, de Frutos R, Martín-Santos R, García-Esteve L, Gornemann I, Gutiérrez-Zotes A, Canellas F, Carracedo A, Gratacos M, Guillamat R, Baca-García E, Robles M (2009) Methods Inf Med 48(3):291. doi:10.3414/ME0562. PMID: 19387507
19. Chang CC, Lin CJ (2011) ACM Trans Intell SystTechnol2(3):27:1.doi:10.1145/1961189.1961199
20. Osuna E, Freund R, Girosi F (1997) Support vector machines: training and applications. Massachusetts Institute of Technology, Cambridge
21. Vapnik VN (1998) Statistical learning theory, 1st edn. Wiley-Interscience, New York
22. Thai-Nghe N, Gantner Z, Schmidt-Thieme L (2010) In: The 2010 international joint conference on neural networks (IJCNN), pp 1–8. doi:10.1109/IJCNN.2010.5596486

Chapter 3

Audit Method Suited for DSS in Clinical Environment

Javier Vicente

Abstract

This chapter presents a novel online method to audit predictive models using a Bayesian perspective. The auditing model has been specifically designed for Decision Support Systems (DSSs) suited for clinical or research environments. Taking as starting point the working diagnosis supplied by the clinician, this method compares and evaluates the predictive skills of those models able to answer to that diagnosis. The approach consists in calculating the posterior odds of a model through the composition of a prior odds, a static odds, and a dynamic odds. To do so, this method estimates the posterior odds from the cases that the comparing models had in common during the design stage and from the cases already viewed by the DSS after deployment in the clinical site. In addition, if an ontology of the classes is available, this method can audit models answering related questions, which offers a reinforcement to the decisions the user already took and gives orientation on further diagnostic steps.

The main technical novelty of this approach lies in the design of an audit model adapted to suit the decision workflow of a clinical environment. The audit model allows deciding what is the classifier that best suits each particular case under evaluation and allows the detection of possible misbehaviours due to population differences or data shifts in the clinical site. We show the efficacy of our method for the problem of brain tumor diagnosis with Magnetic Resonance Spectroscopy (MRS).

Key words Decision support systems, Machine learning, Bayesian decision, Classifier comparison, Model comparison, Brain tumor diagnosis

1 Introduction

Early studies focused on evaluation of predictive models in a Decision Support System (DSS) expected the models to be able to predict "correct" diagnosis by examining the diagnostic accuracy of the DSS functioning in isolation [1, 2]. Recent evaluations, though, balance the value of testing the system and the impact of the DSS on the user's diagnostic plans [3, 4]. This means that the suggestions made by the DSS should positively influence the user's diagnostic reasoning.

In order to make a DSS useful for routine clinical use, a trustworthiness feeling needs to be created in the clinician. Thus, for a clinical diagnosis DSS based on predictive models obtained with

Carlos Fernández-Llatas and Juan Miguel García-Gómez (eds.), *Data Mining in Clinical Medicine*, Methods in Molecular Biology, vol. 1246, DOI 10.1007/978-1-4939-1985-7_3,

inference methods, showing the performance evaluated in laboratory might not suffice.

Let us assume that a DSS contains M models of classification and that every model has been trained with the same data $Z = \{(xj, t_j)\}_1^N$, a set of N samples where x_j is a data vector describing the jth sample and t_j is its associated label. The posterior probability of a model M_i can be expressed as

$$P(M_i \mid \mathbf{Z}) = \frac{P(M_i)P(\mathbf{Z} \mid M_i)}{P(\mathbf{Z})}, \tag{1}$$

Where $P(\mathbf{Z}|M_i)$ is the model likelihood or evidence for $\mathbf{Z}$. The term $P(M_i)$ is a 'subjective' prior over the model space which expresses our prior believe on the basis of experience. This term is typically overwhelmed by the objective term, the evidence [5]. $P(\mathbf{Z})$ is usually ignored since it is assumed that models are compared for the same $\mathbf{Z}$.

Typically, when a trained and evaluated classifier is introduced in a DSS, it is assumed that its predictive performance will remain in the course of time. Such assumption, though, may be unrealistic, especially in biomedical domains where dynamic conditions of the environment may change the assumed conditions in the models: modification of the data distribution, $P(\mathbf{Z})$ (covariate shift [6, 7]), inclusion of new classes through time, which modifies the prior $P(M_i)$, (prior probability shift [7]) or a change in the definition of the classes itself, $P(\mathbf{Z}|M_i)$, (concept shift [7–9]) might take place.

In order to give guidance in the user's diagnostic workflow, we propose a method that, taking as starting point the diagnosis supplied by the clinician, compares and evaluates the predictive skills of those models able to answer to such diagnosis. In addition, this auditing process should also be capable of comparing those predictive models able to answer more general diagnosis (superclasses) since they could serve as a mechanism to reinforce the decisions already taken by the clinician. Analogously, audit of predictive models discriminating subclasses of the initial diagnosis is also desirable since these predictive models might give guidance on the next steps to take in order to refine his/her diagnostic process. Such comparisons can be performed by our method if an ontology describing the relationships among the different diagnosis labels is available.

The Bayesian paradigm offers a model comparison framework that allows it to objectively assess the predictive skills of two or more classifiers by comparing the posterior odds. This approach has been typically followed for model selection under the assumption that all the models are trained with the same data. Nevertheless, this is not the case when deployed classifiers addressing similar and related discriminations have to be compared. The proposed method overcomes the limitation of the model comparison

under the Bayesian paradigm and allows the comparison of predictive models trained with different datasets and answering related questions.

The capabilities of this method are shown for the problem of brain tumor diagnosis with Magnetic Resonance Spectroscopy (MRS). The results obtained reveal the proposed method as able to objectively compare predictive models answering related problems but also to take part in the physicians' diagnostic decisions, contributing to assess the role and potential benefits of the DSS in real clinical setting.

2 Methods

2.1 Bayesian Approach

To compare two models M_m and M_l we form the posterior odds

$$\frac{P(M_m \mid \mathbf{Z})}{P(M_l \mid \mathbf{Z})} = \frac{P(M_m)}{P(M_l)} \frac{P(\mathbf{Z} \mid M_m)}{P(\mathbf{Z} \mid M_l)}. \tag{2}$$

If the odds are greater than one we choose model m, otherwise we choose model l. If we assume a uniform distribution of the prior probabilities $P(M_i)$, models M_i are ranked by evaluating the evidence [10].

$\frac{P(\mathbf{Z} \mid M_m)}{P(\mathbf{Z} \mid M_l)}$ is a ratio of the evidences and is called the Bayes Factor (BF), the contribution of the data toward the posterior odds [11].

Several techniques are available for computing BF. An exhaustive review can be found in [10]. The Bayesian Information Criterion (BIC) gives a rough approximation to the logarithm of $P(\mathbf{Z}|M_i)$ and can be used for calculating Eq. 2 [12]. We consider BIC in this study to calculate an approximation to BF.

2.2 Comparison of Models Adapted to a Clinical Environment

In order to take part in the decision workflow of a clinician, a DSS for diagnosis based on inference models deployed in a clinical environment should inform the user, according to his/her proposed diagnosis, about which of the predictive models available are going to give a useful advise. Let us call L to the set of labels supplied by the clinician as working diagnosis, which is the preliminary diagnosis given by the clinician and is based on experience, clinical epidemiology, and early confirmatory evidence provided by ancillary studies. A sensible mechanism to decide which predictive models are audited is selecting those able to discriminate L or, at least, some of the labels l_j in L.

Additionally, the predictive models may be trained from different sets of data acquired from different patients and centres. Let us call $\mathbf{Z}_i$ to the arbitrary set of samples each model M_i has been estimated with.

In order to apply the Bayesian framework, the M_i models are required to be compared with a common dataset $\mathbf{Z}$. We propose to obtain $\mathbf{Z}$ from the samples of each $\mathbf{Z}_i$ labeled with L.

Furthermore, most of biomedical problems show equivalences and polymorphisms in the classes involved in the discrimination process. This is the case, for example, of brain tumor diagnosis where, depending on the detail level of the addressed question, a tumor can be named with different terms. Thus, depending on the detail in the diagnosis, a glioblastoma can be labeled as aggressive tumor or high-grade glial tumor [13]. This variety of labels for the same concept can be depicted in an ontology. An ontology is a specification of a conceptualisation that consists of a poset (partially ordered set) of concept types, a poset of relations between these concepts and, sometimes, a set of instances of the concepts [14]. Let us call correspondence table, CT, to a tabular structure that reflects the hierarchy between classes where each column indicates an 'is a' relationship. An example of CT for brain tumor types is given in Table 1.

Table 1
Correspondence table (CT) based on the World Health Organization (WHO) classification of tumours of the central nervous system

WHO Label	is an aggressive[a]	is a gII glial[b]	is a grade I–II[c]	is a grade III–IV[d]	is a men[e]
GLIOBLASTOMA	✓	–	–	✓	–
METASTASIS	✓	–	–	✓	–
ANAPLASTIC ASTROCYTOMA	✓	–	–	✓	–
ANAPLASTIC OLIGOASTROCYTOMA	✓	–	–	✓	–
ANAPLASTIC OLIGODENDROGLIOMA	✓	–	–	✓	–
DIFFUSE ASTROCYTOMA	–	✓	✓	–	–
OLIGOASTROCYTOMA	–	✓	✓	–	–
OLIGODENDROGLIOMA	–	✓	✓	–	–
PILOCYTIC ASTROCYTOMA	–	–	✓	–	–
FIBROUS MENINGIOMA	–	–	✓	–	✓
MENINGIOMA	–	–	✓	–	✓
MENINGOTHELIAL MENINGIOMA	–	–	✓	–	✓

[a]Aggressive tumor
[b]Glial tumor grade II
[c]Grade I or II tumor type
[d]Grade III or IV tumor type
[e]Meningioma grade II

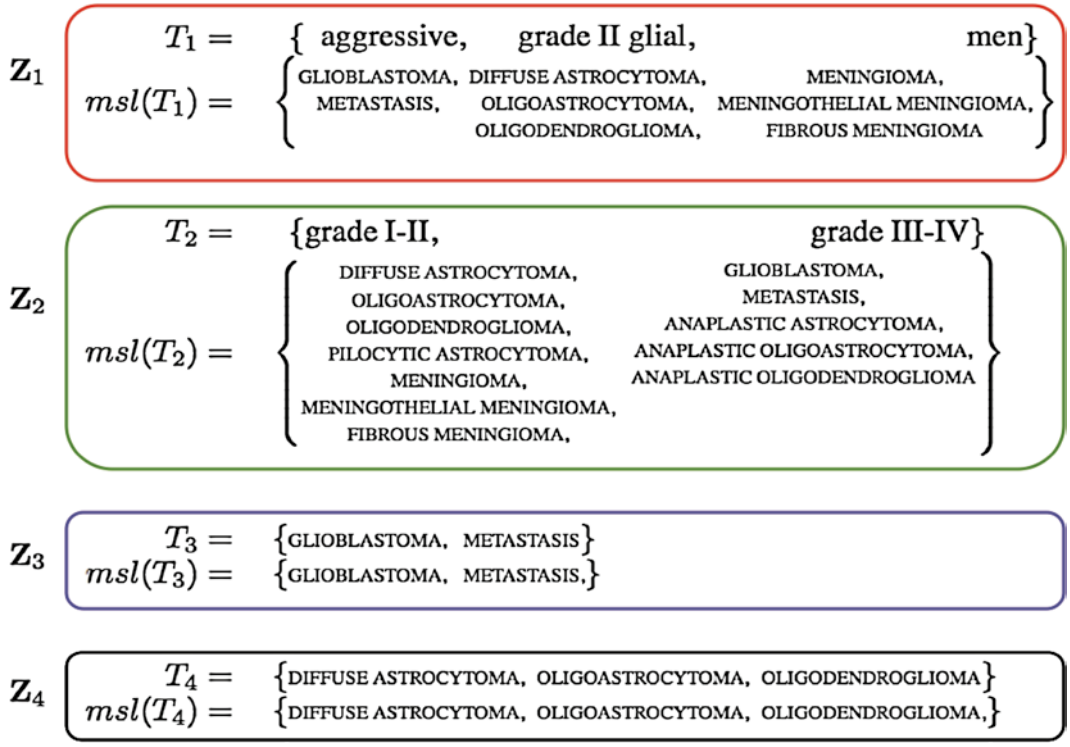

Fig. 1 Example of corpora for brain tumor diagnosis. Each corpus $\mathbf{Z}_i$ has samples labeled as T_i, grouped from the tumor type defined in $msl(T_i)$, according to the CT in Table 1

We propose a method capable of auditing models related through the hierarchy in the classes they discriminate. If we define the function $msl(t)$ that makes reference to the "*most specific label*" of each set of input variables x, we can define $z^* = (x, msl(t))$ as the sample z labeled with its most specific label (the leftmost column of the CT). Figure 1 shows an example of the labels T_i and $msl(T_i)$ of four corpora for brain tumor discrimination.

To apply our method for comparing M_i models estimated from $\mathbf{Z}_i$, we need to know the set of labels T_i each model can discriminate among. Besides, it is assumed that we can express any sample z into its most specific label, z^*. Then, we can produce $\mathbf{Z}$ as the union of the samples in $\mathbf{Z}_i$ having a most specific label equivalent to the most specific label of any element of L. Formally:

$$\mathbf{Z} = \left\{ z^{*(k)} \mid z^{*(k)} = \left(x^{(k)}, msl\left(t^{(k)}\right)\right), msl\left(t^{(k)}\right) \in \left\{ msl\left(l_j\right)\right\}\right\},$$
$$k = 1,\ldots,\left|\mathbf{Z}^{\cup}\right|,$$
$$\mathbf{Z}^{\cup} = \sum_{i=1}^{M} \mathbf{Z}_i,$$
$$j = 1,\ldots,|L|. \tag{3}$$

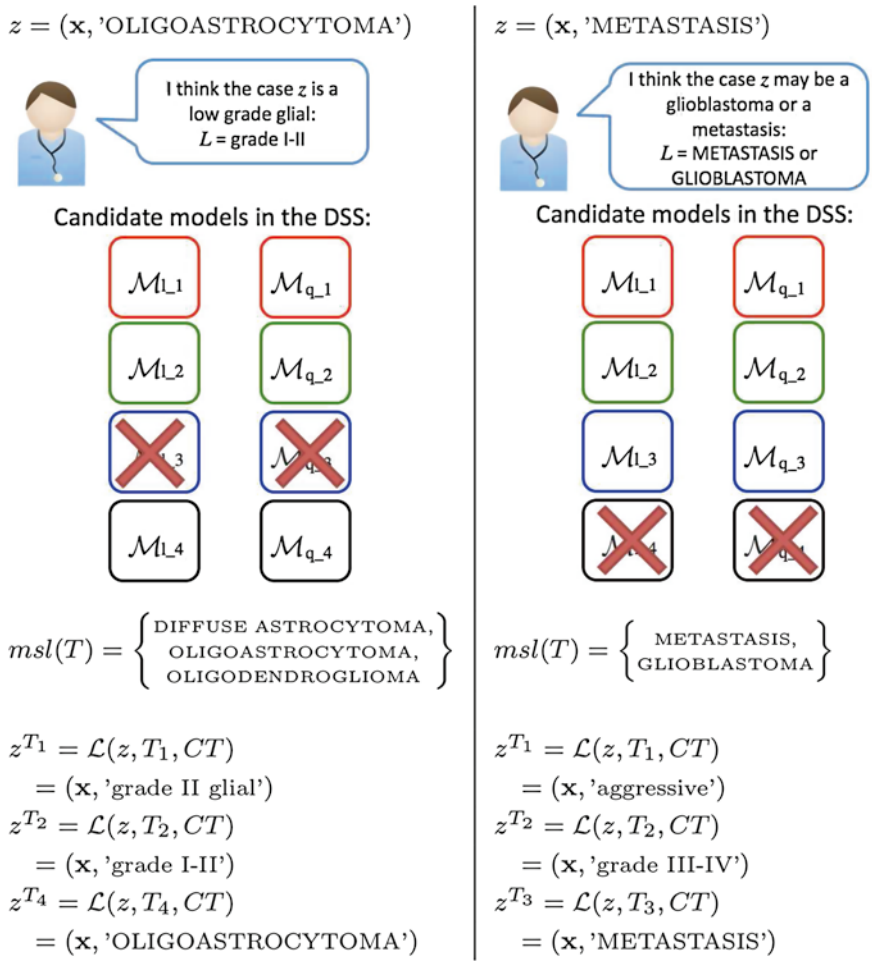

Fig. 2 Example of two scenarios: A clinician, who does not know the diagnostic label of the sample *z*, introduces *z* in the DSS and proposes a working diagnosis. Then, a selection of the models in the DSS is performed attending to that working diagnosis. Models M_{l_i} and M_{q_i} are linear and quadratic Gaussian modeled from the $\mathbf{Z}_i$ defined in Fig. 1. Once the models are selected, the common set of cases labels, *msl*(*T*) is obtained to build $\mathbf{Z}$ to calculate the static odds. Examples of the function *L* that maps the label of a sample *z* into one of the labels T_i (according to the CT) are also given

In order to calculate the evidence from $\mathbf{Z}_i$ we need a mechanism to transform the labels in $\mathbf{Z}$ into the labels T_i that a model M_i understands. Let us define $Z^{T_i} = L(Z, T_i, \text{CT})$ to be the result of applying the function L which transforms the labels of $\mathbf{Z}$ into one of the labels specified in T_i according to CT. This mechanism will allow the models to explain the data according to the labels they discriminate. Formally:

$$P(M_i \mid \mathbf{Z}) = P(M_i \mid \mathbf{Z}^{T_i}). \tag{4}$$

Figure 2 illustrates the process for obtaining the common $\mathbf{Z}$ and shows how the function L works in two different scenarios.

Once $\mathbf{Z}^{T_i}$ has been calculated for each model, we can perform the comparison of two models M_m and M_l in the light of $\mathbf{Z}$ forming the posterior odds

$$\frac{P(M_m \mid \mathbf{Z})}{P(M_l \mid \mathbf{Z})} = \frac{P(M_m)}{P(M_l)} \frac{P(\mathbf{Z}^{T_m} \mid M_m)}{P(\mathbf{Z}^{T_l} \mid M_l)}. \tag{5}$$

2.3 Audit of Dynamic Performances

If the DSS is able to store the data introduced by the users, an audit according to the predictive performances for these data is possible. This dynamic auditing will give a real vision of the predictive skills of the classifiers in the environment of the DSS.

Let us suppose that $\mathbf{Z}_{\text{DSS}}$ is a dataset of new samples introduced into the DSS by the users and that it is composed by a set of samples different to any sample in $\mathbf{Z}$. If we assume that samples in $\mathbf{Z}$ and $\mathbf{Z}_{\text{DSS}}$ are independent and identically distributed (i.i.d.), we can calculate the next posterior odds for comparing M_m and M_i:

$$\begin{aligned} \frac{P(M_m \mid \mathbf{Z}, \mathbf{Z}_{\text{DSS}})}{P(M_l \mid \mathbf{Z}, \mathbf{Z}_{\text{DSS}})} &= \frac{P(M_m)}{P(M_l)} \frac{P(\mathbf{Z}, \mathbf{Z}_{\text{DSS}} \mid M_m)}{P(\mathbf{Z}, \mathbf{Z}_{\text{DSS}} \mid M_l)} \\ &= \frac{P(M_m)}{P(M_l)} \frac{P(\mathbf{Z} \mid M_m)}{P(\mathbf{Z} \mid M_l)} \frac{P(\mathbf{Z}_{\text{DSS}} \mid \mathbf{Z}, M_m)}{P(\mathbf{Z}_{\text{DSS}} \mid \mathbf{Z}, M_l)}. \end{aligned} \tag{6}$$

We have assumed that the samples in $\mathbf{Z}$ are i.i.d. Therefore, we can split $\mathbf{Z} = \mathbf{Z}_i \cup \mathbf{Z}_{\neg i}$, where $\mathbf{Z}_i$ is the data used to design a model M_i, and $\mathbf{Z}_{\neg \mathbf{i}} = (\mathbf{Z} \backslash \mathbf{Z}_i)$ is the set of samples not used in M_i estimation. Notice that the set of samples $\mathbf{Z}_{\neg \mathbf{i}}$ is independent to $\mathbf{Z}_i$ because they are i.i.d. and it is also independent to M_i because its samples has not been used to design M_i. Then, Eq. 6 can be rewritten as

$$\frac{P(M_m \mid \mathbf{Z}, \mathbf{Z}_{\text{DSS}})}{P(M_l \mid \mathbf{Z}, \mathbf{Z}_{\text{DSS}})} = \frac{P(M_m)}{P(M_l)} \frac{P(\mathbf{Z} \mid M_m)}{P(\mathbf{Z} \mid M_l)} \frac{P(\mathbf{Z}_{\text{DSS}} \mid \mathbf{Z}_{\neg m}, \mathbf{Z}_m, M_m)}{P(\mathbf{Z}_{\text{DSS}} \mid \mathbf{Z}_{\neg l}, \mathbf{Z}_l, M_l)}. \tag{7}$$

Hence, Eq. 7 can be simplified to

$$\frac{P(M_m \mid \mathbf{Z}, \mathbf{Z}_{\text{DSS}})}{P(M_l \mid \mathbf{Z}, \mathbf{Z}_{\text{DSS}})} = \frac{P(M_m)}{P(M_l)} \frac{P(\mathbf{Z} \mid M_m)}{P(\mathbf{Z} \mid M_l)} \frac{P(\mathbf{Z}_{\text{DSS}} \mid \mathbf{Z}_m, M_m)}{P(\mathbf{Z}_{\text{DSS}} \mid \mathbf{Z}_l, M_l)} \tag{8}$$

To perform the calculation of the posterior of each model according to $\mathbf{Z}_{\text{DSS}}$, we need to apply our mechanism to translate the labels in $\mathbf{Z}_{\text{DSS}}$ into the labels, T_i, that a model M_i understands as in Eq. 4:

$$P(M_i \mid \mathbf{Z}_{\text{DSS}}) = P(M_i \mid \mathbf{Z}_{\text{DSS}}^{T_i}). \tag{9}$$

Where $\mathbf{Z}_{\text{DSS}}^{T_i} = L(\mathbf{Z}_{\text{DSS}}, T_i, \text{CT})$, the result of applying the function L which transforms the labels of $\mathbf{Z}_{\text{DSS}}$ into one of the labels specified in T_i according to CT.

Thus, applying Eq. 9 to Eq. 8 we obtain

$$\frac{P(M_m \mid \mathbf{Z}, \mathbf{Z}_{\mathrm{DSS}})}{P(M_l \mid \mathbf{Z}, \mathbf{Z}_{\mathrm{DSS}})} = \underbrace{\frac{P(M_m)}{P(M_l)}} \underbrace{\frac{P(\mathbf{Z}^{T_m} \mid M_m)}{P(\mathbf{Z}^{T_l} \mid M_l)}} \underbrace{\frac{P(\mathbf{Z}_{\mathrm{DSS}}^{T_m} \mid \mathbf{Z}_m, M_m)}{P(\mathbf{Z}_{\mathrm{DSS}}^{T_l} \mid \mathbf{Z}_l, M_l)}}, \quad (10)$$

Where $\mathbf{Z} = \mathbf{Z}_i^{T_i} \cup \mathbf{Z}_{\neg i}^{T_i}$. Since $\mathbf{Z}_i^{(T_i)} = L(\mathbf{Z}_i, T_i, \mathrm{CT}) = \mathbf{Z}_i$, the relation between $\mathbf{Z}^{T_i}$ and $\mathbf{Z}_i$ is: $\mathbf{Z}^{T_i} = \mathbf{Z}_i \cup \mathbf{Z}_{\neg i}^{T_i}$.

We call static odds to $\frac{P(\mathbf{Z}^{T_m} \mid M_m)}{P(\mathbf{Z}^{T_l} \mid M_l)}$ because it compares the prediction abilities of both models with $\mathbf{Z}$, which has been produced from the $\mathbf{Z}_i$ used to tune the parameters of each model M_i.

Analogously, we call $\frac{P(\mathbf{Z}_{\mathrm{DSS}}^{T_m} \mid \mathbf{Z}_m, M_m)}{P(\mathbf{Z}_{\mathrm{DSS}}^{T_l} \mid \mathbf{Z}_l, M_l)}$ the dynamic odds because it measures the predictive ratio of the two models with respect to a set of new samples $\mathbf{Z}_{\mathrm{DSS}}$ introduced in the DSS and not previously used during the design of the models.

Finally, to compare M models, Eq. 10 can be generalized:

$$\frac{P(M_m \mid \mathbf{Z}, \mathbf{Z}_{\mathrm{DSS}})}{\sum_{l=1}^{M} P(M_l \mid \mathbf{Z}, \mathbf{Z}_{\mathrm{DSS}})} = \frac{P(M_m) P(\mathbf{Z}^{T_m} \mid M_m) P(\mathbf{Z}_{\mathrm{DSS}}^{T_m} \mid \mathbf{Z}_m, M_m)}{\sum_{l=1}^{M} P(M_l) P(\mathbf{Z}^{T_l} \mid M_l) P(\mathbf{Z}_{\mathrm{DSS}}^{T_l} \mid \mathbf{Z}_l, M_l)}. \quad (11)$$

3 Evaluation

The evaluation of the audit method was performed with a multicenter database of MRS data of brain tumors. Gaussian discriminants were trained and audited according to the formulae described above, simulating the real scenario of a DSS working in a clinical environment.

3.1 Procedure

A database from a multicenter project was divided into two datasets. One dataset was used to create the predictive models that give support in a DSS. To do so, several interesting questions T_i were defined and $\mathbf{Z}_i$ datasets were obtained according to each T_i. Linear and quadratic Gaussian discriminant models ($\mathrm{M}_{\mathrm{l}\_i}$ and $\mathrm{M}_{\mathrm{q}\_i}$, respectively) were fed with each $\mathbf{Z}_i$. The other dataset, $\mathbf{Z}_{\mathrm{DSS}}$, was used as an independent test set. Each sample in $\mathbf{Z}_{\mathrm{DSS}}$ represented a case introduced into the DSS by a clinician in order to obtain support. Each sample in $\mathbf{Z}_{\mathrm{DSS}}$ had an associated proposed diagnosis supplied by the clinician.

To simulate a DSS running in a real clinical setting, we assume that there is an internal order in the N samples of $\mathbf{Z}_{\mathrm{DSS}}$. The procedure is as follows: the clinician introduces into the DSS the nth sample from $\mathbf{Z}_{\mathrm{DSS}}$ for diagnosis support along with his/her proposed diagnosis L_n. Then, the static odds of each predictive

model are calculated from $\mathbf{Z}$, a dataset obtained from the cases in the different $\mathbf{Z}_i$ matching L_n. The dynamic odds are also calculated for each predictive model the set of cases previously introduced in the DSS, $z_{\mathrm{DSS}}{}^1 \ldots z_{\mathrm{DSS}}{}^{n-1}$, that also match L_n. Finally, the posterior odds are calculated by combination of the static and dynamic odds as described in Eq. 11. For this evaluation we assume equal priors for each model.

A randomized algorithm to set the internal order of the samples from $\mathbf{Z}_{\mathrm{DSS}}$ was used, repeating the procedure k times. This repetition procedure prevents from obtaining variance in our results when k is big enough. In this work, k was set to 100.

3.2 Database

The database used for this evaluation consisted of 682 Single Voxel 1H MRS signals at 1.5 T at Short-TE (TE, 20–32 ms) from the European project eTUMOUR [15, 16]. Several studies applying PR-based feature extraction methods in combination with learning strategies to the eTUMOUR spectroscopy database have been previously reported [17–19].

The Peak Integration (PK) selection feature technique was applied to the spectroscopy data. PI is a method described and successfully applied in [20] that allows a reduction from the whole spectra to 15 parameters. PI has proportionality to the concentration of the main metabolites in each spectra.

All the classes considered in this study are described in Table 1 and were based on the histological classification of the central nervous system tumors as described by the WHO Classification [20].

The eTUMOUR spectroscopy database was divided into two datasets. One dataset from where four corpora $\mathbf{Z}_i$ were defined as described in Fig. 1. From the other dataset, $\mathbf{Z}_{\mathrm{DSS}}$ was obtained containing only the samples labeled as "GLIOBLASTOMA," "METASTASIS," "DIFFUSE ASTROCYTOMA," "OLIGOAS TROCYTOMA," or "OLIGODENDROGLIOMA" that corresponds to the elements in T_3 and T_4 described in Fig. 1. With such a $\mathbf{Z}_{\mathrm{DSS}}$, we were able to simulate the two scenarios depicted in Fig. 2: One scenario where the clinician would provide a general working diagnosis L of "grade I–II" tumor when dealing with samples labeled as any of the elements in T_4; and a second scenario where the clinician would express the differential diagnosis L of "METASTASIS or GLIOBLASTOMA" when trying to diagnose samples labeled as any of the elements in T_3. Table 2 shows the number of samples available.

3.3 Classifiers

Gaussian discriminant have been selected to train the predictive models and evaluate the audit method. Parametric Gaussian discriminant functions can describe linear boundaries when the covariance matrices of all the classes are equal and quadratic decision boundaries if a covariance matrix is calculated per class [21].

Table 2
Number of brain tumour samples available for Z_i and Z_{DSS}. Each Z_i discriminates the labels T_i described in Fig. 1

Labell	Z_i	Z_{DSS}
GLIOBLASTOMA	182	126
METASTASIS	67	49
ANAPLASTIC ASTROCYTOMA	17	0
ANAPLASTIC OLIGOASTROCYTOMA	4	0
ANAPLASTIC OLIGODENDROGLIOMA	9	0
DIFFUSE ASTROCYTOMA	52	19
OLIGOASTROCYTOMA	17	7
OLIGODENDROGLIOMA	23	17
PILOCYTIC ASTROCYTOMA	17	0
FIBROUS MENINGIOMA	13	0
MENINGIOMA	40	0
MENINGOTHELIAL MENINGIOMA	23	0
aggressive	279	175
gII glial	92	43
grade I–II	185	43
grade III–IV	279	175
men	76	0

Dataset	Number of samples
$\mathbf{Z}_1$	447
$\mathbf{Z}_2$	464
$\mathbf{Z}_3$	249
$\mathbf{Z}_4$	92
$\mathbf{Z}_{DSS}$	218

Gaussian discriminant were chosen because calculating the complexity of each predictive model, which is required for the BIC criterion, is straightforward.

Complexity in Gaussian discriminant is measured in terms of the cardinality of the mean vector, the covariance matrix and the prior probabilities associated to each class. Linear and quadratic Gaussian discriminants were calculated for each corpus. Table 3 shows the complexities associated to each model.

Table 3
Number of parameters associated to each predictive model

Complexity		GL (Gauss linear)	GQ (Gauss quadratic)
(in parameters)	(2 outputs)	152	272
	(3 outputs)	168	408

4 Results

According to the diagnosis proposed to each sample in $\mathbf{Z}_{\text{DSS}}$, two scenarios can be identified. A set of models is selected as auditing candidates (see Fig. 2) attending to the proposed diagnosis. Then, the audit model calculates the performance of these relevant candidates. Figure 3 shows the average likelihood calculated for both scenarios. The likelihood was averaged by $k = 100$, the times that the experiment was repeated to avoid variances in the results, and was calculated from all the samples in $\mathbf{Z}_{\text{DSS}}$ whose proposed diagnosis L was "grade I–II" (top) or "METASTASIS or GLIOBLASTOMA" (bottom) and from the common $\mathbf{Z}$ obtained from the $\mathbf{Z}_i$ of the eligible predictive models according to L. In general, when the likelihood of the models is measured from $\mathbf{Z}$, better results are reported compared to the likelihood measured from $\mathbf{Z}_{\text{DSS}}$. This behavior is expected because the samples in $\mathbf{Z}$ have been used to tune the parameters of the predictive models, whilst the samples in $\mathbf{Z}_{\text{DSS}}$ remain totally independent. Nevertheless, the differences in the average likelihood are of no importance for each predictive model except in the quadratic Gaussian model that discriminates T_4, where the likelihood from $\mathbf{Z}_{\text{DSS}}$ drops dramatically compared to the likelihood from $\mathbf{Z}_{\text{DSS}}$. This may be due to the complexity inherent to the discrimination of the three classes in T_4.

Another interesting view of the usefulness of our audit method for clinical DSSs is depicted in Fig. 4, which shows the evolution of the posterior odds. The X axis represents the samples of $\mathbf{Z}_{\text{DSS}}$ introduced in the DSS in the course of time. When the clinician introduces the nth sample from $\mathbf{Z}_{\text{DSS}}$ along with L, the posterior odds of each elegible model are calculated from: $\mathbf{Z}$, a dataset obtained from the cases in the different $\mathbf{Z}_i$ matching L (static odds) and from those cases previously introduced into the DSS, $z_{\text{DSS}}{}^1 \ldots z_{\text{DSS}}{}^{n-1}$, which also match L (dynamic odds). Thus, the posterior odds in Fig. 4 is calculated using the Eq. 11 and covering the $\mathbf{Z}_{\text{DSS}}$ so that the first point in the figure corresponds to the static posterior, (since no samples in $\mathbf{Z}_{\text{DSS}}$ are evaluated yet), and the last point corresponds to the scenario where all the samples of $\mathbf{Z}_{\text{DSS}}$ have been introduced in the DSS. Although, in the limit, all the models showed similar posterior odds, those models answering more general questions (T_1 and T_2) obtained a slighly better posterior than

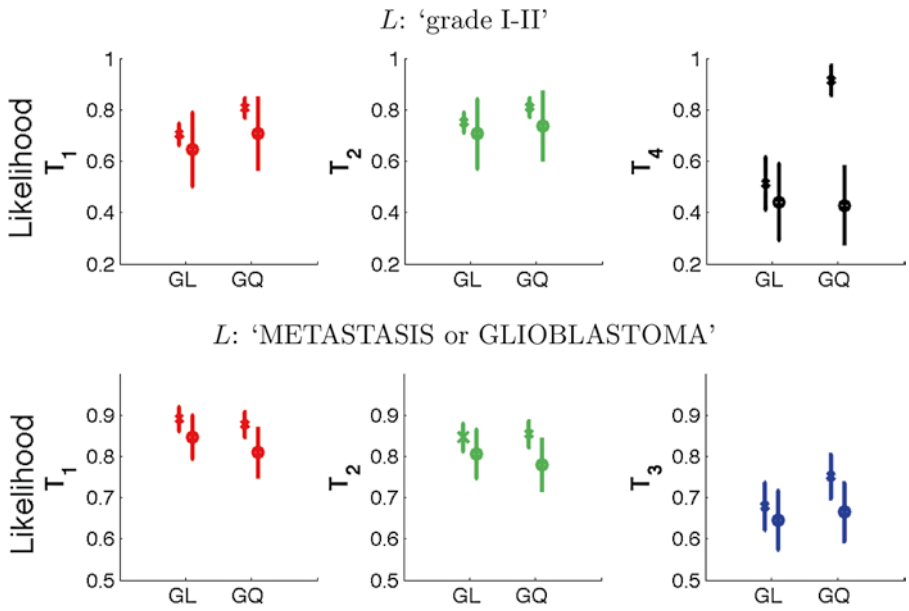

Fig. 3 Average likelihood of the models according to the scenario where the clinician proposes L as "grade I–II" (*top*) and the scenario where the working diagnosis L supplied is "METASTASIS or GLIOBLASTOMA" (*bottom*), as described by Fig. 2. Notice that $T_1, \ldots, T_4$ correspond to the discrimination labels described in Fig. 1. The likelihood of the models are depicted with the error bars for $\mathbf{Z}$ (*crosses*) and for $\mathbf{Z}_{DSS}$ (*circles*) after repeating the experiment $k=100$ times

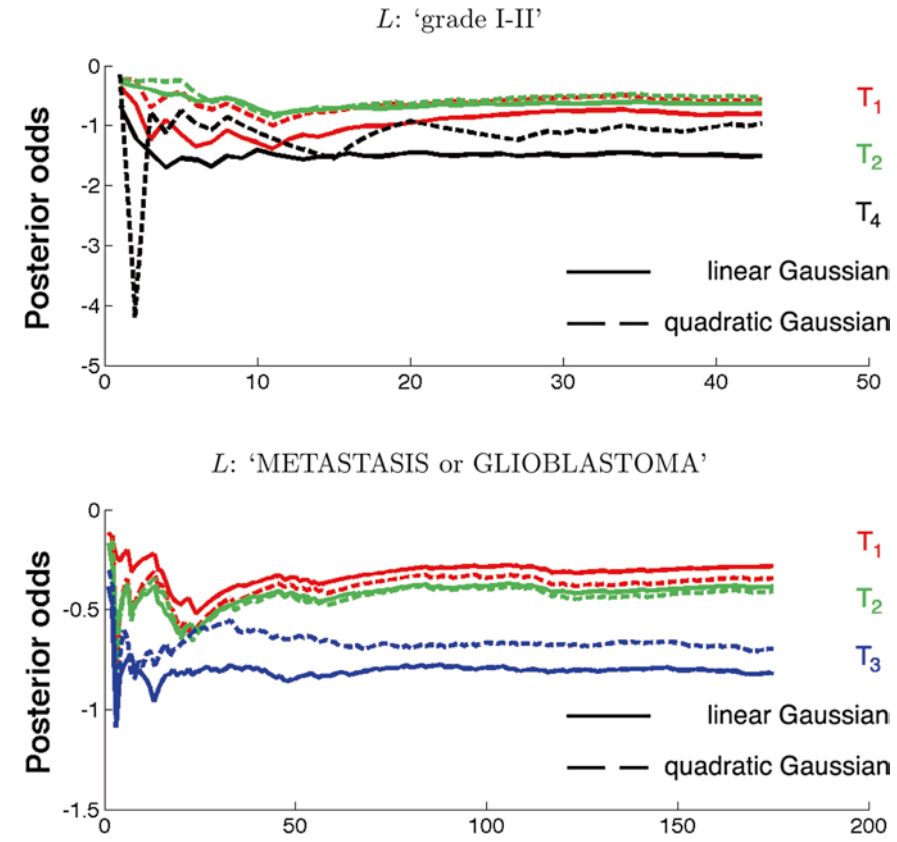

Fig. 4 Evolution of the posterior odds (in logarithmic scale) of the predictive models as the samples of $\mathbf{Z}_{DSS}$ are introduced into the DSS in the course of time

the models solving more specific questions (T_3 and T_4) for the two scenarios. These results are in agreement with previous studies [17], where classifiers discriminating superclasses of tumor types obtained better performances than classifiers discriminating more specific tumor groups. All the predictive models show stationary behaviour in the limit.

5 Discussion

5.1 Design Concerns

The proposed audit method relies on the Bayesian approach, which allows it to objectively select the more adequated model among two or more. When performing model comparison inference under a Bayesian paradigm, the models need to be compared in the light of the same data. Furthermore, the audit method is designed to work on-line, that is, the performance of each eligible model needs to be calculated for each case introduced in the DSS for decision support.

The audit method has been designed to work each time the clinician introduces a case in the DSS. Thus, it is required to have a good time response and optimizations to improve the performance are needed. Precalculation of several operations during low activity periods of the DSS (at nights, Sundays, holidays, …) could positively improve the response time of the audit method. Eligible operations to be precalculated are the calculation of the common $\mathbf{Z}$ from the $\mathbf{Z}_i$ of each model, the complexity of each model in terms of its free parameters, or the calculation of the dynamic performance attending to $\mathbf{Z}_{\mathrm{DSS}}$.

Considering that the scenario of comparison of general and specific models is common in clinical environments, we proposed a mechanism to transform the labels of $\mathbf{Z}$ into the labels T_i each model M_i can understand. By doing so, we can express $P\left(M_i \mid \mathbf{Z}\right) = P\left(M_i \mid \mathbf{Z}^{T_i}\right)$, and calculation of the posterior of each model is properly performed. In addition, we restrict the posterior calculation to only the set of samples $\mathbf{Z}$ that any comparing model can deal with. This might be seen as suboptimal because the models discriminating subclasses are imposing that the samples in $\mathbf{Z}$ should belong to such subclasses. Nevertheless, we consider this design as a compromise solution: if $\mathbf{Z}$ was obtained from the union of all the samples in each $\mathbf{Z}_i$, the situation where a model would try to predict a sample belonging to a class it cannot discriminate upon would arise and misclassifications would occur. Consequently, this would lead to an undesired side effect of punishing those classifiers discriminating very specific labels, since its evidence would dramatically drop.

Taking into account the previous concerns, the audit method can perform a consistent comparison between models in a non-painfully time allowing a "fair" comparison of classifiers discriminating superclasses, which typically will give right answers to a

general discrimination problem, and classifiers devoted to subclasses that might address more challenging questions but be trained with lower number of samples.

The design of the audit method is compatible with the multiple inheritance in an ontology. In the CT described in Table 1, each row with more than one "is a" relationship can be seen as a term that inherits from several superterms. GLIOBLASTOMA, for example, inherits from the group of tumors that can be considered as aggressive but also from those with high grading stage (grade III–IV). With the mechanism defined to obtain the candidates for comparison attending to the working diagnosis, the audit method can effectively select models where multiple inheritance in some of their labels might occur: in case that the user defined his/her working diagnosis as "Glioblastoma," classifiers discriminating this specific tumor type would be selected but, in addition, also those discriminating the label of "grade III–IV" or "aggressive," if they are available in the system.

One possible limitation of the audit method, though, is that it calculates the evidence by using the BIC. The BIC is a rough approximation to the evidence and limits the use of this audit method to the classification techniques where the calculation of the parameters is possible. Implementation of alternatives to the calculation of the evidence using regularization methods should be studied in order to allow the auditing of other classification techniques. Special attention should be taken on the performance of such regularization methods since they usually are computationally expensive and we want them to work in an online scenario where the models will be dynamically audited for each case introduced in the DSS for diagnostic support.

5.2 Decision Support for Diagnostic Confirmation or Further Research

The main purpose of the audit method is to positively influence the user's diagnostic reasoning. Starting from an initial diagnosis, all the models able to discriminate one or more of the label diagnosis supplied by the clinician are selected. Once the models are evaluated, the DSS interface can be designed in order to (a) just show the answer given by the model with the best audit performance, which will be the best suited for the case under evaluation, or (b) to show all the classifiers' answers.

By showing only the classifier with best performance able to deal with the proposed diagnosis, we omit information that can distract the user and focus his/her attention on the specific question he/she wants to be answered. Thus, the adoption of (a) is the most recommended for diagnostic confirmation, where the clinician will not invest more than 30–60 s to obtain reinforcement to his/her decision [15, 16].

Alternatively, the option of showing all the answers, (b), is best suited for clinical research purposes or diagnosis support for a non-trivial case that requires further time investment. The fact of showing

to the user the classifiers able to answer superclasses of the proposed diagnosis can give him/her arguments to decide whether the previous reasoning decision steps were right. In addition, allowing the user to observe the diagnosis support given by classifiers dealing with subclasses, can help on refining the user's reasoning process.

By asking the user at the moment of the data insertion whether he/she considers the case under evaluation as rutinary, a DSS running the proposed audit method would choose to show just the answer of the best audited model. Otherwise, all the answers, sorted by granularity of the question and by the performance obtained, will be shown.

5.3 Complementary Use to Incremental Learning Algorithms

In a clinical or research setting, the gathering, pre-process, and validation of samples is expensive and time-consuming. Usually, the PR-based DSSs rely on classification models obtained from a unique training set and the learning stops once this set has been processed [22]. If the DSS has the ability to store the data introduced by the users, this new data can be used to retrain the models. An incremental learning approach allow us to build an initial classifier with a smaller number of samples and update it incrementally when new data are collected.

Our proposed audit method can help on the use of incremental learning algorithms in two different ways: On one hand, by deciding the moment when a retraining should be performed. A stationary behavior in the evolution of the posterior odds (Fig. 4) usually indicates that the predictive capabilities of the classifiers are stable and no further classification improvement will be reached unless new training samples are introduced. This phenomenon could be used as an indicator to retrain or evolve the predictive models by using the data $\mathbf{Z}_{\mathrm{DSS}}$ already introduced into the DSS.

On the other hand, it can also help on detecting when a learning technique or discrimination problem has reached its "learning roof," that is, when the performance does not improve although new cases are used for retraining. Keeping both the last retrained version of a classifier and the one previous to the last retraining would allow to detect if no performance improvement is achieved in the course of time.

5.4 Detector of Misbehaving Models and Data Shift

Measuring both static and dynamic odds is useful to detect problems in the clinical setting: If a bias between the static and the dynamic odds is detected, it might be due to possible overtraining in the predictive models. The performance evaluated in laboratory might be optimistic for a number of reasons. Here we briefly describe some of them:

- *Low number of training samples*: The number of samples available to train a classifier is a key factor in the success (accuracy) of the results. If a classifier obtains an elevated training performance from few data, an overtraining might be happening: the

classifier has learned the characteristics of each sample and is able to perfectly discriminate them but, due to this excessive training, the classifier has a poor generalization power and limited abilities in discriminating other cases different to the ones used for training. Classifiers trained with a low amount of data might probably obtain a good performance when tested in lab, but chances are high that this discrimination ability might be spurious. That is, it is highly probable that there exists a combination of variables able to perfectly separate the low amount of data. This happens due to a phenomenon called curse of dimensionality [23]: when the amount of cases available is smaller than the number of variables used from each case, the available data becomes sparse and this sparsity makes difficult that any learning method can achieve a good result with statistical significance. In the specific domain of brain tumor diagnosis from MRS signals, if the number of cases available is not significantly greater than the number of variables extracted from the magnetic resonance variables to train the classifier, the curse of dimensionality might occur.

- *Use of a poor evaluation strategy*: Since having a great amount of data to train classifiers is not always possible, a good design of the training and evaluation stage can help to overcome or, at least, reduce the aforementioned problems. Usually, when dealing with small-sample datasets, resampling techniques can be applied when estimating the classifier performance [24]. These techniques try to optimize the use of the available samples in order to give a non-optimistic performance measure and avoiding the overtraining. In addition, the set of training samples might contain unbalanced classes. If so, a proper evaluation metric, like Balanced Accuracy Rate (BAR), that avoids overoptimistic measures, should to be taken into account for the design of a evaluation strategy.
- *Lack of an independent test set to evaluate the classifier*: The use of a test set (independent collection of data not previously seen by the classifier) to prove the generalization capability of the model is a requisite [25]. Although a good evaluation technique can take profit of the available training data and avoid as much as possible any optimistic bias in the evaluation, the use of an independent test is the best practise to establish an objective evaluation of a classifier.

If a classifier suffering from some of these conditions is deployed in a real environment for decision support, its performance will low dramatically. That is, the dynamic odds will be biased with respect to the static odds.

Furthermore, a relevant worsening of the dynamic odds with respect to the static odds might also indicate data shift in the clinical centre when acquiring the biomedical data: classifiers are very sensitive to the input they use to perform a classification. If the clinical center does not follow the MRS acquisition protocol, chances are

high that the classifiers will be unable to give a proper answer. A similar result is to be expected if the test spectra would not be preprocessed following the same steps used with the training samples. In addition, a great bias may also be an indicator of different patient populations if the training cases were gathered from clinical center(s) different to the clinical site where the DSS is deployed.

6 Conclusion

The proposed audit method, suited for DSSs in clinical or research environment, is able to compare predictive models from the initial diagnosis made by the clinician. Comparison is performed using a Bayesian framework and focusing on auditing only those models that are relevant to help on the diagnosis of the current patient.

If an ontology or hierarchy of the labels related to the biomedical domain is available, our audit model compares not only the predictive models able to answer the diagnosis proposed by the clinician, but also those models that can deal with more general labels and those dealing with subclasses of the given diagnosis. Auditing of the models dealing with superclasses can reinforce the decisions already taken by the clinician. Meanwhile, auditing of models focused on subclasses can guide the user on which further step should be taken to address a definitive diagnosis to the current case.

This audit method evaluates both the static and dynamic odds. Static odds tell us how well the predictive models deal with data they were trained with, similarly to an evaluation performed on laboratory, where the models were designed. Dynamic odds evaluation allows to study the effectiveness of the models in a real clinical environment. Measuring both static and dynamic odds is useful to detect problems in the clinical setting: If a bias between the static and the dynamic odds is detected, it might be due to possible overtraining in the predictive models, data shift in the clinical centre when acquiring the biomedical data or even an indicator of different patient populations.

Experiments on real datasets of brain tumor diagnosis with MRS have been performed. We emphasize that this audit method is an effective tool to transmit trustworthiness to the final user and potential benefit of deploying a clinical DSS in real clinical setting.

References

1. Warner HR, Toronto AF, Veasey LG, Stephenson R (1961) A mathematical approach to medical diagnosis: application to congenital heart disease. JAMA 177:75–81
2. Miller RA, Masarie FE (1990) The demise of the "Greek Oracle" model for medical diagnostic systems. Methods Inf Med 29:1–2
3. Ramnarayan P, Kapoor RR, Coren M, Nanduri V, Tomlinson AL, Taylor PM, Wyatt JC, Britto JF (2003) Measuring the impact of diagnostic decision support on the quality of clinical decision making: development of a reliable and valid composite score. J Am Med Inform Assoc 10:563–572

4. Dreiseitl S, Binder M (2005) Do physicians value decision support? A look at the effect of decision support systems on physician opinion. Artif Intell Med 33:25–30
5. MacKay DJC (1992) Bayesian interpolation. Neural Comput 4(3):415–447
6. Shimodaira H (2000) Improving predictive inference under covariate shift by weighting the log-likelihood function. J Stat Plan Inference 90:227–244
7. Moreno-Torres JG, Raeder T, Alaiz-Rodriguez R, Chawla NV, Herrera F (2012) A unifying view on dataset shift in classification. Pattern Recognit 45(1):521–530
8. Street NW, Kim Y (2001) A streaming ensemble algorithm (SEA) for large-scale classification, in KDD'01. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp 377–382, ACM
9. Maloof MA, Michalski RS (2004) Incremental learning with partial instance memory. Artif Intell 154(1–2):95–126
10. Mackay DJC (2003) Information theory, inference, and learning algorithms. Cambridge University Press, Cambridge
11. Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90(430):773–795
12. Hastie T, Tibshirani R, Friedman JH (2001) The elements of statistical learning. Springer, New York
13. Julia-Sape M, Acosta D, Majos C, Moreno-Torres A, Wesseling P, Acebes JJ, Griffiths JR, Arus C (2006) Comparison between neuroimaging classifications and histopathological diagnoses using an international multicenter brain tumor magnetic resonance imaging database. J Neurosurg 105:6–14
14. Gruber T (2008) Ontology (computer science). In: Liu L, Ozsu TM (eds) Encyclopedia of database systems. Springer, New York
15. Tate AR et al (2006) Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra. NMR Biomed 19(4):411–434
16. eTUMOUR Consortium, eTumour: Web accessible MR Decision support system for brain tumour diagnosis and prognosis, incorporating in vivo and ex vivo genomic and metabolomic data. Web site.FP6-2002-LIFESCIHEALTH 503094, VI framework programme, EC, http://www.etumour.net. Accessed 22 Apr 2013; at writing time it was temporarily unavailable
17. Garcia-Gomez JM et al (2009) Multiproject-multicenter evaluation of automatic brain tumor classification by magnetic resonance spectroscopy. MAGMA 22(1):5–18
18. Garcia-Gomez JM et al (2008) The effect of combining two echo times in automatic brain tumor classification by MRS. NMR Biomed 21(10):1112–1125
19. Luts J, Poullet J et al (2008) Effect of feature extraction for brain tumor classification based on short echo time 1H MR spectra. Magn Reson Med 60(2):288–298
20. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK (2007) WHO classification of tumours of the central nervous system. IARC Press, Lyon
21. Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugen 7:179–188
22. Tortajada S et al (2011) Incremental Gaussian Discriminant Analysis based on Graybill and Deal weighted combination of estimators for brain tumour diagnosis. J Biomed Inform 44(4):677–687
23. Bellman R (2003) Dynamic programming. Courier Dover Publication, New York
24. Martin JK, Hirschberg DS (1996) Small sample statistics for classification error rates I: error rate measurements.Tech. Rep. ICS-TR-96-22
25. Altman DG, Royston P (2000) What do we mean by validating a prognostic model? Stat Med 19(4):453–473

Chapter 4

Incremental Logistic Regression for Customizing Automatic Diagnostic Models

Salvador Tortajada, Montserrat Robles, and Juan Miguel García-Gómez

Abstract

In the last decades, and following the new trends in medicine, statistical learning techniques have been used for developing automatic diagnostic models for aiding the clinical experts throughout the use of Clinical Decision Support Systems. The development of these models requires a large, representative amount of data, which is commonly obtained from one hospital or a group of hospitals after an expensive and time-consuming gathering, preprocess, and validation of cases. After the model development, it has to overcome an external validation that is often carried out in a different hospital or health center. The experience is that the models show underperformed expectations. Furthermore, patient data needs ethical approval and patient consent to send and store data. For these reasons, we introduce an incremental learning algorithm base on the Bayesian inference approach that may allow us to build an initial model with a smaller number of cases and update it incrementally when new data are collected or even perform a new calibration of a model from a different center by using a reduced number of cases. The performance of our algorithm is demonstrated by employing different benchmark datasets and a real brain tumor dataset; and we compare its performance to a previous incremental algorithm and a non-incremental Bayesian model, showing that the algorithm is independent of the data model, iterative, and has a good convergence.

Key words Logistic regression, Incremental learning, Brain tumor diagnosis, Bayesian inference, Clinical Decision Support Systems

1 Introduction

During the last decade, a new trend in medicine is transforming the nature of healthcare from reactive to proactive. This paradigm, known as P4 medicine [1, 2], is evolving towards a personalized, predictive, preventive, and participatory medicine, which aims to be cost effective and increasingly focused on wellness. Among other key benefits, P4 medicine aspires to detect diseases at an early stage and introduces diagnosis to stratify patients and diseases to select the optimal therapy based on individual observations. This transformation relies on the availability of complex multi-level

Carlos Fernández-Llatas and Juan Miguel García-Gómez (eds.), *Data Mining in Clinical Medicine*, Methods in Molecular Biology, vol. 1246, DOI 10.1007/978-1-4939-1985-7_4,

biomedical data and the development of new mathematical and computational methods for extracting maximum knowledge from patient records, building dynamic and disease-predictive models from massive amounts of integrated clinical and biomedical data. This requirement enables the use of computer-assisted Clinical Decision Support Systems (CDSSs) for the management of individual patients.

The CDSSs are computational systems that provide precise and specific knowledge for the medical decisions to be adopted for diagnosis, prognosis, treatment, and management of patients. The CDSSs are highly related to the concept of evidence-based medicine [3, 4] since they infer medical knowledge from the biomedical databases and the acquisition protocols that are used for the development of the systems, give computational support based on evidence for the clinical practice, and evaluate the performance and the added value of the solution for each specific medical problem. Many CDSSs are based on the use of predictive models that are inferred from real-world data using statistical learning algorithms.

In this work we focus our attention on automatic diagnostic models. A medical diagnosis is a cognitive process by which a clinician attempts to identify a health disorder or a disease in a patient. The diagnosis is based on a series of data sources that serve as the input information to yield the final result. Therefore, this process can be regarded as a classification problem. Fortunately, the statistical learning methodology provides the mathematical and computational mechanisms to infer knowledge from specific data of a given domain to provide a classification model or classifier that can be useful to make prospective predictions in a clinical environment using such model [5]. However, the development of robust classifiers requires a large dataset to be acquired and at present such condition is unmet in most health organizations irrespective of the clinical problem. Indeed, this condition is often met when cases are accrued from a large number of hospitals over many years and data transferred to a centralized database. This approach has several disadvantages, and ethical approval and patient consent need to be obtained to send and store data. Furthermore, models may learn general patterns of the centralized dataset, instead of particular patterns of each individual centre, leading to overgeneralized and underperforming models. Distributed databases where the data is held at the data collecting hospitals have major advantages [6] and such a system in which classifiers can be trained without moving the data from the hospital at which it was collected would provide a practical solution. The ability to retrain the classifiers as new data accumulates is also an important requirement and to meet these needs, incremental learning algorithms may give a practical optimal solution. Incremental classifiers assume that learning is a continuous process, which goes on each time a new set of data is available. Hence, they despise waiting to gather enough data all in one set, which anyway may be undesirable and/or impractical.

Furthermore, a classification framework with a distributed architecture requires the classification models trained with data from one center to perform well when moved to another center, that is, to *generalize*. A model's performance can be assessed using new data from the same dataset following a hold-out evaluation strategy, but a further assessment of generalization requires evaluation on data from elsewhere [7]. Poor performance in new patients may arise because of deficiencies in the design of the model reaching overfitting, or because the setting of patients between the training and the new samples is different, considering factors such as healthcare systems, patient characteristics, and/or acquisition protocols. As an alternative to the re-calibration of the models [8, 9, 10], we propose the use of incremental learning algorithms for the models to adapt to the center where they are going to be used.

In this work, we review how the Bayesian inference paradigm plainly fits with the design of an incremental learning algorithm by recursively using the posterior probability of the parameters of a model as the prior belief of a new model trained when new data are available. Further, we introduce an application of the Bayesian inference paradigm to develop a logistic regression model for classification and, finally we apply this algorithm to the problem of automatic brain tumor diagnosis and show that its behavior can lead to the customization of a model to the health center where it will be finally used.

Next, we define what an incremental learning algorithm is and introduce the use of Bayesian inference to design such an algorithm. Then, Subheading 2 introduces the incremental algorithm applied for Bayesian logistic regression, describing the likelihood function of our model as well as our prior beliefs about the parameters, and the use of the Laplace approximation for estimating the posterior distribution of the parameters that can be used iteratively to customize a model to the particular patterns of a new center. Subheading 3 introduces the materials and methods used to test the algorithm and some features that the incremental algorithms must accomplish. Subheading 4 shows the results and finally Subheading 5 discusses the advantages and limitations of the incremental approach and its potential clinical interest.

1.1 Incremental Learning Definition

An incremental learning task [11] consists of a set of observations that arrive over time in subsets of samples or batches $\mathscr{S}$. We consider that a sample $\mathscr{S}_t$ arrives at time t. Each sample has n_t observations which are ordered pairs $z_{t,i} = (\mathrm{x}, y)$ where $\mathbf{x} \in \mathscr{X}^D$ are the D-dimensional covariates and $y \in \{0, 1\}$ is the class label, and where the indices denote the i-th observation of sample t. We assume that only the sample $\mathscr{S}_t$ is available in time t for training an incremental model.

We define an incremental learning algorithm as an algorithm that produces a sequence of classifiers $\mathscr{M}_1, \mathscr{M}_2, \ldots, \mathscr{M}_n$ for any

given training set of samples $\mathscr{S}_1,\mathscr{S}_2,\ldots,\mathscr{S}_n$ available at different times such that $\mathscr{M}_t$ is determined by $\mathscr{M}_{t-1}$ and $\mathscr{S}_t$. The main characteristics of this incremental learning algorithm are: (a) it should be able to learn additional information from new data without completely forgetting its previous knowledge; (b) it should not require access to previous data; (c) since each $\mathscr{M}_t$ can be viewed as the best approximation of the target application, the performance should improve over time.

This definition is related to a general problem for classification models called the *stability-plasticity dilemma* [12]. This dilemma reveals that some information may be lost when new information is learned (*gradual forgetting*) and highlights the difference between stable classifiers and plastic classifiers. Thus, the challenge is how to design a learning system that is sensitive to new input without being radically disrupted by such input. A number of authors [11, 13, 14] include in the definition of incremental learning algorithms that it should not require access to previous data. In this work, we will consider that our incremental learning algorithm will not require access to previous data since this condition may be imposed by real-world health organizations.

Another issue to be considered is the problem of the *ordering effects* in incremental learning, which has been addressed by several authors [15, 16, 17]. An incremental learning algorithm suffers from an *order effect* when there exists two or more different ordered sequences of the same instances that lead to different models. This problem would lead to an order bias. We will show later how this incremental learning algorithm shows no order effect.

Hence, we can use these three features to customize a classification model trained with data from a sample coming from the population of one hospital to the population of a different hospital.

1.2 Incremental Learning Using Bayesian Inference

The Bayesian inference introduces conceptual differences in the parameter estimation with respect to the traditional maximum-likelihood estimation [18, 19]. Mainly, the Bayesian subjective interpretation of probability as a degree of belief assumes that the parameters are random variables and therefore the inference results in a distribution of parameters instead of a unique maximum-likelihood estimate. Precisely, one of its main advantages is the use of the prior beliefs about the parameters to estimate their posterior probability. This feature is the basis of our incremental approach since the posterior probability of the parameters of one iteration is used as the prior belief for the next iteration in a straightforward way.

Hence, we will assume in general that a model $\mathscr{M}_t$ is determined by a set of parameters Θ following a distribution in time t. In the Bayesian paradigm the parameters of the model Θ are estimated given the data $\mathscr{S}$ and an overall hypothesis space $\mathscr{H}$ using the Bayes theorem,

$$p_t(\Theta \mid \mathcal{S}_t, \mathcal{H}) = \frac{p_t(\mathcal{S}_t \mid \Theta) p_t(\Theta \mid \mathcal{H})}{p_t(\mathcal{S}_t \mid \mathcal{H})} \quad (1)$$

where $p_t(\Theta \mid \mathcal{S}_t, \mathcal{H})$ is the *posterior probability* distribution of the parameters, $p_t(\mathcal{S}_t \mid \Theta)$ is the *likelihood* function, $p_t(\Theta \mid \mathcal{H})$ is the *prior probability* distribution of the parameters, and $p_t(\mathcal{S}_t \mid \mathcal{H})$ is the *evidence* or the *marginal likelihood*, defined as

$$p_t(\mathcal{S}_t \mid \mathcal{H}) = \int p_t(\mathcal{S}_t \mid \Theta) p(\Theta \mid \mathcal{H}) d\Theta \quad (2)$$

We use this approach to define our incremental algorithm by using the posterior probability estimated in time $t - 1$ as the prior probability of the model estimated in time t. That is, $p_t(\Theta \mid \mathcal{H}) = p_{t-1}(\Theta \mid \mathcal{S}_{t-1}, \mathcal{H})$.

When the model parameters are estimated, a new observation s_{new} can be classified with the *final predictive distribution* expressed as

$$p(s_{\text{new}} \mid \mathcal{S}, \mathcal{H}) = \int p(s_{\text{new}} \mid \Theta) p(\Theta \mid \mathcal{S}, \mathcal{H}) d\Theta \quad (3)$$

where $p(s_{\text{new}} \mid \Theta)$ is computed by the model, for instance a logistic regression model, with parameter Θ and it is weighted by the probability of that parameter using $p(\Theta \mid \mathcal{S}, \mathcal{H})$. Thus, what we have is an ensemble of infinite models weighted by their posterior probability of each model. The expression (3) is often too complex to solve analytically and some approximations have to be applied as we will see later.

2 Incremental Bayesian Logistic Regression

The logistic regression is a generalized linear model that is used as a discriminative model for *binary* classification problems [20, 21]. Usually, the output variable is the membership of the observation into one of two possible classes $\mathcal{Y} = \{0, 1\}$. Taking as a reference the class $y = 0$, the logarithm of the *odds ratio* of the probability of one class $p(y = 1 \mid \mathbf{x})$ and the other $p(y = 0 \mid \mathrm{x}) = 1 - p(y = 1 \mid \mathrm{x})$ can be used as a discriminative function

$$\begin{aligned} g(\mathbf{x}) &= \mathbf{w}^T \phi(\mathbf{x}) \\ &= \log\left\{\frac{p(y = 1 \mid \mathbf{x})}{p(y = 0 \mid \mathbf{x})}\right\} \end{aligned} \quad (4)$$

where $\phi(\mathrm{x})$ is an explicit and general basis function of the input such that $\phi(\mathrm{x}) = [\phi_1(\mathrm{x}), \ldots, \phi_M(\mathrm{x})]^T$, and each $\phi_m(\mathbf{x})$ defines the *m-th* basis function applied to data vector $\mathbf{x}$. By using nonlinear basis functions we allow the model to be a nonlinear function of the input vector. Using the exponential of (4) it is possible to obtain the value of the probability of class $y = 1$ given an observation $\mathbf{x}$

$$p(y=1 \mid \mathrm{x}) = \frac{\exp\{\mathrm{w}^T \phi(\mathrm{x})\}}{1+\exp\{\mathrm{w}^T \phi(\mathrm{x})\}} \tag{5}$$

This expression for the logistic regression can be used to obtain a discriminative classifying method since we can classify an object $\mathbf{x}$ into class $y=1$ if $p(y=1 \mid \mathbf{x}) > 0.5$. This is equivalent to decide that $\mathbf{x}$ belongs to class $y=1$ if $\mathrm{w}^T \phi(\mathrm{x}) > 0$.

To estimate the parameters $\mathbf{w}$ of the discriminative model $g(\mathbf{x})$, instead of using the well-known Maximum Likelihood Estimation to obtain the parameter that maximizes the likelihood, $p(\mathrm{y} \mid \mathrm{X}, \mathrm{w})$, we use a Bayesian inference approach, where we can estimate the posterior probability of the parameters $\mathbf{w}$ using

$$p(\mathrm{w} \mid \mathrm{y}, \mathrm{X}) = \frac{p(\mathrm{y} \mid \mathrm{X}, \mathrm{w}) p(\mathrm{w})}{p(\mathrm{y} \mid \mathrm{X})} \tag{6}$$

where $\mathbf{y}$ is the class vector, $\mathbf{X}$ is the data matrix, and $\mathbf{w}$ is the parameter vector. Therefore, we have to define the likelihood component, the prior probability, the evidence, and the hypothesis $\mathcal{H}$. The overall hypothesis $\mathcal{H}$ assumes that the parameters follow a multivariate Gaussian distribution. Therefore, we assume that the initial prior probability of the parameters follows a multivariate Gaussian distribution with probability $p(\mathbf{w}) = p(\mathbf{w} \mid \beta) = \mathcal{N}(\mathbf{0}, \beta^{-1}\mathbf{I})$, where β is an arbitrarily small precision. Furthermore, the first time a model is built we assume that the posterior probability will also follow a multivariate Gaussian with mean $\bar{\mathbf{w}}$ and covariance matrix $\mathbf{C}$, that is, $\mathbf{w} \sim \mathcal{N}(\bar{\mathbf{w}}, \mathbf{C})$

Now, let $\pi(\phi(\mathrm{x}_n)) = p(y=1 \mid \phi(\mathrm{x}_n))$, which is the model defined in (5) that is parameterized by the vector $\mathbf{w}$. Since the class distribution can be modeled as a Bernoulli distribution then the likelihood function can be expressed as

$$\begin{aligned} p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) &= \prod_{i=1}^{N} \left[\pi(\phi(\mathbf{x}_n))^{y_i} (1-\pi(\phi(\mathbf{x}_n)))^{1-y_i} \right] \\ &= \prod_{i=1}^{N} \frac{\left[\exp(\mathbf{w}^T \phi(\mathbf{x}_n)) \right]^{y_i}}{1+\exp(\mathbf{w}^T \phi(\mathbf{x}_n))}. \end{aligned} \tag{7}$$

We still need to calculate the evidence or marginal likelihood $p(\mathrm{y} \mid \mathrm{X}, \beta)$. The expression for this marginal likelihood is

$$\begin{aligned} p(\mathbf{y} \mid \mathbf{X}, \beta) &= \int p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) p(\mathbf{w} \mid \beta) d\mathbf{w} \\ &= \int \prod_{i=1}^{N} \frac{\left[\exp(\mathbf{w}^T \phi(\mathbf{x}_n)) \right]^{y_i}}{1+\exp(\mathbf{w}^T \phi(\mathbf{x}_n))} \mathcal{N}(\bar{\mathbf{w}}, \mathbf{C}) d\mathbf{w}. \end{aligned} \tag{8}$$

Since there is no analytical solution to this integral, we can obtain a Laplace approximation to the posterior such that our parameter posterior follows a Gaussian and thus we would be able to use it

later as a prior in subsequent computations. At the same time, we avoid the use of sampling techniques, which are computationally more expensive.

2.1 Laplace Approximation

The Laplace approximation is a method that uses a Gaussian distribution to represent a given probability density function. This approximation for Bayesian logistic regression has been used earlier [22]. Assuming that the posterior probability follows a multivariate Gaussian distribution it is possible to apply an analytical method to solve the estimation of $\mathbf{w}$.

$$p(\mathbf{w} \mid \mathbf{y}, \mathbf{X}, \beta) \approx \mathcal{N}(\mathbf{w}_{\text{MAP}}, \mathbf{C}_t), \tag{9}$$

where $\mathbf{w}_{\text{MAP}}$ is the maximum of the posterior and the covariance $\mathbf{C}_t$ is defined as

$$\mathbf{C}_t = -\left(\frac{\partial^2}{\partial \mathbf{w} \partial \mathbf{w}^T} \log p(\mathbf{y}, \mathbf{w} \mid \mathbf{X}, \beta) \right)^{-1}. \tag{10}$$

Furthermore, since the incremental learning algorithm uses the posterior probability estimated in time $t-1$ as the prior probability in the next estimation, it is required that both probabilities belong to the same class of density functions. This is achieved by applying the Laplace approximation. Therefore, we have to estimate the Maximum A Posteriori parameter and compute the curvature of the posterior at that point.

Applying the method and using the iterative Newton–Raphson [23] optimization method, the covariance matrix of the approximate posterior is

$$\mathbf{C}_t = \left(\Phi^T \mathbf{V} \Phi + \mathbf{C}_{t-1}^{-1} \right)^{-1} \tag{11}$$

and the Newton–Raphson step for obtaining the MAP parameter is

$$\mathbf{w}_t = \left(\Phi^T \mathbf{V} \Phi + \mathbf{C}_{t-1}^{-1} \right)^{-1} \left(\Phi^T \left(\mathbf{V} \Phi \mathbf{w}_{t-1} + \mathbf{y} - \mathbf{p} \right) + \mathbf{C}_{t-1}^{-1} \mathbf{w}_{\text{MAP}} \right) \tag{12}$$

where Φ is a $N \times M$ matrix with the applied basis functions to the input, and $\mathbf{V}$ is a variance diagonal matrix where $v_{nn} = p(y = 1 \mid \mathrm{x}_n)(1 - p(y = 1 \mid \mathrm{x}_n))$.

2.2 Classification of New Observations

Once the logistic regression parameters are estimated using the incremental Bayesian approach, it is possible to obtain a class prediction for a new observation $\mathbf{x}_{\text{new}}$ with the following expression

$$P(y = 1 \mid \mathrm{x}_{\text{new}}, \mathrm{X}, \mathrm{y}) = \int P(y = 1 \mid \mathrm{x}_{\text{new}}, \mathrm{w}) p(\mathrm{w} \mid \mathrm{X}, \mathrm{y}) d\mathrm{w}. \tag{13}$$

A Markov Chain Monte Carlo (MCMC) estimate of the above integral can be performed using samples simulated from our

approximate posterior where each $\mathbf{w}_s$ is simulated or drawn from $p(\mathbf{w} \mid \mathbf{y}, \mathbf{X}, \mathscr{H})$ (9), that is, $\mathbf{w}_s \sim \mathscr{N}(\mathbf{w}_{\text{MAP}}, \mathbf{C})$, such that

$$P(y = 1 \mid \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{y}) \approx \frac{1}{N} \sum_{n_s=1}^{N_s} P(y = 1 \mid \mathbf{x}_{\text{new}}, \mathbf{w}_s) = \frac{1}{N} \sum_{n_s=1}^{N_s} \frac{1}{1 + \exp(-\mathbf{w}_s^T \phi(\mathbf{x}_{\text{new}}))}, \tag{14}$$

However, although the MCMC sampling is easy to apply, it has a high computational cost because it needs a minimum number of necessary runs. It also requires the removal of a number of initial runs to guarantee the convergence to the stationary distribution. The alternative to approximating MCMC averaging is to assume that the posterior is sharply peaked around the MAP value. Therefore, the object classification can be carried out using the MAP estimate to approximate the predictive posterior probability with

$$P(y = 1 \mid \mathbf{x}_{new}, \mathbf{X}, \mathbf{y}) \approx P(y = 1 \mid \mathbf{x}_{new}, \mathbf{w}_{MAP}, \mathbf{X}, \mathbf{y}) = \frac{1}{1 + \exp(-\mathbf{w}_{MAP}^T \phi(\mathbf{x}_{new}))}. \tag{15}$$

In our experiments however the behavior of the MCMC and the MAP approximations had the same results, thus we only show the results using the MAP approximation.

3 Materials and Methods

3.1 Stability/ Plasticity Dilemma

3.1.1 Synthetic Datasets

The incremental Bayesian logistic regression (iBLR) learning algorithm proposed has been evaluated with different benchmark databases for dichotomous classification. The first three benchmarks are synthetic datasets with bidimensional data where each of the classes follows a unimodal bivariate Gaussian distribution with different mean vectors and covariance matrices obtaining different decision boundary configurations by selecting the appropriate class distributions. The values of the mean vectors and covariance matrices as well as the theoretical Bayes error, p(error), are shown in Table 1. The fourth benchmark was a synthetic bidimensional dataset where each class belongs to a concentric ring of uniformly distributed data. This benchmark shows the property of having a 0 % Bayes error. Finally, two dichotomous classification problems where each class follows a mixture of two bivariate distributions have been used to evaluate the incremental algorithm in front of non-unimodal Gaussian distributions.

Table 1
True parameters of the distributions of the class-conditional probabilities

Dataset	μ_1	μ_2	Σ_1	Σ_2	*p*(error)
A	$(2\ 0)^{\mathrm{T}}$	$(0\ 2)^{\mathrm{T}}$	$\mathbf{I}$	$\mathbf{I}$	0.08
B	$(0\quad 0)^{T}$	$(0\ 4)^{\mathrm{T}}$	$\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$	0.02
C	$(0\ 1)^{\mathrm{T}}$	$(0\ 3)^{\mathrm{T}}$	$\begin{pmatrix} 1/8 & 0 \\ 0 & 1/4 \end{pmatrix}$	$\begin{pmatrix} 1/4 & 0 \\ 0 & 1/2 \end{pmatrix}$	0.05

A total of $|\mathscr{B}|=15$ incremental training subsets were drawn. Each training set or incremental sample had 20 instances with a different prevalence in each class. Initially, a model is estimated using a prior distribution $p(\mathbf{w}\,|\,\beta)\sim\mathcal{N}(\mathbf{0},\beta^{-1}\mathbf{I})$. Then, an approximated posterior using Laplace approximation is calculated. Therefore, $p(\mathbf{w}\,|\,\mathbf{y},\mathbf{X})\sim\mathcal{N}(\mathbf{w}_{\mathrm{MAP}}^{(1)},\mathbf{C})$ and successive models are estimated using the previous posterior as a prior, obtaining $|\mathscr{B}|$ incremental models where $p(\mathbf{w}\,|\,\mathbf{y},\mathbf{X})\sim\mathcal{N}(\mathbf{w}_{\mathrm{MAP}}^{(b)},\mathbf{C})$. Furthermore, depending on the basis expansion function it is possible to obtain different decision boundaries. Finally, the predictions were obtained using a MAP approach to estimate the new test observations (15). This process was repeated 100 times to avoid any bias.

The results were compared with another incremental algorithm, namely the iGDA algorithm [24]. This algorithm has shown to behave as well as other state-of-the-art incremental algorithms and it fits with most of these synthetic datasets since it assumes that the data follow a unimodal Gaussian distribution.

3.1.2 Vehicle Silhouette Dataset

Two benchmark datasets from the UCI machine learning repository [25] were also used. The purpose of the Vehicle Silhouette dataset is to classify a given silhouette into one of four different types of vehicle using a set of 18 features. Since we have a dichotomous classification algorithm the four original classes were merged into two classes (the first class included Opel and Saab original classes, while the second included Bus and Van original classes). The dataset consisted of 846 instances. It was divided into a training partition (630 instances) and a test partition (216 instances). The training partition was split again into 7 training sets $\mathscr{S}_1,\ldots,\mathscr{S}_7$ of 90 instances with a similar prevalence to the original training dataset for each class. The sequential models obtained were tested with the previous training sets in order to observe if a gradual forgetting appeared, and with the independent test set to observe that the generalization performance increased asymptotically.

3.1.3 Wisconsin Breast Cancer Dataset

The Wisconsin Breast Cancer dataset consists of 569 instances with 30 variables from a digitalized image of a fine needle aspirate (FNA) of a breast mass. The objective in this problem is to classify the instances into a malignant (37.3 %) or a benign (62.7 %) breast tumor. The database was divided into a test partition (169 instances) and a training partition (400 instances) that were also split into five different sets of 80 instances $\mathscr{S}_1,\ldots,\mathscr{S}_5$. Each partition had the same prevalence for each class as the whole dataset.

3.2 Order Effects

A two-class multivariate Gaussian distribution synthetic dataset has been used to assess the instance level order effects for the iBLR algorithm. A training set consisted of 400 instances divided into 20 different batches with 20 instances in each one; whereas the test set consisted of 4,000 instances drawn independently and identically distributed. The samples were used for incremental learning to build consecutive models as explained before. The order effects were evaluated by permuting the training instances in 100 experiments in order to compare the distribution of the accuracies and the decision boundaries yielded by the models.

3.3 Customization to Different Health Centers

3.3.1 Brain Tumor Dataset

Finally, the present algorithm is evaluated with a proton Magnetic Resonance Spectroscopy (^{1}H-MRS) Brain Tumor database. The dataset consists of ^{1}H-MRS of brain tumor tissue that are labeled with one class among three different types of tumour classes. However, since the logistic regression is a two-class classification model, the three classes aggressive (AGG), meningioma (MEN), and low grade glial (LGG) have been transformed into only two classes: AGG and a mixture of non-aggressive (NON) that includes the MEN and the LGG classes.

The dataset used was acquired by six international centers in the framework of the INTERPRET project [26], eight in the eTUMOUR project [27], and four in the HEALTHAGENTS project [6]. The spectra acquired were single-voxel (SV) MRS signals at 1.5 T using Point-Resolved Spectroscopic Sequence (PRESS), using a short Time of Echo (STE) between 30 and 32 ms, or Stimulated Echo Acquisition Mode sequence (STEAM), using a TE of 20 ms. The acquisition was carried out avoiding areas of cysts or necrosis and with minimum contamination from the surrounding non-tumoral tissue. The volume of interest size ranged between $1.5 \times 1.5 \times 1.5$ cm^3, (3.4 mL) and $2 \times 2 \times 2$ cm^3, (8 mL), depending on tumor dimensions. The aim was to obtain an average spectroscopic representation of the largest possible part of the tumor. These signals were acquired with Siemens, General Electric (GE), and Philips instruments. The acquisition protocols included PRESS or STEAM sequences, with spectral parameters: Time of Repetition (TR) between 1,600 and 2,020 ms, TE of 20 or 30–32 ms, spectral width of 1,000–2,500 Hz, and 512, 1,024, or 2,048 data-points for STE, as described in previous studies [28]. Every training spectrum

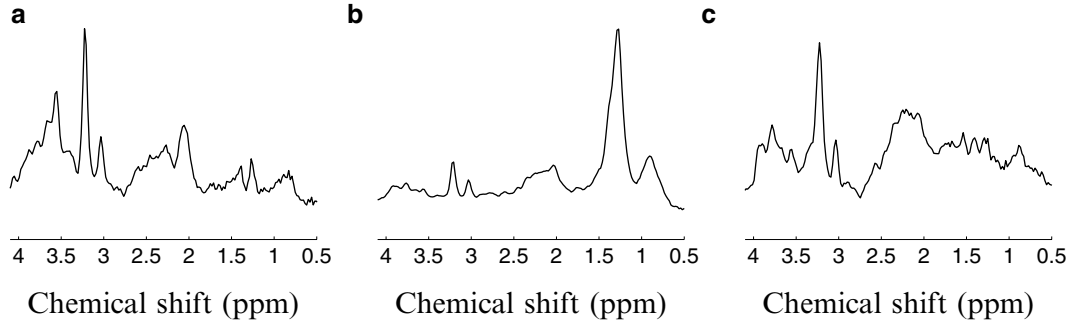

Fig. 1 Different spectra for STE. The *Y*-axis displays arbitrary units and the *X*-axis shows the chemical shift in ppm. The spectra are examples of (**a**) a low grade glioma, (**b**) an aggressive glioblastoma, and (**c**) a meningioma

and diagnosis was validated by the INTERPRET Clinical Data Validation Committee (CDVC) and expert spectroscopists [29]. The classes considered for inclusion in this study were based on the histological classification of the CNS tumors set up by the WHO [30]: glioblastomas (GBM), meningiomas (MEN), metastasis (MET), and low grade gliomas (LGG), which consists of three types of brain tumors: Astrocytoma grade II, Oligoastrocytoma grade II, and Oligodendroglioma grade II (Fig. 1).

A data acquisition protocol was defined to achieve as much as possible the compatibility of the signals coming from different hospitals [31, 32]. As a result, each signal was pre-processed according to the protocol defined in [31]. A fully automatic pre-processing pipeline was available for the training data. Besides, a semi-automatic pipeline was defined for some new file formats of the test cases from GE and Siemens manufacturers. The semi-automatic pipeline was designed to ensure compatibility of its output with the automatic one. Signal quality and the diagnosis associated with each spectrum was validated by the INTERPRET CDVC [29], the eTUMOUR Clinical Validation Committee, and expert spectroscopists. In INTERPRET and eTUMOUR the class of each case was determined by a panel of histopathologists, while in HEALTHAGENTS the class was established by the original histopathologist. Spectral patterns contain resonance peaks related to the concentration of different metabolites in the tissue analyzed which are useful for tumor classification purposes [33, 34]. Spectral peak integration is a knowledge-based feature extraction method that integrates the area under the peaks of the 15 most relevant metabolites (*see* Table 2) as a representation of the significant information contained in the spectra. To obtain the areas under the peaks we have considered an interval of 0.15 ppms from the assumed peak centre.

Table 2
Typical ppm of metabolites and other molecules observed in ^{1}H-MRS. The second resonance frequency at which a metabolite can resonate is indicated with (2) symbol

Metabolite/molecule	Resonance (ppm)
Lipid resonance at 0.92 ppm	0.92
Lipid resonance at 1.29 ppm	1.29
Lactate	1.31
Alanine	1.47
N-Acetyl groups	2.01
Glutamate + Glutamine	2.04
Glutamate + Glutamine (2)	2.46
Creatine	3.02
Choline	3.21
Myo-Inositol/Taurine	3.26
Taurine (2)	3.42
Myo-Inositol	3.53
Glycine	3.55
Alanine (2)	3.78
Creatine (2)	3.92

The dataset has been divided into four different centers in order to simulate the customization of the classification model for a hospital by adapting a general model into the specific distribution of one hospital. Data from three hospitals (CEN_0) were used to train an initial classifier. Three other groups from two hospitals (CEN_1, CEN_2, and CEN_3) were made for testing the algorithm. These groups were chosen to balance the number of samples in each center. In addition, all the centers were grouped together in order to obtain a general behavior of the convergence of the algorithm to compare with. This multicenter dataset is called $CEN_{1\text{-}3}$ and is defined as $CEN_{1\text{-}3} = \bigcup_{i=1}^{3} CEN_i$. Table 3 shows the prevalence of each class in the dataset according to the four data groups used. Each center was divided into a test set and four subsets with 20 random samples in each one. Once the initial classifier was trained, it was used to automatically classify data from the test set of the other centers. Then, the first sample $\mathscr{S}_1$ of CEN_1 was used to update the classifier with the algorithm. The same process

Table 3
The different centers and the number of instances per class. *AGG* aggressive, *NON* non-aggressive, a mixture of low grade gliomas—shown in parenthesis—and meningiomas

Center	Classes		Total
	AGG	NON	
CEN_0	111	73 (44)	184
CEN_1	108	82 (48)	190
CEN_2	114	77 (44)	191
CEN_3	120	75 (26)	195
TOTAL	453	307 (162)	760

was performed with the first sample $\mathscr{S}_1$ of the other two centers, thus obtaining a total of three new incrementally updated classifiers. After incremental updating of the classifier of each center, a new evaluation was carried out using the independent test set of the corresponding center.

A comparison with the iGDA algorithm has been carried out for the two-class classification problem in order to compare the ability of each algorithm to customize the parameters of the corresponding models to the new available data from each new center. A noteworthy fact is that, while the iGDA assumes an underlying Gaussian data distribution, the present algorithm does not assume any specific data distribution.

4 Results

The results of the simulated datasets, the different benchmark datasets, and the Brain Tumor dataset are explained in the subsections below.

4.1 Stability/ Plasticity Dilemma

4.1.1 Synthetic Datasets

The results for the different synthetic bidimensional datasets are shown in Figs. 2 and 3. Each subfigure shows the incremental results for the iBLR, compared with the iGDA algorithm and the results obtained with a discriminative logistic regression algorithm that is trained only with the data of each iteration (top frame); also, an example of the decision boundary extracted by the best iteration of the iBLR algorithm is shown (bottom frame). The results show that each new incremental model outperforms the previous ones. The incremental performance is always better than the performance of a model trained using only each new dataset. An iBLR model $\mathscr{M}_i$ had the same performance than the one obtained

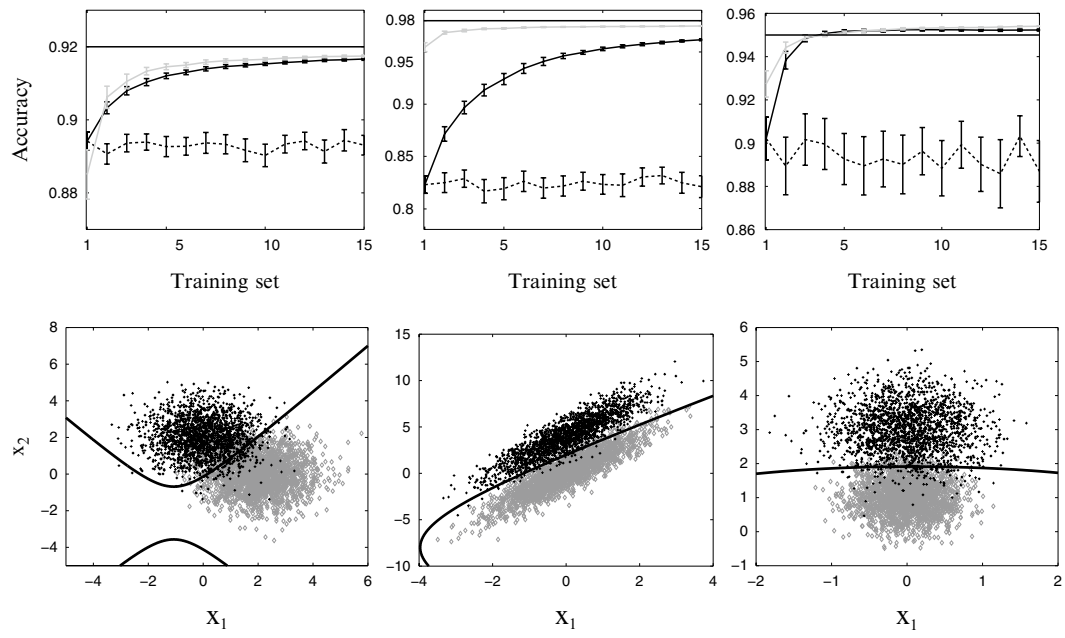

Fig. 2 Results for the first three synthetic bidimensional Gaussian datasets. In the first dataset (*left*) the error converges to the theoretical Bayes error (8 %). In the second synthetic bidimensional dataset (*center*) the iGDA error converges to the theoretical Bayes error (2 %) while the iBLR still shows room to improve. Finally, in the third synthetic bidimensional dataset (*right*), the error converges to the theoretical Bayes error (5 %). The decision boundaries of the iBLR models with the best accuracy of all the repetitions are shown on the *bottom frames*

by creating a new model from scratch with the subset $\bigcup_{j=1}^{i} S_j$ .This curve is not shown because it was overlapped with the incremental one.

The iGDA has a better performance than the iBLR algorithm when the data distributions are Gaussian (*see* Fig. 2). Nevertheless, the latter outperforms the iGDA when the data do not follow Gaussian distributions (*see* Fig. 3). This is consistent with the design of both algorithms since the iGDA assumes that the underlying data distributions follow a multivariate Gaussian while the iBLR does not consider any assumption about the data.

4.1.2 Vehicle Silhouette Dataset

Table 4 shows that there is a gradual loss of accuracy relating to the previous training datasets when new observations are introduced using Bayesian incremental algorithm. We call this effect *gradual forgetting* and it is related to the *stability-plasticity dilemma* [12, 35]. However, the overall performance increases from 89 % to 93 %. This performance is lower than the one obtained by the iGDA algorithm, which increases from 93 % to 97 %.

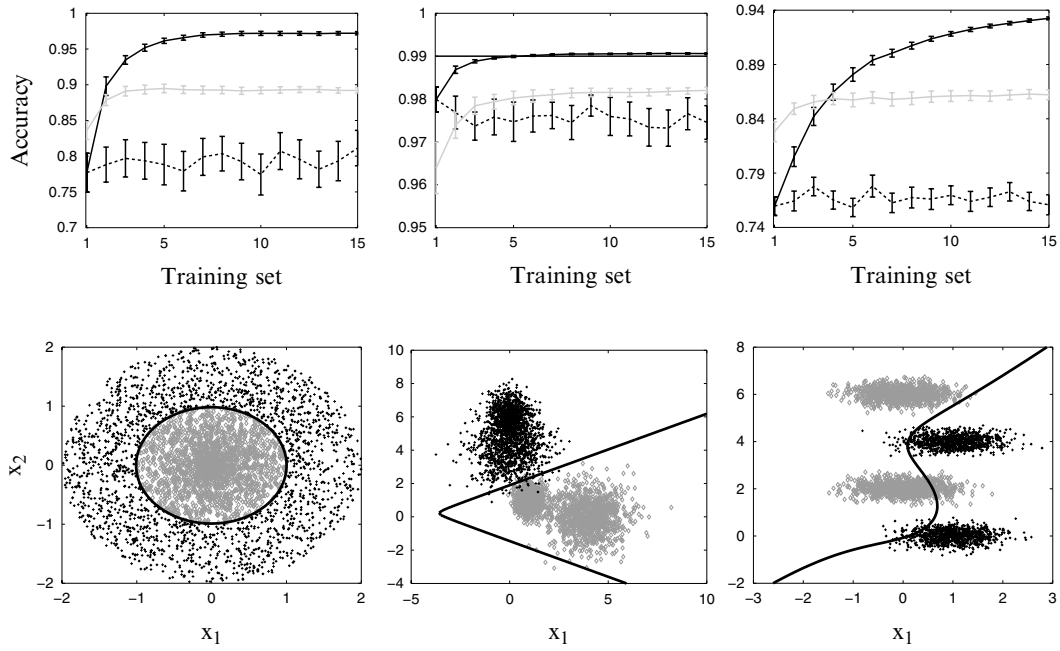

Fig. 3 Results for the last three synthetic bidimensional non-Gaussian datasets. In the first one (*left*) the error converges but the theoretical Bayes error is lower (0 %). In the second dataset (*center*) the iBLR error converges to the theoretical Bayes error (1 %). In the last dataset (*right*) the error converges, but the theoretical Bayes error is lower (0.05 %). However, the iBLR still has room to improve. On the contrary, the iGDA has a higher error rate because it is unable to describe cubic decision boundaries

4.1.3 *Wisconsin Breast Cancer Dataset*

The results are shown in Table 5. There is an improvement in overall classification as the new data are used for incremental learning, but no gradual forgetting is observed with respect to the previous datasets. In this case, the overall performance—an increase from 95 % to 97.5 %—is better than the iGDA results, which increase from 94 % to 97 %.

4.2 Order Effects

4.2.1 *Instance Level Order Effects*

The results for the evaluation of the ordering effects at instance level show that the iBLR algorithm has also a negligible order effect as Fig. 4 shows. The different permutations of the instances show that, after learning from all the available incremental batches, the obtained final models have a convergent accuracy. The convergence of the decision boundaries for the different models obtained for the bidimensional synthetic dataset is shown in Fig. 5. These results prove that the algorithm has no order effect.

4.3 Brain Tumor Dataset

The results for the automatic brain tumor diagnosis show the ability of the Bayesian logistic regression algorithm to improve the performance of the subsequent incremental models when new observations are available to re-adapt its parameters. Figure 6 shows that

Table 4
Training and test accuracy for the Vehicle Silhouette Database using a linear basis function $\phi(\mathrm{x}) = [\mathrm{x}^0, \mathrm{x}^1]$ within the incremental algorithm. The rows indicate the different datasets $\mathscr{S}_1,\ldots,\mathscr{S}_7$ and the columns show the models $\mathscr{M}_t$ built from a previous model $\mathscr{M}_{t-1}$ and the new dataset $\mathscr{S}_t$ using the posterior probability in time $t-1$ as the prior probability in time t; except $\mathscr{M}_1$ which is built from $\mathscr{S}_1$ and assuming that $p(\mathbf{w} \mid \beta) = \mathscr{N}(\mathbf{0}, \beta^{-1}\mathbf{I})$. Each column shows the average performance (%) on the current and the previous training datasets for the current model. The bottom rows (TEST, CI) indicate the evolution of the average accuracy of the models in the course of time evaluated with an independent test set and the confidence interval ($\alpha = 5\,\%$)

Dataset	$\mathscr{M}_1$	$\mathscr{M}_2$	$\mathscr{M}_3$	$\mathscr{M}_4$	$\mathscr{M}_5$	$\mathscr{M}_6$	$\mathscr{M}_7$
$\mathscr{S}_1$	96.44	95.40	95.36	95.27	95.10	95.08	95.03
$\mathscr{S}_2$	–	94.75	94.52	94.33	94.33	94.47	94.25
$\mathscr{S}_3$	–	–	94.95	94.97	94.81	94.80	94.79
$\mathscr{S}_4$	–	–	–	94.78	94.76	94.66	94.67
$\mathscr{S}_5$	–	–	–	–	94.92	94.76	94.60
$\mathscr{S}_6$	–	–	–	–	–	94.34	94.00
$\mathscr{S}_7$	–	–	–	–	–	–	94.53
TEST	89.29	91.85	92.72	93.11	93.40	93.52	93.57
CI ($\alpha = 5\,\%$)	± 0.54	± 0.70	± 0.48	± 0.37	± 0.33	± 0.32	± 0.32

Table 5
Training and test accuracy (%) for the Wisconsin Breast Cancer Database using a linear basis function $\phi(\mathrm{x}) = [\mathrm{x}^0, \mathrm{x}^1]$ within the incremental algorithm

Dataset	$\mathscr{M}_1$	$\mathscr{M}_2$	$\mathscr{M}_3$	$\mathscr{M}_4$	$\mathscr{M}_5$
$\mathscr{S}_1$	97.44	97.84	97.87	98.18	98.15
$\mathscr{S}_2$	–	97.75	98.00	98.14	98.20
$\mathscr{S}_3$	–	–	97.69	97.81	97.85
$\mathscr{S}_4$	–	–	–	97.86	98.00
$\mathscr{S}_5$	–	–	–	–	98.06
TEST	95.69	96.56	97.03	97.33	97.50
CI ($\alpha = 5\,\%$)	± 3.98	± 3.57	± 3.33	± 3.16	± 3.06

the performance of the iBLR is better than the one obtained by the iGDA. This may be due to a non-Gaussian underlying distribution of the brain tumor dataset. The performance difference of the iBLR models compared to the iGDA as we saw in the synthetic experiments supports this. Another reason is that the Bayesian approach

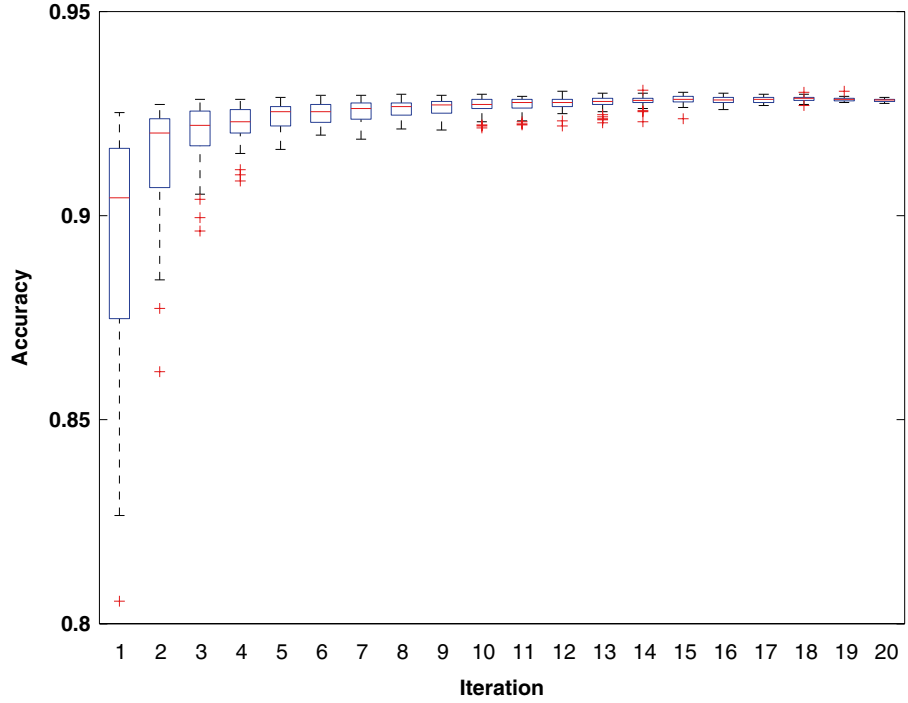

Fig. 4 Boxplots of the accuracy of the models trained with different permutations of the instances. The *X*-axis shows the iterations of the incremental models. The figure shows the results for the two-dimensional synthetic database with 20 iterations

regularizes better than the traditional maximum-likelihood approach which explains the difference in the first iteration where no incremental learning has been performed yet. The CEN_3 shows the better improvement among the other centers, mainly due to the differences in the prevalence of the mixture of non-aggressive tumous where CEN_3 shows more meningiomas and less low grade gliomas than the other centers.

5 Discussion

In this work we introduce an algorithm for updating the parameters of a discriminative logistic regression model by using the posterior probability approximation of iteration t as a prior probability on iteration $t + 1$. Unlike many previous works [36, 37, 38, 39, 40], no access to previous data is allowed, which satisfies a common constraint found in many organizations where the data may be distributed [24]. Hence, we assume that only the data of the current batch t is available in time t. Thus, the classification model is able to classify data of a future batch $t + 1$ only with the knowledge extracted from the previous model $\mathscr{M}_{t-1}$ and the

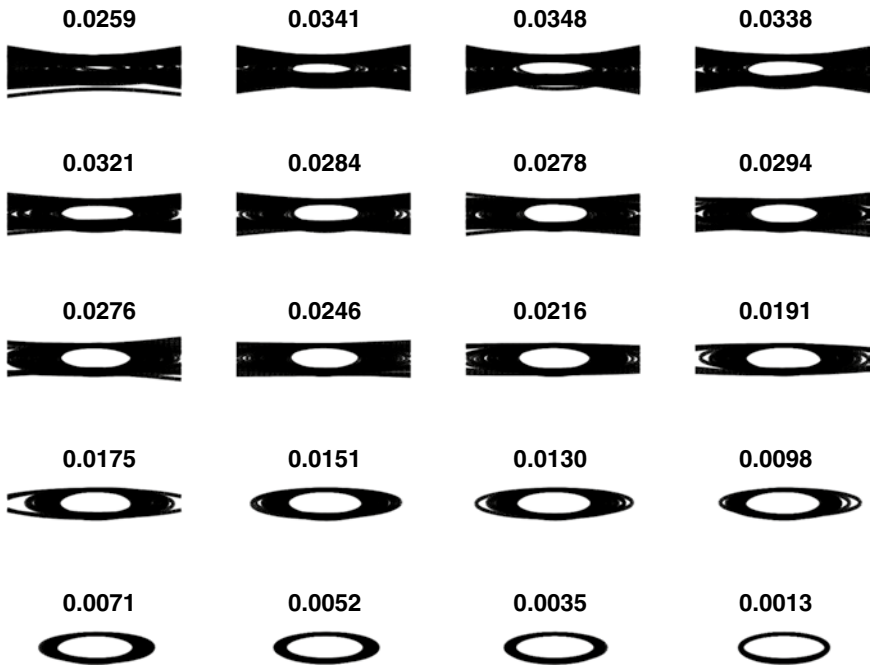

Fig. 5 Convergence of the decision boundaries of each model in 20 iterations for the two-dimensional synthetic database. The variance of the Maximum A Posteriori parameter vector $\mathbf{w}_{\text{MAP}}$ is shown at the *top* of each iteration. The iterations are shown *left-to-right*, *top-to-bottom*. Again, it can be seen that the first models present arbitrary decision boundaries since their parameters are fitted from one single sample. When further samples are used for learning, the decision boundaries and their parameters begin to converge until the final iteration, where the differences are negligible

current data S_t. This is equivalent to having a sliding time window of size 0 [40], which implies that there is no previous data stored to be used for training again new models.

Our solution to the incremental learning problem achieves the following desired properties defined by Street in [36]: it is iterative, processing batches of data at a time instance rather than requiring the whole set at the beginning; the algorithm requires only one pass through each data sample; since the parameters of the model are always the same and the incremental step only changes their values, the model structure requires a constant amount of memory and does not depend on the size of the data; finally, if the evolution of the algorithm is stopped at moment t, the model $\mathscr{M}_t$ provides the best answer at that moment.

Furthermore, the new Bayesian algorithm does not make any assumption over the data. Instead, it assumes that the parameters follow a multivariate Gaussian distribution allowing a good model fitting to the data independently of its distribution. Since the algorithm is based on basis functions $\phi(\mathrm{x})$ it can describe any

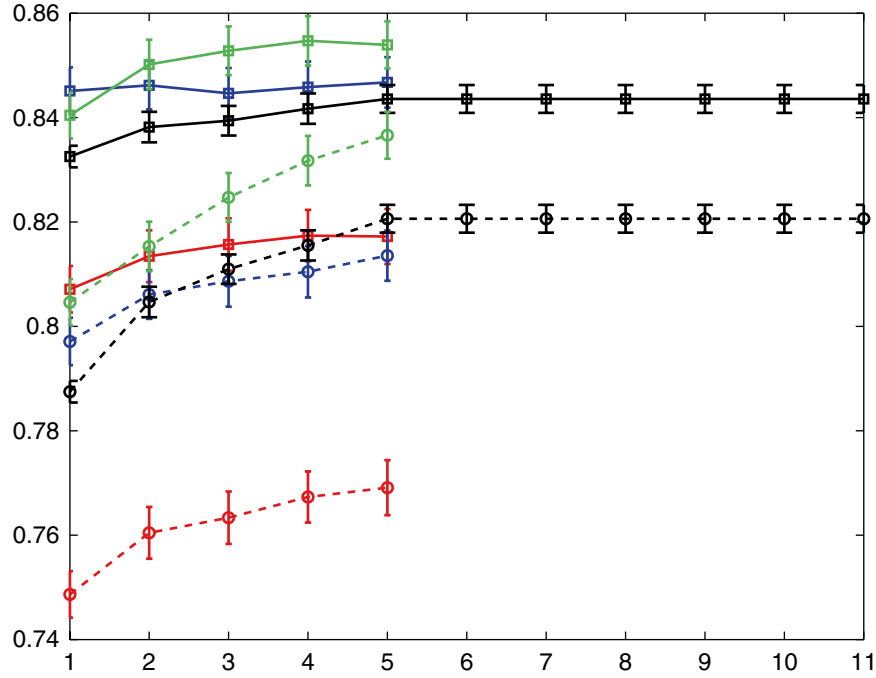

Fig. 6 Comparison of the evolution and improvement of the mean accuracies (*x*-axis) of the iBLR (*solid lines*) and the iGDA (*dashed lines*) and for the new centers: CEN_1 (*blue*), CEN_2 (*red*), CEN_3 (*green*), and the union of the three aforementioned centers (*black*). The confidence intervals ($\alpha = 0.05$) are shown

polynomial decision boundary. One interesting research may be set out to automatically select the degree of the polynomial to best fit the observed data while keeping parsimony.

One obvious limitation of the model is that it is unable to discriminate more than two classes. However, a multinomial logistic regression model can be applied instead to extend its ability to discriminate more classes. This improvement is straightforward if we apply a multinomial distribution to compute the likelihood function and using a multinomial model that gives the probability of a class $y = c$ among a set of possible classes $|\mathscr{C}|$ by using the *softmax* function

$$P(y = c \mid \mathbf{x}) = \frac{\exp(\mathbf{w}_c^T \phi(\mathbf{x}))}{\sum_{k=1}^{|\mathscr{C}|} \exp(\mathbf{w}_k^T \phi(\mathbf{x}))} \tag{16}$$

In addition, the iBLR is only applicable to continuous variables since the Laplace approximation is based on a Gaussian distribution [5]. Another limitation is that such approximation assumes unimodal distributions, capturing only local properties of the true distribution. Hence, if the joint parameter distribution follows a multimodal distribution, then the approximation will be imprecise.

This may be overcome by using a deterministic global approximation, such as the Variational Bayes approach [41, 5] or the Expectation Propagation approach [42], on (9) instead of the Laplace approximation. Finally, the use of the Newton–Raphson method to optimize the MAP value of the distribution may be risky in a situation of extreme shifting distributions, since the old w_{MAP} may be a bad starting point—or a saddle point—for the method which may imply a diverging behavior.

In summary, our contribution proposes an algorithm where the Bayesian approach can be used to develop an incremental learning algorithm for discriminative models by using the knowledge of one model, represented by the posterior probability, as a prior knowledge for the development of a new model given new observations and without making any assumption on the underlying distribution of the data.

The CDSSs that are based on statistical learning techniques have shown promising results for aiding in automatic diagnosis in general, but also for non-invasive brain tumor diagnosis in particular [29, 43]. However, the development of robust classifiers requires acquisition of a large number of cases. Furthermore, in multicenter projects it is usually assumed that the data have similar distributions, however in practice we may expect some differences in data distributions or class assignments. A straightforward application of the incremental method presented here is its ability to customize an already trained classifier to the specific distribution of a particular hospital. In other words, if a hospital has a limited number of samples for a particular class, a classifier trained with data from other hospitals can be used as an initial model and then adapted to the distribution of the patient population. Thus a classifier can be developed that has a customization to the hospital, but without the need for an unachievable acquisition of local data. The development of new models in the course of time as new data is acquired is related to the concepts of temporal and external validation reported by Altman et al. in [7]. Based on the results, our incremental algorithm could enhance the performance of such models when evaluated with subsequent patients coming from new hospitals.

In the framework of a clinical DSS such an algorithm may take advantage of the availability of new information to adapt the knowledge of the current system to the evolution of the data domain and also to extend the lifecycle of the system in a real clinical environment, avoiding possible obsolescence of the models and the system. Assuming that new information is ready for supervised classification at different times, an incremental algorithm can learn from such new data without access to the previously seen data. Since this process may imply also a cost reduction, we may consider that this incremental approach is also in line with some of the main goals of the new trends in medicine such as the P4 medicine.

Acknowledgments

We would like to acknowledge support and fund for this work from the Spanish Ministerio de Ciencia e Innovación through the INNPACTO project 2011 (IPT-2011-1126-900000) and from the European Regional Development Fund (ERDF) (2007–2013).

References

1. Hood L, Heath JR, Phelps ME, Lin B (2004) Systems biology and new technologies enable predictive and preventative medicine. Science 306(5696):640–643
2. Hood L, Friend SH (2011) Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol 8(3):184–187
3. Eddy DM (2005) Evidence-based medicine: a unified approach. Health Aff (Millwood) 24(1):9–17
4. Carney S (2010) Psychiatry: an evidence-based text, chapter introduction to evidence-based medicine. CRC, Boca Raton
5. Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin
6. González-Vélez H, Mier M, Julià-Sapé M, Arvanitis TN, García-Gómez JM, Robles M, Lewis PH, Dasmahapatra S, Dupplaw D, Peet AC, Arús C, Celda B, Van Huffel S, Lluch i Ariet M (2009) Health Agents: Distributed multi-agent brain tumor diagnosis and prognosis. Appl Intell 30(3):191–202
7. Altman DG, Vergouwe Y, Royston P, Moons KG (2009) Prognosis and prognostic research: validating a prognostic model. BMJ 338(b605):1432–1435
8. van Houwelingen HC (2000) Validation, calibration, revision, and combination of prognostic survival models. Stat Med 19(23):3401–3415
9. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD (2004) Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 23(16):2567–2586
10. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y (2008) Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol 61(1):76–86
11. Giraud-Carrier C (2000) A note on the utility of incremental learning. AI Commun 13(4):215–223
12. Grossberg S (1998) Nonlinear neural networks: principles, mechanisms and architectures. Neural Netw 1(1):17–61
13. Polikar R, Udpa L, Udpa SS, Honavar V (2001) Learn++: an incremental learning algorithm for supervised neural networks. IEEE Trans Syst Man Cybern Part C Appl Rev 31(4):497–508
14. Lange S, Zilles S (2003). Formal models of incremental learning and their analysis. Int Joint Conf Neural Netw 4:2691–2696
15. Cornuéjols A (1993) Getting order indepence in incremental learning. In: AAAI Spring symposium on training issues in incremental learning, pp 43–54
16. Langley P (1995) Order effects in incremental learning. In: Reimann P, Spada H (eds) Learning in humans and machines: towards an interdisciplinary learning science. Elsevier, Oxford, pp 1–17
17. Di Mauro N, Esposito F, Ferilli S, Basile TMA (2005) Avoiding order effects in incremental learning. In: Bandini S, Manzoni S (eds) Advances in artificial intelligence (AI*IA05). LNCS. Springer, pp 110–121
18. Bernardo JM, Smith AFM (1994) Bayesian theory. Wiley, Chichester, New Yark
19. Gelman A, Carlin JB, Stern HS, Rubin DB (2003) Bayesian data analysis, 2nd edn. Chapman and Hall/CRC, Boca Raton
20. McCullagh P, Nelder JA (1983) Generalized linear models. Chapman and Hall, London
21. Banerjee A (2007) An analysis of logistic models: exponential family connections and online performance. In SDM
22. MacKay DJC (1992) The evidence framework applied to classification networks. Neural Comput 4(3):448–472
23. Fletcher R (1987) Practical methods of optimization, 2nd edn. Wiley, New York
24. Tortajada S, Fuster-Garcia E, Vicente J, Wesseling P, Howe FA, Julià-Sapé M, Candiota A-P, Monleón D, Moreno-Torres À, Pujol J,

Griffiths JR, Wright A, Peet AC, Martinez-Bisbal MC, Celda B, Arús C, Robles M, García-Gómez JM (2011) Incremental Gaussian Discriminant Analysis based on Graybill and Deal weighted combination of estimators for brain tumour diagnosis. J Biomed Inform 44(4):677–687

25. Asuncion A, Newman DJ (2007) UCI machine learning repository. http://archive.ics.uci.edu/ml/
26. INTERPRET Consortium. INTERPRET. Web site, 1999–2001. IST-1999-10310, EC. http://gabrmn.uab.es/interpret/
27. eTUMOUR Consortium. eTumour: Web accessible MR Decision support system for brain tumour diagnosis and prognosis, incorporating in vivo and ex vivo genomic and metabolomic data. Web site. FP6-2002-LIFESCIHEALTH 503094, VI framework programme, EC, 2009. http://www.etumour.net
28. Julià-Sapé M, Acosta D, Mier M, Arús C, Watson D, (2006) The INTERPRET consortium. A multi-centre, web-accessible and quality control-checked database of in vivo MR spectra of brain tumour patients. MAGMA 19:22–33
29. Tate AR, Underwood J, Acosta DM, Julià-Sapé M, Majós C, Moreno-Torres A, Howe FA, van der Graaf M, Lefournier V, Murphy MM, Loosemore A, Ladroue C, Wesseling P, Luc Bosson J, Cabañas ME, Simonetti AW, Gajewicz W, Calvar J, Capdevila A, Wilkins PR, Bell BA, Rémy C, Heerschap A, Watson D, Griffiths JR, Arús C (2006) Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra. NMR Biomed 19(4):411–434
30. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, Scheithauer BW, Kleihues P (2007) 2007 who classification of tumours of the central nervous system. Acta Neuropathol 114: 97–109
31. van der Graaf M, Julià-Sapé M, Howe FA, Ziegler A, Majós C, Moreno-Torres À, Rijpkema M, Acosta D, Opstad KS, van der Meulen YM, Arús C, Heerschap A (2008) Mrs quality assessment in a multicentre study on mrsbased classification of brain tumours. NMR Biomed 21:148
32. Devos A, Lukas L, Suykens JAK, Vanhamme L, Tate AR, Howe FA, Majós C, Moreno-Torres À, van der Graaf M, Arús C, Van Huffel S (2004) Classification of brain tumours using short echo time 1H MR spectra. J Magn Reson 170(1):164–175
33. García-Gómez JM, Tortajada S, Vidal C, Julià-Sapé M, Luts J, Moreno-Torres A, Van Huffel S, Arús C, Robles M (2008) The effect of combining two echo times in automatic brain tumor classification by MRS. NMR Biomed 21(10):1112–1125
34. García-Gómez JM, Luts J, Julià-Sapé M, Krooshof P, Tortajada S, Robledo JV, Melssen W, Fuster-Garcia E, Olier I, Postma G, Monleón D, Moreno-Torres A, Pujol J, Candiota A-P, Martínez-Bisbal MC, Suykens J, Buydens L, Celda B, Van Huffel S, Arús C, Robles M (2009) Multiproject-multicenter evaluation of automatic brain tumor classification by magnetic resonance spectroscopy. MAGMA 22(1):5–18
35. Muhlbaier M, Topalis A, Polikar R (2009) Learn++.NC: combining ensemble of classifiers combined with dynamically weighted consult-and-vote for efficient incremental learning of new classes. IEEE Trans Neural Netw 20(1):152–168
36. Street NW, Kim Y (2001) A streaming ensemble algorithm (SEA) for large-scale classification. In: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining (KDD '01), pp 377–382. ACM
37. Klinkenberg R (2004) Learning drifting concepts: example selection vs. example weighting. Intell Data Anal 8:281
38. Maloof MA, Michalski RS (2004) Incremental learning with partial instance memory. Artif Intell 154(1 -2):95–126
39. Zico Kolter J, Maloof MA (2007) Dynamic weighted majority: an ensemble method for drifting concepts. J Mach Learn Res 8:2755–2790
40. Scholz M, Klinkenberg R (2007) Boosting classifiers for drifting concepts. Intell Data Anal (IDA) (Special Issue on Knowledge Discovery from Data Streams) 11:3–28
41. Girolami S, Rogers MA (2006) Variational Bayesian multinomial probit regression with gaussian process priors. Neural Comput 18(8):1790–1817
42. Minka TP (2001) propagation for approximate Bayesian inference. In: UAI, pp 362–369
43. Sáez C, García-Gómez JM, Robledo JV, Tortajada S, Fuster-Garcia E, Esparza M, Navarro AT, Robles M (2009) Curiam BT 1.0, decision support system for brain tumour diagnosis. In: ESMRMB 2009: 26th Annual Scientific Meeting, October. Springer

Chapter 5

Using Process Mining for Automatic Support of Clinical Pathways Design

Carlos Fernandez-Llatas, Bernardo Valdivieso, Vicente Traver, and Jose Miguel Benedi

Abstract

The creation of tools supporting the automatization of the standardization and continuous control of healthcare processes can become a significant helping tool for clinical experts and healthcare systems willing to reduce variability in clinical practice. The reduction in the complexity of design and deployment of standard Clinical Pathways can enhance the possibilities for effective usage of computer assisted guidance systems for professionals and assure the quality of the provided care. Several technologies have been used in the past for trying to support these activities but they have not been able to generate the disruptive change required to foster the general adoption of standardization in this domain due to the high volume of work, resources, and knowledge required to adequately create practical protocols that can be used in practice. This chapter proposes the use of the PALIA algorithm, based in Activity-Based process mining techniques, as a new technology to infer the actual processes from the real execution logs to be used in the design and quality control of healthcare processes.

Key words Process mining, Clinical Pathways, Workflows, Process standardization

1 Introduction

The standardization of processes is more and more present in our society. The possibility to have strictly defined protocols allows the creation of highly optimized and traceable processes. This philosophy is quickly outspreading in a great quantity of areas where quality of service is a key factor for the correct functionality of the system.

Healthcare constitutes one of the domains where the importance of process standardization is gradually increasing, realized in the creation of Clinical Pathways [1]. Clinical Pathways (A.K.A Critical Pathways) are care standardized protocols managed by the Medical knowledge where the clinical, support, and management activities are sequentially detailed to allow the coordination of all the stakeholders of the care process: Medical Doctors, Patients, Social Workers, Nurses, etc. To achieve a better accuracy in the

Carlos Fernández-Llatas and Juan Miguel García-Gómez (eds.), *Data Mining in Clinical Medicine*, Methods in Molecular Biology, vol. 1246, DOI 10.1007/978-1-4939-1985-7_5,

process standardization, the processes should be *formally defined and not ambiguous*, to ensure the correct automation of the processes made by humans or computers, enabling both the automatic guidance of the processes and the use of computer assisted techniques in order to evaluate their correctness, accuracy, etc.; *very expressive*, to allow expressing all the different situations that need to be included within a Clinical Pathway; *understandable*, to allow human actors to easily read the process not only to assure their correct and affordable implementation in real life but also to facilitate the design and update of the Clinical Pathway Definition; and *traceable*, in order to help the health professionals to easily identify the specific stage of the process where the patient is.

Traditionally, Clinical Pathways have been described by groups of experts using natural language producing large care manuals available in online digital libraries [2, 3]. Nevertheless, the high ambiguity and complexity of those manuals caused those documents to have less penetration than expected, mainly due to practical deployment difficulties. Other approaches, based on Knowledge Based Systems (KBS) like GLIF [4] or Asbru [5] allow the creation of ontologies and rule-based systems to formally describe the processes. These systems ensure a high expressivity, thanks to their capacity to express a wide set of Clinical Pathways Patterns. Nevertheless, on a normal basis Medical Doctors are not familiarized with these kinds of languages, thus, the creation of easier ways to translate their knowledge into Clinical Pathways becomes a clear necessity for the standardization of healthcare.

Within the process automation paradigm, Workflows are the most used technology in order to enable *non-programming* experts defining processes. A Workflow is defined as *the automation of a business process, in whole or part, during which documents, information, or tasks are passed from one participant to another for action, according to a set of procedural rules* [6]. Workflows are specifically designed to be understandable by experts as a tool to describe processes in a formal way in order to allow said experts to model and automate them. Formal Workflow representation languages are usually associated to specific Workflow Engines. A Workflow Engine is able to automate and trace their execution, thanks to their lack of ambiguity. The main disadvantage of Workflows technology against systems like GLIF, is the expressivity. Usually, KBS are more expressive than Workflows because part of the expressivity of the later is sacrificed in order to achieve a better understandability. Still, there are Workflow representation languages that combine high understandability with high expressivity capabilities like TPA (Timed Parallel Automaton) [7].

In the literature, we can find works that test the use of Workflow technology for defining Clinical Pathways [8, 9, 10]. However, in all approaches, one of the most important barriers when deploying

a Clinical Pathway is the difficulty of their design. For standardizing Clinical Pathways, the consensus of expert groups is needed and achieving this complete consensus requires a large quantity of time. In addition, those Clinical Pathways are not exempt of errors due to the subjectivity of human experts. These problems have caused that the tailoring of protocols to daily practice becomes a very difficult task. To do so in the correct way, there is a need for a high amount of iterative discussions with multiple experts in order to achieve a useful and adequate protocol. Yet it's very common that the protocols do not exactly represent the patient's correct flow due to the high variability of patients and the possible pluripathologies. In those cases, the Clinical Pathway has to be modified only for that concrete patient in order to adequately match her particular needs. Those on execution time modifications are very useful for detecting the usual exceptions that occur in Clinical Pathways. Using that information, the professionals can assess the rationale behind the changes in order to identify new care branches to be added for the next iteration of the Clinical Pathway redesign and increase the reach of the agreed protocol. Nevertheless, that process can be very tedious and takes a long time to achieve good results.

The authors' hypothesis presented in this chapter is that the use of pattern recognition technologies, for automatizing the learning process in order to provide computer assisted tools, can help Clinical Pathways experts in the process of design and deployment of standardized care protocols in real practice. The use of pattern recognition methodology in medicine should be moderated by health experts in order to avoid errors in the application of treatments. Interactive Pattern Recognition technologies have been approached to Clinical pathways [11] in order to provide an iterative methodology based on human experts corrections. However, to allow that it is needed to present to expert, each Pattern Recognition result, in an easy and human understandable way. The use of Process Mining technologies can be a solution to this problem [12]. Process Mining algorithms are created to infer Workflows from real execution samples gathered from workflow execution systems. Applying this paradigm to the Clinical Pathways design problem, it will be possible to infer formally defined Workflows representing the real protocols executed in real environments. These Workflows can be used to help designing Clinical Pathways with more accuracy and with a better alignment with the real execution, enabling more quick and effective design iterations in the creation of clinical pathways.

In this paper, a study of existing process mining algorithms and a methodology for computer assisted iterative design of Clinical Pathways are described, aiming at selecting the best approaches to this research field. A preliminary study of this work was presented on [13].

2 Process Mining

Process Mining technologies use the log of actions performed in previous execution processes in order to infer workflows that explain the whole process in a human understandable way. Traditional Process Mining technologies are based on the use of transactional logs as samples [12, 14, 15]. The information available on those transactional logs is composed by events that represent information about the starting time and, sometimes, their finishing time. This paradigm is called Event-Based Process Mining. Nevertheless, this paradigm does not take into account the information of the result of the actions.

In the case of Clinical Pathways, the result of the performed actions is crucial for understanding the process execution. For instance, let's consider an action called *take temperature* that is able to produce two different results: *Fever* or *Not Fever*. Depending on the result produced by the action, the next action that should be executed is defined according to the Clinical Pathway flow. In our example, if the patient has fever, the next action performed might be *TakePill* or in the case of *Not Fever*, the next action might be *Do Nothing*. In this example, the result of the actions is crucial to understand the cause of the next steps detected on the Clinical Pathway. Traditional Process Mining techniques do not allow such identification because they do not take into account the action result's information. To enable to do so, the authors propose a new Process Mining paradigm that takes into account all this information. This paradigm is called Activity-Based Process Mining.

Activity-Based Process Mining proposes the enrichment of the training corpus by incorporating the results data in the learning corpus and creating algorithms to allow to take advantage of said information and to express it in the resulting Workflow. An example of Activity-Based Process Mining Algorithm is PALIA (Parallel Activity-based Log Inference Algorithm) [13]. PALIA is an algorithm that is able to infer high expressive formal workflow based on TPA [7]. Figure 2 is an example of TPA inferred by authors using PALIA. In this way, authors propose the use of Activity-Based Process Mining Technologies in order to provide a computer assisted methodology for Clinical Pathways design.

3 Activity-Based Process Mining for Iterative Clinical Pathways Design

The application of Activity-Based Process Mining technology to the practical deployment of Clinical Pathways offers multiple advantages and could dramatically empower the adoption and extension of standardization approaches in healthcare. The first reason to sustain this statement is based in the ease in applicability of these tech-

nologies to highly computerized processes. Actually, most healthcare information systems already capture automatically an information log recording all actions, times, and actors that interact in the system, without generating additional consumption of resources to the process or the system. These logs could be easily formatted in order to become an automatically captured and incremental corpus of updated and historical data. The existence of this data will enable that the iterative design process starts with a baseline process that reflects the real process in an understandable way in order that experts can discuss the improvements on the process with the more realistic information possible. Once an improvement is agreed, it could be immediately applied in the practice and a new set of data could be collected in a short period of time to assess the real adoption of said change. In the meantime, the experts could work in producing more improvements of the process.

Furthermore, the passing of time normally causes degeneration in the practice of any standardized process. While some tasks of the process are strongly adopted and become part of the automatic practice of the users, other parts of the process may suffer small changes. These variations can be irrelevant in an isolated way but the accumulation of minor changes usually produces big baseline changes with time if not controlled. The regular application of Activity-Based Process Mining to the execution logs could help detect and correct these deviations when the level of variability increases a pre-set threshold. However, these changes may sometime reflect a change in the reality that needs to be incorporated in the Clinical Pathways, thus this continuous quality control could also help automatically triggering the need for updates or redesigns of the process. By these means, the experts will only be needed when a need for a change or a deterioration of the process is detected, and they will be automatically informed of the part of the process that needs to be revised. This way, the real execution of the process will automatically trigger the need for improvement, increasing the efficiency of the redesign process.

4 Activity-Based Process Mining: Preliminary Experiments

In order to perform a preliminary test to evaluate the accuracy of Activity-Based Process Mining Algorithms, some laboratory experiments have been made. The authors have been involved in several European Funding projects, like Carepaths [8], PIPS [16], and HeartCycle [17] where they have had the opportunity to study the usual characteristics and expressivity of Clinical Pathways. Using that expertise, a formal Clinical Pathways was created using TPA as the representation language. This Clinical Pathway example is show in Fig. 1. This experiment represents a educative simplification of a Clinical Pathway about Heart Failure.

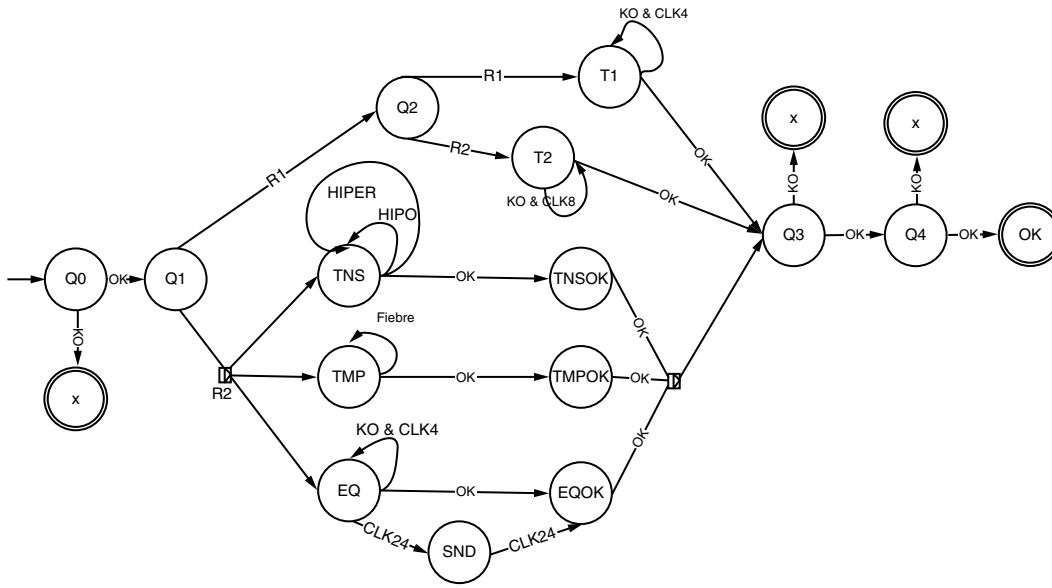

Fig. 1 Clinical Pathways experiment

To test out algorithm, we have selected a set of classic event-based algorithms that are well known in literature. Those Process Mining algorithms are Heuristic Miner (HM) [18], Genetic Process Miner (GPM) [19], α [20], and α_{++} [21].

These Clinical Pathways represent the most usual patterns that are used in normal Clinical Pathways. Using a TPA simulation Engine [22] more than 2,000 executions of those Clinical Pathways have been simulated to create a set of logs for the PALIA Algorithm. An example of execution sample simulated is presented as follows:

```
   12:49:38 => i:0001 Node: Q0 -> StartAction
: Q0.Q0
   12:49:39 => i:0001 EndAction: Q0.Q0 Res: OK
   12:49:39 => i:0001 Node: Q1 -> StartAction
: Q1.Q1
   12:49:40 => i:0001 EndAction: Q1.Q1 Res: R12
   12:49:40 => i:0001 Node: TNS -> StartClock
: CLKTNS_4
   12:49:40 => i:0001 Node: TNS -> StartAction
: TNS.TNS
   12:49:40 => i:0001 Node: TMP -> StartClock
: CLKTMP_8
   12:49:40 => i:0001 Node: TMP -> StartAction
: TMP.TMP
   12:49:40 => i:0001 Node: EQ -> StartClock:
CLKEQ_12
   12:49:40 => i:0001 Node: EQ -> StartClock:
CLKEQ_24
```

```
   12:49:40 => i:0001 Node: EQ -> StartAction
: EQ.EQ
   12:49:41 => i:0001 EndAction: TNS.TNS Res: OK
   12:49:41 => i:0001 Node: TNSOK -> StartActi
on: TNSOK
   12:49:41 => i:0001 EndAction: EQ.EQ Res: OK
   12:49:41 => i:0001 Node: EQOK -> StartActio
n: EQOK
   12:49:42 => i:0001 EndAction: TMP.TMP Res: OK
   12:49:42 => i:0001 Node: TMPOK -> StartActi
on: TMPOK
   12:49:42 => i:0001 EndAction: EQOK Res: OK
   12:49:43 => i:0001 EndAction: TNSOK Res: OK
   12:49:44 => i:0001 -> EndClock: CLKTNS_4
   12:49:44 => i:0001 -> EndClock: CLKEQ_12
   12:49:44 => i:0001 EndAction: TMPOK Res: OK
   12:49:44 => i:0001 Node: Q3 -> StartAction
: Q3.Q3
   12:49:45 => i:0001 EndAction: Q3.Q3 Res: OK
   12:49:45 => i:0001 Node: Q4 -> StartAction
: Q4.Q4
   12:49:46 => i:0001 EndAction: Q4.Q4 Res: OK
   12:49:46 => i:0001 Instance Terminated
```

Using those samples, the PALIA Algorithm has been used to infer the TPA that represents the execution of those samples in order to compare the result of the algorithm with the original designed process. An example of the TPAs inferred by PALIA is shown in Fig. 2.

The results obtained with these experiments have been evaluated via two measurements; Efficacy (Eff) and Relative Understandability coefficient (Cf_{RU}).

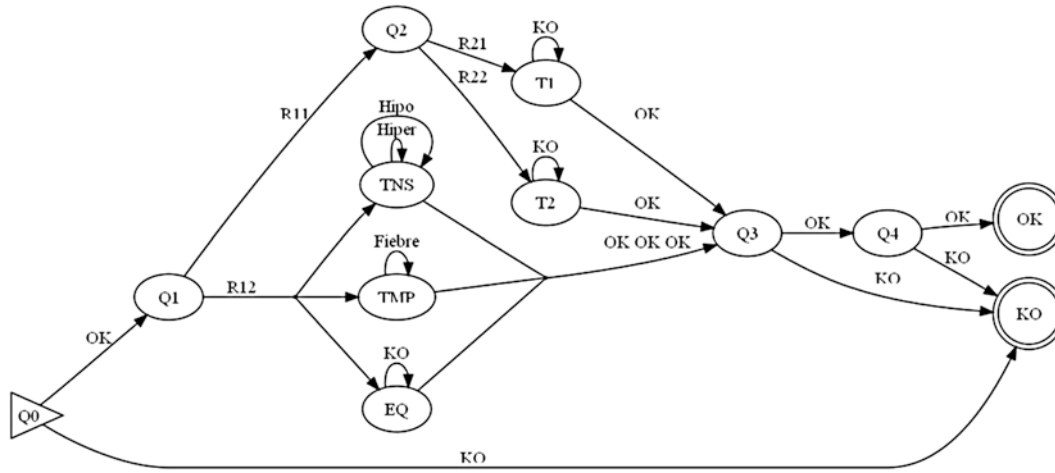

Fig. 2 Example of TPA inferred with PALIA

The Efficacy shows the capability of the algorithm to identify specific Clinical Pathways situations or patterns. For the experiment, human experts have divided the Clinical Pathway in patterns that represent specific situations that can occur in the care protocol. This coefficient is the ratio between the number of patterns correctly identified by the TPA inferred by PALIA and the total of patterns existing in the original Workflow:

$$\text{Eff} = \frac{\text{Patterns}_{\text{Identified}}}{\text{Patterns}_{\text{Total}}}$$

The Relative Understandability coefficient (Cf_{RU}) shows the overhead of structures used to identify the patterns. The lower is the value of the coefficient, the less structures are used to represent each identified pattern. This offers a measure of the easiness of visual understanding of the Workflow. This coefficient is the ratio between the number of Nodes by Arc of inferred Workflow needed divided by the number of identified patterns and the number of Nodes By Arc of the original Workflow divided by the number of original patterns:

$$Cf_{RU} = \frac{\frac{N_{\text{Inferred}} XA_{\text{Inferred}}}{\text{Patterns}_{\text{Identified}}}}{\frac{N_{\text{Original}} XA_{\text{Original}}}{\text{Patterns}_{\text{Original}}}}$$

The Cf_{RU} of the original Workflow is 1. The more closer to 1 this value is, the more efficient the algorithm is. If the value is lower than 1, it's normally due to the fact that the algorithm has sacrificed some patterns to better define others.

The experiments have shown encouraging results. PALIA works better than other algorithms in this experiment. All patterns have been identified by PALIA algorithm and the Cf_{RU} has remained close to the original legibility. Table 1 shows a resume of the experimental results. In this table, the number of nodes and arcs inferred by PALIA is displayed, together with the Efficacy (Eff) and the Relative Understandability coefficient(Cf_{RU}).

5 Discussion and Conclusions

The use of Process Mining methodologies seems be a solution to approach Interactive Pattern Recognition to the optimization of clinical Pathways. The results presented in this chapter show how it is possible present in a graphical way the real execution of process deployed. This information will allow health professionals to understand how the processes works, providing them a view of how the care processes executed are deployed in

Table 1
Experiment results

		Nodes	Arcs	Cf_{NxA}	Patterns	Cf_{Rm}	Efficacy	Cf_{LR}
Real		16	24	384	14	1.75	–	–
PALIA		16	33	544	14	2.44	1	1.38
Heuristic Miner [18]	I/F	29	55	1,479	6	1.9	0.43	9.69
	I	15	29	435	7	1.93	0.50	2.27
	F	14	22	308	7	1.57	0.50	1.60
Genetic Process Mining [19]	I/F	29	61	2.1	5	319	0.36	12.90
	I	15	29	510	8	1.93	0.57	1.98
	F	14	27	406	9	1.93	0.64	1.53
α [20]	I/F	29	67	1,769	10	4.76	0.71	7.08
	I	15	15	240	2	1.2	0.14	4.10
	F	14	11	154	2	0.79	0.14	2.81
α ++ [21]	I/F	–	–	–	–	–	–	–
	I	15	23	345	4	5.47	0.29	3.14
	F	14	19	266	7	4.36	0.50	1.46

reality. Clinicians can detect problems in the deployments of care processes supporting them in the continuous optimization of clinical pathways according to the innovative Interactive Pattern Recognition techniques.

The main limitation of this work is the lack of real experiments. Previous preliminary result shows how classic algorithms using event-based process mining approach applied to clinical pathways generate very big workflows that are difficult to follow (that is commonly called *spaguetti effect*). However, in our experiments we have compared our algorithm with classic algorithms the have drastically decreased the number of elements in the workflow (Cf_{RU}) keeping a very high efficacy in the discovery of workflow patterns of the original flow.

However, it is needed to improve the algorithms to reduce at the maximum this problem as well as provide tools to highlight specific situations in the process executions [23]. This will support experts to achieve a real knowledge about the deployment of their processes.

References

1. Every NR, Hochman J, Becker R, Kopecky S, Cannon CP (2000) Critical pathways. A review. Circulation 101:461–465
2. The Cochrane Collaboration (2010) COCHRANE Library. http://www.cochrane.org/index.htm
3. PubMed Library (2010) National Library of Medicine and The National Institutes of Health PubMed Library. http://www.pubmed.gov
4. Peleg M, Boxwala AA, Bernstam E, Tu SW, Greenes RA, Shortliffe EH (2001) Sharable representation of clinical guidelines in GLIF: relationship to the arden syntax. J Biomed Inform 34(3):170–181
5. Shahar Y, Miksch S, Johnson P (1998) The asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines. Artif Intell Med 14(1–2):29–51
6. WfMC (1999) Workflow management coalition terminology glossary. WFMC-TC-1011, Document Status Issue 3.0
7. Fernandez-Llatas C, Pileggi SF, Traver V, Benedi JM (2011) Timed parallel automaton: a mathematical tool for defining highly expressive formal workflows. In: Fifth Asia modelling symposium (AMS), 2011 IEEE computer society, pp 56–61
8. Naranjo JC, Fernandez-Llatas C, Pomes S, Valdivieso B (2006) Care-paths: searching the way to implement pathways. Comput Cardiol 33:285–288
9. Sedlmayr M, Rose T, Röhrig R, Meister M (2006) A workflow approach towards GLIF execution. In: Proceedings of the European conference on artificial intelligence (ECAI), Riva del Garda
10. Fox J, Black E, Chronakis I, Dunlop R, Patkar V, South M, Thomson R (2008) From guidelines to careflows: modelling and supporting complex clinical processes. Stud Health Technol Inform 139:44–62. PMID: 18806320
11. Fernandez-Llatas C, Meneu T, Traver V, Benedi J-M (2013) Applying evidence-based medicine in telehealth: an interactive pattern recognition approximation. Int J Environ Res Public Health 10(11):5671–5682
12. van der Aalst WMP, van Dongen BF, Herbst J, Maruster L, Schimm G, Weijters AJMM (2003) Workflow mining: a survey of issues and approaches. Data Knowl Eng 47:237–267
13. Fernandez-Llatas C, Meneu T, Benedi JM, Traver V (2010) Activity-based process mining for clinical pathways computer aided design. In: 32th annual international conference of the IEEE engineering in medicine and biology society, pp 6178–6181. PMID: 21097153
14. Cook J, Du Z (2005) Discovery thread interactions in a concurrent system. J Syst Softw 7:285–297
15. van der Aalst WMP (2011) Process mining: discovery, conformance and enhancement of business processes. Springer, Berlin [u.a.]
16. VI Framework Program I S T Project 507019 (2008) PIPS Project. Personalised Information Platform for life and health Services
17. Heart Cycle Consortium (2008) VII Framework Program IST Project 216695: compliance and effectiveness in HF and CHD closed-loop management 2008–2011
18. Weijters AJMM, Ribeiro JTS (2011) Flexible heuristics miner (FHM). In: 2011 IEEE symposium on computational intelligence and data mining (CIDM), pp 310–317
19. de Medeiros AKA, Weijters AJMM, van der Aalst WMP (2007) Genetic process mining: an experimental evaluation. Data Min Knowl Discov 14(2):245–304
20. van der Aalst W, Weijters A, Maruster L (2004) Workflow mining: discovering process models from event logs. IEEE Trans Knowl Data Eng 16:1128–1142
21. Alves de Medeiros AK, Dongen BF, van der Aalst WMP, Weijters AJMM (2004) Process mining extending the alpha algorithm to mine short loops. Technical report, WP113 Beta Paper Series Eindhoven University of Technology
22. Fernandez-Llatas C, Sanchez C, Traver V, Benedi JM (2008) TPAEngine: un motor de workflows basado en TPAs. In: Ciencia y Tecnologia en la Frontera. ISSN:1665-9775
23. Fernandez-Llatas C, Benedi J-M, Garcia-Gomez JM, Traver V (2013) Process mining for individualized behavior modeling using wireless tracking in nursing homes. Sensors 13(11):15434–15451

Chapter 6

Analyzing Complex Patients' Temporal Histories: New Frontiers in Temporal Data Mining

Lucia Sacchi, Arianna Dagliati, and Riccardo Bellazzi

Abstract

In recent years, data coming from hospital information systems (HIS) and local healthcare organizations have started to be intensively used for research purposes. This rising amount of available data allows reconstructing the compete histories of the patients, which have a strong temporal component. This chapter introduces the major challenges faced by temporal data mining researchers in an era when huge quantities of complex clinical temporal data are becoming available. The analysis is focused on the peculiar features of this kind of data and describes the methodological and technological aspects that allow managing such complex framework. The chapter shows how heterogeneous data can be processed to derive a homogeneous representation. Starting from this representation, it illustrates different techniques for jointly analyze such kind of data. Finally, the technological strategies that allow creating a common data warehouse to gather data coming from different sources and with different formats are presented.

Key words Temporal data, Data mining, Hospital information systems, Heterogeneity

1 Introduction

In recent years, data coming from hospital information systems (HIS) and local healthcare organizations have started to be intensively used for research purposes. This rising amount of available data has opened new challenges to the scientific community, and new tools for the representation, storage, and analysis of this data have started to be developed.

Data collected in clinical settings are intended to follow a patient during his hospital stay or his follow-up, which usually takes place outside the hospital. As the patient is monitored during a period of time, the data that are collected are most often temporal in their nature. Moreover, the potential ability to trace all the events that take place during a patient's follow-up enables to record his temporal history, which is usually made up of several types of medical events. Besides temporal data, there is of course also non temporal

Carlos Fernández-Llatas and Juan Miguel García-Gómez (eds.), *Data Mining in Clinical Medicine*, Methods in Molecular Biology, vol. 1246, DOI 10.1007/978-1-4939-1985-7_6,

information (e.g., demographics, environmental data, familiar history) that can be fruitfully exploited for analysis purposes.

The possibility of representing this variety of data in a structured and homogeneous way and of applying appropriate analysis methodologies to it widens the perspectives for knowledge discovery. Understanding complex multivariate clinical histories has many advantages, both at the medical and at the organizational level [1–3]. It can offer new insights into how clinical processes are carried out in a structure, on how patients are commonly treated and on which are the most frequent associations between a patient's clinical status and the care delivery process. The application of Data Mining techniques to this kind of data has been one of the most fruitful approaches to extract new and useful clinical knowledge from it [4].

In this chapter we introduce the new challenges offered to the data mining community by the increasing availability of this kind of data, especially focusing on complex temporal clinical histories. In particular, in Subheading 2 we introduce the peculiarities that characterize complex multivariate temporal histories and in Subheading 3 we propose a framework for a uniform data representation. Subheading 4 is devoted to the introduction of some methodologies for efficient mining of complex temporal data. Finally, in Subheading 5, we describe how heterogeneous temporal data coming from diverse sources can be integrated into a single technological framework.

2 Clinical Time Series: Peculiarities and Heterogeneity

Clinical time series data have some peculiar features that distinguish them from more traditional information recorded in time [5]. First of all, excluding regularly monitored physiological signals (e.g., ECG, respiratory rate), clinical data are often measured without a specific sampling scheme. For example, laboratory tests are executed when needed, and temperature and blood pressure can be taken at different times during the day (as they are manually collected by an healthcare operator). This results in the collection of time series that are most often characterized by an uneven sampling grid. In addition, the number of recorded samples is usually low. For these reasons, traditional signal processing techniques are often not applicable to medical time series data. To cope with these peculiarities, techniques coming from a field of research known as temporal data mining (TDM) [6, 7] have started to be extensively used for analyzing clinical time series. TDM methodologies and tools extend and complement traditional data mining techniques to explicitly take into account the temporal aspects.

In the complex scenario of a patient who is followed inside and outside the hospital, not only strictly clinical information is

recorded. Both hospitals' electronic health records (EHR) and local healthcare services collect some administrative information, which is primarily used for billing purposes. Recently though, such information has started to be also made available for research. If joined to clinical data, administrative data represent an added value to the process of knowledge discovery, as they contribute to build up a complete reconstruction of patients' histories.

Administrative data contain the collection of all the accesses a patient performs to the national healthcare system: hospital admissions, drug purchases, outpatient visits, etc. Given the purposes they are originally collected for, these data do not contain medical information. For example, an administrative record may report that a patient has been admitted to the hospital with a specific diagnosis and that some procedures and lab tests have been performed, but the specific results of such tests are not reported.

The integration of administrative and clinical data is crucial to get the best knowledge out of both. This process in not straightforward though, since the structure of the raw data is naturally different between the two. While clinical time series are stored as time-stamped data associated to a quantitative measurement (temperature, blood pressure, creatinine value, etc.), administrative data are stored as temporal sequences of events. In temporal data mining, an *event* is in general defined as a temporal variable that is associated to a time stamp or interval of occurrence. A *sequence of events* is defined as a list of events, where each event is associated to the same transaction or individual [8, 9]. When dealing with clinical data, the individual is usually the patient.

Differently from time series that contain only time-stamped data (i.e., a measurement is collected at a specific time point), temporal sequences of events can contain both time-stamped events and events with a duration. The duration of an event can be defined on the basis of the temporal *granularity* a specific set of data is collected with. The granularity represents the maximum temporal resolution used to represent an event [10, 11]. According to this definition we can thus state that, on the basis of a specific granularity, sequences of events can contain both zero-length events and events with non-zero duration. In addition to the described differences in the purpose underlying the collection of clinical and administrative data, there are thus also some differences in the way they are represented.

Figure 1 shows a simple patient's history, made up by merging a set of clinical data to a set of administrative information related to him. Data are collected with a granularity of 1 day. In the interval [t1, t2], the considered patient is undergoing a therapy with ACE inhibitors. At time t3 some clinical data are collected for the patient. In particular, his systolic blood pressure (SBP), triglycerides and blood glucose are measured. The patient is hospitalized in the interval [t4, t5]. The hospitalization is recorded into the

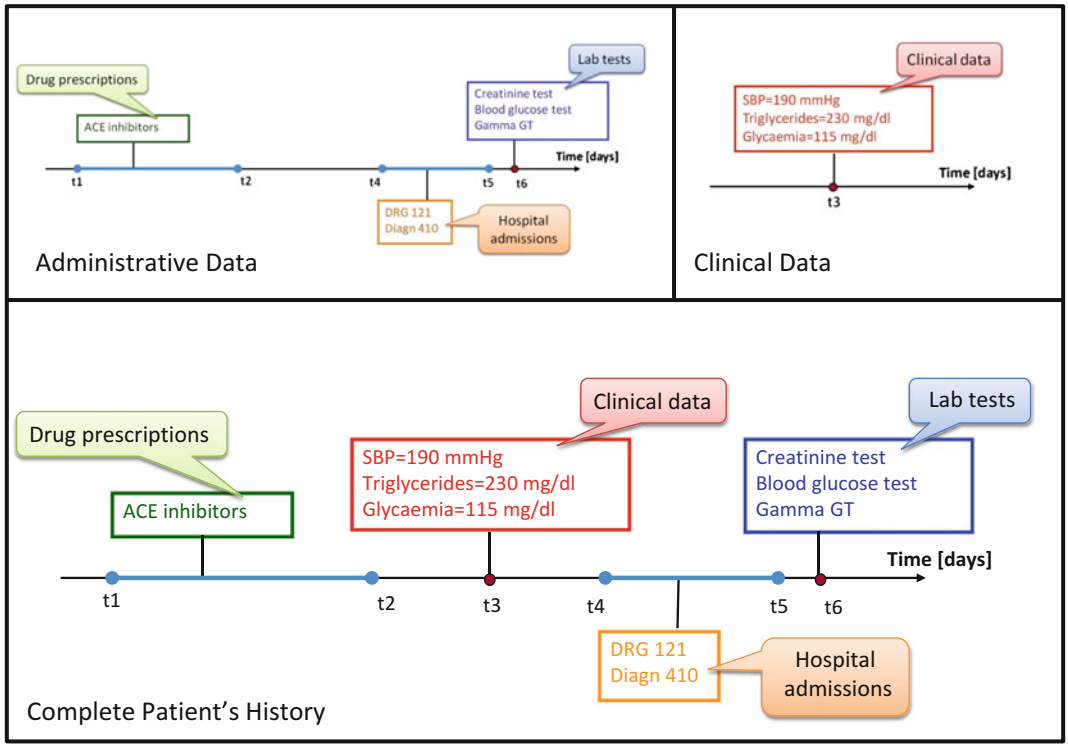

Fig. 1 Example of a patient's clinical history made up by jointly considering clinical and administrative data

administrative repository through the DRG and the main diagnosis associated to the stay. After being discharged, the patient performs some control blood tests at t6. As shown in Fig. 1, both the therapy administration and the hospitalization are events with duration, as they last more than 1 day. On the other hand, events like taking a measurement or performing a blood test are typically lasting less than 1 day and are thus represented as time points (duration = 0) in the data.

As already underlined, the joint analysis of clinical and administrative data might provide a deeper insight into the temporal mechanisms underlying patients' histories. Given the very nature of the different types of data though, it is first necessary to integrate them accordingly to reach a common representation format enabling researchers to treat them in a uniform way. In the next section we illustrate some methodologies for performing this step.

3 Sharing a Homogeneous Temporal Representation: Knowledge-Based Temporal Abstractions

One of the first steps to jointly analyze clinical and administrative data is to be able to share a common representation for them. As raw administrative data contain time points and time intervals associated

to clinical events (Fig. 1), the idea is to be able to transform raw clinical time series data into a similar representation.

In the literature, the framework that allows performing this step is known as *knowledge-based temporal abstraction* (KBTA or TA). Initially theorized by Shahar in the late 1990s [12], TAs have become more and more popular in the clinical data mining community [13–18], especially in recent times when the need for data integration has become a primary issue in the research. In general, temporal abstractions are a way to perform a shift from a quantitative time-stamped representation of raw data to a qualitative interval-based description of time series, with the main goal of abstracting higher level concepts from time-stamped data.

Shahar set up the theoretical framework by introducing the basic time primitives: *time stamps*. Time stamps are associated to specific time units, which are referred to as *granularity units*. *Time intervals* are defined as ordered pairs of time stamps, which represent the interval's end points. In Fig. 1, for example, the pair [t1, t2] represents a time interval. *Time points* are introduced as time intervals of length zero according to the granularity unit data are represented with.

When applied in practice, TAs often rely on variants of the framework originally presented by Shahar. Post et al. introduce a TA ontology containing raw data and abstractions definitions [19]. In this ontology, data represented by time stamps (such as time series) are called *observations*, while data that can be either associated to a time point or a time interval (such as administrative data) are called *events*. The ontology proposed in [19] is set in a framework devoted to heterogeneous data integration, representation and retrieval. For this reason, it also includes data that are not temporal. Such data can be very useful for the analysis, as they might allow stratifying the patients and understanding behaviors related to specific groups of individuals (e.g. related to age, living area or educational level).

Another TA model, totally oriented to the representation of time series data, is the one proposed in Ref. [18]. Similarly to Shahar's model, this model is based on the definition of time points and intervals. According to the models specified in Refs. [18, 19], it is in general possible to identify two types of abstraction tasks, which lead to the definition of two types of TAs: a basic (or low level) and a complex abstraction task.

Basic TAs take as input the raw time series of time-stamped clinical data and output a series of intervals where specific behaviors hold. Basic TAs identify simple behaviors, such as states and trends, in the data. *State TAs* extract the intervals in which a variable is within a predefined set of interesting values. State abstractions correspond to expressions like "*high* blood glucose values between September 22nd and September 24th 2013" or "*normal* blood pressure in the last week." *Trend* TAs represent increase, decrease, and stationary patterns in a numerical time series.

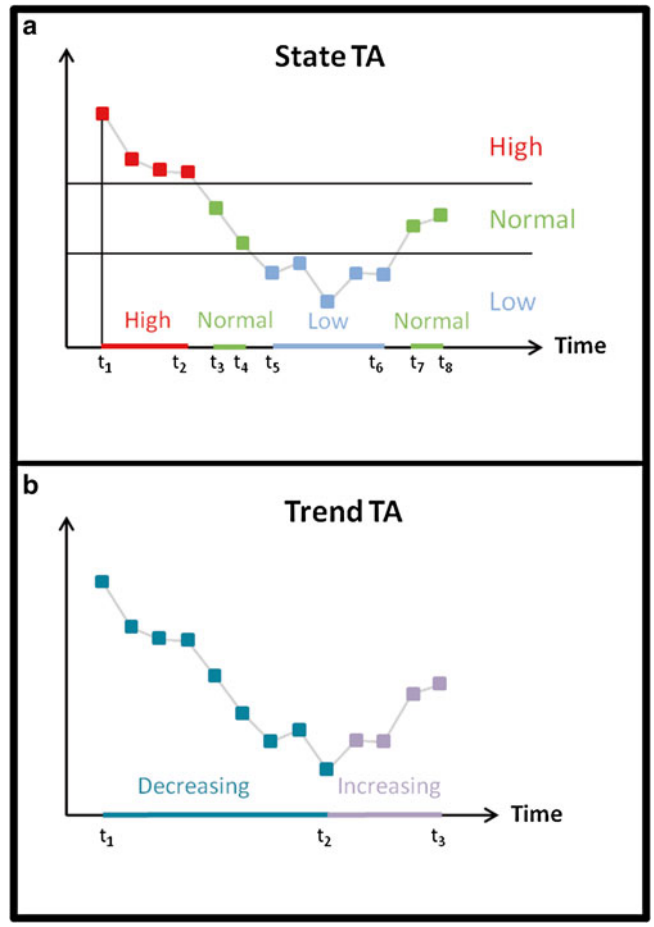

Fig. 2 (**a**) A raw time series represented through basic state temporal abstractions. (**b**) A raw time series represented through basic trend temporal abstractions

Figure 2 represents an example of a time series transformed using basic State and Trend TAs. Figure 2a shows how the labels *Low*, *Normal* and *High* have been associated to the corresponding intervals on a raw time series. Both the levels of validity of the three labels and the final interval-based state TA representation are shown. Figure 2b shows a raw time series processed to extract the two trend abstractions *Increasing* and *Decreasing* (with the corresponding time intervals).

As shown in Fig. 2, Basic TAs offer a way to represent quantitative variables through qualitative labels holding on time intervals. Time series processed using basic TAs are thus already represented by sequences of (interval-based) events, exactly as happens with administrative data. For example, the time series in Fig. 2a can be represented by the following sequence: *High* $[t_1, t_2]$ *Normal* $[t_3, t_4]$ *Low* $[t_5, t_6]$ *Normal* $[t_7, t_8]$.

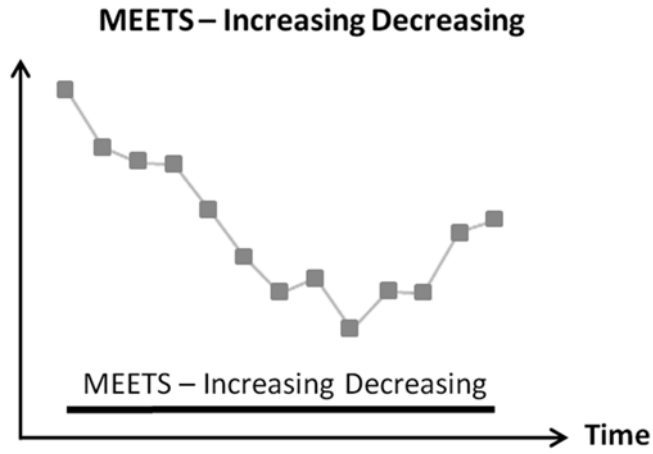

Fig. 3 Complex temporal abstraction

Complex TAs, also referred to as temporal patterns [14, 20], are used to extract those complex behaviors that can't be represented by basic TAs. Complex TA applications work on sets of intervals rather than on raw time series. They require two intervals sets as inputs, associated to two TAs, and provide an interval set as output. Such interval set is associated to a new TA, which is evaluated according to a pattern specified between the intervals of the two composing sets. The definition of a complex TA pattern is based on temporal relationships defined among the TAs on which it is built up. The temporal relationships that are usually exploited to detect temporal patterns are the temporal operators defined in Allen's algebra [21]. Allen has defined 13 operators able to exhaustively describe the basic relations that can occur between a pair of intervals. These include *BEFORE*, *FINISHES*, *OVERLAPS*, *MEETS*, *STARTS*, *DURING*, their corresponding inverse relations, and the *EQUALS* operator. Figure 3 shows how a complex TA can be derived using the basic trend abstractions shown in Fig. 2b and applying the *MEETS* temporal operator. Given two time intervals $I_1 = [i_1.\text{start}, i_1.\text{end}]$ and $I_2 = [i_2.\text{start}, i_2.\text{end}]$, we say that I_1 *MEETS* I_2 if $i_1.\text{end} = i_2.\text{start}$.

To focus only on specifically interesting scenarios, some authors have restricted the set of operators they consider to build up temporal patterns. For example, Refs. [22, 23] introduce the temporal operator *PRECEDES* to represent the notion of precedence in the data. PRECEDES synthesizes 6 out of the 13 Allen's operators. Batal and co-authors [24] use the *BEFORE* and the *CO-OCCURS* operators to identify the mutual position of the two intervals involved in a temporal pattern.

The increasing popularity TA has recently gained in clinical data analysis is due to the several advantages that this representation

offers. First of all, in clinical practice, medical experts often have in mind a qualitative idea of the abstract patterns they would be interested to detect in the data. TAs are able to automatically translate such pattern into a formal representation and to search for it in the data. This is particularly useful when large sets of multivariate time series are collected and the manual inspection of the medical expert is no longer sufficient to explore the data set. The second advantage of using this framework is that it is particularly well suited for dealing with short and unevenly spaced time series. Representing data at the abstract level allows overcoming such limitations, which would instead heavily impact on a data analysis performed on the raw quantitative time series.

Methods that have recently been proposed in the literature on the application of the TA framework to clinical data devote particular attention to the visualization and query of temporal patterns. A recent work [25] proposed a framework for visually defining and querying clinical temporal abstraction, with particular attention to the representation of different temporal granularities in the data. In this work, the definition of the TAs is based on a technique, the paint strip metaphor, which had already been shown as one of the methods preferred by users to specify TAs [26]. The methodology has been evaluated by involving clinical experts in the study and by asking them to use the system to represent some complex scenarios expressed only in natural language and to consequently query the database.

Shahar's group presented several tools to explore and visualize time-oriented data. KNAVE-II allows the use of a set of computational mechanisms to derive and visualize temporal patterns in a real-time fashion [27]. While KNAVE—II is mainly oriented to a single-patient perspective, the VISualizatIon of Time-Oriented RecordS (VISITORS) system is focused on patients' populations [28]. This tool provides aggregate views of time-oriented data and abstractions of groups of patient records. It allows exploring the data and getting a first insight into associations among the temporal patterns that characterize the selected group of patients.

Once time series data are abstracted using TAs, their representation becomes uniform to the one characterizing data represented as sequences of events. Once data are represented using a common strategy it is thus possible to start extracting information from it [29]. In the next section, we describe some of the methodologies that allow mining useful information from this kind of representation.

4 Mining Complex Temporal Histories

Among the variety of TDM techniques proposed in the literature, those that resulted more suitable to analyze complex patients' temporal histories are the ones related to the efficient mining of

frequent temporal patterns from data. Such patterns may identify the most common sequences of events in the data (e.g., sequences of hospitalizations, drug prescriptions).

The very first introduction of algorithms for mining sequential patterns was naturally developed as an extension of traditional association rule (AR) mining techniques [30]. From then on, many efforts have been made to develop more efficient search strategies to maximize computational performance. Moreover, attention has been devoted to developing methodologies to mine frequent temporal patterns either on time point or on time intervals sequences [31–35]. Within these techniques, an interesting group of papers has been published on how to derive temporal association rules (TARs) from sequential data [36, 37].

Temporal Association Rules are association rules of the kind A→C, where the antecedent (A) is related to the consequent (C) by some kind of temporal operator. As also mentioned when introducing complex TAs, the most commonly used operators are the ones derived from Allen's algebra. As in traditional ARs mining, TARs mining algorithms are aimed at extracting *frequent* associations, where frequency is evaluated on the basis of suitable indicators, the most used being support and confidence. The *support* gives an indication of the proportion of cases verifying a specific rule in the population (P(A,C)). *Confidence* instead represents the probability that a subject verifies the rule given that he verifies its antecedent (P(C|A)). In the case of TARs such indicators need to be properly extended to take into account the temporal nature of the data [22]. Recently, techniques for temporal patterns and TARs mining have been applied also to clinical data [38].

In Refs. [22, 23, 39], the authors present a framework for mining TARs based on TAs. These papers show how different types of data can be analyzed to derive interesting information. Initially [23], the authors present a method to derive temporal association rules using basic TAs in the antecedent and in the consequent and apply this method for quality assessment in hemodialysis services. Basic TAs are though not enough to represent the complex qualitative behaviors that clinical experts want to retrieve when reasoning about temporal data representation. In Ref. [22], the methodology has been extended to take into account complex patterns both in the antecedent and in the consequent of the rule. Patterns of interest can be specified on the basis of domain knowledge into a set called *Abstractions of Interest* (AoI), and rules containing such pattern in the antecedent and in the consequent are extracted. The definition of the set AoI is very useful, since it allows taking into account the clinical knowledge on the specific problem, and limiting the set of TARs extracted by the mining procedure. The method has been validated on the extraction of temporal relationships between heart rate and blood pressure variables recorded during hemodialysis sessions.

After the development of a TARs mining framework mainly oriented to the analysis of clinical data, the authors extended their perspectives to incorporate also administrative healthcare information into the data set. This required the methodological effort to be focused on two aspects: on the one hand, also events with a zero-length duration had to be taken into account in the mining process [39], and on the other, clinical and administrative information had to be integrated [40]. This second task was easily performed using the TA framework set up in Ref. [22]. The first task was instead performed by extending Allen's operators and the definitions of confidence and support, in order to take into account events with zero length (time points). The joint use of clinical and administrative information allowed carrying out interesting analyses, such as exploring the changes in patients' health status that lead to specific drug prescriptions or the associations between patients' clinical conditions and accesses to the national healthcare service [41]. Interestingly, the methodologies introduced so far can be also applied to evaluate the economic impact of particular treatments, as it has been shown in Ref. [42]. Focusing on TARs showing a drug prescription in their consequent, a way to associate a specific estimated cost to each derived TAR is presented. Rules related to a specific treatment can then be clustered together, so that the overall cost of that therapy can be computed. This procedure is possible thanks to the cost-related information contained in administrative records.

As the application of pattern mining methods to EHR data started spreading, particular attention has been devoted to reduce the number of output patterns, which often make the results interpretation process hard to perform and to be interpreted by a clinician. In Ref. [39] a methodology to mine disease-specific TARs is presented. Rules are mined in a population of diabetic patients and in a group of controls, and only significant disease-specific patterns are extracted. In addition, the mining process includes additional filtering strategies that are coupled to the more traditional cutoffs based on confidence and support. Ref. [43] presents recent temporal pattern mining. This framework is based on the assumption that the patterns that are closer in time to an event of interest have the highest predictive power with respect to that event. Having lower support, less recent patterns are discarded and do not take part in the mining phase. The authors propose also further strategies to efficiently reduce the search space, based on the definition of incoherent patterns (i.e., those patterns that never show an occurrence in the data) and improvements in the candidate generation phase.

As mentioned, methods for extracting TARs from clinical histories are able to detect behaviors of the type A→C. Rules of this type can contain arbitrarily complex patterns in the antecedent and in the consequent, but are usually able to capture histories with limited length. As a matter of fact, frequent clinical histories

made up of longer chains of events are not straightforward to be mined through traditional TARs extraction techniques. To overcome this limitation, the reconstruction of the so-called *clinical pathways* is becoming nowadays one of the most challenging fields in data mining in health care. The opportunity of developing these novel algorithms is primarily offered by the availability of HISs, which allow collecting large amounts of data related to complete clinical histories.

Since their first introduction, approaches for clinical pathways mining have been often borrowed from business process analysis. In analogy to this discipline, in clinical pathways mining the sequences of events occurring to each patient during his clinical history are commonly referred to as *event logs*.

Differently from clinical process modeling where workflows are manually constructed on the basis on some kind of medical evidence (e.g., clinical guidelines), clinical pathways mining has the advantage of exploiting the huge amount of data collected through the hospitals information systems to reconstruct the most frequent histories that took place in a particular medical center. This gives the possibility of detecting anomalous pathways or site-specific behaviors that can be operated in a hospital for specific reasons possibly not stated in the current clinical guidelines. On the other end though, given the high heterogeneity and variability of the processes of care, the interpretation of the results is not straightforward.

The techniques that have been most often exploited to analyze clinical histories [44–47] are the ones coming from process mining (PM), a general method used in business process analysis [48–50]. To cope with the variability of healthcare processes, techniques to help interpreting and synthesizing PM results have been developed. Ref. [51] proposes a methodology to summarize process mining results to facilitate interpretation by the clinical experts. The method has been evaluated by a set of clinical experts and hospital managers.

Clinical pathways mining usually works on event logs made up of data coming from administrative data streams. Interestingly, only a few works in the literature deal with the exploitation of clinical data into the mining process. One example of these methods is given by Ref. [52], where the authors present a workflow mining algorithm able to work on event logs enhanced with clinical data represented as sequences of events (e.g., TAs).

5 Collecting Data in a Common Framework

As stated in the previous paragraphs, one of the main challenges in collecting a data set that aims at representing the whole medical history of a chronic patient is related to data sources' diversity in terms of data structures and acquisition purposes. The last impor-

tant effort we illustrate in this chapter is the one related to the integration of the data in a common repository.

Complex multivariate temporal data sets have been defined as data sets where data instances are traces of complex behaviors characterized by multiple time series [20]. Ideally, for each patient there are at least two complex time series, one linked to his clinical history, and another collecting the succession of his contacts with the National Healthcare Service. As a matter of fact, healthcare organizations have become aware that a new level of data aggregation and reporting is needed to fulfill existing and future requirements. The most common solution that has been proposed to this kind of problem is to build a data warehouse where it is possible to integrate, visualize and query data in an informative way, considering both their temporal nature and complexity.

The state-of-the-art open-source tool available to collect multidimensional data possibly coming from different sources, and aggregate them in a format suitable for temporal analysis is the Informatics for Integrating Biology and the Bedside (i2b2) Data Warehouse [53]. I2b2 is one of the seven centers funded by the US National Institute of Health (NIH) Roadmap for Biomedical Computing [54]. The mission of i2b2 is to provide clinical investigators with a software infrastructure able to integrate clinical records and research data [55].

Being an open source tool, i2b2 gives the possibility to be highly customized for specific medical applications and to develop ad hoc plug-ins for data extraction and visualization. In addition, i2b2 is characterized for being a tool developed specifically to deal with clinical data management and translational research applications. For these reasons it has become widely popular among healthcare organizations and it is usually preferred to other commercial tools, such as Business Objects (SAP) [56] or Cognos (IBM) [57].

One of i2b2 key features, which makes it highly suitable to handle complex multivariate temporal data, is that it interlocks medical record data and clinical data at a person level so that diseases, medical events, and outcomes can be related to each other.

Within the i2b2 model, data are stored in a *star* relational database. The star architecture is based on a central table where each row represents a single fact. Since in i2b2 a fact is an observation about a patient, this table is referred to as the "observation fact table". The fact table contains all the quantitative or factual data coming from observations about each visit (or contact with the NHS in the case of administrative data) related to each patient, and it is the table where all the values of each observation are stored. Each row of the fact table identifies one observation about a patient (described in the Patient Dimension table) made during a visit (stored in the Visit Dimension table). All the observations derived from different events about a patient are recorded in a specific time

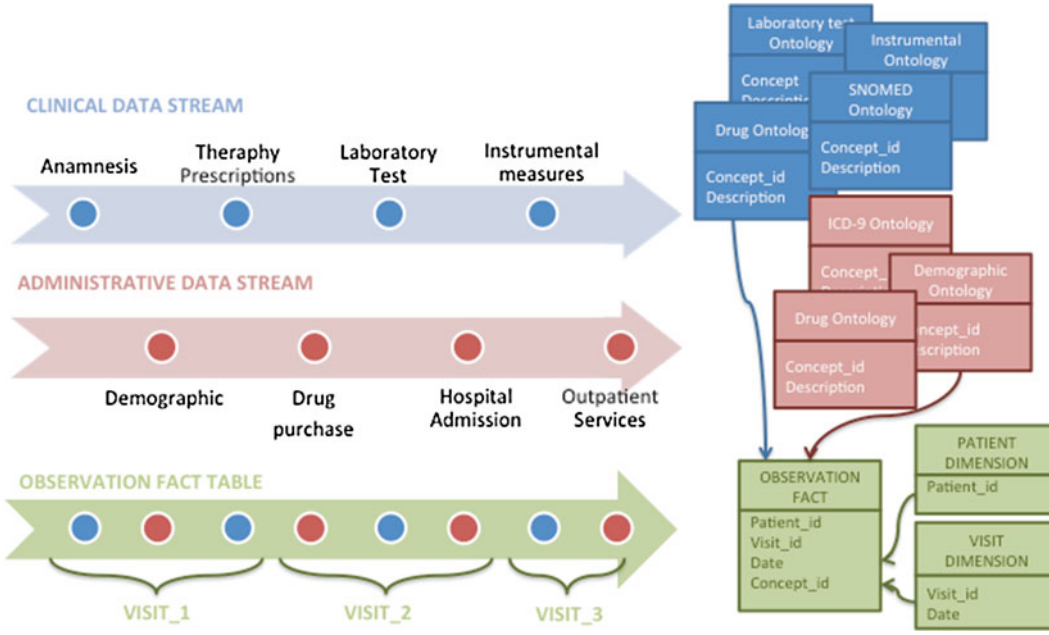

Fig. 4 The i2b2 data warehouse to integrate clinical and administrative data streams

range, defined by start and end dates, and are related to a specific concept. The concept can be any coded attribute, such as an ICD9 code for a certain disease or a medication or a specific test result. Figure 4 shows how the "Observation Fact Table" collects for each patient categorized data on a single timeline, divided into consecutive visits.

To facilitate the query process for the user, data are mapped to concepts organized in an ontology-like structure. I2b2 ontologies aim at organizing concepts related to each data stream in a hierarchical structure. This solution allows separately managing different data sets and informatively formalizing their content, since each ontology contains the necessary fields to associate each patient's observation with a specific concept. Figure 4 shows the clinical and administrative data streams for an example patient, how they are merged together in the fact table and how they can be described through a set of ontology tables. In this example, drug prescriptions are represented through their ATC drug codes in the Drug Ontology and the subset of Laboratory test from Anatomy Pathology are linked to the SNOMED (Systematized Nomenclature of Medicine) Ontology.

Thanks to this approach, data entered in the fact table can be saved with a shared structure that allows storing complex observations with different granularity and extract them in a common format in which each event (identified as an observation on the patient) is associated with specific visits, prescriptions, or hospital services identified by a precise time frame and a coded description.

An interesting example on how the i2b2 system has been used to integrate administrative and clinical data in a research framework is presented in Ref. [19]. In this paper the authors propose The Analytic Information Warehouse (AIW), a platform for specifying and detecting phenotypes of patients' characteristics to support healthcare data analysis and predictive modeling. Interestingly, the considered patients' features that make up a phenotype can be of different nature: administrative codes, numerical test results and temporal patterns. This is possible thanks to the fact AIW includes a configurable system for transforming data from its source to a common data model. Moreover, the TA extraction system embedded into AIW allows specifying the phenotypes of interest in the TA ontology. AIW is a remarkable example of how the architectures and the methodologies presented in this chapter can be joined into a common framework offering potentially very innovative data mining opportunities.

Another interesting application of the i2b2 system is offered by Ref. [58], where the authors describe the implementation of an i2b2 framework to support clinical research in oncology, called ONCO-i2b2. In this work, the main efforts were related to an integration process dedicated to retrieving and merging data from a biobank management software and from the HIS of the Fondazione Salvatore Maugeri in Pavia, Italy. The authors illustrate the steps required to integrate data from heterogeneous sources. For example, they introduce the design of a Natural Language Processing software module to extract information from unstructured text documents relevant to the clinical characterization of cancer patients, and the SNOMED and ICD9-CM ontologies integration on the basis of the NCBO BioPortal web services [59]. Even if one of the primary goals of the implementation was to allow researchers to exploit i2b2 query capabilities relying on a user-friendly web interface, the availability of such a robust system, which integrates clinical, administrative and research data, allows supporting the application of advanced temporal data mining techniques. Thanks to the possibility to query a system where administrative data had been coded and related to clinical procedures, it is possible to identify patients with an history of malignant breast cancer and to retrieve from the fact table structured data logs suitable to perform temporal and process mining analyses to highlight meaningful clinical careflow patterns.

6 Conclusions

This chapter has tackled the major challenges faced by TDM researchers in an era when huge quantities of complex clinical temporal data are becoming available. We have focused on the peculiar features of this kind of data to describe the methodological and

technological aspects that allow managing such kind of complex framework. In particular, we have explained how heterogeneous data can be processed to derive a homogeneous representation. Starting from this representation, we have described different techniques for jointly analyze such kind of data. Finally, we have described the technological strategies that allow creating a common data warehouse to gather data coming from different sources and with different formats.

References

1. Parsons A, McCullough C, Wang J, Shih S (2012) Validity of electronic health record-derived quality measurement for performance monitoring. J Am Med Inform Assoc 19(4): 604–609
2. Mouttham A, Peyton L, Kuziemsky C (2011) Leveraging performance analytics to improve integration of care. Proceedings of the 3rd workshop on software engineering in health care (SEHC '11). pp 56–62. ACM New York, NY, USA, 2011
3. Kahn MG, Ranade D (2010) The impact of electronic medical records data sources on an adverse drug event quality measure. J Am Med Inform Assoc 17(2):185–191
4. Brown DE (2008) Introduction to data mining for medical informatics. Clin Lab Med 28(1):9–35
5. Benin AL, Fenick A, Herrin J, Vitkauskas G, Chen J, Brandt C (2011) How good are the data? Feasible approach to validation of metrics of quality derived from an outpatient electronic health record. Am J Med Qual 26(6):441–451
6. Mitsa T (2010) Temporal data mining. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. ISBN:1420089765 9781420089769
7. Post AR, Harrison JH Jr (2008) Temporal data mining. Clin Lab Med 28(1):83–100
8. Mannila H, Toivonen H, Verkamo AI (1997) Discovery of frequent episodes in event sequences. Data Min Knowl Discov 1:259–289
9. Kam PS, Fu AWC (2000) Discovering temporal patterns for interval-based events. In: Kambayashi Y, Mohania M, Tjoa AM (eds) 2nd International conference on data warehousing and knowledge discovery. Springer, London, UK, pp 317–326
10. Combi C, Franceschet M, Peron A (2004) Representing and reasoning about temporal granularities. J Log Comput 14(1):51–77
11. Bettini C, Wang XS, Jajodia S (1998) A general framework for time granularity and its application to temporal reasoning. Ann Math Artif Intell 22(1–2):29–58
12. Shahar Y (1997) A framework for knowledge-based temporal abstraction. Artif Intell 90: 79–133
13. Stacey M, McGregor C (2007) Temporal abstraction in intelligent clinical data analysis: a survey. Artif Intell Med 39:1–24
14. Post AR, Harrison JH Jr (2007) PROTEMPA: a method for specifying and identifying temporal sequences in retrospective data for patient selection. J Am Med Inform Assoc 14(5): 674–683
15. Verduijn M, Sacchi L, Peek N, Bellazzi R, de Jonge E, de Mol BA (2007) Temporal abstraction for feature extraction: a comparative case study in prediction from intensive care monitoring data. Artif Intell Med 41:1–12
16. Combi C, Chittaro L (1999) Abstraction on clinical data sequences: an object-oriented data model and a query language based on the event calculus. Artif Intell Med 17(3):271–301
17. Bellazzi R, Larizza C, Riva A (1998) Temporal abstractions for interpreting diabetic patients monitoring data. Intell Data Anal 2(1–4): 97–122
18. Shahar Y, Musen MA (1996) Knowledge-based temporal abstraction in clinical domains. Artif Intell Med 8:267–298
19. Post AR, Kurc T, Cholleti S, Gao J, Lin X, Bornstein W, Cantrell D, Levine D, Hohmann S, Saltz JH (2013) The Analytic Information Warehouse (AIW): a platform for analytics using electronic health record data. J Biomed Inform 46(3):410–424
20. Batal I, Sacchi L, Bellazzi R, Hauskrecht M (2009) Multivariate time series classification with temporal abstractions. Int J Artif Intell Tools 22:344–349
21. Allen JF (1984) Towards a general theory of action and time. Artif Intell 23:123–154
22. Sacchi L, Larizza C, Combi C, Bellazzi R (2007) Data mining with Temporal

Abstractions: learning rules from time series. Data Min Knowl Disc 15(2):217–247

23. Bellazzi R, Larizza C, Magni P, Bellazzi R (2005) Temporal data mining for the quality assessment of hemodialysis services. Artif Intell Med 34(1):25–39
24. Batal I, Valizadegan H, Cooper GF, Hauskrecht M (2011) A pattern mining approach for classifying multivariate temporal data. IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp 358–365
25. Combi C, Oliboni B (2012) Visually defining and querying consistent multi-granular clinical temporal abstractions. Artif Intell Med 54(2):75–101
26. Chittaro L, Combi C (2003) Visualizing queries on databases of temporal histories: new metaphors and their evaluation. Data Knowl Eng 44(2):239–264
27. Shahar Y, Goren-Bar D, Boaz D, Tahan G (2006) Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions. Artif Intell Med 38(2):115–135
28. Klimov D, Shahar Y, Taieb-Maimon M (2010) Intelligent visualization and exploration of time-oriented data of multiple patients. Artif Intell Med 49(1):11–31
29. Bellazzi R, Sacchi L, Concaro S (2009) Methods and tools for mining multivariate temporal data in clinical and biomedical applications. Conf Proc IEEE Eng Med Biol Soc, pp 5629–5632
30. Agrawal R, Srikant R (1995) Mining sequential patterns. In: Yu PS, Chen ALP (eds) Proceedings of the 11th international conference on data engineering. IEEE Comput Soc, pp 3–14
31. Zaki MJ (2001) SPADE: an efficient algorithm for mining frequent sequences. Mach Learn 42(1–2):31–60
32. Ayres J, Flannick J, Gehrke J, Yiu T (2002) Sequential PAttern mining using a bitmap representation. In: Hand D, Keim D, Ng R (eds) Proceedings of the 8th ACM SIGKDD International conference on knowledge discovery and data mining. ACM, Edmonton, pp 429–435
33. Mörchen F, Ultsch A (2007) Efficient mining of understandable patterns from multivariate interval time series. Data Min Knowl Disc 15(2):181–215
34. Patel D, Hsu W, Lee ML (2008) Mining relationships among interval-based events for classification. In: Lakshmanan L, Ng R, Shasha D (eds) Proceedings of the 2008 ACM SIGMOD International conference on management of data. ACM, New York, NY, pp 393–404
35. Zhang L, Chen G, Brijs T, Zhang X (2008) Discovering during-temporal patterns (DTPs) in large temporal databases. Expert Syst Appl 34(2):1178–1189
36. Höppner F, Klawonn F (2002) Finding informative rules in interval sequences. Intell Data Anal 3(6):237–256
37. Winarko E, Roddick JF (2007) ARMADA—an algorithm for discovering richer relative temporal association rules from interval-based data. Data Knowl Eng 63(1):76–90
38. Bellazzi R, Ferrazzi F, Sacchi L (2011) Predictive data mining in clinical medicine: a focus on selected methods and applications. Wiley Interdiscip Rev Data Min Knowl Discov 1(5):416–430
39. Concaro S, Sacchi L, Cerra C, Fratino P, Bellazzi R (2011) Mining health care administrative data with temporal association rules on hybrid events. Methods Inf Med 50(2): 166–179
40. Concaro S, Sacchi L, Cerra C, Fratino P, Bellazzi R (2009) Mining healthcare data with temporal association rules: improvements and assessment for a practical use. In: Combi C, Shahar Y, Abu-Hanna A (eds) Proceedings of the 12th conference on artificial intelligence in medicine, AIME 2009. Springer, Verona, Italy, pp 16–25
41. Concaro S, Sacchi L, Cerra C, Bellazzi R (2009) Mining administrative and clinical diabetes data with temporal association rules. In: Studies in health technology and informatics. 150:574–8
42. Concaro S, Sacchi L, Cerra C, Stefanelli M, Fratino P, Bellazzi R (2009) Temporal data mining for the assessment of the costs related to diabetes mellitus pharmacological treatment. AMIA Annu Symp Proc 2009:119–123
43. Batal I, Fradkin D, Harrison J, Moerchen F, Hauskrecht M (2012) Mining recent temporal patterns for event detection in multivariate time series data. Proceedings of the international conference on knowledge discovery and data mining (SIGKDD). pp 280–288
44. Rebuge A, Ferreira DR (2012) Business process analysis in healthcare environments: a methodology based on process mining. Inf Syst 37(2):99–116
45. Huang Z, Lu X, Duan H (2012) On mining clinical pathway patterns from medical behaviors. Artif Intell Med 56(1):35–50
46. Yang W, Hwang S (2006) A process-mining framework for the detection of healthcare fraud and abuse. Expert Syst Appl 31(1):56–68
47. Lin F, Chen S, Pan S, Chen Y (2001) Mining time dependency patterns in clinical pathways. Int J Med Inform 62(1):11–25

48. van der Aalst WMP, Weijters AJMM, Maruster L (2004) Workflow mining: discovering process models from event logs. IEEE Trans Knowl Data Eng 16(9):1128–1142
49. Agrawal R, Gunopulos D, Leymann F (1998) Mining process models from workflow logs. In: Schek HJ, Saltor F, Ramos I, Alonso G (eds) Sixth international conference on extending database technology. Springer, London, Uk, pp 469–483
50. Cook JE, Wolf AL (1998) Discovering models of software processes from event-based data. ACM Trans Softw Eng Methodol 7(3): 215–249
51. Huang Z, Lu X, Duan H, Fan W (2013) Summarizing clinical pathways from event logs. J Biomed Inform 46(1):111–127
52. Fernandez-Llatas C, Meneu T, Benedi JM, Traver V (2010) Activity-based process mining for clinical pathways computer aided design. Conf Proc IEEE Eng Med Biol Soc 2010:6178–6181
53. i2b2: Informatics for integrating biology & the bedside. https://www.i2b2.org/. Accessed 10 Oct, 2013
54. National Institute of Health Roadmap for Biomedical Computing. http://www.ncbcs.org. Accessed 10 Oct, 2013
55. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I (2010) Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 17(2): 124–130
56. Business Intelligence Software. http://www54.sap.com/pc/analytics/business-intelligence.html. Accessed 10 Oct, 2013
57. IBM Cognos Software. http://www-01.ibm.com/software/analytics/cognos/. Accessed 10 Oct, 2013
58. Segagni D, Tibollo V, Dagliati A, Perinati L, Zambelli A, Priori S, Bellazzi R (2011) The ONCO-I2b2 project: integrating biobank information and clinical data to support translational research in oncology. Stud Health Technol Inform 169:87–91
59. BioPortal web services. http://bioportal.bioontology.org. Accessed 10 Oct, 2013

Part II

Mining Medical Data Over Internet

Chapter 7

The Snow System: A Decentralized Medical Data Processing System

Johan Gustav Bellika, Torje Starbo Henriksen, and Kassaye Yitbarek Yigzaw

Abstract

Systems for large-scale reuse of electronic health record data is claimed to have the potential to transform the current health care delivery system. In principle three alternative solutions for reuse exist: centralized, data warehouse, and decentralized solutions. This chapter focuses on the decentralized system alternative. Decentralized systems may be categorized into approaches that move data to enable computations or move computations to the where data is located to enable computations. We describe a system that moves computations to where the data is located. Only this kind of decentralized solution has the capabilities to become ideal systems for reuse as the decentralized alternative enables computation and reuse of electronic health record data without moving or exposing the information to outsiders. This chapter describes the Snow system, which is a decentralized medical data processing system, its components and how it has been used. It also describes the requirements this kind of systems need to support to become sustainable and successful in recruiting voluntary participation from health institutions.

Key words Distributed health data network, Medical informatics, Methods, Organization and administration

1 Introduction

Multipurpose reuse of electronic health record (EHR) data has been a vision for a long time. Such reuse could provide knowledge about the outcome of current medical practice and be a great source for generating new knowledge. Some even claim such reuse has the potential to "transform the current health care delivery system" [1, 2]. Also, many medical conditions have no guidelines that could help clinicians make a decision about treatment. Large scale reuse of EHR data could provide guidance in form of statistics based on matching cases of health condition, treatment and outcomes [3]. The General Practice Research Database (GPRD) in the UK [4], established as far back as in 1987, is an excellent example of the value such a resource can create. However, in many

Carlos Fernández-Llatas and Juan Miguel García-Gómez (eds.), *Data Mining in Clinical Medicine*, Methods in Molecular Biology, vol. 1246, DOI 10.1007/978-1-4939-1985-7_7,

countries, large-scale reuse of EHR data from many institutions on a large scale has not reached common use. The reasons for not achieving reuse are many and complex, including legal obstacles, privacy issues, technical security constraint, data quality issues, and organizational access, to name some.

In principle three possible data storage solutions exist to enable reuse of data. (1) Establish a data warehouse based on extractions from distributed clinical systems, (2) use centralized storage, and (3) reuse data where they are stored, in decentralized systems. Of these types the most common ones are data warehouses that are based on extractions from many EHR systems, systems similar in architecture to GPRD. As many clinical systems were established before the use of networks became commonplace, few examples of centralized system exists. Some examples of such systems exist in Scotland, Sweden, Denmark, and in the UK with varying experiences and degree of success. Also, as administrative boundaries always exist at some level, spanning from departments in health institutions, to regions and countries, the majority of clinical system must be regarded as distributed, decentralized and heterogeneous. In the latest years examples of systems and approaches for the decentralized architectures have appeared [1]. In these systems and approaches some of the obstacles to reuse are solved by aggregating health data [2]. As exchange of aggregated (anonymous) data is less problematic, compared to de-identified and identified data, many of the obstacles to reuse of EHR data are solved using this approach. De-identified data about one specific person (potentially identifying attributes are removed) always have the risk of re-identification, while aggregated data, summarized over a population, is less likely to identify patients. However, even aggregated data, that typically include a lot of variables, could become an identifier, destroying the anonymity of the data. Also, only having access to aggregated data does not enable establishing casual relationships or performing record linkage between data elements about the same person across institutional boundaries. The ideal system would therefore enable reuse of EHR data without introducing the fear of violating the privacy of the patients, as the professional secrecy is a fundamental requirement that all health personnel are bound by.

In the following section we look at requirements for the decentralized approaches, category (3) above, as these approaches has the potential to become ideal systems for reuse of EHR data. A centralized EHR system that covers several countries is very hard or impossible to build; even a national one is difficult. Research or reuse based on EHR data from several countries on the other hand is very easy to imagine and also necessary to generate knowledge about both rare and common diseases. Research on how to build systems that enables large-scale reuse of EHR data is therefore important.

2 Requirements

As access to information about health is strictly regulated, any system for reuse of health data must satisfy legal and regulatory requirements. In Norway the regulations for such systems are very strict, limiting the possible solutions. However, the solutions that satisfy these strict regulations, potentially also become solutions to most countries legal and regulatory requirements. Of these, the most limiting one is the requirement that all network connections must be initiated from the secure zone, that is, where the patient data is stored. This requirement blocks most solutions that allow for use or reuse of clinical data that crosses administrative boundaries. This requirement limits the solutions to message-based asynchronous solutions, using e-mail like mechanisms like POP/SMTP and the XMPP protocols for transport of messages.

Other requirements, regulated through the Norwegian research act [5], are that research on patient identified data must be based on patient consent. Also, even knowing that a patient have a record in a health institution, is regarded as sensitive data. Also, an ethical committee must approve all medical research using patient identified data. Fortunately, the same act open up for research where data is "anonymous on the researchers hand." If the data that the researcher get access to can be kept anonymous, like the output of statistical computations, no consent from the patients is possible or needed. This aspect creates the potential for ideal systems for reuse of EHR data.

Coverage is maybe the most important aspect of a system for reuse of EHR data. 100 % coverage will make it possible to provide data from the whole population, leading to more reliable results. For rare diseases it may also be necessary to reuse EHR data from several countries. Also, reuse of EHR data for disease surveillance purposes for instance is of course dependent of such coverage. Also, having access to data from the whole population reduces the problems related to avoiding identification of specific patients, as the number of cases will be larger. In health systems with both private and public health services, participation would need to be voluntary. Brown et al.'s study of data holders identified the following list of requirements for voluntary participation [6]: (1) *complete control of, access to, and uses of, their data*, (2) *strong security and privacy features,* (3) *limited impact on internal systems*, (4) *minimal data transfer*, (5) *auditable processes*, (6) *standardization of administrative and regulatory agreements*, (7) *transparent governance*, and (8) *ease of participation and use*. A system for large-scale reuse of EHR data would also need to be as *automatic and independent* of manual work or manual configuration of the system as possible, seen from a system administration point of view. If such a system would require a lot of local system administration,

then it will be viewed as a hassle that the owners of the systems and the data would try to avoid. *Easy and hassle-free installation, software update, system monitoring, and administration* are therefore requirements that support both the coverage and sustainability of such a system. With regard to use of such systems, some people argue that institutions may want to *manually validate the computations requests* on their local data [7]. For example: "Local policies will determine whether the query is automatically executed or manually reviewed for approval. Query results can be automatically encrypted and returned to the central website, or they can be queued for manual data holder approval before being returned."

A consequence of the connection establishment requirement mentioned above is that a system for reuse necessarily becomes *asynchronous*. A system that enables local computations of data (because the data cannot be moved) may need to run for a long period of time, necessarily needs to be asynchronous and independent of the requester. A decentralized system for reuse of EHR data must also support *two-way communication*, as it would be the only way to support local selection of data for processing and aggregation of the computation results. This aspect is in contrast to data warehouse approaches where one way transport of data may be sufficient.

3 Methods for Computations on Decentralized Data

Fundamentally there are two ways to compute decentralized data. The first is to move the data to a repository where the computations can be done, and the second is to move the computations to where the data is stored. For the first case many solutions for performing computations exists. Data warehouse, grid computing, cloud computing, to name some, are examples of solutions aimed at exploiting the potential of data that can be moved. If the decentralized data is sensitive, too large, or need to be protected for other reasons, few alternatives exist for exploiting the data for secondary use. One approach, invented in the early 1990s [8], is the mobile agent approach. The basic idea is to enable computations by moving the program to the data, instead of moving the data. This approach can perform data aggregation or computations locally where the data is stored to avoid privacy, legislative and storage related problems. Later on attempts to perform the computations on the encrypted data, also known as homomorphic encryption [9, 10], have appeared. Currently, one of the most promising approaches is exploiting the techniques within secure multiparty computations. These methods have matured the last decade and practical applications of these approaches are starting to appear, *see* ref. [11] for a detailed overview.

The Snow system utilizes the mobile agent approach to enable computations on sensitive decentralized data. This chapter describes how sensitive decentralized data is made available for computations by the Snow system. Our current effort is focused in the direction of exploiting distributed secure multiparty computations to enhance our ability to exploit the potential in sensitive decentralized data.

4 History of the Snow System

The Snow project [12, 13], initiated more than a decade ago, has been working in the direction of providing a decentralized and distributed alternative for medical data processing. The ideas and concepts used in the system dates back to the virtual secretary project [14, 15] initiated in the years 1994–1997 at the Department of Computer Science at University of Tromsø. The decentralized and distributed approach was selected as this approach could avoid many of the obstacles involved in reuse of EHR data. The Snow project has continuously been specifying, implementing and deploying the Snow system and services generated by the approach since 2003. In 2010 the first Snow server were put into production at the Department of Microbiology at the University Hospital of North Norway, providing access to aggregated microbiology data for Troms and Finnmark county in Northern Norway. Based on the Snow platform we have implemented a number of end user services that take advantage of the Snow systems ability to make medical data available for processing. If deployed to a larger number and types of institutions, a lot of new services can be generated. This chapter briefly presents the system and its practical implementations.

5 The Snow System

The Snow system is built on the assumption that the institutions participating in a peer-to-peer system own their data and the computing resources. As the philosophy behind the system is a peer-to-peer network, each server in the system has the same capabilities. To satisfy the requirements outlined above we have extended the open source XMPP server OpenFire [16–18], with a number of plug-ins and components. End user clients, agents, internal and external components are implemented as XMPP clients, enabling communication between the entities using the XMPP protocol. Each user, agent, internal and external components has an address, a jabber ID (JID), for sending and receiving XMPP messages.

The first main responsibility of a Snow server is to protect the institutions data and computing resources. The second responsibility is to make the data available for processing and knowledge

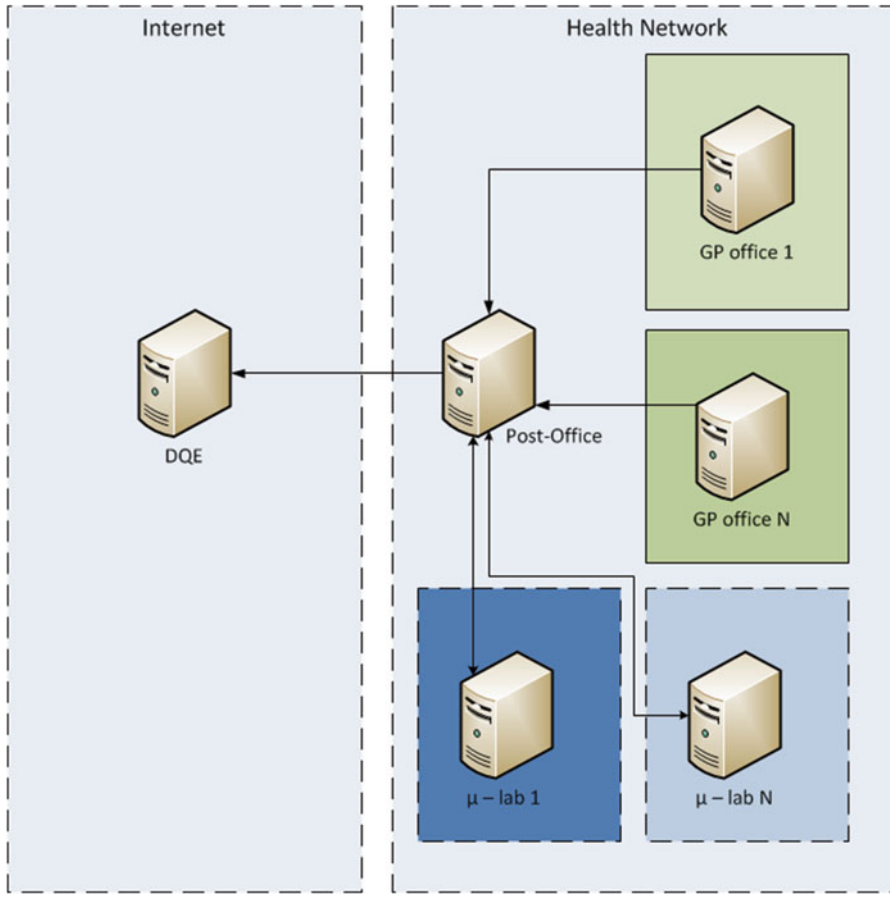

Fig. 1 High-level architecture for the Snow system

generation. Enabling coordinated computations on potentially all servers in the system ensures this second responsibility. To achieve this we organize sets of agents (computing processes) into "missions." A mission specifies what kind of computation (mission type) that should be performed, a list of targets (Snow servers) where the agents should run and the recipient of the computation result. A mission is specified by creating a "mission specification."

As the missions normally are performed periodically, we use a "Mission Scheduler" (MS) on a Snow server dedicated to coordination, the "Post office" (or PO for short) in Fig. 1, to maintain initiation of periodic missions. The computation times are specified using a Cron-like specification, as seen in Fig. 2.

When the time for performing a mission arrives, the Mission Scheduler sends the Mission specification to the Mission Controller (MC). The Mission Controller can receive mission specifications from the Mission Scheduler, clients of the Snow system and agents. Requests for creation of remote missions are sent to the Mission Controller using an XMPP message. The mission controller negotiates

Mission configuration

Select	Mission id	Description	Max duration (minutes)	Application type	Cron expression
	1	19-epidemio-all	60	main-agent:epidemio	0 0 3 ? * *
	2	19-epidemio-respiratory	60	main-agent:epidemio	0 4 3 ? * *
	3	19-epidemio-respiratory-influensa-A	60	main-agent:epidemio	0 2 3 ? * *
	4	19-epidemio-respiratory-influensa-B	60	main-agent:epidemio	0 6 3 ? * *
	5	19-epidemio-respiratory-RS-virus	60	main-agent:epidemio	0 8 3 ? * *
	6	19-epidemio-respiratory-Forkjolelsesvirus	60	main-agent:epidemio	0 10 3 ? * *
	7	19-epidemio-respiratory-Atypiske-luftveisagens	60	main-agent:epidemio	0 12 3 ? * *
	8	19-epidemio-respiratory-Andre-bakterier	60	main-agent:epidemio	0 14 3 ? * *
	9	19-epidemio-gastrointestinalt	60	main-agent:epidemio	0 16 3 ? * *
	10	19-epidemio-gastrointestinalt-bakterie	60	main-agent:epidemio	0 18 3 ? * *

Fig. 2 Screenshot from the list of periodic missions maintained by the Mission Scheduler

executions of the agents on all mission targets on behalf of the requester. It also manages the missions by monitoring its creation, execution and result generation. It can also abort ongoing missions by instructing each individual agent to abort its execution.

A system component is dedicated to protect and preserve the computing resources for local usage. This system component is named the Agent Daemon or AgD. Its responsibility is to make sure that the computing resources is not over exploited, reducing the responsiveness of the system towards the local users. It also controls what computations that is allowed on the local data, by authenticating the processing requests, before the processes are instantiated, and maintaining a list of supported agent types. The owner of the server controls the list of the supported agent types. The Mission controller and the Agent Daemon negotiate execution of agents by the Mission Controller forwarding its mission specification to one or more Agent daemons. If the mission specification is acceptable, the agent type supported, and the requester is authenticated, the agent is scheduled for local execution. *When the execution time arrives, the agent is instantiated by the Agent Daemon.*

When an agent has been instantiated by the Agent Daemon, it informs its Mission controller that it is operational. Agents normally

are of two kinds, coordination agents or computing agents. Coordination agents typically initiate sub-missions (to coordinate computation of data on many targets). This is done by the agent sending a mission specification to a Mission Controller, specifying what computation agents to execute on each target. The coordination agent then act as recipient of the computation results done by the set of computation agents, by receiving the computation results from the computation agents as XMPP messages, *or by using a file transfer sub mission to transfer the result file to another location*. Normally the coordination agent forwards its computation results to a component called "Publisher." The responsibility of the Publisher is to make available the computation result via a report database. In the current version of the Snow system we make the data in the report database available via web service interfaces for different kinds of clients.

File transfer in the Snow system is provided by an agent mission organized into three separate stages as shown in Fig. 3. In Fig. 3 the Mission Scheduler at the coordination server (PO for post office)

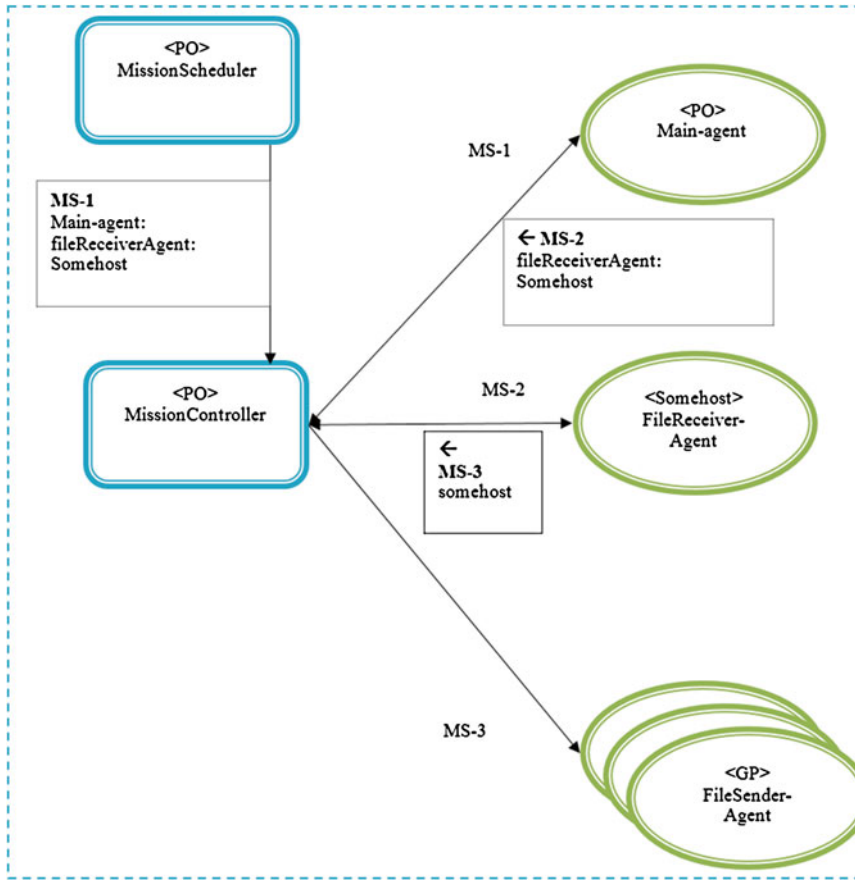

Fig. 3 Mission stages for file transfer missions in the Snow system where files from a number of GP offices are transferred to "somehost"

send a mission specification (MS-1) to the Mission Controller. The Mission Controller negotiate MS-1 with the local Agent Daemon that instantiate the Main Agent. When operational, the Main Agent requests a sub-mission (MS-2), a fileReceiverAgent on the file transfer target host (somehost). When operational the fileReceiverAgent request a number of fileSenderAgent's to transfer files to "somehost." When all files are transferred, the three agent missions are completed, in the order MS-3, MS-2, and finally MS-1.

6 Security Manager

The authentication and encryption functionality in the Snow system [19] ensures that users of snow clients (doctors and disease prevention officers etc.) can be authenticated independent of the institution and Snow server they are connected to. This is done to enable GP's to roam between health institutions, which they normally do, and still receive messages and alarms of disease outbreaks. The Security manager ensures this authentication and enables end-to-end encryption of messages between users of the Snow system. Each Snow user is given a unique Jabber Identity (JID) in the form <health personnel number>@<domain>. The health personnel number (HPN) is a globally unique identity assigned to each health worker by the Norwegian health authorities. The domain contains the hostname of the health institution where the user connects. The JID facilitates XMPP message routing within the Snow substrate. The security manager ensures message routing, authentication, certificate revocation and end-to-end encryption based on its role either at the post office (*see* Fig. 1) or as a server within a health institution. In the role as a post office it keep track of users located within the area covered by the server and health personnel visiting from other geographical areas. *See* ref. [19] for additional details.

7 The Export and Import Managers

As data is the major resource driving the need for distributed computations, the Snow system provides Export Manager and the Import Manager components for making data available for distributed computations. Both components run on each site to export data from EHR, pre-process and import to the Snow database. Only agents running locally on a Snow server can request exports of data. This limitation ensures that no external entity is able to request data export from the Export Manager, as only supported agent types can be instantiated by the Agent Daemon. As the sources of the contributed data may be located on a host remote to the Snow server, but able to connect to the snow server, the Export Manager is implemented as an XMPP client that runs as a service

on the EHR server. When a data export is requested by a local agent, the Export Manager ensures that the export is made and made available for the Import Manager on the Snow server at the same site, a fileSender-agent or some other agent. When the export is completed, the Export Manager notifies the requesting agent about it. The agent then potentially request import of the data by the Import Manager. The Import Manager then import the data files into the local Snow system dedicated database. The Import Manager use rule files (and Drools) to transform, validate and/or perform validation of the data before inserting or updating data in the Snow database. Which rule file to use is normally specified in the configuration of the Import Manager. When the import is completed the requesting agent is notified about the success or failure of the import request. If the import was successful, the agent may initiate computations, like performing aggregations or computing descriptive statistics about the newly available data resource.

8 Software Update and System Monitoring

As the Snow system is a decentralized system supporting arbitrary large networks of health institutions that contribute data and computing resources for the benefit of our community, software update and system monitoring are major concerns with regard to managing such a system with minimal system administration resources. To ensure that the software, utilizing the data made available for computations by the health institutions, has sufficient quality and satisfy the necessary security requirements, only software certified and signed by the Snow project can be used and distributed to snow servers. All software added to the software repositories utilized by Snow servers need to be rigorously tested outside and within our test environment before it is distributed through our software update service. All software must be signed on a standalone computer with no network connection before it can be distributed to the snow servers. Every night all Snow servers connect to check whether updates to any Snow software module is available. If updates to some software modules are available they are downloaded, verified to be valid (signed by the Snow project) and installed. As the Snow system covers large geographical areas, centralized system monitoring of correct system behavior is crucial for several reasons. (1) As data is reused to provide services to health workers and the population, the representativeness and coverage of the data is essential to document to avoid misinterpretation of the data. (2) Avoiding local manual system administration. (3) Knowing when and why system components on any snow server fail. To support these aspects the Snow Monitor, log manager and mission log are used.

9 Application of Reuse Using the Snow System

The Snow disease surveillance service performs daily extractions of data from microbiology laboratory and general practitioners' (GP) offices. Data from the GPs' offices provide an indication of the activity level in primary care within the symptom groups monitored by the service. It is also used to check when a GP office passes the threshold of an epidemic situation, as done by the national sentinel surveillance system. If the amount of consultations labelled with the diagnosis code, like "R80 Influenza" in the ICPC2 terminology, an epidemic situation is signalled, for the GP office in question. Data from the microbiology laboratory at the University Hospital of North Norway provide a solid foundation for a number of services offered to health personnel and the general public. These data are used to provide a weekly email service to all primary care providers, hospital doctors, and other interested, with an overview of infectious diseases in the municipalities of Troms and Finnmark counties. The data are also presented on a graphical user interface in the Norwegian healthnet and on Internet at http://snow.telemed.no.

Based on daily updated data about communicable diseases in the municipalities of Troms and Finnmark counties, we have implemented a prototype of a disease query engine, aimed at helping the general public to find reliable self-care information about likely diseases matching a set of symptoms. The motivation for doing so is to enable an increasing degree of self-help and empower people to have appropriate health service utilization. The disease query engine prototype is available at http://www.erdusyk.no.

Almost every year the seasonal influenza floods our health service with ill patients. Our national sentinel surveillance system is aimed at being prepared for the outbreak. As an alternative to the sentinel system we have developed two components aimed at the same objective, an automatic outbreak detection and disease forecast service. The automatic outbreak detection service is based on the cSiZer algorithm [20], which signals when a significant increase in number of cases is detected. Secondly, we use retrospective case data to give a prediction about the future number of cases. The forecast service uses the retrospective microbiological data to predict the spatiotemporal spread of multiple diseases including Influenza A across municipalities in northern Norway. Inspired by the weather forecast service, the system makes routine forecasts 1 week ahead. The forecasts are visualized through a web-based client on a table and map.

The provided examples are only a few of the potential services the Snow system can deliver, if deployed on a large scale. We are currently working towards enabling health professionals to compare their medical practice, without needing to worry about patient

privacy or exposing their treatment profiles to outsiders. This work is done in collaboration with Mediata, which is a company that provides tools to extract data from primary care EHR systems, and for analysis of practice. We are also working on methods and tools for privacy preserving statistical processing of health data that will allow for ad hoc statistical processing of data stored in the distributed EHR systems.

By implementing and deploying the services described above we have proved the distributed approach's ability to make medical data available for processing. As a result of this approach all municipalities covered by the Snow disease surveillance system now fulfill the Infectious Disease Control Act [21], where the municipal medical officer in paragraph 7.2b is imposed to "keep continuous overview of communicable diseases in the municipality" without any investments from the municipalities. When our disease query engine becomes fully operational, the municipal medical officer will also fulfill the responsibility described in paragraph 7.2e "provide information, knowledge and advice to the public," without any investments being necessary by the municipalities. These examples demonstrate the potential benefits of making data, currently hidden away in medical systems, available for processing and sharing for the benefit of our society.

10 Discussion

Systems for large-scale reuse of EHR data need to meet a challenging set of requirements. They must meet the legal and regulatory requirements for privacy for the entities covered by the system. Enable good coverage by supporting requirements for voluntary participation [6], be automatic, support easy installation, update and system monitoring, be asynchronous, and support two-way communication. Others and ours efforts [1, 22] have proved that a decentralized system, like the Snow system, is able to meet these requirements and provide access to the data in a way that both protect the privacy of the patients while enabling reuse of clinical data for the benefit of our society and future patients. The drawback of using the decentralized approach is the resources that need to be used to manage and maintain such a system. The alternative, the data warehouse approach to enable reuse of data, is based on moving data. Moving data has implications for what kind of data that can be moved, ensured by laws as The Health Insurance Portability and Accountability Act (HIPA) in the US and Norwegian Research ACT in Norway [5]. These laws and regulations necessarily lead to reduced value of the data, as data need to be less detailed to ensure privacy. Another issue is that moving data require patient consent to reuse of data that potentially can create bias. Compared to the alternative where data can be used anonymously, the burden

for the patient is larger as the patient's data become more visible (more eyes see it) in the data warehouse approach.

Few examples of large-scale centralized storage approaches exist. Such systems would enable easy reuse of data. However, these systems do not scale beyond the administrative border they support, raising the need for a distributed system to support studies that need datasets beyond the population served by the system. Research on rare diseases, rare combinations of conditions, and large-scale studies would suffer from this limitation. Another issue that affects the usefulness of such systems is the support for heterogeneous health systems. Countries that have a combination of public and private health services may not be able to build the necessary coverage based on a centralized storage solution, as private health institutions may be reluctant to share all of their data. History has also shown us that it is difficult to establish centralized systems. Some efforts in this direction have failed, leading to wasted resources and damage reputation.

A decentralized system, where each participant contribute their data and computing resources, have the benefit of being able to scale beyond the administrative borders that exists for data warehouse and centralized systems. As long as the privacy of the patients can be ensured, computations on the data can be performed, independent of the size of the computation network.

11 Conclusion

Our conclusion is that decentralized system for reuse of health data may scale beyond the limits of data warehouse and centralized systems approaches. As each participant own his or her data and computing resources, the benefit of joining such a network increases as the heterogeneity and size of the network grows. The drawback of using this approach is the costs necessary to establish and manage such a system. The benefits of the decentralized approach is the value that can be generated without trading the privacy of the patients for the benefit of having access to the data, we can have both, but need to invest the necessary resources. A lot of problems still need to be solved to reduce the drawbacks of the decentralized approach. More research is also needed in privacy preserving computations methods that can be used to extract knowledge from the vast data resource that is currently underexploited.

References

1. Sittig DF, Hazlehurst BL, Brown J, Murphy S, Rosenman M, Tarczy-Hornoch P et al (2012) A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogenous clinical data. Med Care 50(Suppl):S49–S59
2. Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S et al (2009)

The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc 16(5):624–630

3. Frankovich J, Longhurst CA, Sutherland SM (2011) Evidence-based medicine in the EMR era. N Engl J Med 365(19):1758–1759
4. Chen Y-C, Wu J-C, Haschler I, Majeed A, Chen T-J, Wetter T (2011) Academic impact of a public electronic health database: bibliometric analysis of studies using the general practice research database. PLoS One 6(6):e21404
5. omsorgsdepartementet H (2009) Helseforskningsloven. Available from: http://www.regjeringen.no/nb/dep/hod/dok/lover_regler/forskrifter/2009/helseforskning-sloven.html?id=570542
6. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R (2010) Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. Med Care 48(6 Suppl):S45–S51
7. Brown J, Syat B, Lane K, PLatt R. Blueprint for a distributed research network to conduct population studies and safety surveillance. AHRQ Agency for Healthcare Research and Quality. Available from http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=465
8. Johansen D, Renesse R van, Schneider FB (1995) An introduction to the TACOMA distributed system version 1.0. Technical Report No. 95-23, University of Tromsø
9. Gentry C (2009) Fully homomorphic encryption using ideal lattices. Proceedings of the 41st annual ACM symposium on theory of computing. ACM, New York, NY, pp 169–78. Available from http://doi.acm.org/10.1145/1536414.1536440
10. Gentry C (2009) A Fully homomorphic encryption scheme. Stanford University. Available from http://crypto.stanford.edu/craig/craig-thesis.pdf
11. Bogdanov D (2013) Sharemind: programmable secure computations with practical applications. Thesis. Available from http://dspace.utlib.ee/dspace/handle/10062/29041
12. Bellika JG, Sue H, Bird L, Goodchild A, Hasvold T, Hartvigsen G (2007) Properties of a federated epidemiology query system. Int J Med Inform 76(9):664–676
13. Bellika JG, Hasvold T, Hartvigsen G (2007) Propagation of program control: a tool for distributed disease surveillance. Int J Med Inform 76(4):313–329
14. Hartvigsen G, Johansen S, Helme A (1995) A secure system architecture for software agents: the virtual secretary approach. University of Bologna
15. Bellika JG, Hartvigsen G, Widding RA (1998) The virtual library secretary—a user model based software agent. Pers Technol 2(3):162–187
16. Saint-Andre P (2004) Extensible messaging and presence protocol (XMPP): instant messaging and presence. RFC 3921
17. Saint-Andre P (2004) Extensible messaging and presence protocol (XMPP): core. RFC 3920
18. Ignite Realtime: Openfire Server. Available from http://www.igniterealtime.org/projects/openfire/
19. Bellika JG, Ilebrekke L, Bakkevoll PA, Johansen H, Scholl J, Johansen MA (2009) Authentication and encryption in the Snow disease surveillance network. Stud Health Technol Inform 150:725–729
20. Skrøvseth SO, Bellika JG, Godtliebsen F (2012) Causality in scale space as an approach to change detection. PLoS One 7(12):e52253
21. regjeringen.no. LOV-1994-08-05-55 Smittevernloven (2007). Available from: http://www.regjeringen.no/nb/dok/lover_regler/lover/Smittevernloven.html?id=448170
22. Weber GM (2013) Federated queries of clinical data repositories: the sum of the parts does not equal the whole. J Am Med Inform Assoc 20(e1):e155–e161

Chapter 8

Data Mining for Pulsing the Emotion on the Web

Jose Enrique Borras-Morell

Abstract

The Internet is becoming an increasingly important part of our lives. Internet users share personal information and opinions on social media webs expressing their feelings, judgments, feelings or emotions easy. Text mining and information retrieval techniques allow us to explore all this information and discover what the authors' opinions, claims, or assertions are. A general overview of sentiment analysis' current approaches and its future challenges, providing basic information on their current trends, is made throughout this chapter.

Key words Sentiment analysis, Natural language processing, Data mining, Web 2.0, Social web

1 Introduction

In recent years there has been a growing interest in text analysis; written data such as individual words, sentences, or documents give information about the writer's opinion, the feeling of the blogger or the sensation of the journalist. Every day the Internet becomes more accessible for users, making possible for them to express their opinions on any topic that matches their interests. The web has changed and the rise of social media webs help users to interact and have a close relation with other users. In that way, every day, more users before buying a product search for other users' product opinions, or open a new forum or create a post to discuss about what products match better their needs.

Web written data is easily available for the users and the sellers. The powerful combination of information retrieval systems and text mining techniques allows us to detect the users' opinions. Thus new social media trends can be determined, improve the performance of personalization systems, and get users more comfortable with the site layout.

The language analysis of words, expressions and abbreviations, say a lot about who we are. However, the overwhelming amount of information present on the Internet makes hard for the users to analyze this information. Natural language processing (NLP) deals

Carlos Fernández-Llatas and Juan Miguel García-Gómez (eds.), *Data Mining in Clinical Medicine*, Methods in Molecular Biology, vol. 1246, DOI 10.1007/978-1-4939-1985-7_8,

with the automatic treatment of natural language speech. Through the use of statistics, machine learning methods, etc. the NLP tools allow us to identify what sentiments are expressed in texts, and whether the expressions indicate positive neutral or negative opinions. It is possible a relation between positive words and positive texts, blogs or reviews? Sentiment analysis is the way to discover what the answer is.

Sentiment analysis approaches allow users to determine the customer's product feelings such as positive or negative or their emotions: love, joy, surprise, anger, sadness and fear [1]. However sentiment is a more completed concept, just imagine the vastness of ways people expresses their opinions, attitudes and emotions and how hard is the task to determine what is the writer' feeling.

2 Opinion Mining or Sentiment Analysis

Sentiment analysis or Opinion mining can be defined as the classification of documents based on the overall sentiments expressed by opinion holders [2]; Both terms are widely used in the data mining literature with a common goal: to identify and determine subjective information [3]. The sentiment analysis accuracy is defined as how well it agrees with human judgments.

Classification is the fundamental technology for sentiment analysis developments; recognize and extract subjective information usually attempt three independent classification goals: *opinion detection*, *polarity classification* and *opinion's target* [4]. The first goal, opinion detection, classifies the text as objective or subjective. The second goal, polarity classification, classifies the opinion as one of two, or more, opposing sentiment polarities (positive and negative). Finally, as a complementary task, identify the opinion's target helps to improve the final results. However the sentiment analysis tasks are not easy and present a complex variety of problems that need to be considered in each approach; either the domain dependency; which makes that algorithm predict well texts from one domain instead of other domains, the different words meaning or the users' subjective definition [5, 6].

Traditionally sentiment analysis classification relied on two approaches machine learning models and semantic orientation. While the machine learning models rely on a feature vector which is made by n-grams, occurrence frequencies, etc. The semantic orientation approach uses a dictionary or lexicon of pre-tagged words, which are classified into different classes and then system match each word to the dictionary ones. Finally the word polarity value is added to final text score [7]. Finally, using the correct sentiment analysis approach is key to success, where the performance depends on the problem, data and features.

Detect the subjectivity of the opinion, the importance of the domain; slang or vernacular abbreviations, determine the irony, or the rhetorical issues are some of the key challenges for sentiment analysis developments that new approaches should deal.

2.1 Classification

Classification is perhaps the most widely studied topic in sentiment analysis [2, 3] and maybe one of the most complicates. There are lots of situations where different approaches will get different results, for example[1]: the sentence "just read the book" contains no explicit sentiment word and it is highly depending on the context. If it appears in a movie review it means that the movie is not good. However, if it is in a book review it delivers a positive sentiment in the lexical approach. Thus the definition of sentiment is based on the analysis of individual words and/or phrases and the context where they are.

Sentiment analysis methodologies are not a perfect algorithm. It is not possible to get an insight based on an isolated instance. If a human could not provide a useful answer based on a single context much less a mathematical method will. This is why it is needed to get a huge quantity of opinions and analyzed them to get a light idea about the sentiment on one topic. Given a sufficiently large corpus any language or dialect can be analyzed.

2.2 Sentiment Analysis Approaches

Machine learning and lexical approach are the two most common sentiment analysis approaches, although many algorithms have elements of both [8] (for a deep literature review [2] and [3] articles). Both approaches are applied to classify different text levels (documents, simple sentences, etc.). Computers perform automated sentiment analysis of digital texts, using elements from machine learning such as latent semantic analysis, support vector machines, and semantic orientation. This section provides an overview of each task.

Each approach has its own advantages and disadvantages. Thus the lexical approach does not need for labeled data and previous text training. Although it needs a powerful dictionary, which it is usually hard to find, and usually has difficulties in adapting to different contexts. On the other hand machine learning approach does not need a dictionary and, usually, it adapts better to new contexts. However the features' selection needs to be done carefully and demands an initial manual training [9].

Machine Learning Approach

Machine learning approaches have been widely successful for various text classification tasks in the past. According to A. Ethem machine learning' goal is to program computers to use example data or past experience to solve a given problem [10]. The machine

[1] http://stackoverflow.com/questions/4199441/best-algorithmic-approach-to-sentiment-analysis

learning algorithms implemented in Sentiment analysis, usually, are organized as supervised learning and unsupervised learning:

- Supervised learning: Correct results are part of the training data. The training data must be labeled, the algorithm learns text features that associate with positive or negative sentiment [11].
- Unsupervised learning: The model is not provided with the correct results during the training. It usually starts with a simple sentiment rule. Turney uses an unsupervised sentiment analysis method using only the words *excellent* and *poor* as a seed set [12].

A significant disadvantage of this approach, for social science uses, is about the sentiment features extraction, because it can extract non sentiment features that associate with sentiment because they are frequently used in sentences of a particular polarity.

The machine learning classification key is the engineering of a set of effective feature. The features are typically based on [3]:

- Terms and their frequency: words appearance (n-grams: sets of n consecutive words in texts that normally associate with sentiment) and their frequency in the text.
- Part of speech: The part-of-speech (POS) or linguistic category of words.

Lexicon Approach

The lexicon approaches are focused on building a collection of sentiment words. There are some words such as: *like*, *love hate*, *bad*, or phrases and idioms whose indicators, in sentimental meaning, are highly significant. A sentiment lexicon approach starts with a lexicon dictionary of positive and negative words and phrases. Each word in the text is compared against the dictionary and the word polarity value is added to the global sentence polarity value. Finally, with the total polarity value, the system classifies the sentences (i.e.: positive domain or negative domain).

The lexicon approach predicts the sentiment of text based on the orientation of words or phrases in the document. However use these words is not enough to get good results. Lexicons usually apply a set of rules to predict the sentiment, these rules deal with aspects such as negation, the use of boosters/intensifiers [13], adjectives etc. Dictionaries of lexicon words can be created manually or automatically using a seed of words.

2.3 Sentiment Analysis Web Resources

The Internet offers many opportunities for developing sentiment analysis projects. We have separated these resources between the lexicon dictionaries—a dictionary of pre-tagged words with affective ratings and datasets, where the sentences are labeled regarding to different sentiment polarity approaches.

Lexicon Dictionaries

- *General Inquirer*: a computer-assisted approach for content analyses of textual data. Database of words and manually created semantic and cognitive categories, including positive and negative connotations [14].
- *LIWC* (Linguistic Inquiry and Word Count): a text analysis software program. Counts words belonging to categories such as positive and negative words [15].
- *Wordnet*: a large lexical database of English nouns, verbs, adjectives, and adverbs. They are grouped into sets of cognitive synonyms (synsets), each one express a distinct concept. Words are linked according to lexical conceptual relations [16].
- *SentiWordNet*: a lexical resource for opinion mining. It assigns to each synset of WordNet three sentiment scores: positivity, negativity and objectivity [17].
- *Whissell's Dictionary of Affective Language*: an emotional text recognition based on the Cynthia Whissell's Dictionary of Affective Language [18].
- *Sentic API*: a semantics database with more than 14,000 common sense concepts [19].

Datasets Classified

- *Movie Review Data*: collections of movie-review documents labeled with respect to their overall sentiment polarity (positive or negative) or subjective rating (e.g., "two and a half stars"). Besides sentences are labeled in relation to their subjectivity status (subjective or objective) or polarity [20].
- *Multi-Domain Sentiment Dataset*: based on reviews from Amazon.com, includes star rating and reviews divided into positive/negative [21].
- *The MPQA Opinion Corpus*: contains articles from a wide variety of news sources manually annotated for opinions and other private states (i.e., beliefs, emotions, sentiments, speculations, etc.) [22].

2.4 Creating a Sentiment Analysis Model

As sentiment analysis popularity increases new software solutions are created. These applications can be easily used by users to build their own sentiment analysis models, for example: the *Google prediction API*: a sentiment analysis model used to analyze a text string and classify it with one of the labels that the user provide [23], or *LingPipe*: a language classification framework that can implement two classification tasks, separate subjective from objective sentences, or separate positive from negative reviews [24].

3 The Web 2.0, Social Media and Sentiment Analysis

Kaplan defines the Web 2.0 as *a group of Internet-based applications that is built on the ideological and theological foundations of Web 2.0 and that allows the creation and exchange of user-generated content* [25]. The web 2.0 has changed the way information is accessed. Nowadays the Internet lets users to communicate easily with each other or exchange information. People generate their own content, post real time messages and respond to opinion on a whole variety of topics.

A late 2012 survey by the Pew Research Center's Internet and American Life Project [26] showed that 67 % of all Internet users use social networks: opinion sharing, discussions, micro blogging, etc. The data exchanged over social media helps users make better decisions. Thus, sentiment analysis' tools can analyze automatically this amount of data, dealing with the tedious users' task of detecting subjectivity in texts and the subsequent extraction, classification, and summarization of the opinions available on the users' interest topics.

On the companies' side, nowadays, they are not able to control all the information available about them. However trough sentiment analysis techniques they can manage it, for example: the positive and negative brands' reviews can be tracked, and then measures their overall performance, besides is an easy way for brands to interact with their followers and engage them, or simply know what a new product consumer reaction is.

Sentiment analysis is the most efficient and effective way to understand a customer and their social media interactions. Companies can track social media conversations to predict new users' trends, identify new opportunities, and manage company's reputation (online brand reputation). Moreover, for the users, it can be used for spam detection (looking for anomalous language patterns) or find similar users.

4 Future

As the Internet expands new forms of communication are created, increasing the amount of emotion we are able to pass on. Sentiment analysis approaches are evolving from simple rules-based till statistical methods, machine learning etc. Sentiment analysis techniques detect opinions, feelings, and emotions in online, social, if we are able to compare our behavior with millions of other events and records, computationally talking, the limit is the energy required to do such big analysis. However, differing contexts make this analysis extremely difficult to turn a string of written text into a positive or negative sentiment.

Where sentiment analysis is heading next? Sentiment analysis has a big future; probably, the major efforts will be aimed at online and social media usages, special mention to micro-blogging, as well as enterprise feedback in surveys. However, other targets as ehealth will be developed, for example: through sentiment analysis health organizations can take a picture of the patient feelings.

Technology is advancing every day, making users' life easier, we are getting closer and closer to pulse the emotion on the web and, as we have seen, sentiment analysis is the key to success.

References

1. Parrott W (2001) Emotions in social psychology: essential readings. Psychology Press
2. Bo P, Lee L (2008) Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. pp 1–135
3. Liu B (2012) Sentiment analysis and opinion mining. Morgan & Claypool, p 167
4. Mejova Y (2009) Sentiment analysis: an overview [Online]. [Cited: 08 08, 2013]. http://www.cs.uiowa. edu/~ymejova/publications/CompsYelenaMejova.pdf
5. Anthony A, Gamon M (2005) Customizing sentiment classifiers to new domains: a case study. Proceedings of the International Conference on Recent Advances in Natural Language Processing, vol 1. pp 2–1
6. Songbo T, Wu G, Tang H, Cheng X (2007) A novel scheme for domain-transfer problem in the context of sentiment analysis. ACM proceedings of the sixteenth ACM conference on information and knowledge management. pp 979–982
7. Whitelaw C, Navendu G, Shlomo Argamon (2005) Using appraisal groups for sentiment analysis. Proceedings of the 14th ACM international conference on information and knowledge management. ACM. pp 625–631
8. Thelwall M, Kevan B (2013) Topic-based sentiment analysis for the social web: the role of mood and issue-related words. J Am Soc Inform Sci Technol 64:1608–1617. doi:10.1002/asi.22872
9. He Y (2012) Incorporating sentiment prior knowledge for weakly supervised sentiment analysis. ACM Transactions on Asian Language Information Processing (TALIP) 11(2): 979–982
10. Alpaydin E (2004) Introduction to machine learning. MIT Press
11. Bo P, Lee L, Vaithyanathan S (2002) Thumbs up?: sentiment classification using machine learning techniques. Association for Computational Linguistics. Proceedings of the ACL-02 conference on empirical methods in natural language processing, vol 10. pp 79–86
12. Turney, PD (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, pp 417–424
13. Maite T, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. Computational linguistics. pp 267–307
14. Hurwitz R (2012) General Inquirer home page [Online]. [Cited: August 08, 2013]. http://www.wjh.harvard.edu/~inquirer/
15. Pennebaker JW, Booth RJ, Francis ME (2007) Linguistic inquiry and word count [Online]. [Cited: August 08, 2013]. http://www.liwc.net/
16. Princeton. Wordnet a lexical database for english [Online]. [Cited: Agosto 08, 2013]. http://wordnet.princeton.edu
17. Andrea E, Fabrizio S (2010) SentiWordNet [Online]. [Cited: August 08, 2013]. http://sentiwordnet.isti.cnr.it/
18. Whissell (2013) Emotional text recognition using Whissell's dictionary of affective language [Online]. [Cited: August 08, 2013]. http://sail.usc.edu/dal_app.php
19. MIT Media Laboratory (2013) Sentic API [Online]. [Cited: August 08, 2013]. http://sentic.net/api/
20. Bo P, Lee L (2005) Movie review data [Online]. [Cited: August 08, 2013]. http://www.cs.cornell.edu/people/pabo/movie-review-data/
21. Dredze M, Blitze J (2009) Multi-domain sentiment dataset [Online]. [Cited: August 08, 2013]. http://www.cs.jhu.edu/~mdredze/datasets/sentiment/
22. Wilson T (2013) MPQA opinion corpus [Online]. [Cited: August 08, 2013]. mpqa.cs.pitt.edu

23. Google (2013) Google prediction API [Online]. [Cited: August 08, 2013]. https://developers.google.com/prediction/docs/sentiment_analysis
24. Ling P (2005) Sentiment tutorial [Online]. [Cited: August 08, 2013]. http://alias-i.com/lingpipe/demos/tutorial/sentiment/read-me.html
25. Kaplan AM, Michael H (2010) Users of the world, unite! the challenges and opportunities of social media. Business horizons, vol 1. pp 59–98
26. Duggan M, Brenner J (2013) The demographics of social media users — 2012. PewInternet [Online]. Pew Research centers. [Cited: August 08, 2013] http://pewinternet.org/Reports/2013/Social-media-users/The-State-of-Social-Media-Users.aspx

Chapter 9

Introduction on Health Recommender Systems

C.L. Sanchez-Bocanegra, F. Sanchez-Laguna, and J.L. Sevillano

Abstract

People are looking for appropriate health information which they are concerned about. The Internet is a great resource of this kind of information, but we have to be careful if we don't want to get harmful info. Health recommender systems are becoming a new wave for apt health information as systems suggest the best data according to the patients' needs.

The main goals of health recommender systems are to retrieve trusted health information from the Internet, to analyse which is suitable for the user profile and select the best that can be recommended, to adapt their selection methods according to the knowledge domain and to learn from the best recommendations.

A brief definition of recommender systems will be given and an explanation of how are they incorporated in the health sector. A description of the main elementary recommender methods as well as their most important problems will also be made. And, to finish, the state of the art will be described.

Key words Health recommender systems, Information research, Health information, Recommenders, Web 2.0

1 Introduction

Information and communication technologies (ICT) provide new ways of searching and gathering health information. Health consumers have access to a vast amount of different kinds of resources which are disseminated through the Word Wide Web [1].

Health content creators are continuously overloading the Internet with information. This makes the search for trusted health information more complicated, yet necessary.

People are demanding accurate and trustworthy health information. Search engines are in charge of this task and many studies analyze how these search engines determine health information to be trustworthy [2].

What users demand is trusted information selected according to their user profiles. Therefore, the so-called health recommenders find trustworthy health information and adapt it to the user profile, which can be obtained from their personal health record [3], in a process that contributes the empowerment of the patient [4–6].

Carlos Fernández-Llatas and Juan Miguel García-Gómez (eds.), *Data Mining in Clinical Medicine*, Methods in Molecular Biology, vol. 1246, DOI 10.1007/978-1-4939-1985-7_9,

Recommender methods depend mainly on user profiles, health information (also known as items) to search, and domains. Health recommender systems experience several difficulties [7] mainly cold-star, serendipity, sparsity, and spam (all these concepts will be explained in the following sections).

Furthermore, recommender systems still require improvements to produce more effective choices. In the next section we will present a brief description of health recommender systems.

2 Recommender Elements

Domains, user profile, and items are the main elements of a recommender system.

First, we describe a domain as the environment where all elements (items, users and their relationships) interact. For instance, if we are talking about diabetes, all documents, research, and relationships between clinicians and patients belong to the diabetes domain. Recommender methods usually fit best in a particular domain and may be useless in another; they are mostly focussed on and built for specific domains.

We can define a user profile as all those properties that identify a unique person within several domains. For example, a 28-year-old patient from Ireland suffering type 1 diabetes who measures his blood sugar levels and obtains 11.2 mmol/L. The recommender gathers all this information (place of birth, date of birth, diabetes type, blood sugar level) and includes it in the system. All of these properties determine one recommended choice or another.

Items are those elements that users are searching according to their needs. In many systems these items are documents, but generally an item is just a piece of information (for instance, drugs indications, dosing, use instructions, videos describing a disease).

Some recommenders collect those health information choices, process them all, and improve future recommendations in order to offer the best one. These interactions represent the essence of recommendations [8].

3 Basic Methods on Recommender Systems

3.1 Collaborative Approach

One of the two main options in recommender systems is the so-called collaborative approach, focussed on the users' social networks and the items they have selected in the past. The recommender processes them all in order to make future proposals. This approach analyzes the users' interactions in the past and shapes a recommendation. This model is also known as memory-based approach.

Table 1
User/item interaction matrix

User/item	Item 1	Item 2	…	Item *N*
User 1				
User 2				
…			(a) Integer value (b) Like/dislike	
User *M*				

Table 2
Recommended items

Recommender items	Item 3	Item 2	…	item *N*

In this approach, a table similar to Table 1 may be used to represent users in rows and items in columns. The intersections are filled with an integer or Boolean value. This helps us to determine the behavior among users.

We can also represent it as a matrix of items recommended for users (*see* Table 2).

The basic consideration of collaborative recommendation is "if users shared the same interests in the past, they would have similar tastes" [8] (*see* Fig. 1). The Collaborative Filtering (CF) answers these questions: How do users find others with similar taste? What about new users? What about new items?

In this method, the rating of an item by a user is computed basically as a weighted sum of the ratings of other users. The weight in this sum is a measure of the similarity of the users, and there are several ways to compute it [7].

Pearson's correlation is one of these ways to measure the users' nearness. But this simple method is usually not enough to achieve the optimal recommendation of the recommendation; that is why it is used only in a few subsets of domains.

An alternative is the Cosine similarity measure, where users are considered as vectors in a n-dimensional space (with n being the number of items). Then we define a standard metric that measures the angle between them. The result may vary between 0 (orthogonal vectors, that is, no similarity) and 1 (equal) [7].

Both methods, Pearson's Correlation and Cosine similarity, can be improved if we subtract the average rating behavior of the user. This way the fact that different users may rate differently is taken into account. The resulting method is called Adjusted Cosine

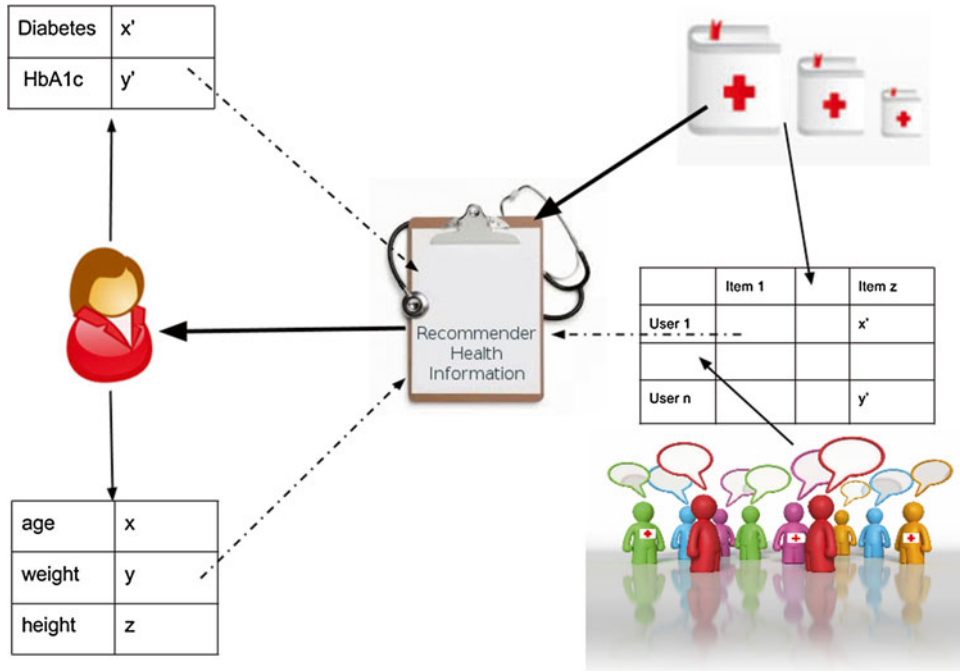

Fig. 1 Collaborative approach

measure. Other options include Spearman's rank correlation coefficient and mean square difference measure [8].

Other more elaborate solutions found in the literature are inverse user frequency, significance weighting, and case amplification [7]:

- Inverse user frequency: Reduce the relative importance of those cases that receive universal agreement (such cases are rated more frequently but their rates are not very useful).
- Significance weighting: Two users may be highly correlated based on too few items, so a linear reduction of the similarity weight can be used in these cases. If a minimum of, say, 50 co-rated items is imposed, the prediction improvement is significant.
- Case amplification: Gives more weight to highly similar users through an amplification factor.

A different method is to try to reduce the computational complexity. For instance, when the nearest active neighbor with positive correlations is selected. After several selections, this method tends to select the same recommended items. However, it suffers from sparsity (insufficient available data leading to poor recommendations), so a surrogate method can be used which selects the K nearest neighbors (using the K nearest neighbor algorithm) [9, 10]. Both methods decrease the quality of the choice.

Many of above mentioned methods (correlation, cosine) can be adapted to compute similarities between items instead of users, and it has been shown that this approach may provide advantages in terms of computational performance [7].

3.2 Item-Based Nearest Neighbor Recommendation

Defines the similarity between items. Tere is an offline processing that reduces subsequent real-time recommendations.

First we have to consider users as vectors in a n-dimensional space. Then we define a standard metric that measures the angle between them. The result may vary between 0 (no similarity) and 1 (equality) [7].

A better adaptation is the Adjusted Cosine measures that considers the average rating behavior of the user [10, 11].

Processing the entire matrix of users and items relations in real time could consume a lot of time and resources. To prevent that, all this data may be previously processed to obtain an item ratings' matrix for the community. This preprocessing is a usual approach in "learning machines." However, obtaining this matrix is very difficult because of the required minimum number of rates on the same item by different users, as well as the limited number of neighbors [12].

Finally, it is worth mentioning the "Slope One" method, which precomputes the average difference between the ratings of different users on one item [13]. This approach is easier to implement.

3.3 Content-Based Approach

This approach considers item properties and user profiles as the essence of the recommendation (*see* Fig. 2). The idea is that if we know more about the item and/or the user the recommendation will be more accurate, even if the number of previous recommendations or the size of the user community is small. Many authors recognize as the model-based approach [8].

Each item may be represented by a list of properties (*see* Table 3) which contains all main characteristics to be identified.

First we need an off-line learning-mode phase. Only learned models are used to make recommendations online, so the system needs regular offline phases in order to improve its algorithms and make better recommendations each time [14].

Content-based approach looks for items similar to those that the user liked in the past, matching item attributes with the user profile to fit the best choice. Sometimes the entries in Table 3 are not meta-data, but simply keywords or terms appearing in the item (for instance, a document). In this case, content-based methods usually get documents (or text descriptions of other types of items) and simply filter the occurrences of these keywords. In the same manner this idea we can find the following basic methods.

3.4 The Vector Space Model and TF-IDF

The vector space method analyses all terms that appear in an item, and encode items as vectors with the dimension of the number of terms. Two measures are used: Term Frequency (TF, or how often

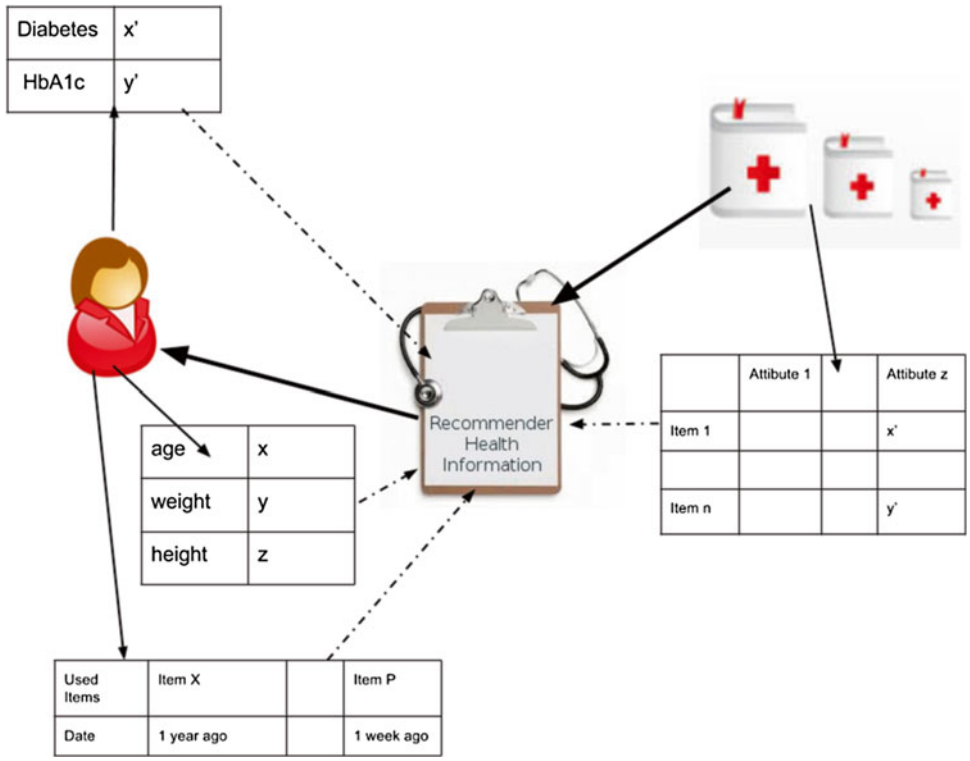

Fig. 2 Content-based approach

Table 3
Items with explicit properties

Item	Size	Height	Price
Item 1	X	Υ	Z
Item 2	A	B	C
…	…	…	…
Item X	X'	Υ'	Z'

the term appears in the item) and Inverse Document Frequency (IDF, a measure intended to reduce the weight of terms that appear too often, which are not very useful) [8]. Note that although originally this method was used with keywords (terms) in documents (items), it could be used in other contexts (for instance, with movies if associated textual contents are available). There are several improvements of this method, mainly in the sense of reducing the size of these vectors and/or the required amount of information such as stopping words and stemming, size cutoff, phrases, and context [8].

Similarity-based retrieval is based on two main measures: item similarities and user likes/dislikes on previous items. Then, if a given number of similar items were liked before, a new item will be recommended. It uses k-nearest neighbor (kNN) algorithm and is simple to implement [15], but the prediction accuracy is lower than that of more complex methods.

The Rocchio method basically allows a user to rate the items (documents), and these ratings are incorporated into the user's profile. This feedback is used by the system to improve the query [16].

Probabilistic methods use an approach similar to that of a classification task, labeling the items according to previous user's ratings. Simple classification algorithms such as Naive Bayes classifier have been successfully used [8].

Machine learning is created to separate relevant and non-relevant items. The model may be based on machine learning techniques such as clustering, decision trees, neural networks, etc. One example is the Widrow-Hoff algorithm [17].

3.5 Knowledge-Based Approach

This method is a subset of the content-based approach. It creates a knowledge base of the items characteristics in order to improve the recommendation (*see* Fig. 3). In some cases, it can solve content-based approach weaknesses, for instance knowledge-based

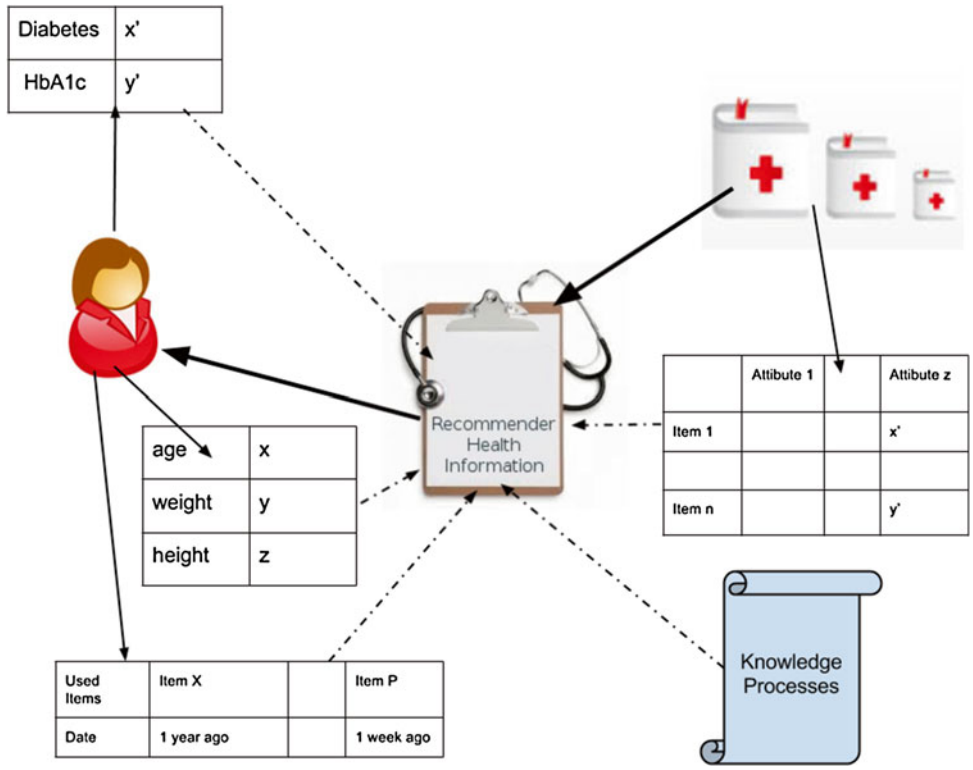

Fig. 3 Knowledge approach

recommendation may be used even with few available or useful ratings, as it can make recommendations based on the knowledge about the user and or/the item and not on previous explicit ratings.

Knowledge about the users can be obtained by asking them to specify their requirements. If the outcomes of the recommender are not suitable, the user may be asked to change the requirements. There are two basic types of knowledge-based recommenders: constraint based and case based. Constraint-based recommenders search for a set of items that fulfil the recommendation rules (or requirements). The process is similar to that of a Constraint Satisfaction Problem—CSP, that is, user's requirements and items characteristics form a set of variables and the recommender tries to obtain a solution for all these variables [8]. On the other hand, case-based recommenders try to find items similar to the user's requirements. To check this similarity a complete description (set of features) of the items is required [18].

4 Further Recommender Models

Recommender systems are taking advantage of some advancements and better models, several of which we will discuss next.

4.1 Matrix Factorization/Latent Factor Models

One of the common problems with Information retrieval consists of finding data in documents. Given a user's query, a possible solution is to find data that coincides with that query, a solution that can be used in recommended systems. However, this method has a set of latent hidden factors (for instance, synonyms or polysemous words in textual documents) [8]. Possible solutions include:

- Singular Value Decomposition (SVD), which algebraically reduces the dimension of the associated matrices (formed by the users and items) [19].
- Latent Semantic Analysis (LSA) or latent semantic indexing (LSI), discovers the latent factors and reduces the size of these matrices by merging semantically similar elements, sometimes in conjunction with SVD [20, 21].
- LSI-based retrieval: make it possible to retrieve relevant documents even if it does not contain many words of the user's query [22].

4.2 Association Rule Mining

This method finds useful patterns in a transaction dataset. The association rules are following the form $X \rightarrow \Upsilon$ *where* X *and* Υ are two disjoint subsets of all available items that satisfy constraints on measures of the significance and interestingness [23]. This method can be used in health recommender systems. For instance, consider a user looking for information on diabetes. If he/she downloads a video explaining the possible complications of the disease, a recommender may offer him/her videos on healthy lifestyles.

Table 4
True/false transactions for each item

# Transactions	Item 1	Item 2	…	Item *N*
1	True	False		True
2	False	True		False
…				
n	True	True		False

This method is represented in an iteration item matrix; each intersection cell contains a boolean value: True means presence and False means absence (*see* Table 4).

The matrix contains almost thousands of transactions. The method finds rules that correlate the presence of one set of items with another set.

This method may be evaluated as successful if given X, Y as items, it finds rules X and $Y \rightarrow Z$ with minimum support (probability that a transaction contains the three items) and confidence (conditional probability that a transaction with X and Y also contains Z) [24].

4.3 Ontologies and Semantic web Recommendations

The Resource Description Format (RDF) is a standard that represents information modeled as a "graph." RDF is one of the pillars of the so-called Semantic Web, where it can be used to serialize information represented using graphs. Together with languages like OWL (Ontology Web Language), these developments allow an enhanced representation of the information which can be used to obtain improved recommendations [25, 26].

Ontologies support coding systems both in electronic healthcare records [27] or social media [28], gathering the precise meaning of one term, and determining relationships between terms. A term is made of words that represent an item. Either terms or relations help the system to improve the recommendations it makes [29].

4.4 Hybrid Methods

Hybrid recommender systems combine two or more recommendation approaches to gain better performance. Most commonly, collaborative filtering is combined with some other techniques (as content based or knowledge based) in order to minimise their respective weaknesses [30].

5 Challenges on Recommender Systems

In this section we briefly describe some of the challenges that research on recommender systems have to address.

5.1 Implicit and Explicit Ratings

Some recommenders ask for explicit item ratings to obtain users' opinion more precisely (5-points, 7-points, like/dislike). Using these ratings allows more precise user recommendations (for example, a 10-points scale is better accepted in movies recommendations) though it requires additional effort from users [8].

On the other hand, implicit ratings collect external properties from the environment. For example, if someone buys food in the city of Málaga (southern Spain) in summer at 13 pm, the place and time are considered implicit ratings to provide healthy recommendations adequate for the usually hot and humid weather. Therefore, we cannot be sure whether this user behavior is correctly interpreted with these ratings.

Obtaining useful ratings without additional efforts from users is one of the challenges of recommender systems.

5.2 Data Sparsity

Another typical problem of recommender systems, particularly important with those using a collaborative approach, is that data and ratings tend to be sparse. A good solution is to use user profile information when calculating user similarity [7]. That is, if two users suffer from the same disease they could be considered similar even if they have not rated the same item (video, document) similarly. Other characteristics like gender, age, education, and interests could help in classifying the user. Additionally, methods like matrix factorization or latent factor models can be used to reduce the size and dimension of the rating matrix [9].

5.3 Cold-Start

It takes time to include new users and items that have not been rated yet, so several questions arise: How can recommendations be made to new users that have not rated yet? And how to deal with items that have not been rated yet either?

Some approaches avoid the new-user problem by asking the user for a minimum number of ratings before the service can be used. Others exploit the supposed "transitivity" between them through relationship graphs [8]. But the most usual approach is adopting hybrid methods, that is, a combination of content-based and collaborative methods [7].

5.4 Serendipity/ Overspecialization

Some recommenders (mainly content based) rely on the similarity of items, and the user's interest is taken for granted. However, in many contexts this assumption is wrong. For instance, when the system can only recommend items that score highly against a user's profile, the recommendation may be too obvious to be useful. This undesirable effect is called overspecialization. On the contrary, increasing the serendipity is usually useful, that is, sometimes the recommender can offer unsearched but maybe useful items, even with a certain randomness [31].

5.5 Latency

This problem appears on new items (mainly with the collaborative approach), the system is unable to select recently added items which need to be reviewed before they can be recommended [32].

6 Health Recommender Systems

Nowadays more than 80 % of the Internet users have searched for health information [33]. Patients can identify the reasons of their disease, find treatments, learn healthy habits or contact other people suffering from the same disease, and as a result they are continuously increasing their knowledge and empowerment.

But the amount of information is so huge that sometimes it is difficult for the users to find trusted health information. Identifying relevant resources is very difficult for the user; they may find confusing and misleading information about their disease, most of them with a very low quality. Additionally they want to access this information as quickly as possible. Therefore, the role of recommender systems is important, and required characteristics are usefulness, trustworthiness, performance, and ability to adapt the outcomes according to users' profiles and needs [34].

Collaborative recommendations were the most used recommenders on health sector [7]. However, health decisions depend on the knowledge of the patient's needs, and this knowledge is often incomplete. This results in errors or omissions and eventually adverse outcomes.

6.1 Personal Health Record and Recommender Systems

Personal health record (PHR) is a health record that contains information related to the care of a patient [35]. PHRs are usually maintained by the patient or a carer, so a useful approach is having a health recommender system include PHR as a part of the user profile [34]. Of course, there are many privacy issues related to this approach, but still it seems clear that a reduced version of the user's PHR, which carefully excludes the most sensible information, would be very useful to define the user's profile of a health recommender system.

All approaches discussed so far, collaborative, content based, and knowledge based, can be used together with information from personal health records to provide useful recommendations, with the limitations mentioned in Subheading 5. Sometimes hybrid models are used together with the personal health record as part of the user profile (*see* Fig. 4). Furthermore, mobile recommender systems are becoming more widespread as users increase the access to health information through mobile systems [36].

A way to exploit recommendation based on the personal health record is through semantic networks like Wikipedia [34]. A semantic network is a network that represents relations between concepts as a knowledge representation. Normally such a network is

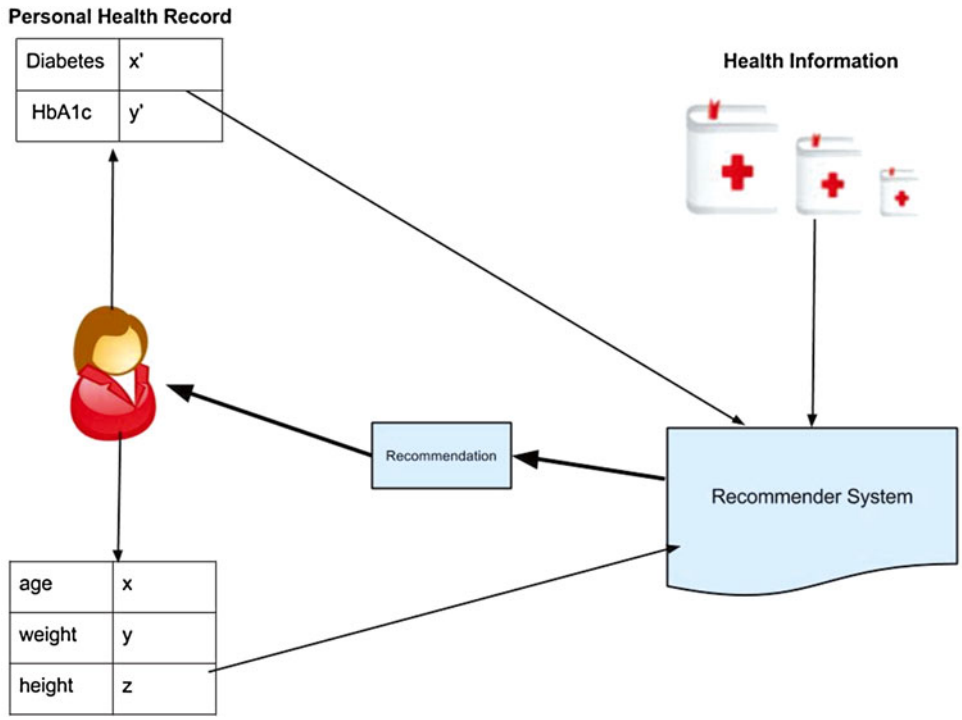

Fig. 4 Health recommender system

represented by a graph where each node represents the concept (subject) and the edges represent the relations (verb) [37].

Figure 5 represents an example of a semantic network. Nodes contain the health concept (subject and object) and the edges associated with these nodes represent their relations. This example, extracted from JMIR [38], includes a brief representation on SQL tables. The resulting recommender system uses the distances among nodes as its main criterion.

One recent concept related to the collaborative nature of recommended systems was introduced by Eysenbach, who named it "apomediation," a "socio-technological term" that identifies "trustworthy and credible information and services" [39]. Apomediation refers to the process of guiding users to information or services through different "agents" that are not clearly identified as intermediaries; instead, they may be other users and/or automatic systems (recommenders) that filter the information collaboratively. As a result, the credibility and trustworthiness cannot be directly guaranteed. The recommender systems need to include a way to ensure trusted health information, fitting the information according to the user needs, which is the main challenge for health recommender systems.

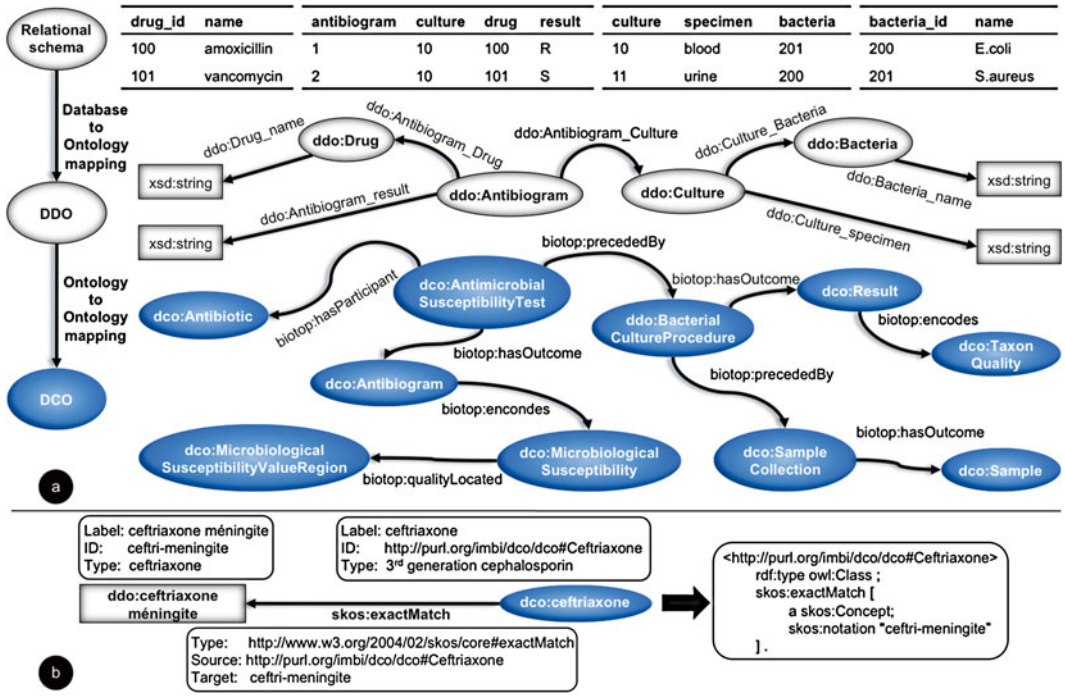

Fig. 5 Example of semantic network. *Source*: JMIR (Example of a semantic network: the hybrid ontology-driven interoperability mapping model. JMIR. http://www.jmir.org/2012/3/e73/) [38]

6.2 Social Media and Recommender Systems

Health professionals and patients are taking advantage of social media, using the Internet as a tool to find and share health information. They can benefit from using the Internet to connect with each other, and to provide other helpful insights, but it is essential to be aware of possibly unwanted consequences.

One way to avoid these effects is to take advantage of health educational resources. Recommender systems and personalized health education work together to filter the overwhelming information and focus on the actual user needs [5]. The explanation-driven approach to a personalized content recommendation for patients is an example of this thread [40]. The information is accompanied with detailed explanations that educate patients to make informed medical decisions.

6.3 Sentiment Analysis

Machine learning increases the potential of recommender systems; the system learns with the interaction of the users and adapts its decisions according to previous choices. The understanding increases with the recommender process.

Machine learning is effective in sentiment classifications [41]. Sentiment analysis obtains subjective information from the context. People are continuously communicating their feelings; we can

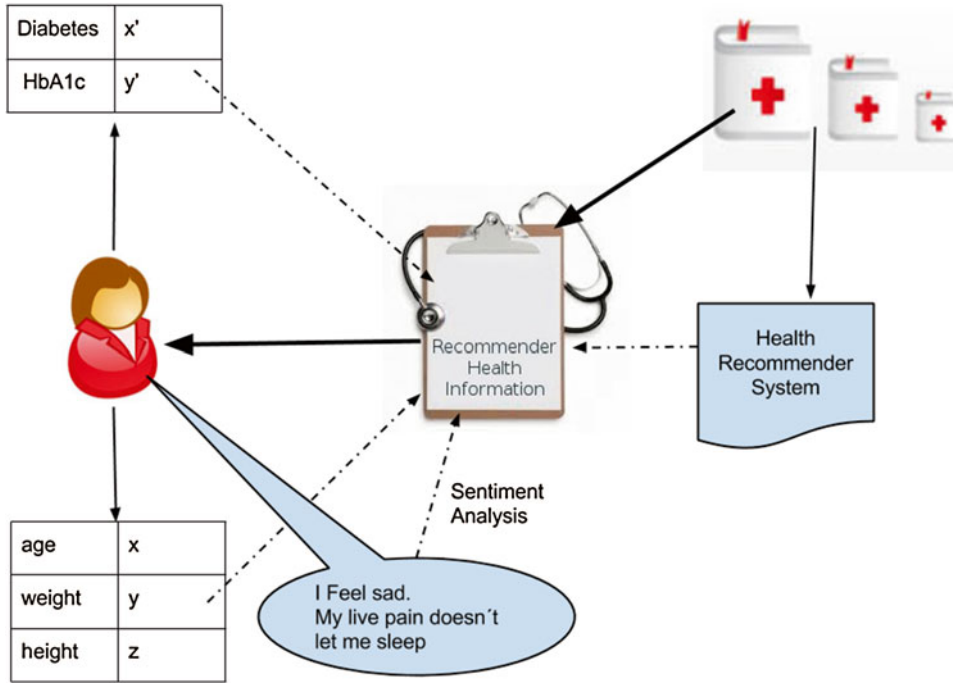

Fig. 6 Sentiment analysis and health recommender system

for instance look for any verbs expressing sentiments (to feel, to want, to love) or nouns (pain, loss, feelings) and try to adapt the recommendation following the sentiment context (*see* Fig. 6).

The treatment of opinions, sentiments, and subjective texts is one of the main research interest of current recommender systems [42]. Feelings are one of the implicit values to be included in content-based systems. Social networks (like Facebook, Twitter, Youtube) are examples where these messages can be retrieved to be processed on the health recommender system. Moreover there are other health dedicated channels that contain sentiment messages (like Patientlikeme,[1] Crochane,[2] Webicina[3]). Information from the abovementioned social media can be retrieved and linked with trusted health sources. This process increases the possibilities in recommenders.

[1] Web Patientlikeme http://www.patientslikeme.com/
[2] Web Chrocane http://www.cochrane.org/
[3] Web Webicina http://www.webicina.com/

References

1. Fox S, Jones S (2009) The social life of health information. Pew Internet & American Life Project, Washington, DC, 2009-12
2. Fernandez-Luque L, Karlsen R, Melton GB (2012) Healthtrust: a social network approach for retrieving online health videos. J Med Internet Res 14(1):e22
3. Cline RJW, Haynes KM (2001) Consumer health information seeking on the Internet: the state of the art. Health Educ Res 16(6): 671–692
4. Khan SA, McFarlane DJ, Li J, Ancker JS, Hutchinson C, Cohall A et al (2007) Healthy Harlem: empowering health consumers through social networking, tailoring and web 2.0 technologies. AMIA Annu Symp Proc 2007 Oct 11:1007
5. Fernandez-Luque L, Karlsen R, Vognild LK (2009) Challenges and opportunities of using recommender systems for personalized health education. MIE, pp 903–907
6. Pattaraintakorn P, Zaverucha GM, Cercone N (2007) Web based health recommender system using rough sets, survival analysis and rule-based expert systems. In: An A, Stefanowski J, Ramanna S, Butz CJ, Pedrycz W, Wang G (eds). Rough sets, fuzzy sets, data mining and granular computing [Internet]. Berlin: Springer. Available from: http://link.springer.com/chapter/10.1007/978-3-540-72530-5_59
7. Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 17(6):734–749
8. Zanker M, Felfernig A, Friedrich G (2011) Recommender systems: an introduction. Cambridge University Press, New York, NY
9. Sarwar B, Karypis G, Konstan J, Riedl J (2000) Application of dimensionality reduction in recommender system - a case study. Proc. ACM workshop on web mining for e-commerce--challenges and opportunities -WebKDD, Boston MA. ACM, New York, NY
10. Melville P, Sindhwani V (2010) Recommender systems. In: Sammut C, Webb G (eds) Encyclopedia of machine learning. Springer, New York, NY
11. Symeonidis P, Nanopoulos A, Papadopoulos AN, Manolopoulos Y (2008) Collaborative recommender systems: combining effectiveness and efficiency. Exp Syst Appl 34(4): 2995–3013
12. Linden G, Smith B, York J (2003) Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Computing 7(1): 76–80
13. Wang P, Ye H (2009) A personalized recommendation algorithm combining slope one scheme and user based collaborative filtering. International Conference on Industrial and Information Systems, 2009 IIS'09, pp 152–154
14. Lops P, Gemmis M de, Semeraro G (2011) Content-based recommender systems: state of the art and trends. In: Ricci F, Rokach L, Shapira B, Kantor PB (eds). Recommender systems handbook [Internet]. New York, NY: Springer. Available from: http://link.springer.com/chapter/10.1007/978-0-387-85820-3_3
15. Suchal J, Návrat P (2010) Full text search engine as scalable k-nearest neighbor recommendation system. In: Bramer M (ed). Artificial intelligence in theory and practice III [Internet], Berlin: Springer, pp 165–73. Cited September 26 2013. Available from: http://link.springer.com/chapter/10.1007/978-3-642-15286-3_16
16. Pazzani MJ, Billsus D (2007) Content-based recommendation systems. In: Brusilovsky P, Kobsa A, Nejdl W (eds). The adaptive web [Internet]. Berlin: Springer, pp 325–341. Cited September 26 2013. Available from: http://link.springer.com/chapter/10.1007/978-3-540-72079-9_10
17. Lewis DD, Schapire RE, Callan JP, Papka R (1996) Training algorithms for linear text classifiers. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval [Internet], New York, NY: ACM, pp 298–306. Cited September 26 2013. Available from: http://doi.acm.org/10.1145/243199.243277
18. Smyth B (2007) Case-based recommendation. In: Brusilovsky P, Kobsa A, Nejdl W (eds) The adaptive web, vol 4321, LNCS. Springer, Berlin, pp 342–376
19. Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. Computer 42(8):30–37
20. Landauer TK, Foltz PW, Laham D (1998) An introduction to latent semantic analysis. Disc Process 25(2–3):259–284
21. Hofmann T (1999) Probabilistic latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval [Internet]. New York, NY: ACM, pp 50–57. Available from: http://doi.acm.org/10.1145/312624.312649
22. Dumais ST (1991) Improving the retrieval of information from external sources. Behav Res Methods Instrum Comput 23(2):229–236

23. Hipp J, Güntzer U, Nakhaeizadeh G (2000) Algorithms for association rule mining—a general survey and comparison. SIGKDD Explor Newsl 2(1):58–64
24. Sarawagi S, Thomas S, Agrawal R (1998) Integrating association rule mining with relational database systems: alternatives and implications. Proceedings of the 1998 ACM SIGMOD international conference on Management of data [Internet]. New York, NY: ACM, pp 343–354. Available from: http://doi.acm.org/10.1145/276304.276335
25. Carroll JJ, Dickinson I, Dollin C, Reynolds D, Seaborne A, Wilkinson K (2004) Jena: implementing the semantic web recommendations. Proceedings of the 13th international World Wide Web conference on alternate track papers & posters [Internet]. New York, NY: ACM, pp 74–83. Available from: http://doi.acm.org/10.1145/1013367.1013381
26. Tomanek K, Wermter J, Hahn U (2007) Efficient annotation with the Jena ANnotation Environment (JANE). Association for computational linguistics, pp 9–16. Available from: http://dl.acm.org/citation.cfm?id=1642059.1642061
27. Rector AL, Qamar R, Marley T (2009) Binding ontologies and coding systems to electronic health records and messages. Appl Ontol 4(1):51–69
28. Bedi P, Kaur H, Marwaha S (2007) Trust based recommender system for the semantic web. Proceedings of the 20th international joint conference on artificial intelligence [Internet]. San Francisco, CA: Morgan Kaufmann Publishers Inc., pp 2677–2682. Cited November 14 2013. Available from: http://dl.acm.org/citation.cfm?id=1625275.1625706
29. Peis E, Morales-del-Castillo JM, Delgado-López JA (2008) Analysis of the state of the topic [Internet]. "Hipertext.net", num. 6
30. Burke R (2002) Hybrid recommender systems: survey and experiments. User Model User Adap Inter 12(4):331–370
31. Iaquinta L, de Gemmis M, Lops P, Semeraro G, Filannino M, Molino P (2008) Introducing serendipity in a content-based recommender system. Eighth international conference on hybrid intelligent systems, 2008 HIS'08, pp 168–173
32. Sollenborn M, Funk P (2002) Category-based filtering and user stereotype cases to reduce the latency problem in recommender systems. In: Craw S, Preece A (eds). Advances in case-based reasoning [Internet]. Berlin: Springer, pp 395–405. Cited October 17 2013. Available from: http://link.springer.com/chapter/10.1007/3-540-46119-1_29
33. Pew Internet & American Life Project (2013) Health topics—use of internet users look for health information online. http://pewinternet.org/Reports/2013/Health-online.aspx
34. Wiesner M, Pfeifer D (2010) Adapting recommender systems to the requirements of personal health record systems. Proceedings of the 1st ACM international health informatics symposium [Internet]. New York, NY: ACM, pp 410–414. Cited October 25 2013. Available from: http://doi.acm.org/10.1145/1882992.1883053
35. Kimmel Z, Greenes RA, Liederman E (2004) Personal health records. J Med Pract Manage 21(3):147–152
36. Ricci F (2010) Mobile recommender systems. Inf Tech Tour 12(3):205–231
37. Spitzer M, Braun U, Hermle L, Maier S (1993) Associative semantic network dysfunction in thought-disordered schizophrenic patients: direct evidence from indirect semantic priming. Biol Psychiatry 34(12):864–877
38. Teodoro D, Pasche E, Gobeill J, Emonet S, Ruch P, Lovis C (2012) Building a transnational biosurveillance network using semantic web technologies: requirements, design, and preliminary evaluation. J Med Internet Res 14(3):e73
39. Eysenbach G (2008) Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. J Med Internet Res 10(3):e22
40. Roitman H, Messika Y, Tsimerman Y, Maman Y (2010) Increasing patient safety using explanation-driven personalized content recommendation. Proceedings of the 1st ACM international health informatics symposium [Internet]. New York, NY: ACM, pp 430–434. Cited September 22 2013. Available from: http://doi.acm.org/10.1145/1882992.1883057
41. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol 10 [Internet]. Stroudsburg, PA: Association for Computational Linguistics, pp 79–86. Cited October 29 2013. Available from: http://dx.doi.org/10.3115/1118693.1118704
42. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1–2):1–135

Chapter 10

Cloud Computing for Context-Aware Enhanced m-Health Services

Carlos Fernandez-Llatas, Salvatore F. Pileggi, Gema Ibañez, Zoe Valero, and Pilar Sala

Abstract

m-Health services are increasing its presence in our lives due to the high penetration of new smartphone devices. This new scenario proposes new challenges in terms of information accessibility that require new paradigms which enable the new applications to access the data in a continuous and ubiquitous way, ensuring the privacy required depending on the kind of data accessed. This paper proposes an architecture based on cloud computing paradigms in order to empower new m-Health applications to enrich their results by providing secure access to user data.

Key words Cloud computing, m-Health, Smartphone, Electronic Health Record, Personal Health Record

1 Introduction

The deep penetration of smartphones in current society has exponentially increased the presence of applications that are used on mobile devices. The existing applications available in mobility are not only focused on leisure and entertainment but also, health management and m-Health applications [1] has recently increased dramatically its presence in smartphone application market. As a proof of that, the offer for m-Health applications in smartphones has increased a 700% in 2011 [2]. In this scenario, m-Health applications demand more storage capabilities, higher reliability, scalability, sustainable QoS in order to be competitive. Therefore, the user need to have the information at any point to take the right decision wherever he/she is.

The access to health information is a problem that has been deeply discussed in the literature [3, 4]. Since the early 1990s, the need for a common representation of medical records in order to provide access from anywhere to health centers and professional

Carlos Fernández-Llatas and Juan Miguel García-Gómez (eds.), *Data Mining in Clinical Medicine*, Methods in Molecular Biology, vol. 1246, DOI 10.1007/978-1-4939-1985-7_10, © Springer Science+Business Media New York 2015

was detected. As a result of that, the creation of an Electronic Health Record (EHR) [3, 4] is a common goal not only in European Union countries but also at a world scale.

However, the appearance of new personal devices and the new role of the patient as a continuous data provider [5, 6] offer new opportunities to health applications in order to build more personalized applications and approach more wide environments that usually take into account the current legacy Health Systems. Those applications need information that is not available in EHR systems. For that, it is needed to find new information sources for those systems. Personal Health Records (PHR) are called to fill this gap.

PHR [7] is an electronic record intended to provide patients a complete and accurate summary of their clinical staff, adding important personal parameters for the care of their health and well-being, such as nutritional aspects, family history, and others, involving more people in their care.

In addition to provide the data, it is needed to provide a tool to allow a deployment of this data making that information available continuously and ubiquitously. Cloud Computing [8] is a powerful framework that provide infrastructure to store and deploy PHRs and EHRs through the internet.

PHR as well as EHR has important legal restrictions. While EHR only can be accessed by Authorized Health professionals, PHR is owned by the patient and only him/her should have the power to enable the availability of that information to third-party applications outside the public health authorities. For that, the access to this information must be authorized and authenticated. In this way, it is a key point the fact that the deployed platform was role-based, allowing the patient to take the total control of the applications, as concerns sensible information of patients.

In this paper, an architecture that uses a cloud computing based architecture for providing a combined and secured access to EHR, PHR, and contextual information for third-party applications is presented. The paper is structured as follows. First in Related Work section a state of art on EHR and PHR existing works is made. After that, a Health information model that summarizes the author's vision of the combination of data that should be available for m-health smartphone applications is presented. Then, the Cloud based architecture is proposed, and finally the paper is concluded.

2 Related Work

The ICT are more and more present in Health Technologies. The basis of the creation of e-Health technologies is to allow an access to the data available about the patients. EHRs are the former solution available called to deal with this. EHR is formally defined as

the systematic collection of electronic health information about individual patients or populations [9]. In literature, there are lots of studies about EHR available. These studies are models that should cover that technologies in different fields [4]; the definition of different standards of EHR [10] for the unification of different existing models; the creation of systems to interoperate among different EHR [11]; fact standards defined by autonomous governments [12] and studies about the practical deployment of EHR [13], and the impact of the use of EHR by the Health professionals [3].

As it is defined, the EHR store all relevant data, in order to know the health status of the patient in real time, but only from the medical point of view. Nevertheless, with the penetration of m-Health technologies, it is more and more common that the health application not only made use of data from the medical point of view, but also they use information about the personal actions of the patients (Activity, Diet, etc.) and the Context information (Weather, temperature, etc.).

Context information can be usually gathered from data available on smart spaces or through internet, but the gathering of personal health data is a more complex problem. In this way, a new approach health record where health data and information related to the care of a patient are entered by himself is needed [7]. This approach is called Personal Health Record (PHR). There are available emerging commercial technologies for PHR population. HealthVault [14] is a Microsoft initiative that allows to connect monitoring devices and to develop services related to the data. Other example is INDIVO [15] which is an extension of Personal Health Folder. It incorporates health data from different sources and it is free and open. Moreover, it has been designed to be improved and personalized by the users. For example, they can connect their records to a third-party application which improves the management and analysis of the health data.

In order to provide access to those storage services, it is needed an infrastructure that facilitate a scalable, ubiquitous, and continuously accessible way to deploy them. Cloud computing [8] is a technology though to approach this problem.

Cloud computing describes a novel consumption and delivery model for IT services enabling dynamic and scalable environments for the sharing of virtual resources. Cloud computing is Web-based processing, whereby shared resources, software, and information are provided to computers and other devices on demand over the Internet. Details are abstracted from the users that are supported by the technology infrastructure “in the cloud.” The term “cloud” is used as a metaphor for the Internet, based on the cloud drawing used in the past to represent the telephone network, and later to depict the Internet in computer network diagrams as an abstraction of the underlying infrastructure it represents. Most cloud

computing infrastructures consist of services delivered through common centers and built on servers. Clouds often appear as single points of access for consumers' computing needs. The last generation of Cloud Technology allows users to feel remote resource and software remotely running as part of its own computation resource. Cloud vision enables several innovative business scenarios that assume customers do not own the physical infrastructure. Cloud Computing has become a scalable services consumption and delivery platform in the field of Services Computing.

In [16] some unsolved problems of cloud computing in health are explored. Recent researches that has been focused on use PHR in cloud computing [17, 18] do not consider the inherent problems of security issues from a role based computing point of view.

3 Context-Aware Health Record

The common concepts of EHR and PHR can be extended by using context information. This integration significantly extends the potential application field of the records since it provides rich information about the record owner behavior. Context information can be managed by end-user applications in order to provide extended and improved capabilities in terms of data analysis.

3.1 CA-HR Modules

Merging EHR and PHR, as well as the context-aware understanding of the health records, advise a modular structure for the resulting record (CA-HR). CA-HR implicitly defines a semantic understanding of the context meaning. The context provides stand-alone information as well as links to EHR and PHR.

Figure 1 shows the model proposed for the CA-HR. In this model, there are different data repositories that are responsible of different kind of information. Each one of the repositories could offer access to different modules depending on the scope of the application. The three repositories share a common semantic layer in order to provide a common way of the access. A common semantic framework ensures the extensibility of the model allowing the incorporation of new modules to fulfill the future new needs.

In that framework, m-Health applications are supposed that can work independently of the permission gathered. Nevertheless, the more access the application have, the more personalized the result is. For example, supposing a diet manager as third-party application, it can propose general recipes to a user if it does not has permission to access to any part of the CA-HR model. If the application have permission to access the context, the recipes could be refined according to the ingredients available at the smart space (i.e., Patient's Home). If the patient has enabled the access to the PHR for the application, the application can use the activity per-

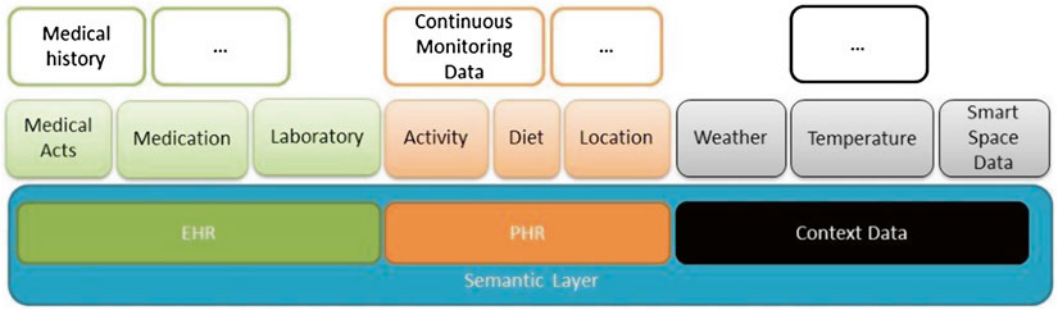

Fig. 1 CA-HR model

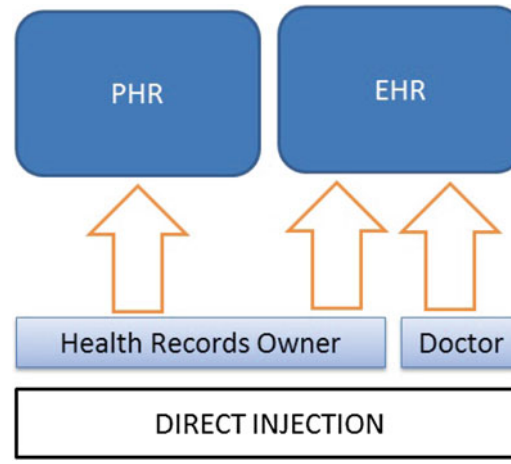

Fig. 2 Direct data injection

formed by the user in order to calculate the amount of user calories wasted in order to build the most adequate recipe for each day. Finally, if, in addition, the application has access to EHR, the recipes can be selected avoiding the ingredients that user is not able to ingest due to allergy problems.

An extended analysis of the model is out of paper scope.

3.2 Context-Aware Data Injection

In the common meaning, the data injection in the PHRs is an action always directly driven by the owner. The data injection in EHRs can be driven by the owner or explicitly authorized stakeholders (e.g., medical doctors). The normal action on health records as previously described (Fig. 2) is referred as direct data injection.

The CA-HR enables a further interaction mode (Fig. 3) in which an explicitly authorized application can inject data in PHR and/or EHR both with the related context data. This is the typical case in which the records owner is part of some kind of ecosystem

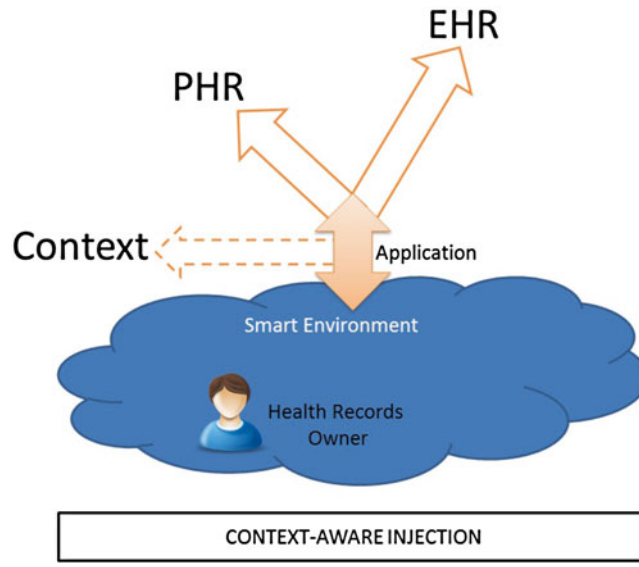

Fig. 3 Context-aware data injection

that assumes the existence of active smart environments able to monitorize and detect personal and/or environmental situations.

4 A Cloud Platform for Context-Aware Enhanced m-Health Services

4.1 Overview

The Cloud Platform model is shown in Fig. 4. An exhaustive overview at the architecture design (as well as a detailed analysis of functional layers) is out of the paper scope. A short description of the main functional actors composing the platform is provided. The proposed model is composed of four main components:

- **CA-HR Virtual Repository:** CA-HR is understood as a shared resource among heterogeneous third-party applications. Regardless by the physical storage location (the different modules could be distributed, as well as the context information), the platform assumes a virtual repository that makes CA-HR records available for the third-party applications.
- **Enablers:** They have the critical role of allowing the access to the modules information. In practical case, a refinement of the enablers could be required: the information related to a single module could not be atomic and so the enabler could allow the access to a part of the fields, denying the access to other fields. A concrete combination of the enablers defines, at the same time, the rights of a concrete application on the information, the role of the application user, as well as the privacy contract between the CA-HR owner and the application.

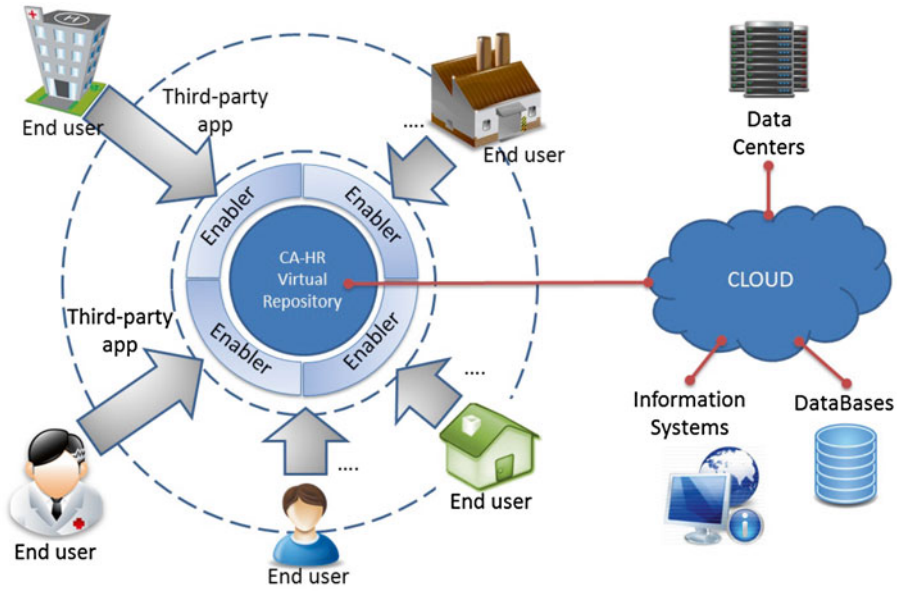

Fig. 4 Cloud platform

- **Third-Party Applications:** Applications that explicitly sign a rights contract (enabled by the Enablers) with the end-user and so they are allowed of using and/or managing CA-HRs information (or part of it).
- **Application Users:** Users of third-party applications. In practice they are the final users.

4.2 Cloud Approach

In order to allow an effective and efficient use of shared resources (CA-HRs) in a context of economic competitiveness and sustainability, the platform is implicitly referring to cloud infrastructures. This approach also provides a significant support in order to assure law restrictions about the privacy are respected, as well as a high security. The interaction with CA-HRs has different phases or modes (data collection, data delivery, data analysis). For each of these modes, the cloud approach plays a fundamental role:

- Data collection can happen as a direct data injection or as a context-aware data injection (as described in the previous sections). In both cases, cloud infrastructures allow a virtual understanding of the information: distributed information sources are available in the platform as a unique virtual repository. Furthermore, sensor devices (as well as any other kind of active actor able to associate observations or measurements to the record owner) just need an authorized gateway (e.g., a mobile device) to load new information or updates.

- Data delivery is driven by Enablers. Enablers are for instance a centralized concept that can be effectively applied to distributed environments by using virtualized resources. This virtual view is implemented on large scale by using cloud technologies.
- Data analysis could be a critical issue for the privacy and other related issues. The cloud approach fully supports "code injection" and similar techniques able to assure that third-party applications just access the output data and not the intermediate information produced in the process. Furthermore, in the case of complex analysis, additional resources for data processing in the cloud could be integrated allowing the execution of final applications in smart devices.

4.3 Role-Based Computation

The modular and extensible features of CA-HRs both with the dynamism assured by the Enablers provide a powerful environment for role-based computation. For example, a hospital could have, for instance, basic rights on the CA-HRs. These rights could be automatically extended to be full rights if the situation requires it (an urgent situation for example). This dynamic switching of rights has to be carefully monitored and logged in order to assure the privacy of the owner is respected according to the law restrictions. The cloud approach, by this point of view, enables a centralized view of the activity in the platform and on the platform. Role-based computation facilitates both the application deployment and the platform management.

5 Conclusions

The use of cloud computing technologies to ensure m-Health applications a ubiquitous and continuous access to users data can be the solution to new need in this research field. Nevertheless, to ensure the privacy of that information allowing the user to have the total control over their data a role-based computation approach should be deployed. In this paper, a reference architecture for defining CA-HR model that involvers EHR, PHR, and context data is presented proposing a unified, scalable, and extensible solution for a role-based access to user data ensuring their privacy.

References

1. Istepanian RSH, Pattichis CS (2006) M-health: emerging mobile health systems. Birkhäuser, Boston
2. Research2Guidance. Mobile Market Health Report 2011 http://www.research2guidance.com/the-market-for-mhealth-application-reached-us-718-million-in-2011/
3. Poissant L, Pereira J, Tamblyn R, Kawasumi Y (2005) The impact of electronic health records on time efficiency of physicians and nurses: a systematic review. J Am Med Inform Assoc 12(5):505–516
4. DesRoches CM, Campbell EG, Rao SR, Donelan K, Ferris TG, Jha A, Kaushal R, Levy DE, Rosenbaum S, Shields AE, Blumenthal D (2008) Electronic health records in ambulatory care—a national survey of physicians. N Engl J Med 359(1):50–60

5. Merilahti J, Pärkkä J, Antila K, Paavilainen P, Mattila E, Malm E-J, Saarinen A, Korhonen I (2009) Compliance and technical feasibility of long-term health monitoring with wearable and ambient technologies. J Telemed Telecare 15(6):302–309
6. Rashvand HF, Salcedo VT, Sanchez EM, Iliescu D (2008) Ubiquitous wireless telemedicine. IET Commun 2(2):237–254
7. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ (2006) Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc 13(2):121–126
8. Buyya R, Yeo CS, Venugopal S, Broberg J, Brandic I (2009) Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility. Future Gener Comput Syst 25(6):599–616
9. Gunter TD, Terry NP (2005) The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions. J Med Internet Res 2005; 7(1):e3
10. Eichelberg M, Aden T, Riesmeier J, Dogac A, Laleci GB (2005) A survey and analysis of Electronic Healthcare Record standards. ACM Comput Surv 37(4):277–315
11. Maldonado JA, Moner D, Boscá D, Fernández-Breis JT, Angulo C, Robles M (2009) LinkEHR-Ed: a multi-reference model archetype editor based on formal semantics. Int J Med Inform 78(8):559–570
12. Cerdà-Calafat I, Continente-Gonzalo M, García-López C, Guanyabens-Calvet J (2010) Personal health folder. Medicina Clínica 134(Suppl 1):63–66
13. Maldonado JA, Costa CM, Moner D, Menárguez-Tortosa M, Boscá D, Giménez JAM, Fernández-Breis JT, Robles M (2012) Using the ResearchEHR platform to facilitate the practical application of the EHR standards. J Biomed Inform 45(4):746–762
14. Microsoft Health Vault (2014). http://www.healthvault.com/
15. IndivoHealth (2012) The personally controlled health record. Last access: 28th March 2012. http://indivohealth.org/
16. Klein CA (2011) Cloudy confidentiality: clinical and legal implications of cloud computing in health care. J Am Acad Psychiatry Law Online 39(4):571–578
17. Kikuchi S, Sachdeva S, Bhalla S (2012) Applying cloud computing model in PHR architecture. In: Proceedings of the 2012 joint international conference on human-centered computer environments, HCCE '12. ACM, New York, pp 236–237
18. 2010 Ayia Napa MobiHealth. 2 (2011) Wireless mobile communication and healthcare: second international ICST conference; revised selected papers. Springer, Berlin

Part III

New Applications of Data Mining in Clinical Medicine Problems

Chapter 11

Analysis of Speech-Based Measures for Detecting and Monitoring Alzheimer's Disease

A. Khodabakhsh and C. Demiroglu

Abstract

Automatic diagnosis of the Alzheimer's disease as well as monitoring of the diagnosed patients can make significant economic impact on societies. We investigated an automatic diagnosis approach through the use of speech based features. As opposed to standard tests, spontaneous conversations are carried and recorded with the subjects. Speech features could discriminate between healthy people and the patients with high reliability. Although the patients were in later stages of Alzheimer's disease, results indicate the potential of speech-based automated solutions for Alzheimer's disease diagnosis. Moreover, the data collection process employed here can be done inexpensively by call center agents in a real-life application. Thus, the investigated techniques hold the potential to significantly reduce the financial burden on governments and Alzheimer's patients.

Key words Alzheimer's disease, Speech analysis, Support vector machines

1 Introduction

Alzheimer's disease (AD) is becoming more widespread with the aging population in the developed countries. Thus, it is a significant economic burden on the governments as well as the patients and their families. Simplifying the healthcare processes and reducing the costs through the use of technology for this disease can make a significant economic impact.

Diagnosis of the disease is not easy. Even in the later stages, recognition or evaluation of the disease by clinicians fail 50 % of the time [1]. Moreover, even if the disease is diagnosed correctly, monitoring the progression of the disease by a clinician over time is costly. Thus, patients cannot visit the clinicians frequently and what happens between the visits is largely unknown to clinicians.

Telephone-based automated measures for detection and/or monitoring of the disease can be a low-cost solution to the diagnosis problem. Patients who do not feel comfortable visiting a doctor, or cannot afford a doctor visit, can do private self-tests.

Carlos Fernández-Llatas and Juan Miguel García-Gómez (eds.), *Data Mining in Clinical Medicine*, Methods in Molecular Biology, vol. 1246, DOI 10.1007/978-1-4939-1985-7_11,

Moreover, diagnosed patients can be monitored frequently by the system with minimal cost and convenience for the patients.

Typically, clinicians use tests such as Mini-Mental State Examination (MMSE) and linguistic memory tests. Linguistic memory tests are based on the recall rates of word lists and narratives and they are typically more effective than the MMSE tests. None of those typical practices, however, consider the speech signal in diagnosing the disease. Moreover, they are hard to do over the telephone line because of problems in user-interface design, the need for high accuracy speech recognition systems which still are not good enough to meet the demands of such an application. Furthermore, both patients and elderly people often fail to use such sophisticated technology.

Analysis of speech signal has been considered for Alzheimer's detection in [2, 3]. However, in those works, speech signal is recorded during the standard clinical tests. Moreover, most of the focus is on the spoken language itself, which requires manual transcription, rather than the speech signal. A more limited study with one patient and a focus on the prosodic features of speech, which determines stress, intonation, and emotion, is reported in [4]. Problems of speech production that are related to central nervous system problems are also noted in [5]. Speech-based features are investigated in [6] to detect frontotemporal lobar degeneration with promising results.

Here, we propose a system where a spontaneous conversation is carried with the patient. Thus, contents of the conversations are not predetermined. The goal of this approach is to keep the subject comfortable during speech without constraining the conversation. Moreover, lack of structure is appealing to the subjects since this requires minimal effort in terms of cognition. That way, subject's speech can be recorded in the most natural and effortless way by a person with minimal technical or clinical skills. For example, the conversation can be carried by a person at a call center and recorded automatically which is substantially lower-cost compared to a hospital visit. Such conversational data has been investigated in [7, 8]; however only linguistic features are analyzed and speech features are not considered in [7, 8]. Similarly, conversational data has been investigated in [6, 9] for linguistic and speech dysfluency features by measuring the correlation of those features with the disease and without an attempt to do diagnosis using them. Correlation of linguistic capability with the Alzheimer's disease was also shown in [10].

Data that is used in this study has been collected in nursing homes in Istanbul. 13 speech features are extracted automatically from the recordings and disease detection is done with linear discriminant analysis (LDA), support vector machines (SVMs), and decision tree classifiers. Proposed speech features, especially using the SVM algorithm, seem to be particularly good at distinguishing between healthy and sick people.

2 Materials

In this research, conversational speech recordings of 27 patients (17 male, 10 female) with late stage Alzheimer's disease, and 27 healthy elderly people (12 male, 15 female) have been used. All subjects were native speakers of Turkish. The age range is between 60 and 80 in both healthy and elderly subjects. For each subject, approximately 10 min of conversation have been recorded using a high-quality microphone with 16 kHz sampling rate. Subjects were directed casual/conversational questions which are not fixed between different patients. The data has been collected in health-care facilities at Istanbul and then hand-labeled to split question and response parts. The labels are later refined using an automated voice activity detector (VAD) to accurately mark the beginning and end times of phonations in the conversations. Patients refused to use noise-cancelling microphones which require installation on their clothes. Therefore, there is some background noise in the recordings but the signal-to-noise ratio is high and speech features could be extracted reliably.

3 Methods

3.1 Voice Activity Detection (VAD)

Because of the amount and nature of background noise in the recorded files, finding a robust VAD was an important task. The VAD used here is based on the distribution of the short-time frame energy of the speech signal. Because there is both silence and speech in the recordings, energy distribution has two modes both of which can be modeled with Gaussian distributions. Thus, based on the energy of the speech frames, they are classified as either speech or silence (Fig. 1).

3.2 Speech Features

The features that are used to detect Alzheimer's disease are extracted from the conversational speech recordings. A total of 13 features have been extracted and evaluated for detecting Alzheimer's disease. A description of the speech features are given below.

Voice Activity-Related Features

Silence and speech segments are labeled in each recording for feature extraction as discussed in Subheading 3.1. Using the voice activity information, following features are extracted from each recording:

(A) *Logarithm of the response time*: When the interviewer asks a question, it takes some time before the subject gives an answer. It is hypothesized that this time can be an indicator of the disease since it is expected to be related to the cognitive processes such as attention and memory. Logarithm of the average time it takes the subject to answer a question is calculated in each recording to extract this feature.

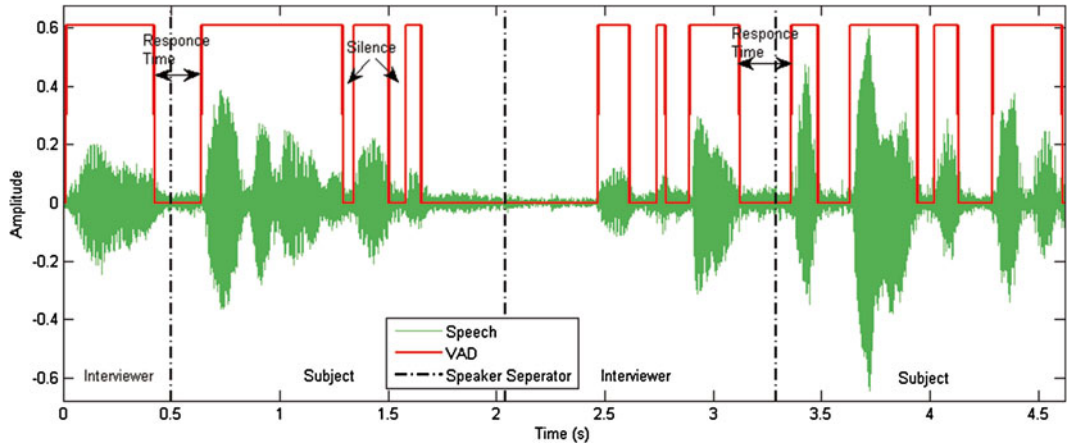

Fig. 1 Sample speech waveform from the interviews and the voice activity detector (VAD) output is shown. Response time indicates the amount of time it takes the subject to answer a question. Silence segments indicate the level of noise in the signal

(B) *Silence-to-speech ratio*: Ratio of the total silence time over the total amount of speech is a measure of the hesitations during speech.

(C) *Logarithm of continuous speech*: Duration of each continuous speech segment indicates how long the subject can talk without pausing.

(D) *Logarithm of continuous silence*: Duration of each continuous silence segment indicates how long the subject pauses in the middle of the speech.

(E) *Logarithm of pauses per second*: Besides the length of hesitations during speech, frequency of them can also be an indicator of the disease. Logarithm of the average number of pauses per seconds is used in this feature.

(F) *Average absolute delta energy*: Similar to pitch, energy variations also convey information about the mood of the subject. Changing speech energy significantly during speech may indicate a conscious effort to stress words that are semantically important or changes of mood related to the content of the speech. Average of the energy changes are used in this feature.

Articulation-Related Features

The voice activity related features discussed above are related cognitive thought processes. However, it is also important to measure how the subject uses his/her voice articulators during speech. For example, if the subject gets too emotional, significant changes in the fundamental frequency (pitch) can be expected. Similarly, changes in the resonant frequencies (formants) of speech can be a

strong indicator of the subject's health. If the formants do not change fast enough or are not distinct enough, sounds may become harder to identify which can indicate mumbling in speech. To see the effects of these in classification of the disease, pitch and formant trajectories are extracted and following features are derived or each recording:

(A) *Average absolute delta pitch*: Average of absolute delta pitch indicates the rate of variations in pitch. This feature is has high correlation with the communication of emotions through the speech signal.

(B) *Average absolute delta formants*: Average of absolute delta formant frequencies indicates the rate of change in the formant features. Formants are related to the positions of the vocal organs such as tongue, lips etc. Reduction of control over these organs because of a damage in the brain, such as AD, can create speech impairments such as mumbling. In this case, formants do not change quickly and speech becomes less intelligible.

(C) *Logarithm of voicing ratio*: Another speech impairment is the loss of voicing in speech which results in smokers' voice. In this case, the subject loses the ability to vibrate the vocal cords which results in breathy and noisy speech. Average duration of voiced speech is compared with the unvoiced speech to detect any potential impairment in the vocal cords due to AD.

(D) *Logarithm of voicing per second*: This feature measures the ratio of voiced speech to unvoiced speech per second.

(E) *Logarithm of the mean-duration of continuous voiced speech*: Besides the ratio of voiced and unvoiced speech, for how long the subject maintains voicing without pausing is measured (Fig. 2).

Rate of Speech-Related Features

Using an automatic Turkish phoneme recognizer trained with a Turkish broadcast speech database, recordings are transcribed into phonemes. Following features are extracted using the phonemic transcriptions:

(A) *Phonemes per second*: number of phonemes generated per second is used to represent the rate of speech of the subject.

(B) *Logarithm of variance of phoneme frequency*: 42 phonemes are recognized by the automatic system in this study. However, distribution of the recognized phonemes change from subject to subject. Higher variance in phoneme frequency distribution may indicate clarity in speech and may be correlated with using a bigger dictionary during speech. Because the dictionary size and speech clarity is related to cognition and articulatory organs, this feature may be useful in detecting AD.

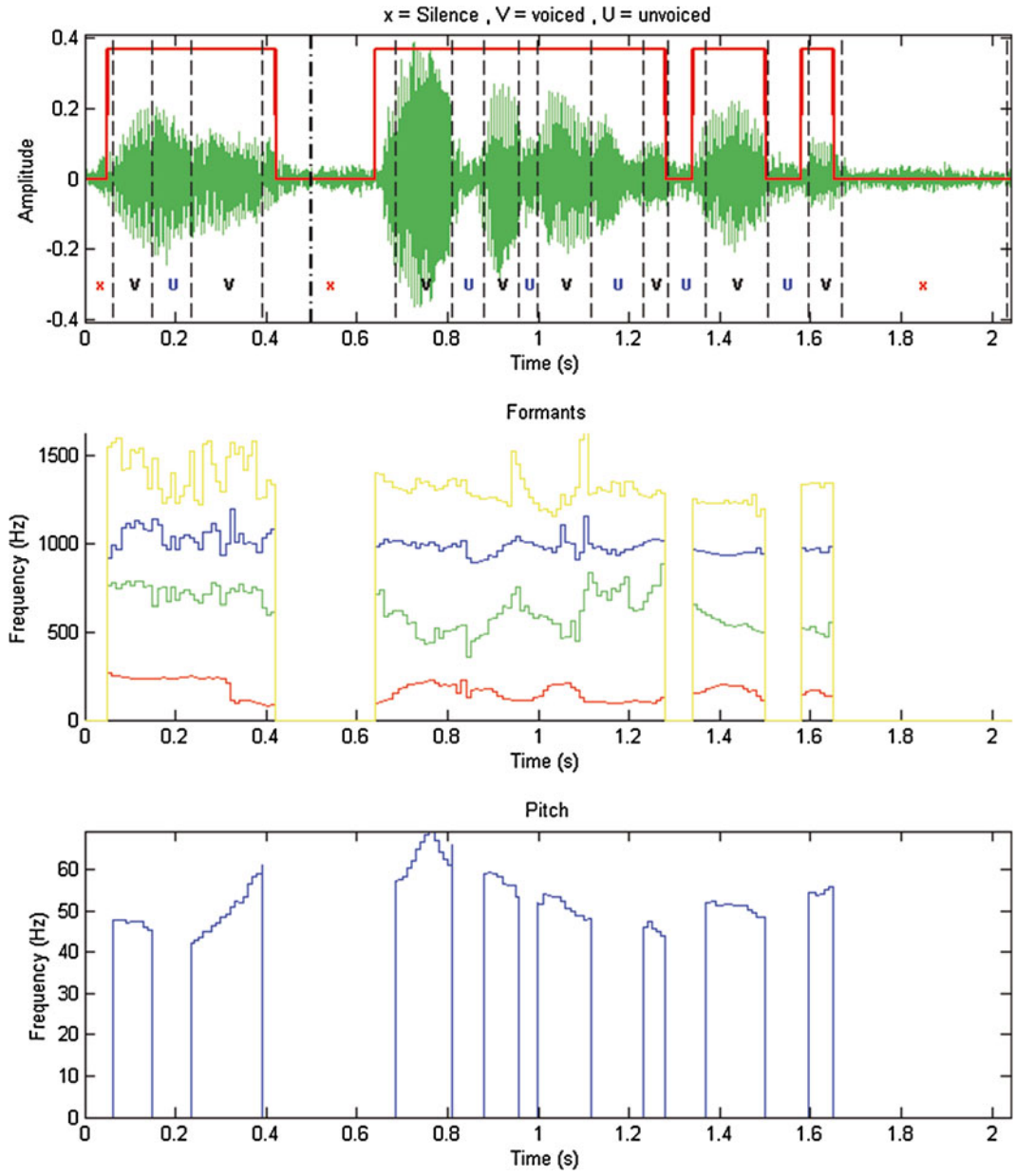

Fig. 2 Output of the voicing detector is shown in the *top* figure. The *middle* figure shows the formant tracks for the first three formants extracted from the same speech sample. The *bottom* figure shows the output of the pitch extractor

3.3 Evaluation Methods

For classification of patients and healthy subjects, three classifiers are used: linear discriminant analysis (LDA), support vector machines (SVM), and decision trees. Quadratic kernel is used in the SVM classifier.

In the first phase of testing, each feature is tested separately to assess the classification power of individual features. Then, combinations of features are used to increase the classification power of the algorithms. Increasing the number of features used in classification, brute force search is done using all possible combinations of features. Feature sets that give the best results for each feature number are found using such brute force approach.

Because there is limited number of subjects in the test, leave-one-out strategy is used where one of the subjects is left out and the classifier is trained with the rest of the subjects. Then, testing is done on the left-out subject. All subjects are tested using this strategy. 95 % confidence intervals of the classification are scores are also computed.

4 Notes

1. Classification performance of the individual features is shown in Table 1. Performance results are reported in terms of detection rate (probability of detecting the disease in a patient), false alarm (probability of diagnosing the disease in a healthy subject), and the total accuracy (deciding between healthy and ill subjects correctly). 95 % confidence intervals are also shown in Table 1. Individual features are not particularly strong at diagnosing the disease. Highest performance is obtained with the logarithm of voicing ratio, average absolute delta feature of the first formant, and average absolute delta pitch feature. Logarithm of voicing ratio is lower in the patients compared to healthy subjects as shown in Fig. 3. Similarly, average absolute delta formant feature is lower in the patients as shown in Fig. 4. Lack of voicing and smaller variations in the formants may indicate a correlation between ability/desire to control articulatory organs and the disease. In contrast, average absolute delta pitch feature has a higher variance in the patients as shown in Fig 5. That indicates more variations of emotion and emphasis during speech in the patients' speech.
2. After testing the predictive power of each individual feature, combinations of features are used to improve the classification performance. All possible pairs of the 13 features are tested with the three classifiers. Highest scoring features are shown in Table 2. Using feature pairs improved the performance of LDA and SVM algorithms. However, performance of the decision tree algorithm is degraded with more features. Decision tree could not generate enough leaf nodes given the fixed set of features and therefore could not perform well in this task.
3. Similar to single feature case, SVM algorithm outperformed the LDA and decision tree algorithms for the feature pairs.

Table 1
Performance of each feature individually using different classification algorithms sorted by SVM performance

	LDA			SVM quadratic			Decision Tree		
Feature	**Total**	**Detection**	**False alarm**	**Total**	**Detection**	**false alarm**	**Total**	**Detection**	**False alarm**
Logarithm of voicing ratio	69.8 57:82	55.6 37:74	15.4 2:29	71.7 60:84	51.9 33:71	7.7 0:18	71.7 60:84	74.1 58:91	30.8 13:49
Average absolute delta formant(1)	69.8 57:82	63 45:81	23.1 7:39	69.8 57:82	74.1 58:91	34.6 16:53	73.6 62:85	74.1 58:91	26.9 10:44
Average absolute delta pitch	60.4 47:74	40.7 22:59	19.2 4:34	69.8 57:82	48.1 29:67	7.7 0:18	49.1 36:63	55.6 37:74	57.7 39:77
Logarithm of continuous silence	64.2 51:77	51.9 33:71	23.1 7:39	67.9 55:80	51.9 33:71	15.4 2:29	54.7 41:68	48.1 29:67	38.5 20:57
Logarithm of response time	62.3 49:75	70.4 53:88	46.2 27:65	67.9 55:80	85.2 72:99	50 31:69	60.4 47:74	55.6 37:74	34.6 16:53
Logarithm of pauses per second	66 53:79	51.9 33:71	19.2 4:34	66 53:79	44.4 26:63	11.5 0:24	45.3 32:59	48.1 29:67	57.7 39:77
Logarithm of continuous speech	67.9 55:80	59.3 41:78	23.1 7:39	62.3 49:75	63 45:81	38.5 20:57	60.4 47:74	59.3 41:78	38.5 20:57
Phonemes per second	58.5 45:72	44.4 26:63	26.9 10:44	62.3 49:75	40.7 22:59	15.4 2:29	50.9 37:64	59.3 41:78	57.7 39:77
Logarithm of the mean duration of continuous voiced speech	50.9 37:64	48.1 29:67	46.2 27:65	60.4 47:74	48.1 29:67	26.9 10:44	62.3 49:75	63 45:81	38.5 20:57
Average absolute delta energy	62.3 49:75	29.6 12:47	3.8 0:11	56.6 43:70	18.5 4:33	3.8 0:11	52.8 39:66	55.6 37:74	50 31:69
Logarithm of variance of phoneme frequency	49.1 36:63	33.3 16:51	34.6 16:53	56.6 43:70	25.9 9:42	11.5 0:24	50.9 37:64	44.4 26:63	42.3 23:61
Silence-to-speech ratio	49.1 36:63	33.3 16:51	34.6 16:53	54.7 41:68	33.3 16:51	23.1 7:39	37.7 25:51	37 19:55	61.5 43:80
Logarithm of voicing per second	45.3 32:59	33.3 16:51	42.3 23:61	49.1 36:63	14.8 1:28	15.4 2:29	64.2 51:77	66.7 49:84	38.5 20:57

In each cell, mean value is shown on the top and the lower and upper limits of confidence is shown on the bottom

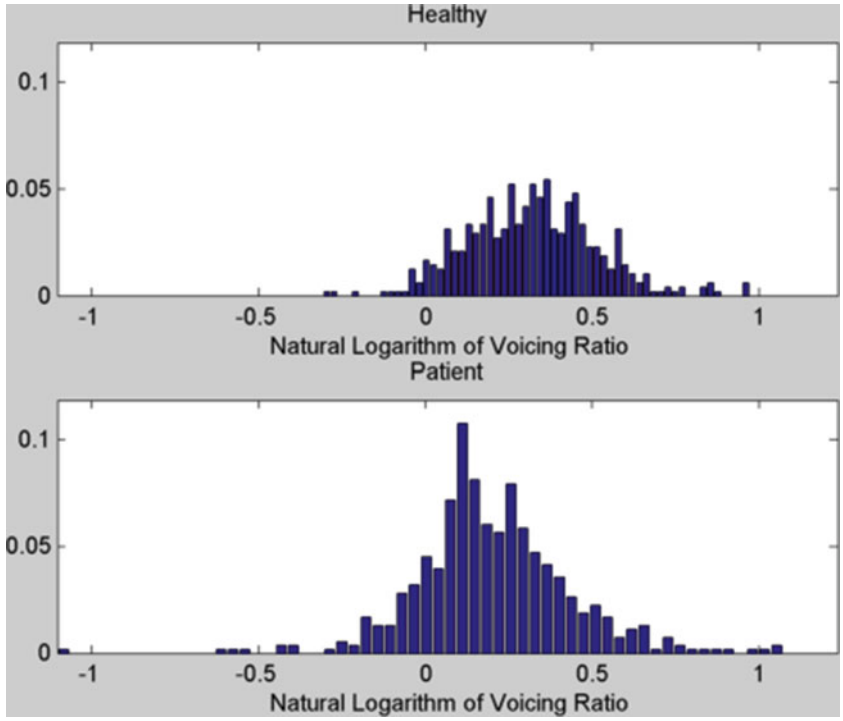

Fig. 3 Logarithm of voicing ratio distribution for healthy subjects and patients

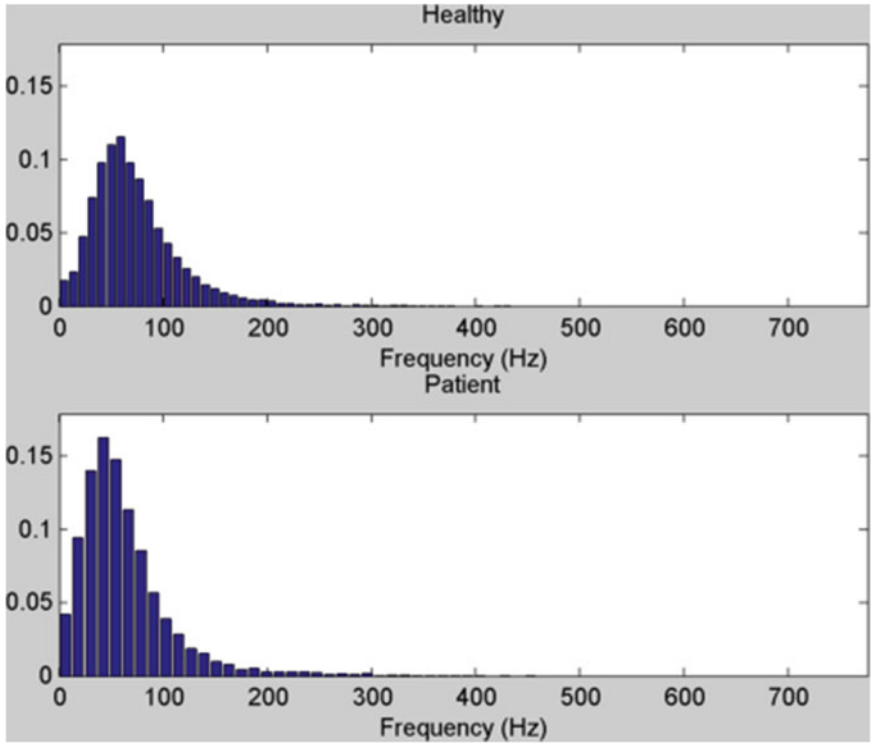

Fig. 4 Average absolute delta formant distribution for healthy subjects and patients

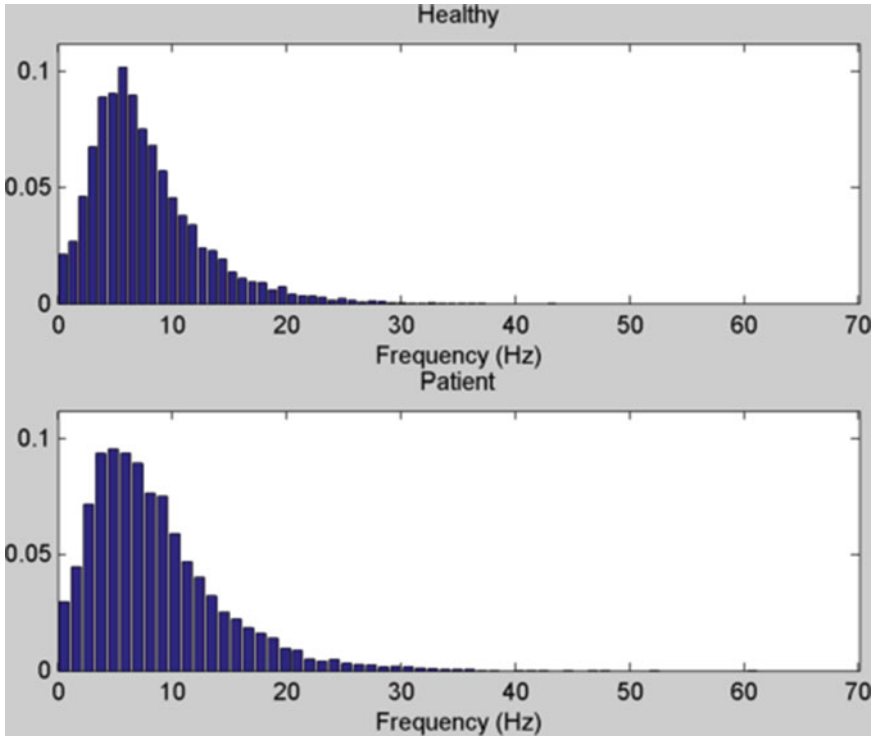

Fig. 5 Average absolute delta pitch distribution

Table 2
Scores of paired features using different classification algorithms. Systems are sorted by SVM performance

Features	LDA			SVM quadratic			Decision Tree		
	Total	Detection	False alarm	Total	Detection	False alarm	Total	Detection	False alarm
Average absolute delta pitch+ Logarithm of voicing ratio	75.5 64:87	66.7 49:84	15.4 2:29	79.2 68:90	74.1 58:91	15.4 2:29	67.9 55:80	74.1 58:91	38.5 20:57
Logarithm of continuous silence+ Average absolute delta formant(1)	69.8 57:82	66.7 49:84	26.9 10:44	75.5 64:87	77.8 62:93	26.9 10:44	58.5 45:72	59.3 41:78	42.3 23:61
Logarithm of continuous speech+ Logarithm of voicing ratio	75.5 64:87	70.4 53:88	19.2 4:34	75.5 64:87	70.4 53:88	19.2 4:34	67.9 55:80	70.4 53:88	34.6 16:53

(continued)

Table 2
(continued)

Features	LDA			SVM quadratic			Decision Tree		
	Total	Detection	False alarm	Total	Detection	False alarm	Total	Detection	False alarm
Logarithm of pauses per second+ Average absolute delta formant(1)	77.4 66:89	77.8 62:93	23.1 7:39	75.5 64:87	74.1 58:91	23.1 7:39	73.6 62:85	74.1 58:91	26.9 10:44
Logarithm of pauses per second + Logarithm of voicing ratio	73.6 62:85	66.7 49:84	19.2 4:34	75.5 64:87	63 45:81	11.5 0:24	62.3 49:75	74.1 58:91	50 31:69
Average absolute delta formant(1)+ Logarithm of voicing ratio	77.4 66:89	81.5 67:96	26.9 10:44	75.5 64:87	77.8 62:93	26.9 10:44	62.3 49:75	70.4 53:88	46.2 27:65
Logarithm of variance of phoneme frequency+ Logarithm of pauses per second	64.2 51:77	55.6 37:74	26.9 10:44	73.6 62:85	55.6 37:74	7.7 0:18	52.8 39:66	48.1 29:67	42.3 23:61
Phonemes per second+ Average absolute delta pitch	71.7 60:84	63 45:81	19.2 4:34	71.7 60:84	59.3 41:78	15.4 2:29	54.7 41:68	63 45:81	53.8 35:73
Logarithm of continuous speech+ Average absolute delta formant(1)	71.7 60:84	66.7 49:84	23.1 7:39	71.7 60:84	77.8 62:93	34.6 16:53	67.9 55:80	66.7 49:84	30.8 13:49
Logarithm of pauses per second+ Average absolute delta pitch	67.9 55:80	55.6 37:74	19.2 4:34	71.7 60:84	55.6 37:74	11.5 0:24	67.9 55:80	63 45:81	26.9 10:44
Logarithm of pauses per second+ Logarithm of the mean duration of continuous voiced speech	73.6 62:85	70.4 53:88	23.1 7:39	71.7 60:84	66.7 49:84	23.1 7:39	62.3 49:75	55.6 37:74	30.8 13:49

Combinations with scores less than 70 % are not shown. In each cell, mean value is shown on the top and the lower and upper limits of confidence is shown on the bottom

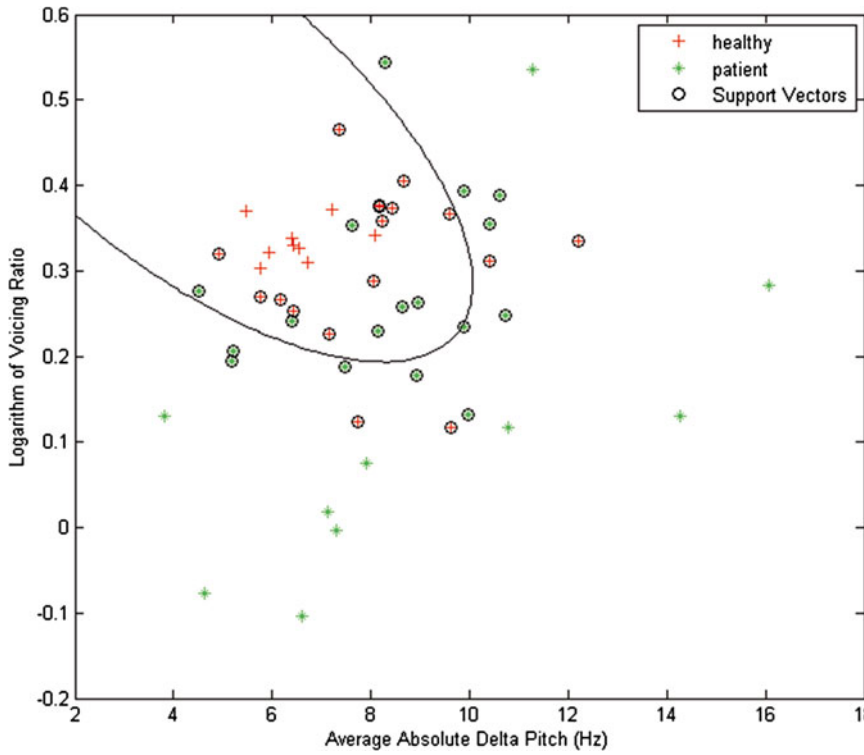

Fig. 6 Scatter plot of the logarithm of voicing rate and average absolute delta pitch for the patients and healthy subjects. Quadratic decision surface is found by the support vector machine. Support vectors are *circled*

Moreover, combination of the average absolute delta pitch and logarithm of voicing ratio feature has the best performance. In Fig. 6, those two features are plot together with the decision surface and the support vectors found by the SVM algorithm. Higher pitch variation and lower voicing ratio can clearly separate the data into two classes which can be classified with a quadratic decision surface in most cases. Moreover, the data points that are misclassified are close to the decision surface which indicates that those can also be easily recovered if more features are available.

4. All possible combinations of three features are also evaluated and the best performing 3-tuples are shown in Table 3. Performance of the best performing systems with increasing number of features is shown in Table 4. Performance of all systems increase with the number of features until a point where overfitting occurs and classification performance gets lower with more features. Average accuracy of the SVM system goes up to 94 % for the 7-feature case with confidence interval having a lower-bound of 88 %. Thus, in the worst case, the SVM system can achieve a total accuracy of 88 % which is quite high.

Table 3
Scores of combinations of three features using different classification algorithms. Scores are sorted by the SVM performance

	LDA			SVM quadratic			Decision Tree		
Features	**Total**	**Detection**	**False alarm**	**Total**	**Detection**	**False alarm**	**Total**	**Detection**	**False alarm**
Phonemes per second+ Logarithm of response time+ Average absolute delta formant(1)	71.7 60:84	77.8 62:93	34.6 16:53	83 73:93	88.9 77:100	23.1 7:39	73.6 62:85	77.8 62:93	30.8 13:49
Logarithm of continuous speech+ Logarithm of pauses per second+ Logarithm of voicing ratio	75.5 64:87	81.5 67:96	30.8 13:49	83 73:93	74.1 58:91	7.7 0:18	62.3 49:75	70.4 53:88	46.2 27:65
Logarithm of continuous silence+ logarithm of continuous speech+ Average absolute delta formant(1)	67.9 55:80	66.7 49:84	30.8 13:49	81.1 71:92	85.2 72:99	23.1 7:39	50.9 37:64	51.9 33:71	50 31:69
Logarithm of continuous silence+ logarithm of pauses per second+ Average absolute delta formant(1)	71.7 60:84	70.4 53:88	26.9 10:44	81.1 71:92	88.9 77:100	26.9 10:44	62.3 49:75	59.3 41:78	34.6 16:53
Logarithm of response time+ Average absolute delta pitch+ Logarithm of voicing ratio	79.2 68:90	77.8 62:93	19.2 4:34	79.2 68:90	70.4 53:88	11.5 0:24	67.9 55:80	59.3 41:78	23.1 7:39
Silence-to-speech ratio+ Logarithm of continuous speech+ Average absolute delta formant(1)	73.6 62:85	74.1 58:91	26.9 10:44	79.2 68:90	88.9 77:100	30.8 13:49	66 53:79	66.7 49:84	34.6 16:53

(continued)

Table 3
(continued)

Features	LDA			SVM quadratic			Decision Tree		
	Total	Detection	False alarm	Total	Detection	False alarm	Total	Detection	False alarm
Logarithm of continuous speech+ Average absolute delta energy+ Logarithm of voicing ratio	71.7 60:84	59.3 41:78	15.4 2:29	79.2 68:90	81.5 67:96	23.1 7:39	50.9 37:64	55.6 37:74	53.8 35:73
Logarithm of pauses per second+ Average absolute delta pitch+ Average absolute delta formant(1)	71.7 60:84	70.4 53:88	26.9 10:44	79.2 68:90	81.5 67:96	23.1 7:39	69.8 57:82	70.4 53:88	30.8 13:49
Logarithm of pauses per second+ Average absolute delta formant(1)+ Average absolute delta energy	67.9 55:80	51.9 33:71	15.4 2:29	79.2 68:90	81.5 67:96	23.1 7:39	67.9 55:80	70.4 53:88	34.6 16:53
Logarithm of pauses per second+ Average absolute delta formant(1)+ Logarithm of voicing ratio	75.5 64:87	70.4 53:88	19.2 4:34	79.2 68:90	77.8 62:93	19.2 4:34	56.6 43:70	70.4 53:88	57.7 39:77
Logarithm of pauses per second+ Logarithm of voicing ratio+ Logarithm of the mean duration of continuous voiced speech	75.5 64:87	74.1 58:91	23.1 7:39	79.2 68:90	66.7 49:84	7.7 0:18	58.5 45:72	70.4 53:88	53.8 35:73

Systems with scores less than 80 % are not shown. In each cell, mean value is shown on the top and the lower and upper limits of confidence is shown on the bottom

Table 4
Scores with combinations of different number of features using LDA, SVM, and decision tree classifiers

Feature		4	5	6	7	8	9	10
LDA	Total	83	81.1	81.1	83	84.9	84.9	84.9
		73:93	71:92	71:92	73:93	75:95	75:95	75:95
	Detection	77.8	74.1	77.8	85.2	81.5	85.2	85.2
		62:93	58:91	62:93	72:99	67:96	72:99	72:99
	False alarm	11.5	11.5	15.4	19.2	11.5	15.4	15.4
		0:24	0:24	2:29	4:34	0:24	2:29	2:29
SVM quadratic	Total	86.8	88.7	92.5	94.3	90.6	88.7	84.9
		78:96	80:97	85:100	88:100	83:98	80:97	75:95
	Detection	92.6	88.9	96.3	92.6	92.6	88.9	85.2
		83:100	77:100	89:100	83:100	83:100	77:100	72:99
	False alarm	19.2	11.5	11.5	3.8	11.5	11.5	15.4
		4:34	0:24	0:24	0:11	0:24	0:24	2:29
Decision tree	Total	84.9	84.9	83	81.1	81.1	79.2	75.5
		75:95	75:95	73:93	71:92	71:92	68:90	64:87
	Detection	85.2	85.2	85.2	88.9	85.2	85.2	81.5
		72:99	72:99	72:99	77:100	72:99	72:99	67:96
	False alarm	15.4	15.4	19.2	26.9	23.1	26.9	30.8
		2:29	2:29	4:34	10:44	7:39	10:44	13:49

For each number of features, best performing features are found using a brute force search method. In each cell, mean value is shown on the top and the lower and upper limits of confidence is shown on the bottom

References

1. Roark B, Mitchell M, Hosom J, Hollingshead K, Kaye J (2011) Spoken language derived measures for detecting mild cognitive impairment. IEEE Trans Audio Speech Lang Process 19(7):2081–2090
2. Hoffmann I, Nemeth D, Dye CD, Pákáski M, Irinyi T, Kálmán J (2010) Temporal parameters of spontaneous speech in Alzheimer's disease. Int J Speech Lang Pathol 12:29–34
3. Leea H, Gayraudb F, Hirsha F, Barkat-Defradas M (2011), Speech dysfluencies in normal and pathological aging: a comparison between Alzheimer patients and healthy elderly subjects, 17th International Conference on Phonetic Sciences, Hong Kong, Aug 2011
4. Bucks RS, Singh S, Cuerden JM, Wilcock GK (2000) Analysis of spontaneous, conversational speech in dementia of Alzheimer type: evaluation of an objective technique for analysing lexical performance. Aphasiology 14: 71–91
5. Thomas C, Cercone N (2005) Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In Proc of IEEE ICMA, 2005.
6. Roark B, Hosom J, Mitchell M, Kaye J (2007) Automatically derived spoken language markers for detecting mild cognitive impairment. In Proc 2nd Int Conf Technol Aging (ICTA), 2007
7. Boise L, Neal M, Kaye J (2004) Dementia assessment in primary care: results from a study in three managed care systems. J Gerontol 59(6):M621–M626
8. Snowdon D, Kemper S, Mortimer J, Greiner L, Wekstein D, Markesbery W (1996) Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. Findings from the Nun Study. J Am Med Assoc 275(7):528–532
9. Tosto G, Gasparini M, Lenzi GL, Bruno G (2011) Prosodic impairment in Alzheimer's disease: assessment and clinical relevance. J Neuropsychiatry Clin Neurosci 23:E21–E23
10. Vassiliki I, Stergios K (2003) Clinical psychoacoustics in Alzheimer's disease central auditory processing disorders and speech deterioration. Annal Gen Hosp Psychiat 2:12

Chapter 12

Applying Data Mining for the Analysis of Breast Cancer Data

Der-Ming Liou and Wei-Pin Chang

Abstract

Data mining, also known as Knowledge-Discovery in Databases (KDD), is the process of automatically searching large volumes of data for patterns. For instance, a clinical pattern might indicate a female who have diabetes or hypertension are easier suffered from stroke for 5 years in a future. Then, a physician can learn valuable knowledge from the data mining processes. Here, we present a study focused on the investigation of the application of artificial intelligence and data mining techniques to the prediction models of breast cancer. The artificial neural network, decision tree, logistic regression, and genetic algorithm were used for the comparative studies and the accuracy and positive predictive value of each algorithm were used as the evaluation indicators. 699 records acquired from the breast cancer patients at the University of Wisconsin, nine predictor variables, and one outcome variable were incorporated for the data analysis followed by the tenfold cross-validation. The results revealed that the accuracies of logistic regression model were 0.9434 (sensitivity 0.9716 and specificity 0.9482), the decision tree model 0.9434 (sensitivity 0.9615, specificity 0.9105), the neural network model 0.9502 (sensitivity 0.9628, specificity 0.9273), and the genetic algorithm model 0.9878 (sensitivity 1, specificity 0.9802). The accuracy of the genetic algorithm was significantly higher than the average predicted accuracy of 0.9612. The predicted outcome of the logistic regression model was higher than that of the neural network model but no significant difference was observed. The average predicted accuracy of the decision tree model was 0.9435 which was the lowest of all four predictive models. The standard deviation of the tenfold cross-validation was rather unreliable. This study indicated that the genetic algorithm model yielded better results than other data mining models for the analysis of the data of breast cancer patients in terms of the overall accuracy of the patient classification, the expression and complexity of the classification rule. The results showed that the genetic algorithm described in the present study was able to produce accurate results in the classification of breast cancer data and the classification rule identified was more acceptable and comprehensible.

Key words Genetic algorithm, Data mining, Breast cancer, Rule discovery

1 Introduction

Breast cancer is a major cause of concern in the USA today. For instance, it affects one in every seven women in the USA [1]. At a rate of nearly one in three cancers diagnosed, breast cancer is the most frequently diagnosed cancer in women. The American Cancer

Carlos Fernández-Llatas and Juan Miguel García-Gómez (eds.), *Data Mining in Clinical Medicine*, Methods in Molecular Biology, vol. 1246, DOI 10.1007/978-1-4939-1985-7_12, © Springer Science+Business Media New York 2015

Society projected that 211,300 invasive and 55,700 in situ cases would be diagnosed in 2003 [2]. Furthermore, breast cancer the second leading cause of death for women in the USA, and is the leading cause of cancer deaths among women aged 40–59. According to the American Cancer Society 39,800 breast cancer-related deaths are expected in 2003 [3]. Even though in the last couple of decades, with increased emphasis towards cancer related research, new and innovative methods for early detection and treatment have been developed, which helped decrease the caner-related death rate [4], cancer in general and breast cancer in specific is still a major cause of concern in the USA.

The mammography is the traditional method for breast cancer diagnosis. However, the radiologists show considerable variability in how they interpret a mammogram [5]. Moreover, Elmore [6] indicated that 90 % of radiologists recognized fewer than 3 % of cancers and 10 % recognized about 25 % of the cases. The fine needle aspiration cytology is another approach adopted for the diagnosis of breast cancer with more precise prediction accuracy. However, the average correct identification rate is around 90 % [7]. Generally, the purpose of all the related research is identical to distinguish between patients with breast cancer in the malignant group and patients without breast cancer in the benign group. So, the breast cancer diagnostic problems are in the scope of binary classification problems [5, 8, 9]. In spite that some literature shows the artificial intelligence approaches can be successfully applied in order to predict the breast cancer [5, 10].

Since the early 1960s, researchers involved in artificial intelligence have developed the predictive models for various areas of science including chemistry, engineering, finance, and medicine [11, 12]. The medical predictive models are designed to aid physicians to overcome health problems, routine classification tasks which should otherwise be referred to the specialist in a particular area of medicine. These models are built from a "data warehouse," which constitutes the data acquired from actual cases. The data can be preprocessed and expressed in a set of rules, such as is often the case in the knowledge-based expert systems, or serve as the training data for statistical and medical learning models. However, the effort required to retrieve relevant knowledge from the databases has increased significantly. As a consequence, there has been a growing interest in data mining which is capable of facilitating the discovery of interesting and useful knowledge from a large database [13].

Classification is one of the major tasks in the data mining field. Among the available options in the data mining field, the most popular models in medicine are logistic regression (LR), artificial neural networks (ANNs), and decision tree. However, since the medical domain classification problem is highly nonlinear in nature, it is difficult to develop a comprehensive model to take into account all the independent variables using conventional statistical modeling

techniques. Furthermore, traditional ad hoc mixtures of statistical techniques and data management tools are no longer adequate for analyzing the vast collection of data [14].

Recently, biologically inspired genetic algorithms have also been used successfully to learn the concepts described by the attribute-value pairs. Genetic algorithms (GA) are search algorithms developed from the principles of natural selection and natural genetics [15, 16]. The biological background makes genetic algorithms robust, i.e., capable of a good performance in a variety of environments [15]. Genetic algorithms can be introduced easily into new problem domains, due to their operation requiring only a very small amount of problem-specific knowledge [17]. A drawback of the domain independence is that a genetic algorithm sometimes achieves only a near optimal performance level; however, this problem can be tackled by exploiting the problem knowledge [17]. Moreover, genetic algorithms can be readily comprehensively and noise tolerant, which is an important characteristic for a search algorithm when real-world problems are solved. Genetic algorithms have been applied to a considerable number of hard problems in different problem domains [15, 18]. In medicine, genetic algorithms have been utilized, for example, to model immune systems [19] and to identify individuals who are at risk of coronary artery disease [20].

Data mining finds patterns and relationships in data by using sophisticated techniques to build models—abstract representations of reality. A good model is a useful guide to understanding your domain and making decisions. However, data mining applies many older computational techniques from statistics, machine learning and pattern recognition [21, 22].

There are two main kinds of models in data mining: predictive and descriptive. Predictive models can be used to forecast explicit values, based on patterns determined from known results. For example, from a database of customers who have already responded to a particular offer, a model can be built that predicts which prospects are likeliest to respond to the same offer. Descriptive models describe patterns in existing data, and are generally used to create meaningful subgroups such as demographic clusters [23]. This knowledge discovery process has several distinct steps or sub-processes that begin with data gathering, followed by data cleaning, then aggregation, and integration. At this point, the data is ready to be utilized for data visualization and finally data mining. Rather than being sequential, sub-processes in the data mining process are iterative i.e. movement from data visualization back to data cleaning if irregularities are discovered in the data set [24–26].

We applied various data mining model, including decision tree, logistic regression, neural network, and genetic algorithm, to establish the prediction model. It is the combination of the serious

Table 1
Prediction variables used in breast cancer classification modeling

1. Clump thickness	6. Bare nuclei
2. Uniformity of cell size	7. Bland chromatin
3. Uniformity of cell shape	8. Normal nucleoli
4. Marginal adhesion	9. Mitoses
5. Single epithelial cell size	10. Class (benign/malignant)

effects of breast cancer, the promising results for prior related research, the potential benefits of the research outcomes and the desire to further understand the nature of breast cancer that provided the motivation for this research effort.

2 Materials

To verify the feasibility and effectiveness of the proposed GA-based modeling approach for predicting breast cancer, a data set containing 699 patient's records has been used. The dataset we used is provided by the University of Wisconsin Hospitals, Madison from Wolberg and also can be obtained from the website (ftp://ftp.ics.uci.edu/pub/machine-learning-databases/breast-cancer-wisconsin). These diagnostic results of each patient's record in above dataset consist of ten variables that are summarized in Table 1. Each instance consists of nine measurements without considering the patient number, namely clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. The measurements are assigned an integer value between 1 and 10, with 1 being the closest to benign and 10 the most anaplastic. One of the ten variables is the response variable representing the diagnostic status of the patient with or without breast cancers (i.e., malignant or benign). The training data are selected from the whole dataset randomly and directly fed into the proposed mining approach.

In the original dataset, the sample patient number has been excluded in this study. In the 699 patients' records of the dataset, there are 241 patients (34.5 %) with breast cancers (malignant) and the remaining 458 patients (65.5 %) without breast cancers (benign).

3 Methods

We adopt attribute selective procedures to reduce unnecessary properties, making the model construction more efficient and to figure out the simplified classification rules.

3.1 Attribute Selection with Information Gain Ranking

This is one of the simplest (and fastest) attribute ranking methods and is often used in text categorization applications where the sheer dimensionality of the data precludes more sophisticated attribute selection techniques. If A is an attribute and C is the class, Eqs. 1 and 2 give the entropy of the class before and after observing the attribute:

$$H(C) = -\sum_{c \in C} p(c) \log_2 p(c), \tag{1}$$

$$H(C \mid A) = -\sum_{a \in A} p(a) \sum_{c \in C} p(c \mid a) \log_2 p(c \mid a). \tag{2}$$

The amount by which the entropy of the class decreases reflects the additional information about the class provided by the attribute and is called information gain. Each attribute A_i is assigned a score based on the information gain between itself and the class:

$$\begin{aligned} IG_i &= H(C) - H(C \mid A_i) \\ &= H(A_i) - H(A_i \mid C) \\ &= H(A_i) + H(C) - H(A_i, C) \end{aligned} .$$

This method was used to evaluate the importance of each variable that predict the breast cancer model. Information gain increases with the average purity of the subsets that an attribute produces. We will depend on the importance of variables to increase or decrease attributes with artificial method. Next step is to train and to test the network.

3.2 Design GA Model

Before the extracting the decision rule(s), the most significant predictors of the best subset have to be decided; otherwise, the insignificant predictors become the noise which may worsen the genetic learning or even mislead the In addition to the best subset of the significant predictors considered in this study, multiple rules are explored for increasing the prediction accuracy. Unlike those approaches used in literature [27–29], the proposed approach is not to select the first n best fit rules within a genetic searching process as the mined rules for building the prediction model. As mentioned previously, those best n rules are the rules not converged to be the best one. In our proposed approach, additional new decision rule is to be explored when the previous rule(s) failed to classify all the sample data correctly. In other words, if one rule is generated but not with good enough prediction accuracy, an additional new rule is going to be extracted using those data which cannot be classified correctly. The computation procedures of the proposed rule mining approach contain two major processes, which are data preprocess, GA mining process, respectively (*see* Fig. 1).

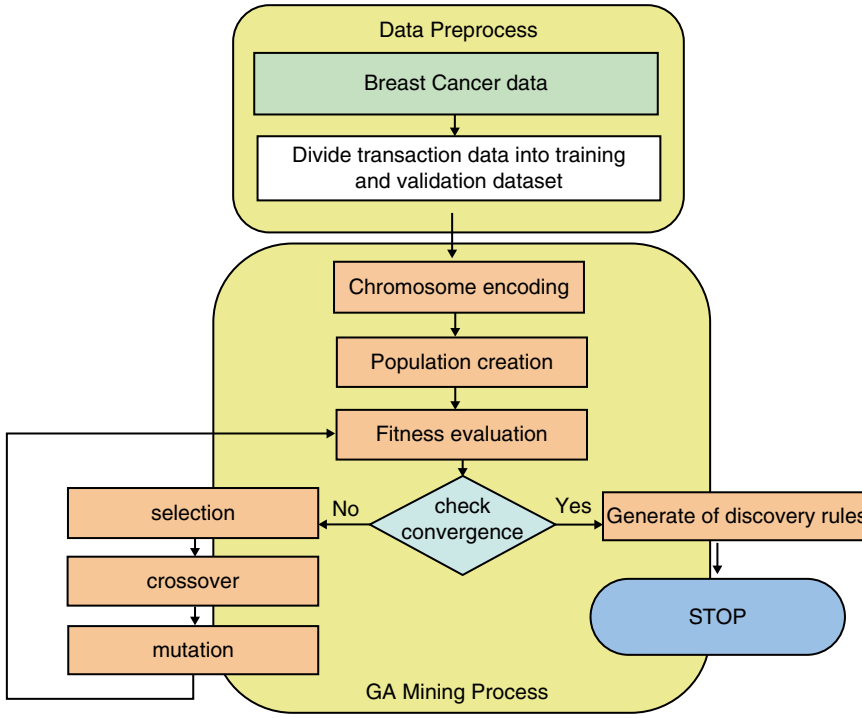

Fig. 1 The GA-based mining processes

3.3 GA Processes

A genetic algorithm is an iterative procedure until a pre-determined stopping condition (usually the number of generation). Genetic algorithm involves a population of individuals, each one represented by a finite string of symbols, known as the genome, encoding a possible solution in a given problem space. The simple genetic algorithm as a pseudo code is:

Step 1. Generate an initial population of strings randomly.

Step 2. Convert each individual of the current population into If–Then rule.

Step 3. Evaluate each of If–Then rules from training dataset.

Step 4. Select parents for the new population.

Step 5. Create the new population by applying selection, crossover and mutation to the parents.

Step 6. Stop generation if a stopping condition is satisfied, otherwise go to step 3.

3.4 Chromosome Encoding

A GA manipulates populations of chromosomes, which are string representations of solutions to a particular problem. A particular position or locus in a chromosome is referred to as a gene and the letter occurring at that point in the chromosome is referred to as the allele value or simply allele. Any particular representation used

for a given problem is referred to as the GA encoding of the problem. The classical GA uses a bit-string representation to encode solutions. The hybrid of the integral and real number encoding systems is used for encoding the genes in the chromosomes of the present study. The chromosomes are represented by the final diagnostic outcomes (0: normal; 1: breast cancer). The classification variables in Table 1 are encoded with integral numbers and the numerical variables are encoded with real numbers. The two types of variables are encoded as follows:

- Classification variables: integral number encoding

$$v_\kappa = \begin{cases} * & \text{if } p < 0.1 \\ \text{Int}(p*10)\,\text{Mod}(\text{UB}-\text{LB}+1) & \text{if } p \geqq 0.1 \end{cases}$$

- Numerical variables: real number encoding
- $v\kappa_{\_\text{LB}} = \text{LB} + (\text{UB}–\text{LB})*p$; $v\kappa_{\_\text{UB}} = \text{LB} + (\text{UB}–\text{LB})*p$

$$v_{\kappa-\text{LB}} = \begin{cases} * & \text{if } p < 0.1 \\ \text{LB} + (\text{UB}-\text{LB})*p & \text{if } p \geqq 0.1 \end{cases}$$

$$v_{\kappa-\text{UB}} = \begin{cases} * & \text{if } p < 0.1 \\ \text{LB} + (\text{UB}-\text{LB})*p & \text{if } p \geqq 0.1 \end{cases}$$

$$v_\kappa = \begin{cases} * & \text{if both} = * \\ \geqq v_{\kappa-\text{LB}} & \text{if } v_{\kappa-\text{UB}} = * \text{ and } v_{\kappa-\text{LB}} \neq * \\ < v_{\kappa-\text{UB}} & \text{if } v_{\kappa-\text{LB}} = * \text{ and } v_{\kappa-\text{UB}} \neq * \\ \geqq v_{\kappa-\text{LB}}, v_{\kappa-\text{UB}} & \text{if } v_{\kappa-\text{LB}} < v_{\kappa-\text{UB}} \\ \geqq v_{\kappa-\text{UB}}, v_{\kappa-\text{LB}} & \text{if } v_{\kappa-\text{LB}} \geqq v_{\kappa-\text{UB}} \end{cases}$$

Note: p: a random number between 0 and 1; int: integer function; Mod: remainder function; v_κ: genetic content after encoding; LB: lower bound of variable range; UB: upper bound of variable range; $v_{\kappa-LB}/v_{\kappa-UB}$ are the genetic constituents of numerical variables.

During the random formation of genetic contents, the * (don't care) operation is added to serve as the basis for the formation of the original matrix as the first GA model for the early diagnosis of disease. The chromosomal contents of the initial matrix randomly produced according to the rules described above are illustrated in Table 2. Rules represented by the above chromosomes are as follows: (3.2≦Clump thickness < 4.7) and (5.14≦cell shape < 8.33) and (1.81≦Marginal adhesion < 3.12) and (4.22≦epithelial cell size < 6.17) and (2.25≦Mitoses < 4.1) and … then breast cancer.

3.5 Selection

Selection is a process for choosing the rules with high fitness value as parents for reproduction. The mating selection of the proposed GA method is restricted within the same species; that is, the parents for reproduction are selected from the rules with the same class

Table 2
Rules of the random production of chromosomes

Clump thickness	Cell size	Cell shape	Marginal adhesion	Epithelial cell size	Mitoses	…	Result
≥3.2,<4.7	*	≥5.14,<8.33	≥1.81, <3.12	≥4.22,<6.17	≥2.25,<4.1	…	1

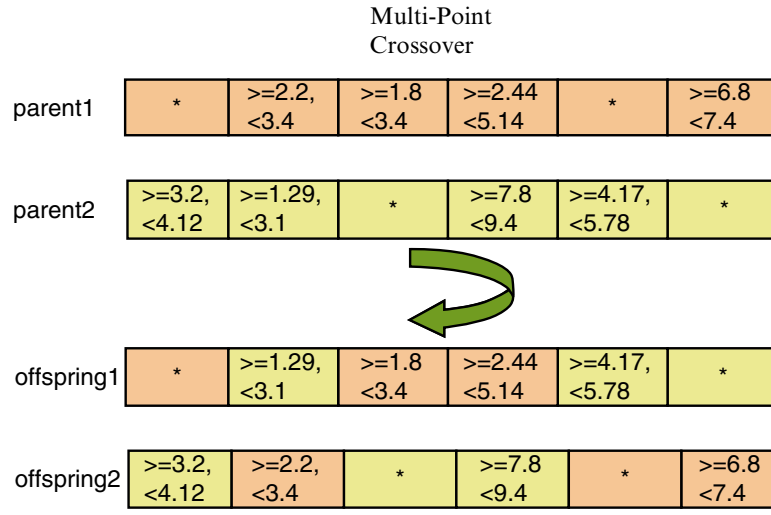

Fig. 2 Multi-point crossover operation

because the genetic operation between two rules of different classes tends to generate low-performance offsprings [16].

3.6 Crossover

The crossover operator represents the mixing of genetic material from two selected parent chromosomes to produce one or two child chromosomes. Once the two parent chromosomes are selected, a random number between 0 and 1 is produced. If the random number is greater than a certain crossover rate, the exchange between two chromosomes is required. In the present study, the crossover rate is set as 0.5. In general, the single-point crossover is used, i.e., only one specific chromosome is selected for the exchange at a time. However, the multi-point crossover is used in the present study where the random numbers are generated first when the exchange is required and the chromosomes are exchanged based on the actual situation. The crossover of three chromosomes is illustrated in Fig. 2.

3.7 Mutations

After a crossover operation is performed, the mutation process is the next step. The step is to prevent all solutions in the population from falling into a local optimum of solved problems. Figure 3 illustrates that the individual genes of new offspring are changed randomly with probability p ($p = 0.05$)

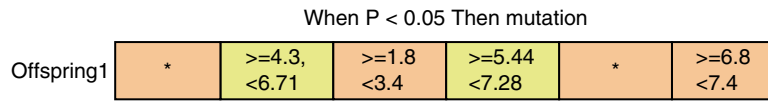

Fig. 3 Mutation operation

3.8 Fitness Evaluation of Rules

The role of the fitness function is to encode the performance of the rule numerically. In our study, the objective of the GA method is to find the accurate and general rules among all the rules in the population. Thus, the GA method uses the composite fitness function consisting of accuracy and coverage. To measure the accuracy and coverage of the rule, we use the following definitions: when a rule is used to classify a given training instance, one of the four possible concepts can be observed: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The true positive and true negative are correct classifications, while false positive and false negative are incorrect classifications. Using these concepts, we present a very simple fitness function defined as

$$\text{Maximize Fitness Function} = \frac{TP}{TP + FP}$$

3.9 Procedures of Establishing Model

At first, we collected clinical, demographic data for 699 cases of breast cancer. After data pre-processed which needed includes cleaning, transforming, and standardizing, and merging data into one table. Secondly, we used data mining tool WEKA establishing C4.5, ANNs, and logistic regression classification predict model (Fig. 4). There are nine collective attributes at the start, but those adopt attribute selective procedures to reduce unnecessary properties, making the model construction more efficient and to also figure out the simplified classification rules. This study used information gain to decrease attributes. Finally, we reserved all variables, because all variables have their own contribution.

4 Results

We used three performance measures: accuracy (Eq. 3), sensitivity (Eq. 4) and specificity (Eq. 5) where TP, TN, FP and FN denotes true positives, true negatives, false positives, and false negatives, respectively. True positives denote the correct classifications of positive examples. True negatives are the correct classifications of negative examples. False positives represent the incorrect classifications of negative examples into class positive and false negatives are the positive examples incorrectly classified into class negative. Table 3 lists:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \qquad (3)$$

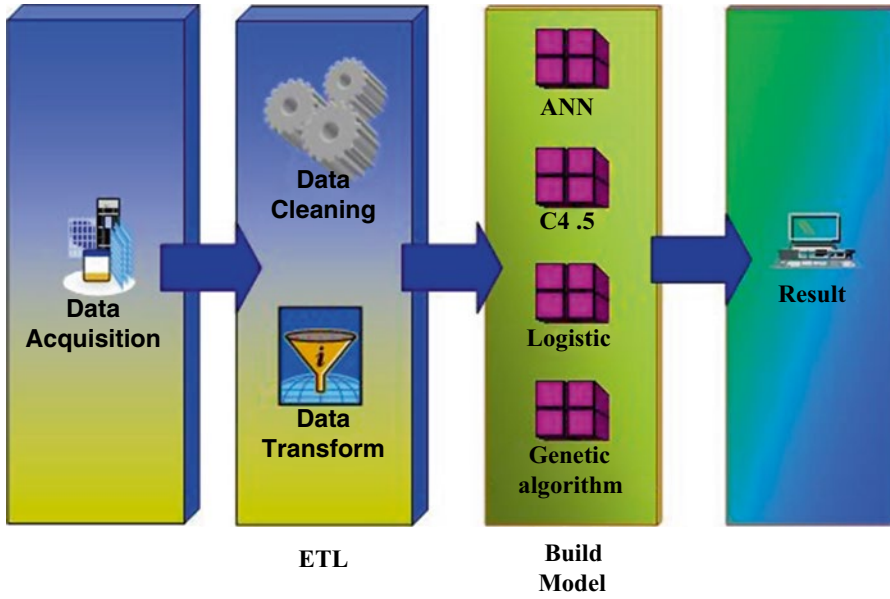

Fig. 4 Procedures of establishing model

$$Sensitivity = (TP) / (TP + FN) \tag{4}$$

$$Specificity = (TN) / (TN + FP) \tag{5}$$

In order to minimize the bias associated with the random sampling of the training and holdout data samples in comparing the predictive accuracy of two or more methods, researchers tend to use *k*-fold cross-validation. In *k*-fold cross-validation, also called rotation estimation, the complete dataset (D) is randomly split into k mutually exclusive subsets (the folds: $D_1, D_2, \ldots, D_k$) of approximately equal size. The classification model is trained and tested k times. Each time ($t \, \varepsilon(1, 2, \ldots, k)$), it is trained on all but one folds (Dt) and tested on the remaining single fold (Dt). The cross-validation estimate of the overall accuracy is calculates as simply the average of the k individual accuracy measures $\mathrm{CVA} = \sum_{i=1}^{k} A_i$ where CVA stands for cross-validation accuracy, k is the number of folds used, and A is the accuracy measure of each folds.

Since the cross-validation accuracy would depend on the random assignment of the individual cases into k distinct folds, a common practice is to stratify the folds themselves. In stratified *k*-fold cross-validation, the folds are created in a way that they contain approximately the same proportion of predictor labels as the original dataset. Empirical studies showed that stratified cross-validation tend to generate comparison results with lower bias and lower variance when compared to regular *k*-fold cross-validation [30].

Table 3
Decision confusion matrix table

		Actual class	
		I (without breast cancer)	II (with breast cancer)
Classified class	I (without breast cancer)	*A*	*B*
	II (with breast cancer)	*C*	*D*

A is the number of correct predictions that an instance is negative
B is the number of incorrect predictions that an instance is positive
C is the number of incorrect predictions that an instance negative
D is the number of correct predictions that an instance is positive

In this study, to estimate the performance of classifiers a stratified tenfold cross-validation approach is used. Empirical studies showed that 10 seem to be an optimal number of folds (that optimizes the time it takes to complete the test while minimizing the bias and variance associated with the validation process) [30, 31]. In tenfold cross-validation the entire dataset is divided into ten mutually exclusive folds with approximately the same class distribution as the original dataset. Each fold is used once to test the performance of the classifier that is generated from the combined data of the remaining nine folds, leading to ten independent performance estimates.

4.1 Neural Network

Neural networks can be classified into two different categories, feed-forward and feedback networks. The feedback networks contain nodes that can be connected to themselves enabling a node to influence other nodes as well as itself. Kohonen self-organizing network and the Hopfield network are examples of this type of network [32–34].

We used ANNs parameter setting to adjust the classification to optimum level. Establishing of Classification model and examination of condition are described as follows: In Parameter, two of the parameters have to be adjusted accordingly. One is the learning rate, and the other is momentum factor. As Rumelhart [35] concluded that lower learning rates tend to give better network result, we decided to incorporate the lower learning rates in our test models, thus, our study applied each learning rate in 0.3, 0.35, 0.4, 0.45, and 0.5 of hidden layer, and set momentum factor from 0.1 to 0.9 (all ANNs parameter setting in Table 4). This experimentation applied 50 kinds of model in total to evaluate the efficiency of the prediction. According to the prediction result from different parameters setting, this study used Accuracy and Sensitivity to be the evaluate criterion and set training times as 300 in order to choose the best parameters of ANNs, as shown in Table 4.

We put the Accuracy, SEN, SPC, and Roc of different network type in order. As illustrated in Table 5, the (9,6,1) topology with a learning rate of 0.5 gives the best result.

Table 4
ANNs network structure parameter setting

Input variables	9
Hidden layers notes	(Attributes + classes)/2 = (9 + 2)/2 = 6
Output variables	1
Hidden layers	Default hidden layer is 1 in WEKA.
Learning rate	From 0.3 adjust to 0.5 Each time increase 0.05
Momentum	From 0.1 adjust to 0.9 Each time increase 0.1
Training time	300 times

Table 5
Accuracy, TP rate, and Roc of different network type (V)

Learning rate	Momentum	Accuracy	SEN	SPC	Roc
0.5	0.1	0.971	0.968	0.977	0.991
	0.2	0.967	0.962	0.977	0.992
	0.3	0.967	0.962	0.977	0.992
	0.4	0.967	0.962	0.977	0.990
	0.5	0.971	0.981	0.943	0.989
	0.6	0.964	0.975	0.966	0.994
	0.7	**0.975**	**0.981**	**0.966**	**0.994**
	0.8	0.955	0.968	0.932	0.992
	0.9	0.971	0.975	0.966	0.986

TP rate True Positive rate, *ROC* Receiver Operating Characteristic, *SEN* Sensitivity, *SPC* specificity

4.2 Genetic Algorithm Model

The proposed GA based modeling approach is implemented in Evolver 4.0 software which is an optimized add-in for Microsoft EXCEL. Moreover, the determination of GAs' parameters is a significant problem for the GA implementation. However, no formal methodology can be used to solve the problem because various value-combinations of the parameters result in different characteristics as well as different performances of GAs. Therefore, one should note that the best values for the GAs' parameters are case-dependent and based upon the experience from preliminary runs. The parameter settings for genetic algorithm in the present study are illustrated in Table 6.

Table 6
Setting of system parameters

Parameters	Value
Population size	30
Chromosome length	10
Selection	Roulette wheel and tournament
Crossover	Multi-points
Crossover rate	0.5
Mutation rate	0.05
Stop criteria	Until 100 generations

Table 7
Stability test results of GA model

Group	1–5				6–10					
Evolution generation	1	2	3	4	5	6	7	8	9	10
Mean fitness value	0.93	0.83	0.89	0.81	0.86	0.85	0.77	0.90	0.96	0.90
Optimal fitness value	1	0.976	0.983	0.988	1	0.988	0.984	1	1	1

When the same parameter settings are used, the genetic algorithm operation was performed ten times and the fitness values were recorded in Table 7. The numbers in Table 7 are the average fitness values from each experiment where the evolution generations 1–5 and 6–10 were divided into two groups. SPSS10.0 was used for the t testing and the testing of the differences between two population means (confidence interval 95 %) to validate the difference in the means of the two groups.

The results show t-statistics = −0.293; one-tailed p value = $0.392 > 0.05$. Therefore, H_0 was accepted since there was no difference between the means of two groups confirming the stability of the GA model. The convergence graph of the three best out of ten experiments is illustrated in Fig. 5. A starts to converge from the 25th generation, B from the 37th generation, and C from the 32nd generation. Although the convergence generations are different, they are still close to each other indicating the stability of this model. Furthermore, the fitness value of 1 is still approached at the end in this prediction model.

The parameter settings described above were used for the investigation of the breast cancer data. The results obtained were further used for the evaluation of the proposed models. The rules

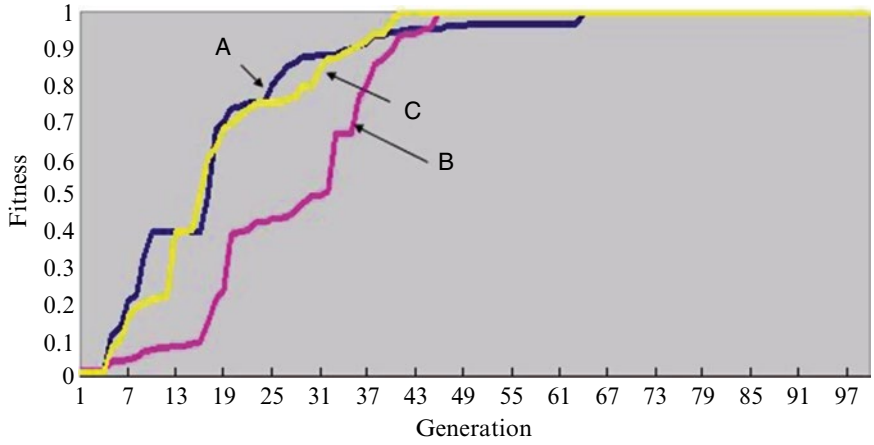

Fig. 5 Convergence graph of GA Model

Table 8
Rules with the corresponding accuracy

Decision rules	Fitness value (train data, 453 cases)	Accuracy (train data, 453 cases)	Accuracy (test data, 246 cases)
IF 5.6 < Clump thickness < 7.2 AND 1.8 < Marginal adhesion < 4.0 AND 3.2 < Single Epithelia < 8.6 AND 2.1 < Normal nucleoli < 3.1 THEN Class = benign	1	0.993	0.9878

Table 9
Classification results using genetic algorithm

		Actual class	
		I (without breast cancer)	II (with breast cancer)
Classified class	I (without breast cancer)	149	0
	II (with breast cancer)	3	94

obtained from the investigation are illustrated in Table 8. The model proposed by the present study is comprehensible for the user in a very plain and simple way. A high degree of accuracy is also achieved in this model.

The results of the classification of breast cancer data using genetic algorithm are illustrated in Table 9 where the average classification accuracy is 0.9878, the sensitivity 0.9802, and the specificity 1. In class I, three breast cancer negative cases are incorrectly classified as breast cancer positive. In class II, there is no incorrect classification.

References

1. Wingo PA, Tong T, Bolden S (1995) Cancer statistics, 1995. CA Cancer J Clin 45(1):8–30
2. Calle J (2004) Breast cancer facts and figures 2003–2004. Am Cancer Soc 2004:1–27
3. Jerez-Aragones JM et al (2003) A combined neural network and decision trees model for prognosis of breast cancer relapse. Artif Intell Med 27(1):45–63
4. Edwards BK et al (2002) Annual report to the nation on the status of cancer, 1973–1999, featuring implications of age and aging on U.S. cancer burden. Cancer 94(10):2766–2792
5. Pendharkar P, Rodger J, Yaverbaum G (1999) Association, statistical, mathematical and neural approaches for mining breast cancer patterns. Exp Syst Appl 17:223–232
6. Elmore JG et al (1994) Variability in radiologists' interpretations of mammograms. N Engl J Med 331(22):1493–1499
7. Fentiman IS (1998) Detection and treatment of breast cancer. Martin Duntiz, London
8. Anderson TW (1984) An introduction to multivariate statistical analysis. Willey, New York, NY
9. Johnson RA, Wichern DW (2002) Applied multivariate statistical analysis. Prentice-Hall, Upper Saddle River, NJ
10. Kovalerchuk B et al (1997) Fuzzy logic in computer-aided breast cancer diagnosis: analysis of lobulation. Artif Intell Med 11(1):75–85
11. Barr EAF (1982) The handbook of artificial intelligence, vol 1–3. William Kaufmann, Los Altos, CA
12. Laurikkala J, Juhola M (1998) A genetic-based machine learning system to discover the diagnostic rules for female urinary incontinence. Comput Methods Programs Biomed 55(3):217–228
13. Myoung-Jong K, Ingoo H (2003) The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms. Exp Syst Appl 25:637–646
14. Chen TC, Hsu TC (2006) A GAs based approach for mining breast cancer pattern. Exp Syst Appl 30:674–681
15. Goldberg DE (1989) Genetic algorithm in search, optimization, and machine learning. Addison-Wesley, Reading, MA
16. Holland JH (1975) Adaption in natural and artificial systems. The University of Michigan Press, Ann Arbor, MI
17. Goldberg DE (1994) Genetic and evolutionary algorithms come of age. Comm ACM 37:113–119
18. Mitchell M (1996) An introduction to genetic algorithms. MIT Press, Cambridge, MA
19. Forrest S (1993) Genetic algorithms: principles of natural selection applied to computation. Science 261(5123):872–878
20. Congdon CB (1995) A comparison of genetic algorithms and other machine learning systems on a complex classification task from common disease research. Department of Computer Science and Engineering, University of Michigan
21. Bali RK et al (2005) Introduction to the special issue on advances in clinical and health-care knowledge management. IEEE Trans Inf Technol Biomed 9(2):157–161
22. Gurbaxani BM et al (2006) Linear data mining the Wichita clinical matrix suggests sleep and allostatic load involvement in chronic fatigue syndrome. Pharmacogenomics 7(3):455–465
23. Berger AM, Berger CR (2004) Data mining as a tool for research and knowledge development in nursing. Comput Inform Nurs 22(3):123–131
24. Hobbs GR (2001) Data mining and healthcare informatics. Am J Health Behav 25(3):285–289
25. Obenshain MK (2004) Application of data mining techniques to healthcare data. Infect Control Hosp Epidemiol 25(8):690–695
26. Koh HC, Tan G (2005) Data mining applications in healthcare. J Healthc Inf Manag 19(2):64–72
27. Bauer RJ (1994) Genetic algorithm and investment strategies. Willey, New York, NY
28. Kim YS et al (2003) Screening test data analysis for liver disease prediction model using growth curve. Biomed Pharmacother 57(10):482–488
29. Shin KS, LEE YJ (2002) A genetic algorithm application in bankruptcy prediction model. Exp Syst Appl 23(3):321–328
30. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. The fourteenth International Joint Conference on Artificial Intelligence 1995. San Francisco, CA.
31. Breiman L, Friedman JH, Qlshen RA (1984) Classification and regression trees. Wadsworth & Brooks/Cole Advanced Books, Pacific Grove, CA
32. Kim EK et al (1993) Comparison of neural network and k-NN classification methods in medical image and voice recognitions. Med J Osaka Univ 41–42(1–4):11–16
33. Richardson CJ, Barlow DJ (1996) Neural network computer simulation of medical aerosols. J Pharm Pharmacol 48(6):581–591
34. Eghbaldar A et al (1996) Identification of structural features from mass spectrometry using a neural network approach: application to trimethylsilyl derivatives used for medical diagnosis. J Chem Inf Comput Sci 36(4):637–643
35. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. MIT Press, Cambridge, MA

Chapter 13

Mining Data When Technology Is Applied to Support Patients and Professional on the Control of Chronic Diseases: The Experience of the METABO Platform for Diabetes Management

Giuseppe Fico, Maria Teresa Arredondo, Vasilios Protopappas, Eleni Georgia, and Dimitrios Fotiadis

Abstract

This chapter provides an overview of how healthcare institution could benefit from the usage of technologies and personal health systems. Clinical, Usage and Technical data are mined in different ways and with different methods to support users (patients, health professionals and informal caregivers) in taking decisions. As a case study, the solutions and the techniques adopted in a research project focused on the delivery of technologies to improve diabetes management are described.

Key words Data mining, Diabetes, Disease management and modeling, Personal health systems

1 Introduction

The established healthcare systems in European countries are well suited to the treatment of acute diseases, but are mostly inadequate for dealing with chronic diseases. In contrast to acute diseases, where care professionals treat patients in a physician-centred fashion, with short appointments and limited patient instruction, the treatment of chronic diseases is rather a long-term management program aiming to first stabilise the patient's health condition and subsequently to prevent long-term complications. The focus must be moved from acute disease management of patients, reacting on critical episodes, to health maintenance of individuals at home.

Health technologies, Personal Health Systems, and Medical Informatics may be facilitators to enable better diagnosis, treatment, and management at points of need. However, how these technologies can be endurably used by patients, and how they can enable collection of information in a way that meets care provider needs remain

Carlos Fernández-Llatas and Juan Miguel García-Gómez (eds.), *Data Mining in Clinical Medicine*, Methods in Molecular Biology, vol. 1246, DOI 10.1007/978-1-4939-1985-7_13,

open questions. Research is focusing on understanding how to collect information to support behavioural monitoring [1], how information needs to be presented and aggregated in order to provide personalization, adaptation, exploration, and visualization [2, 3].

According to a Cochrane review of 86 trials issued in 2009 [4], the use of Decision Support Systems (DSS) leads to better knowledge, risk awareness, self-management and shared decision making. This has been confirmed in the Salzburg statement [5], representing patient advocacy groups from about 20 countries. How these good intentions are transformed into evidence is still an issue when it comes to put individuals at the center of the process.

With the emerging mass adoption of smart technologies and personalized health applications, the research is moving from proving the feasibility and acceptability of telemedicine systems to validations focused on the effectiveness of the information that patients are managing through these systems: the way of mining this huge amount of data, combining and aggregate their great variety and the different dimensions is still a challenge.

In this chapter, we provide an example of the approach used in a research project whose aim was to deliver technologically based solutions to support diabetes disease management.

First we introduce the current state of the art in data mining applied to diabetes and then we present the METABO project.

On November 2011, a systematic review of the applications of data-mining techniques applied to diabetes have been made [6], showing that they are useful for generating new hypothesis for further experimentations, extracting knowledge and improving the healthcare for diabetes patients.

A great part of the methods are focused on blood glucose level interpretation and prediction [7–9], trying to identify which are the best predictors, rules and trends that are associated with glycemic control. In other cases, feature selection techniques are applied to identify the most important factors that influence blood glucose control [10] or predict complications [11]. Data-mining methods have been also applied for genomic analysis related to diabetes [12, 13], showing encouraging theoretical results but missing in most of the cases an experimental validation, generating new hypothesis for research. Healthcare flow analysis, data pre-processing/cleaning, adverse drug events, detection of insurance frauds, clinical guideline enrichment, and prediction of early mortality have been the subject of data-mining methods too [14–17].

The conclusion of the review is that the research is limited to the dataset used and is lacking universal prediction rules that could be applied to multiple datasets. Yet, the existing datasets should be better exploited, as they may contain useful information for every patient, including personal, clinical, and social information.

Information about the most relevant factors determining an optimal glucose control should be supported and enhanced by

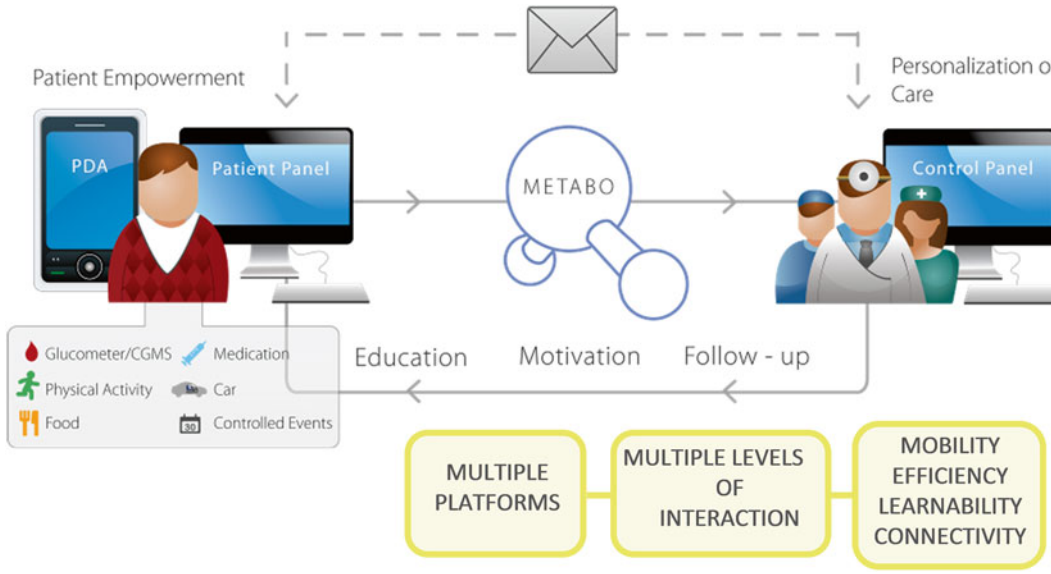

Fig. 1 The METABO platform

actions that aim to guide patients on controlling these factors on a daily basis and to help professionals on having a better picture of what is happening to their patients, from both an individual and a global perspective.

The following sections provide an example of which methods can be applied to produce information that can be used for different purposes (supporting patient and professionals in practical scenarios, in a personalized fashion, on understanding what is affecting an optimal control and what can be avoided in the future) and on different time scales (short-term, medium-term, and long-term assessments).

The METABO project was a research project aiming at designing, building and testing an IT platform, serving to monitor glucose values and lifestyle/pharmacological factors affecting blood glucose concentrations in patients with diabetes mellitus in real-life situations in order to provide structured information and therapeutic decision support to diabetic patients and their care givers [18] (Fig. 1).

The IT platform included a desktop application for the physician, the Control Panel, and a set of personal monitoring devices for the patients.

The Control Panel is a client-based patient-management platform through which the care providers can access in an ordered way all the relevant information of their patients and receive feedback from the METABO Decision Support System. It provides a standard framework of care, based on the American Diabetes Association guidelines [19], that structures the patient follow-up in four major milestones: diagnosis, education, treatment, and

complications, providing this way a unique and standardized care pathway. This structure is flexible to permit its adaptation to the needs and requirements of the care providers and the Health Care Centers, bringing efficiency to the daily care processes. The tool allows the care providers to access reliable data, tailor the treatments, and define care plans.

The *Patient Monitoring Device* (PMD) allows collecting metabolic data, including food and drug intake, physical activity (through the combined use of a pedometer and a wearable "metabolic holter"), discrete glycemic values (by fingersticks), and continuous subcutaneous glucose concentration (by Continuous Glucose Monitoring System, CGMS). Different PMD applications were developed following a user-centred design with the identification of two archetypes of patients, type 1 and type 2.

The platform was tested in a small-scale pilot study aiming at providing an exploratory analysis of the usability of the METABO system (PMD + Control Panel) and its acceptability by the users, as well as its impact on clinically relevant parameters in type 1 and 2 diabetic patients in comparison with the standard diabetes care (without the use of METABO) for four consecutive weeks in real-life scenarios (Table 1).

The data produced within the platform (biosensors, lifestyle subjective data, interactions with the interfaces and clinical parameters) have been used, modeled, and mined at different levels and for different purposes:

1. For creating an individualized metabolic model for prediction of glucose excursions
2. Knowledge extraction and clustering tools for professionals
3. Aggregated data to present relevant information to users
4. Assessing system performances through a combined clinical-behavioural-technical perspective.

2 The Dataset

The information has been produced and managed based on the following:

- Current medications, including time of administration and dose
- Glucose fingerstick measurements by glucometer and inserted by the patient through the PMD and paper diary
- Food intake data through the PMD and a paper diary
- Physical activity and walking data
  - Pedometer: OMRON Walking Style Pro pedometer
  - Metabolic holter: Sensewear Armband

Table 1
Overview of indicators used for the METABO platform assessment

Indicator	Data recorded	Measurement method
Compliance to treatment		
Self-monitoring	Monitoring of Blood Glucose, Food intake, physical activity, education, blood pressure, weight, drug intake	Personal Health Records and Electronic Diaries (in Smartphones) and paper diary. Depending on what the physicians has prescribed, compliance to self-monitoring is measured as number of performed actions respect to prescribed total actions over a certain period.
Food intake suggestion	Recommended Daily Calories, recommended Calories and CHO intake per meal (breakfast, snacks, lunch, dinner)	Personal Health Records and Electronic Diaries
Drug intake	drug (ATC dode), strength, time and dosage	Personal Health Records and Electronic Diaries
Physical activity	Duration, intensity and frequency	Personal Health Records and Electronic Diaries
Education	Education level	Quiz and questionnaires
Goal achievement	Goal to achieve and progression over time	Weight loss and education level improvement
Questionnaires provided to patients		
Quality of life	Questionnaire score	ADS questionnaires, before and at the end of pilot [20]
Patient motivation	Questionnaire score	Custom questionnaires, before and at the end of pilot
Good practice adoption	Questionnaire score	Custom questionnaires, end of pilot
Perceived usefulness	Questionnaire score	Davis Scale, end of pilot [21]
User satisfaction	Questionnaire score	AttrakDiff questionnaire, end of pilot [22]
Perceived importance of diabetes knowledge	Questionnaire score	Custom questionnaires, before and at the end of pilot
Self-estimation of diabetes knowledge	Questionnaire score	Custom questionnaires, before and at the end of pilot
User knowledge of ICT	Questionnaire score	CLS questionnaire, before pilot

(continued)

Table 1
(continued)

Indicator	Data recorded	Measurement method
Healthcare personnel indicators		
User satisfaction toward ICT system	Questionnaire score	AttrakDiff questionnaires, end of pilot
Attitude towards HIS	Questionnaire score	Part B of Boy's scale, before pilot [23]
Good practices adoption	Questionnaire score	Custom questionnaires, before and at the end of pilot

  - Physical activity log sheet through the PMD application
  - Physical activity described through the paper diary
- Subcutaneous continuous glucose: Medtronic Guardian RT
- Events and notes from patient's diary in the PMD

It is worth to mention that the information has been produced in a duplicated (and in some cases triplicated) way, as it came from sensors or a paper diary or a smart electronic diary (the PMD): the implementation of the data-mining methods has been realized, when possible, independently of the data source with the aim of making possible their reproducibility in multiple settings and circumstances.

Individualized Metabolic Modeling

Prediction of glucose is useful in that it can provide immediate critical feedback to the diabetic patients about how the glucose is affected by their lifestyle and treatment. In addition, it offers the means of making real-time suggestions regarding modifications to diet and activity related profile as well as diabetes medications in order to avoid critical events. In METABO, predictions of glucose values and of glucose-related events (hypo/hyper events) are provided to the patients through the PMD based on the recent history of all collected contextual information.

The system responsible for making glucose predictions and generation of decision support is termed the Metabolic Modeling System: it accepts as input the collected glucose, insulin, dietary, and lifestyle data under free-living conditions and consists of a data pre-processing module, compartmental models that describe the glucoregulatory mechanisms and a final subsystem responsible to learn patient's behaviour and thereafter predict future glucose profiles. The overall components of system are shown in Fig. 2: several features are extracted from the input data and thereafter food,

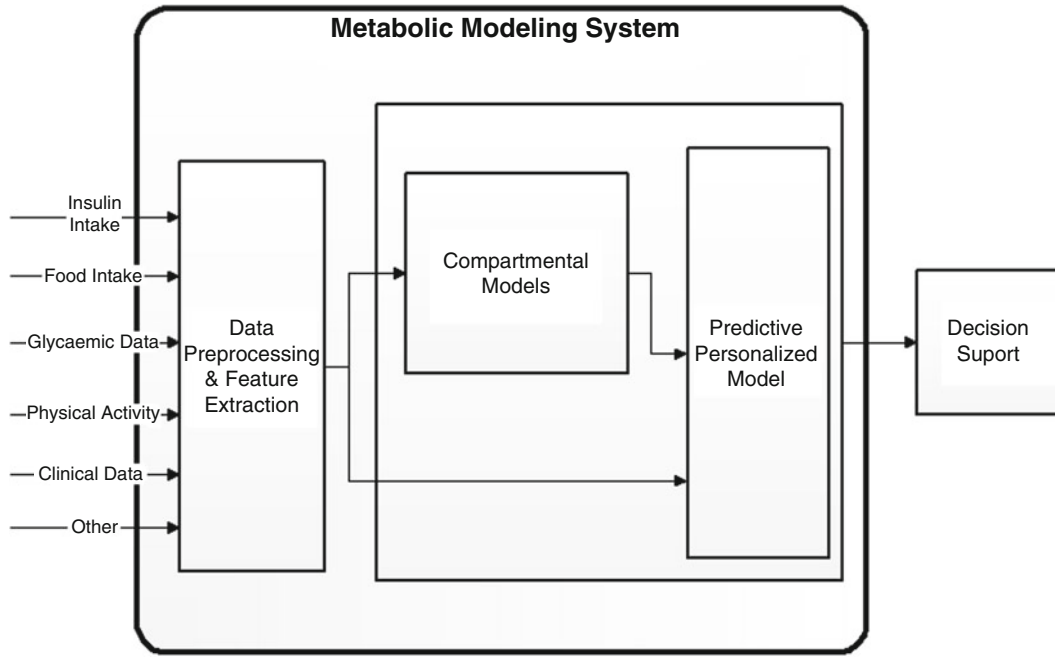

Fig. 2 Diagram of the overall metabolic modeling system

insulin, and physical activity-related information pass through compartmental models so as to compute glucose production, absorption, and insulin secretion rates and concentrations. The final stage is to use all relevant information so as to train a machine-learning system to make suitable predictions.

2.1 Dataset for System Implementation and Validation

The dataset used to train and test the Metabolic Modelling and Prediction System has been based on the data collected from 22 diabetic patients under free-living conditions under the supervision of five pilot centers in Madrid, Parma, Paris, Prague, and Modena. The observation period of the pilot study was on average 10 days (range from 5 to 14 days). The study was designed as an observational study meaning that the data were first collected during the study period and thereafter model training and testing was performed based on the collected data.

2.2 Input Data

All patients wore the Guardian Real-Time CGM system (Medtronic Minimed) that monitors the subcutaneous glucose concentrations every 5 min. The glucose sensor calibration requires at least four blood glucose measurements to be made daily using a standard blood glucose meter.

The patients were also using the SenseWear body monitoring system (BodyMedia Inc.) which continuously monitors their daily physical activities. The SenseWear armband collects data using five sensors: heat flux, skin temperature, near body temperature,

galvanic skin response, and a two-axis accelerometer. At the same time, exercise events were also manually described (time, duration, type of activity, intensity) by the patients using a suitably designed paper diary.

Information regarding the food intake (i.e., type of food, serving sizes, and time), the blood glucose pricks (value, time) and the insulin injections (type, dose, and time) was recorded by the patients using a specially designed paper diary. The food composition (i.e., CHO, calories, carbohydrates, fat) was post-analyzed by the dieticians of the corresponding pilot centers.

From all the collected data, the following signals and parameters were used as input features to the metabolic model:

- Meal (time and CHO)
- Insulin (time, type, and dose)
- Physical activity: time, duration, and parameters from sensor data (heat flux, skin temperature, MET)
- Blood glucose (time, value)
- Continuous glucose values

From the above data, some are directly used from the Predictive Personalized Model (e.g., glucose recordings) and some other are pre-processed (e.g., segmentation of physical activity into intense-activity periods) and some other pass through the compartmental models to compute glucose and insulin rates and concentrations in blood (Fig. 3).

2.3 Description of Predictive Modeling System

The method for the prediction of the s.c. glucose concentration is presented schematically in Fig. 2. It comprises compartmental models of the glucose—insulin regulatory system and a predictive model of glucose. The compartmental models are used to simulate (a) the ingestion and absorption of carbohydrates (the Meal Model), (b) the absorption and the pharmacokinetics/pharmacodynamics of subcutaneously administered insulin (the Insulin Model), as well as (c) the impact of exercise on glucose—insulin metabolism (the Exercise Model). In addition, support vector machines for regression (SVR) are employed to provide individualized glucose predictions. The input variables of the proposed model include the rate of glucose appearance in plasma after a meal, the plasma insulin concentration, the s.c. glucose measurements, as well as a set of physical activity related variables. Two different approaches to investigate the physical activity's effects on diabetes have been assumed. In the first, the Metabolic Equivalent of Task (MET), the heat flux and the skin temperature variables, which are recorded by the SenseWear armband, were used as inputs in the model. The second approach utilizes the alterations in circulating glucose and insulin concentrations during and shortly after exercise as computed by the Exercise Model.

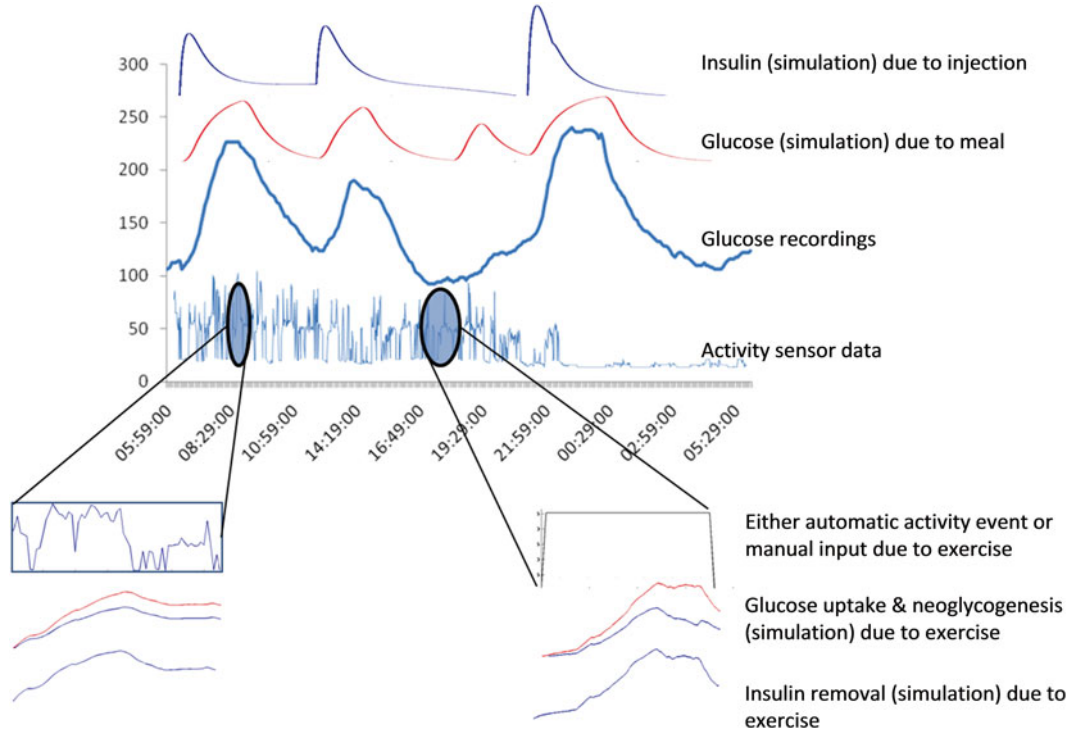

Fig. 3 Illustration of the different input data used in the Metabolic Modeling System coming from recording from sensors or computed within the Metabolic System, e.g., from various the compartmental models

3 Knowledge Extraction for Care providers

Knowledge extraction refers to the discovery of useful information related to a patient or to multiple patients using data-mining techniques applied to the collected and profiling data. The extracted knowledge allows the physicians through the Control Panel (CP) to interpret changes in patients' status, assess the progress in response to specific treatment prescriptions, discover new associations among clinical, dietary and lifestyle parameters, cluster patients, and manage clinical pathways.

Three categories of knowledge discovery have been provided, namely (a) association rules aiming at finding relations among lifestyle, treatment, and metabolic data and (b) clustering which groups patients according to similarities hidden in the data, and (c) classification with which patients can be assigned to a cluster according to their characteristics.

3.1 Association Analysis

Association (or explanation) analysis which belongs to the rule-mining category of the data-mining field allows the treating physician to find and explain dependencies and similarities that are observed significantly often among the collected and profiling

data. The discovered knowledge can refer to a specific patient or for a group of patients.

Rule-mining is useful and effective when working with large datasets in which relationships are hidden among the data. The discovered relationships are mined by making if—then queries with given support and confidence values. In this direction, the CP provides effective GUIs which permit the physicians to "build" such statements and thereafter overview the produced results.

3.2 Clustering Analysis

Clustering refers to the creation of sub-groups of patients according to the similarities of their characteristics. Clustering is performed using unsupervised machine learning techniques that calculate clusters characterized by their centroid in which each member (patient) shares common characteristics. With clustering, groups are made based not only on established medical knowledge, such as type 1 vs. type 2, or insulin-dependent vs. insulin-independent, but mainly by identifying (high-dimensional) similarities hidden in the collected and profiling data that might be new to the medical professionals.

3.3 Classification Analysis

In classification analysis, the aim is to assign a patient (a new or an existing) to a category/class of medical status. In this respect, any treatment to be prescribed can be initialized or modified based on the knowledge already available for his/her class. A first approach to classification problem adopted in METABO is to assign a new patient to an existing cluster by calculating the distance of his/her characteristics from the clusters' centroids.

Knowledge extraction is based on the analysis of the recorded and profile data and information. In this direction, the different types of data were categorized according to their nature or to their source prior to be used as input parameters into the knowledge extraction. This categorization is helpful regarding (a) the functionality of the knowledge extraction tools since it facilitates the physicians who can select the suitable inputs for each type of knowledge extraction in a structured and effective manner and (b) the execution of pre-processing that is needed for each data type prior to their input to the analysis.

The types of input parameters are listed below:

- Glycemic parameters, such as glucose values and their relation with respect to other events (post-prandial, after exercise, etc.)
- Dietary parameters, such as CHO, calories, time, and type of meal
- Physical activity parameters, such as type, intensity, duration of exercise events, number of exercise events per time period, daily steps, energy expenditure
- Medication parameters related to insulin dosages, site of injection, time of intake, missed registrations, etc.

- Demographic and clinical parameters, such as age, type of diabetes, comorbidities, etc.
- Lifestyle parameters, such as working schedules, working days, hours of sleep.
- Other, such as periods of stress, and emotional state (this category is still under investigation).

It has to be noted that not all the abovementioned types of input are applicable to all categories of knowledge extraction. For instance, the demographic and clinical parameters (which are static or quasi-static) are not used in association analysis performed for a single patient but are extremely useful for the same type of analysis when performed to extract knowledge within multiple patients. More specifically, the types of input used for each category of knowledge extraction are summarized as follows:

- *Association analysis for a single patient*: Glycemic, physical activity, dietary, medication, medical.
- *Association analysis for group of patients*: Glycemic, physical activity, dietary, medication, clinical, demographic, and lifestyle.
- *Clustering analysis*: There are two classes of input, namely static containing demographic, lifestyle, and clinical parameters and acquired (from the sensors or manually through the PMD GUIs) containing glycemic, dietary, and physical activity parameters.
- *Classification*: The same as in clustering analysis.

In order to determine what type of pre-processing is needed for the various data types, such as calculation of daily averages, and segmentation of values into ranges and provide clinically meaningful input data, detailed questionnaires were prepared and sent to the medical partners. Their feedback was evaluated and then incorporated into the pre-processing stage.

4 Description of Knowledge Extraction Application

The Knowledge extraction application is integrated into the CP and consists of four different GUIs corresponding to each category of knowledge. "Association analysis for a single patient" and "Classification" are accessible through the GUI that is presented after a specific patient has been selected as shown in Fig. 4.

On the other hand, "Association analysis for group of patients" and "Clustering" through the main menu (Fig. 5), since these two approaches do not refer to a specific patient but rather to any patients that will be selected.

Fig. 4 Shortcut to association analysis and classification for a single patient under the button Pattern Recognition

All the above-mentioned types of input parameters are provided in the corresponding GUIs in the form of distinct tabs with each tab containing the list of parameters (as shown in the figures in the next sections). This structured representation concept was found effective and was followed in each GUI for allowing the physicians to stepwise "build" the IF- and THEN-parts of the statements, select input parameters for clustering, etc.

4.1 Association Analysis

Association analysis is applied to uncover relationships from large datasets and is represented in the form of association rules (IF-THEN statements) or sets of frequent items. For the performance of association analysis in METABO, either for one or for multiple patients, the "a-priori" algorithm was used. It uses support-based pruning to systematically control the exponential growth of candidate itemsets.

The knowledge to be extracted is based on queries in the form of "IF variable(s) THEN variable(s)." In association analysis terminology, these lifestyle and medical variables with specific ranges of values correspond to items, whereas the IF-THEN statements are

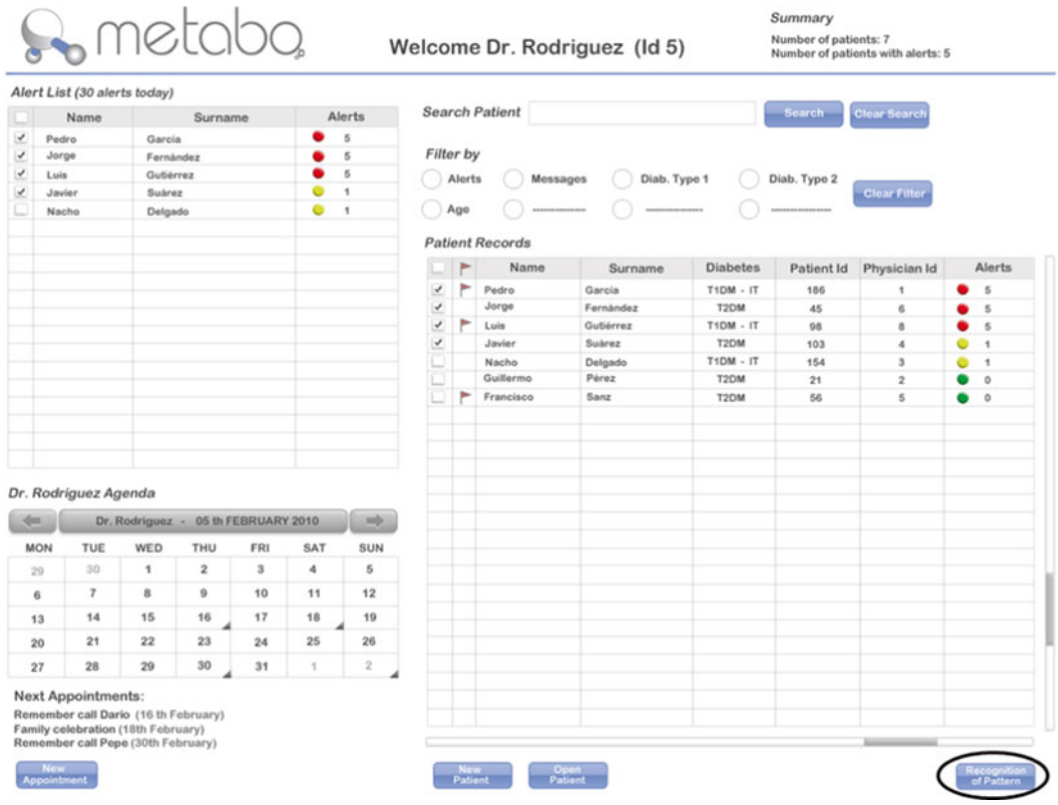

Fig. 5 Shortcut to association analysis and clustering for multiple patients under the button Knowledge Extraction for a Group

transactions for which associations are sought. The strength of an association rule can be measured in terms of its support and confidence. Support determines how often a rule is applicable to a given dataset (which represents frequency, i.e., number of times that IF and THEN variables appear over the given dataset), while confidence determines how frequently items in THEN appears in transactions that contain IF parameters (i.e., the conditional probability of the number of occurrences of the "THEN variables" given the occurrence of the "IF variables" within the given dataset). In this sense, a valid rule is generated when the thresholds of support and confidence are satisfied.

4.2 Association Analysis for a Single Patient

Figure 6 shows the initial screen in which the physicians can create IF-THEN statements and carry out association-rule mining. The physician is allowed to select variables from predefined lists, specify logical operators and choose the observation period. The support and confidence thresholds are user-specified or left-blank (in which case weak rules can be returned as a result). Moreover, the physicians are free to construct and search for multiple rules at a time.

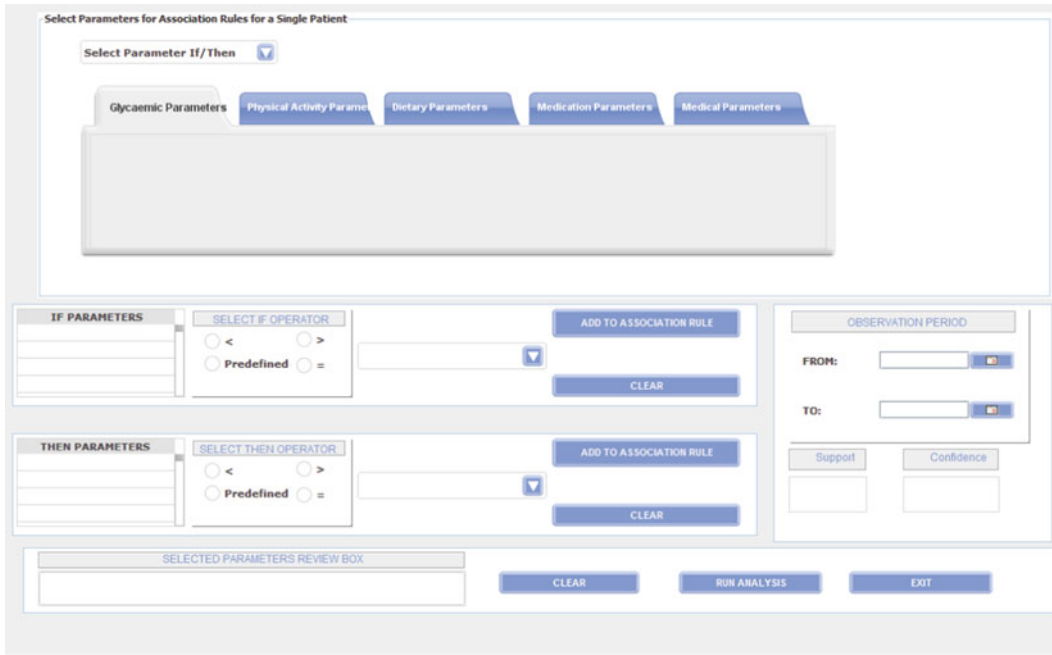

Fig. 6 Initial screen of association analysis for a patient

Figure 7 depicts how the physician can select one or more IF parameters, and thereafter using suitable operators to manually specify ranges of values, select predefined ones (according to the feedback received from the medical partners through the questionnaires) or leave blank (unspecified) in which case all possible combinations are sought. Additionally, in case of more than one IF parameters, the physicians can select and apply logical (Boolean) operators between them, such as AND, and NOT.

The physicians can select "THEN" parameters in a similar way and specify values, operators as shown in Fig. 8.

The physicians can also specify the observation period for which the analysis will be executed in the form of "from – to" time periods as shown in Fig. 9.

After the IF-THEN statement is constructed, the clinician is ready to run the analysis by pressing the "Run Analysis" button. The discovered rules (if any) are represented in a textual way in a new pop-window as shown in Fig. 10.

Some examples of IF-THEN statement for different cases of thresholds, support and confidences values along with the obtained results are provided below.

- IF "Physical Activity">5 METs AND "Glucose Before Activity"<100 mg/dl THEN "Glucose 2 h After Exercise"<60 mg/dl: confidence=70 %, support=5 %

  Result: Rule found (Yes/No)

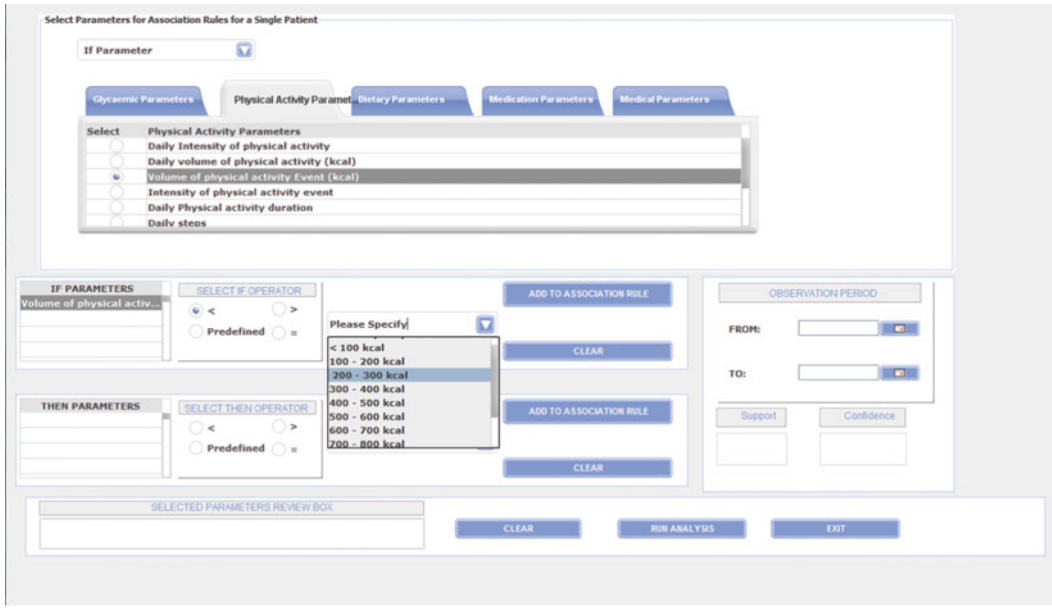

Fig. 7 Snapshot of selecting parameters from the physical activity tab for constructing the IF-part of the statement

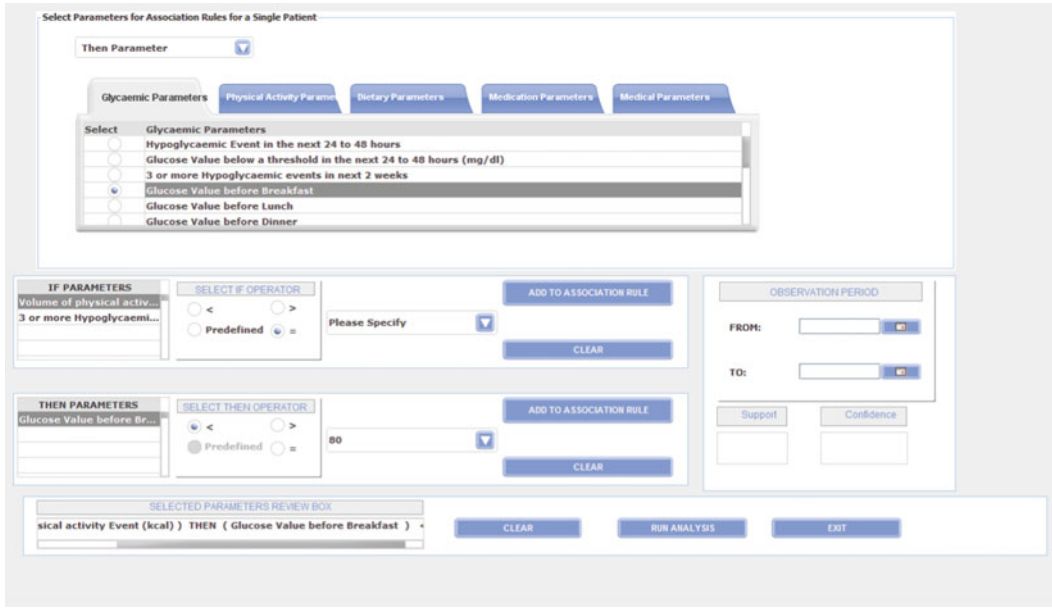

Fig. 8 Selection of THEN parameters

- IF "Lunch CHO">150 g THEN "Glucose Value 2 h after Lunch">130 mg/dl

  Result: Rule found (confidence 20 %, support 0.1 %)
- IF "Daily Steps" THEN "Hypoglecaemia": confidence 70 %, support 5 %

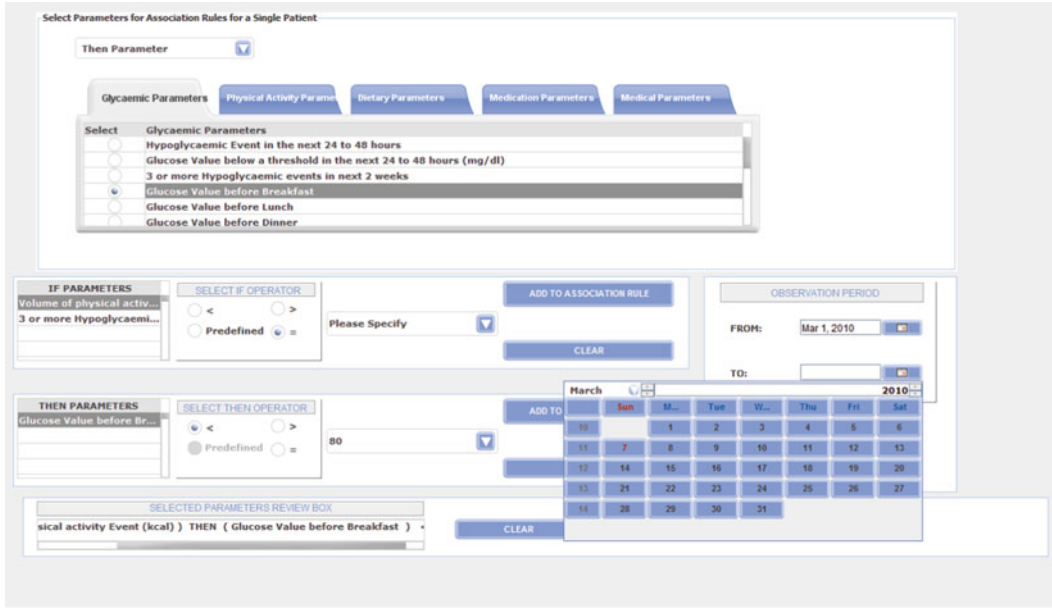

Fig. 9 Specification of observation period within which the association analysis will be executed

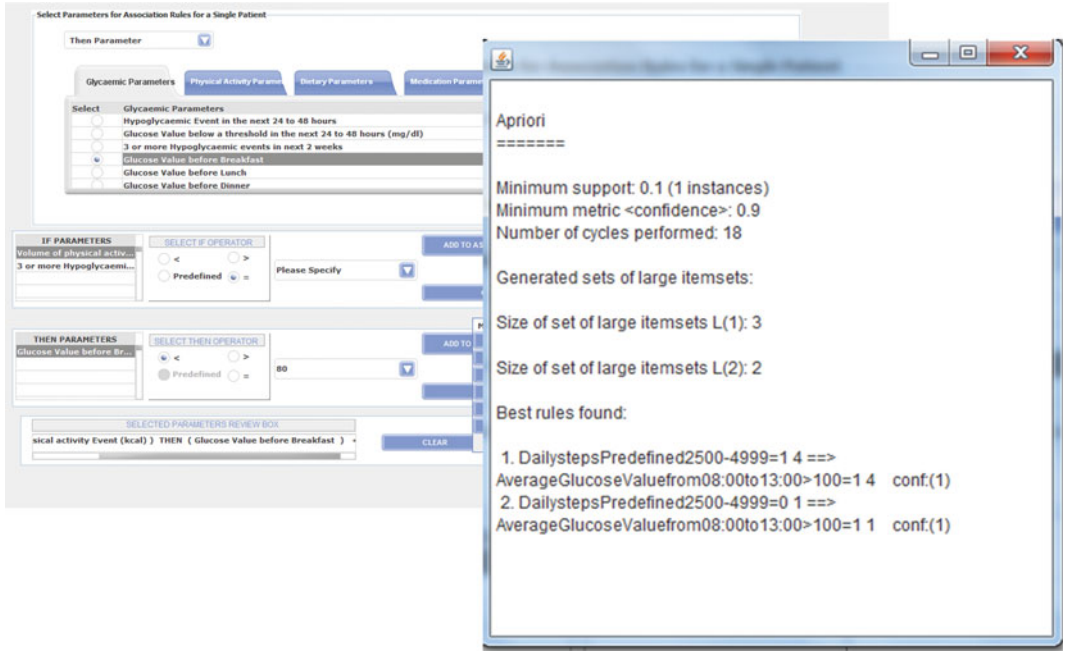

Fig. 10 Representation of the discovered rules along with their confidence and support values

```
Result: Best Rules found:
0-2,500, Hypoglycemia—No
2,500-4,999, Hypoglycemia—No
5,000-8,000, Hypoglycemia—No
```

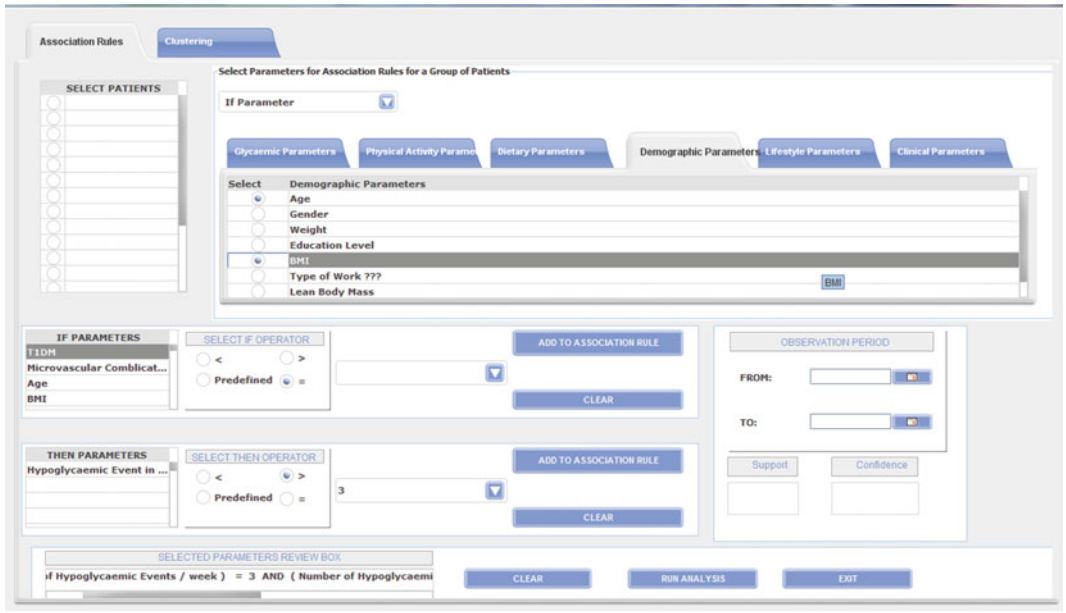

Fig. 11 The GUI for association analysis for multiple patients

8,000–9,999, Hypoglycemia—No

10,000–14,999, Hypoglycemia—No

>15,000 steps→Hypoglycemia (confidence 82 %, support 6 %)

4.3 Association Analysis for Multiple Patients

The functionality of the association analysis for multiple patients is similar to that of a single patient. As mentioned before, the types of parameters, i.e., demographic, clinical, and lifestyle are additional provided (Fig. 11). Also, there exists a list in which the patients to be included in the analysis are specified.

Two examples of IF-THEN statement for different cases of thresholds (with and without threshold for the IF parameter "Physical Activity Events/week") along with the obtained results are provided below.

- IF "Age">50 AND "Type 2" AND "BMI">30 AND "Lunch" AND "Physical Activity Events/week">3 THEN "Hyperglycaemic Events/Month"<3: confidence 70 %, support 5 %

  Result: Rule found (Yes/No)

- IF "Age">50 AND "Type 2" AND "BMI">30 AND "Lunch" AND "Physical Activity Events/week" THEN "Hyperglycaemic Events/Month"<3: confidence 70 %, support 5 %

  Result: Best Rules found:

```
Physical Activity Events/week: between 4 and
   5 (conf. 80 %, support 7 %)
Physical Activity Events/week: between 2 and
   3 confidence 72 %, support 5 %
```

5 Clustering Analysis

Clustering analysis allows the physicians to specify static (demographics, clinical, and lifestyle) and/or acquired (dietary, physical activity, and glycemic-related) features upon which clusters will be computed. The members of the created cluster share similar characteristics as described by the clusters' centroids. As opposed to association analysis that can be applied to any dataset that contains even few itemsets (records), clustering has clinical meaning only if sufficiently enough patients and data are obtained and for a medium/long term period.

As shown in Fig. 12, the physicians can select features from the corresponding tabs. In Clustering, no value ranges or thresholds are calculated and for this reason the logic of how these features were specified by the physicals through the questionnaires as well as the required pre-processing differs from the association analysis. Finally, the physicians can specify the number of clusters (and as such the computation of clusters is forced) or leave it unspecified. The specification of the number of clusters is something that can

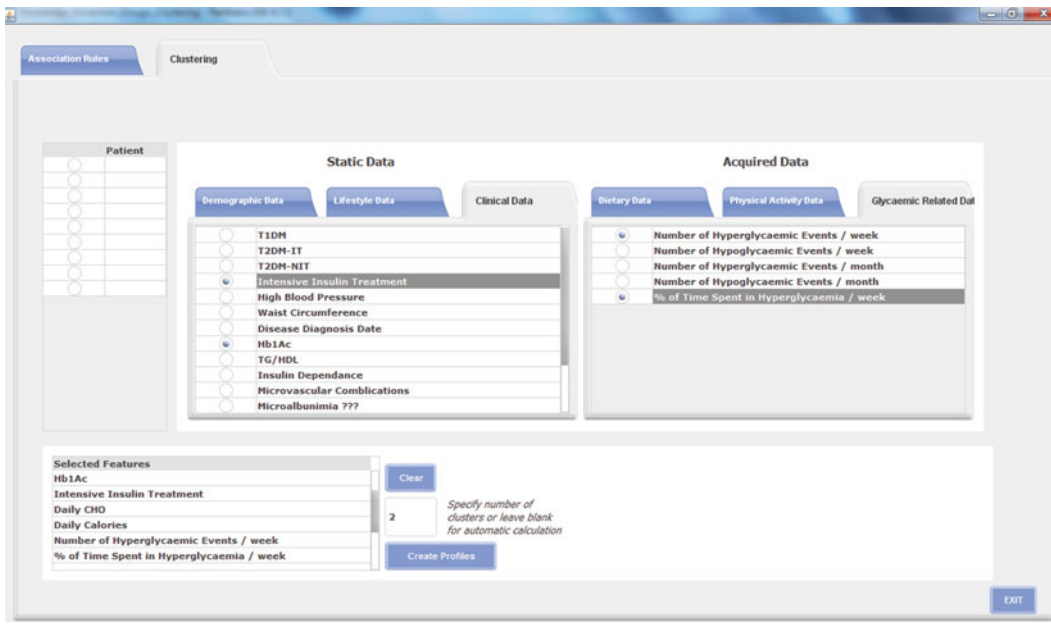

Fig. 12 GUI for specifying the features with which clusters will be created. The ability to specify or not the desired number of clusters is shown

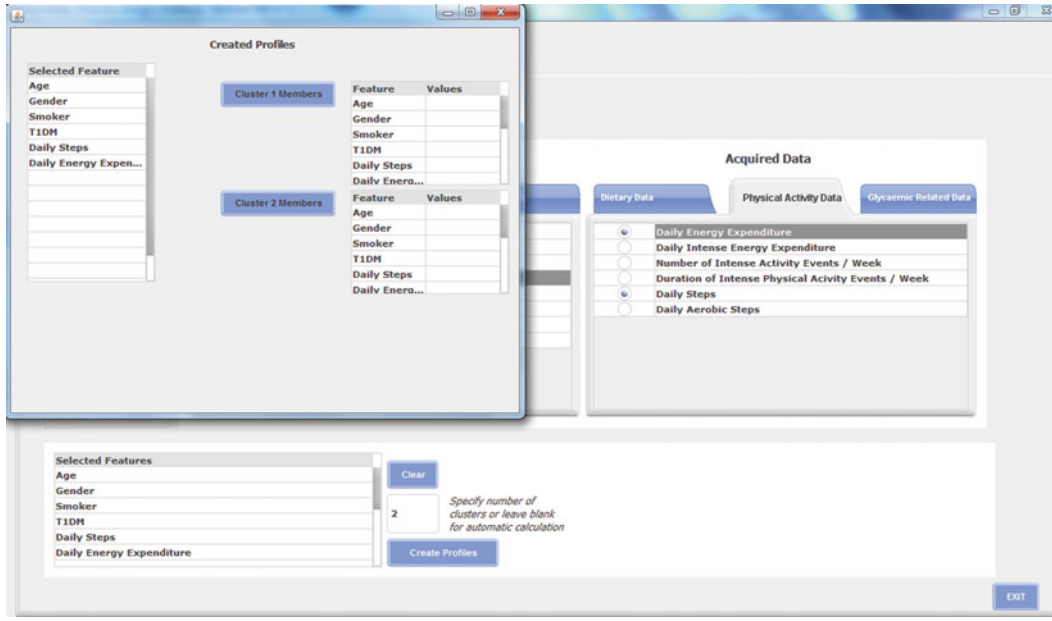

Fig. 13 Representation of the clustering report in which the members of each cluster are shown and the values of the input features

be performed in a trial-and-error manner or can be based on a priory medical knowledge.

Figure 13 shows the report that is generated after the clustering analysis is performed in which the computed clusters are shown together with their members and the values of their features.

6 Displaying Information to Users

6.1 Comprehensive View of Health Data

Data have been visualized in modular and multiple forms. This information is usually more relevant for T1DM than T2DM. T2DM users are more "guided" to insert data, receiving feedback if the treatment plan is not adhered, while T1DM are provided with specific graphs and tables for each module, as they are more aware about their disease and the support they need is focused on prompting them information in a way it helps them better understanding what is affecting their optimal glucose control. For this reason, information is displayed and combined in different ways (some examples are provided Fig. 14):

1. Physical Activity: A graph summarizing information about frequency, energy expenditure, total and aerobic steps over a certain period of time;
2. Blood Glucose: Values are presented depending on the period the user wants to visualize and against the thresholds set up by the doctor;

Fig. 14 Example of how information has been displayed in the PMD for each module: (**a**) bar graphs for physical activity levels; (**b**) food intake table displaying calories or CHO consumption versus daily and type of meal prescriptions; (**c**) drugs per time of intake; (**d**) diary-based presentation of information and cake graph presentation of blood glucose values depending on hypo-, hyper-, and normal glycemias

3. Drug Intakes: Insulin and medication intakes related to the moment of the day are shown in a table fashion.
4. Food Intakes: Bromatological composition of each meal is calculated and prompted when the patient is inserting the food intakes (available also before for simulating future meals). Meals can be reviewed in terms of calories and carbohydrate (CHO) content and compared to what doctors have prescribed for each day and for each kind of meal.

But the most important visualization is probably the comparative one, provided to both type of patients and displayed in Fig. 15: by selecting the date and the appropriate tabs, patients can visualize the evolution of different parameters on a daily/weekly/monthly basis. Although for a very novice user this graph might look complicated, for average users the options and the way of representation are friendly and comprehensive. In the comparative view, major events like insulin injections, physical activity events and CHO/calories of the meals are simultaneously represented and superimposed with the continuous glucose monitored ones: in this way, the patient can understand why he has experienced hypo/hyper/glycemic events or high variations, by associating them with

Fig. 15 Aggregation of physical activity, drug intake, continuous and discrete glucose reading, and food intake into a unified picture

the other parameters. This information is also displayed as a traditional diary (Fig. 14).

In the case of T2DM the feedback should support the user to insert the information more than for T1DM. Given the typical high complexity of their drug regimes, the insertion is done displaying only the drugs that are foreseen for the selected time, as agreed with doctors. Physical activities, food insertion and reading educational contents are promoted through the achievement of goals (reducing weight in a certain period of time and improving knowledge levels through quizzes) agreed with the treating professional. For this reason, food, physical activity are combined and represented with respect to the achievement of weight reduction goals, as displayed in Fig. 16.

6.2 Doctor Prescription and Views

In the same screen, the treating professional can check the current status of the patient and refine the treatment consequently. The treatment has been defined as the combination of multiple sub-prescriptions that may range from drug regimens to self-monitoring of vital parameters (blood glucose monitoring that can be discrete or even continuous for T1DM and the most complex cases of T2DM), diet suggestions, physical activity prescriptions and education. The health professionals participating in the usability tests pointed out the need of:

1. Splitting data displayers in raw-data format (sorted by timestamp),
2. Grouping them according to the source (laboratory data, sensor data, manually recorded data),

Fig. 16 Goal quest achievements for type 2 subjects

3. Combine them in tables, charts and
4. Complex graphs providing statistical-based glycemic trends, meal-oriented parameters, continuous glycaemia during nights.

7 Mining Data for the Evaluation of an Ehealth System

But how data can be grouped for extracting knowledge about the assessment of the performances an e-health system when it is used?

Different approaches exists for validating technologies but latest research in health technology assessment is demanding for more emphasis on including human factors in evaluating these systems. Moreover, especially when dealing with chronic diseases, concepts like treatment and adherence are complex and should not be reduced to measure and monitor drug regimes only, as it happens in the current clinical practice and health information systems, where lifestyle information is present, when it is present, in the "comments" or "notes" sections.

In our experience with diabetes, we defined treatment as the combination of multiple sub-prescriptions that may range from drug regimes, of course, to self-monitoring of vital parameters (depending on the disease), diet suggestions, physical activity prescriptions, and education.

These are generic concepts that can be applied to other chronic diseases. We hereby list the indicators used in diabetes, but the same methodology has been used for other chronic diseases such as cardiovascular, Parkinson, and cognitive diseases.

We have assessed the value of prescribing medical treatment as the multifactorial combination of:

- *Drug* intake prescription (insulin and/or oral antidiabetic drugs)
- *Diet* (usually focused in terms of CHO, Kcal, Glucose Indexes composition)
- *Physical Activity* (number of times per week, duration and intensity of exercise sessions)
- *Measurements* (number of times per day/week/month the patient is asked to perform self-monitoring activities like BG readings, BP measurements, weight assessment, insulin/OAD intakes, etc.)
- *Education*, organized in topics and difficulty level (depending on the person scholar level and her knowledge about the disease)

The result is that treatment compliance can be measured, taking into account the support of ICT tools, as the combination of the compliance to each of the prescribed sub-treatment. With this flexible method one can use specific sub-indicators, part of them or all of them. In our pilot we have been assessing diabetic patients' compliance as indicated below. Also, we are measuring subjective parameters related with user experience, through literature-based and customized questionnaires (Fig. 17).

How this data can be mined is the challenge,[1] for instance, compliance can be evaluated properly if it is combined with at least usage and data sent (Fig. 18).

In the case of the Control Panel application, Fig. 19, the learning curve shows that after a first usage of 10 min per patient, the average decreased to 5 min. This is a positive value if one takes into account that this minutes corresponds to the desired behavior that doctors expressed in the Good Practices Adoption questionnaire.

8 Conclusions

Using data mining to deal with the avalanche of clinical data collected from patients and generated from management of diabetes is a valuable asset that can help researchers and clinicians to provide better health care for the patients affected by this modern-society disease. Data-mining techniques are becoming more widely used in the field of diabetes, confirming that has a good future and will be used more and more in the area of diabetes in particular.

[1] Results are currently under analysis.

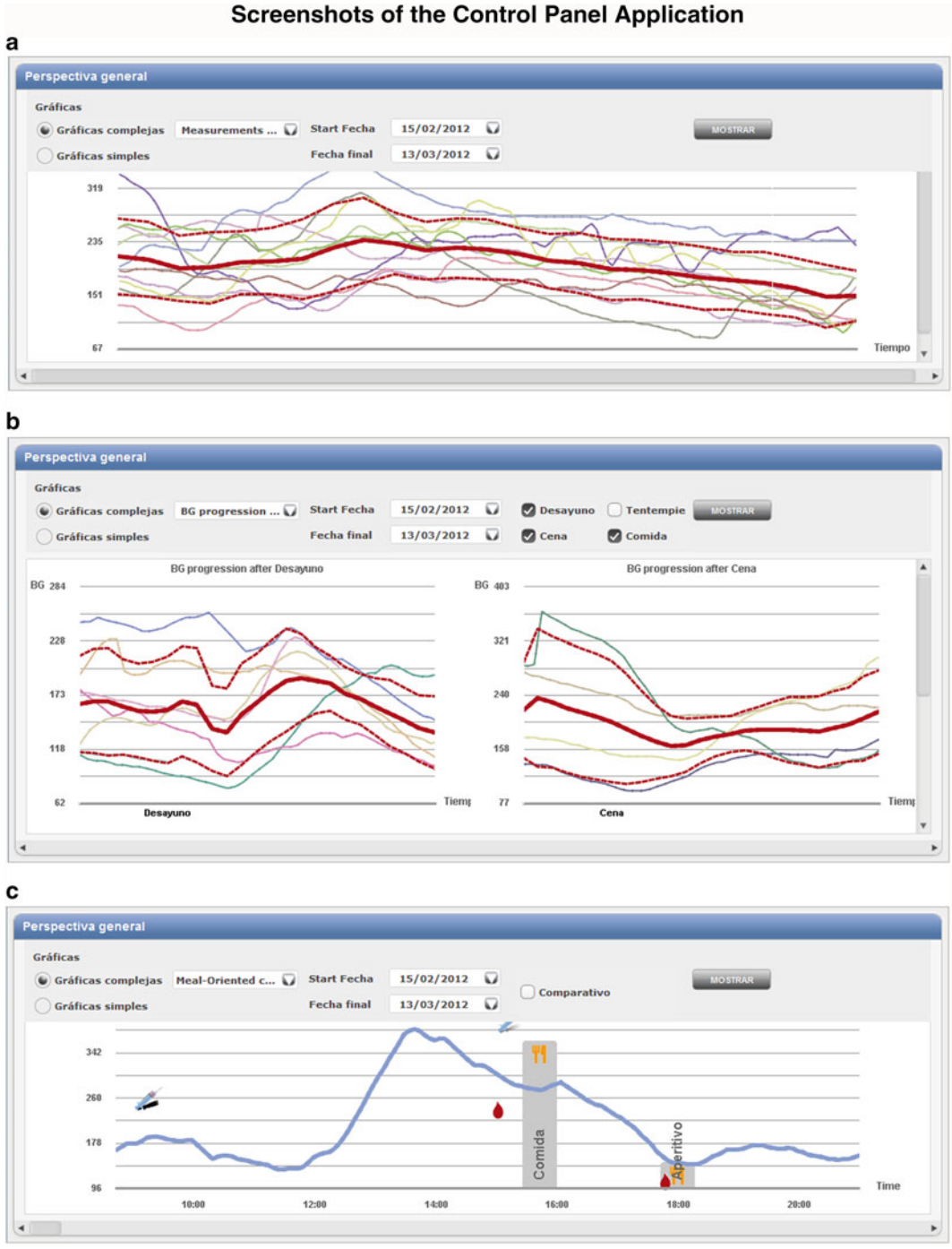

Fig. 17 CP data visualization module. (**a**) Continuous glucose progressions during nights, (**b**) around meals and (**c**) aggregated with food and drugs intakes depending on the patient, TP can visualize the information in multiple ways. For instance, in the case of T1DM she needs to visualize complex graphs, while in the case of T2DM she will use the tables, looking for indications about the compliance to drug, food, and the input frequencies and more focus on feedback and messages

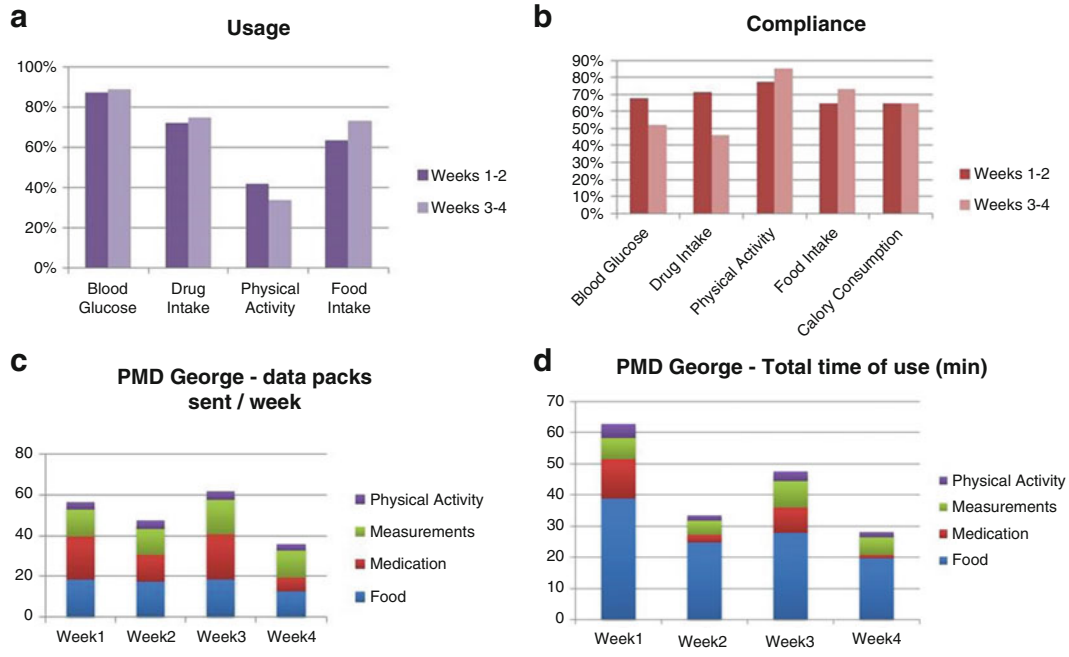

Fig. 18 Analysis of patient adherence, (**a**) based on system usage, (**b**) compliance, (**c**) data sent, and (**d**) time of use per module

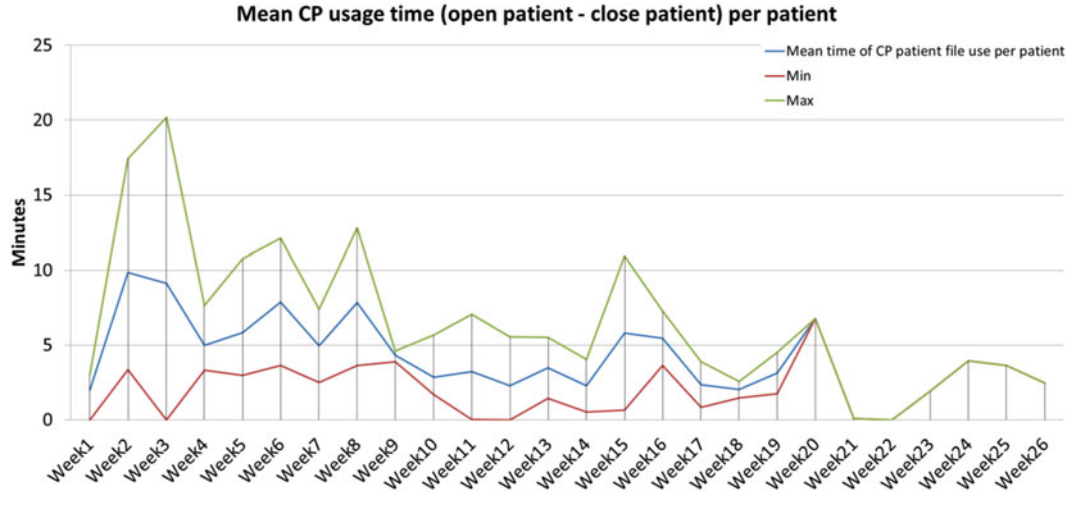

Fig. 19 Control panel learning curve

References

1. Abdelsalam H et al (2009) Smart home-based health. J Diabetes Sci Technol 3(1):141–148
2. Synnot J (2012) Flexible and customizable visualization of data generated within intelligent environments. Conf Proc IEEE Eng Med Biol Soc 2012:5819–5822. doi:10.1109/EMBC.2012.6347317
3. Shahar Y et al (2003) Interactive visualization and exploration of time-oriented clinical data using a distributed temporal-abstraction architecture. AMIA Annu Symp Proc 2003: 1004
4. Stacey D, Bennett CL, Barry MJ et al. (2011) Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev 10: CD001431
5. The salzburg statement on shared decision making, Salzburg Global Seminar, 7th February 2011
6. Marinov M, Mohammad Mosa AS, Yoo I, Austin BS (2011) Data-mining technologies for diabetes: a systematic review. J Diabetes Sci Technol 5(6):1549–1556
7. Bellazzi R, Abu-Hanna A (2009) Data mining technologies for blood glucose and diabetes management. J Diabetes Sci Technol 3(3):603–612
8. Bellazzi R, Magni P, Larizza C, De Nicolao G, Riva A, Stefanelli M (1998) Mining biomedical time series by combining structural analysis and temporal abstractions. Proc AMIA Symp 1998: 160–164
9. Breault JL, Goodall CR, Fos PJ (2002) Data mining a diabetic data warehouse. Artif Intell Med 26(1–2):37–54
10. Huang Y, McCullagh P, Black N, Harper R (2007) Feature selection and classification model construction on type 2 diabetic patients' data. Artif Intell Med 41(3):251–262
11. Miyaki K, Takei I, Watanabe K, Nakashima H, Omae K (2002) Novel statistical classification model of type 2 diabetes mellitus patients for tailor-made prevention using data mining algorithm. J Epidemiol 12(3):243–248
12. Brown AC, Olver WI, Donnelly CJ, May ME, Naggert JK, Shaffer DJ, Roopenian DC (2005) Searching QTL by gene expression: analysis of diabesity. BMC Genet 6:12
13. Covani U, Marconcini S, Derchi G, Barone A, Giacomelli L (2009) Relationship between human periodontitis and type 2 diabetes at a genomic level: a data-mining study. J Periodontol 80(8):1265–1273
14. DuMouchel W, Fram D, Yang X, Mahmoud RA, Grogg AL, Engelhart L, Ramaswamy K (2008) Antipsychotics, glycemic disorders, and life-threatening diabetic events: a Bayesian data-mining analysis of the FDA adverse event reporting system (1968-2004). Ann Clin Psychiatry 20(1):21–31
15. Liou FM, Tang YC, Chen JY (2008) Detecting hospital fraud and claim abuse through diabetic outpatient services. Health Care Manag Sci 11(4):353–358
16. Richards G, Rayward-Smith VJ, Sönksen PH, Carey S, Weng C (2001) Data mining for indicators of early mortality in a database of clinical records. Artif Intell Med 22(3): 215–231
17. Toussi M, Lamy JB, Le Toumelin P, Venot A (2009) Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes. BMC Med Inform Decis Mak 9:28
18. European Commission. Information Society Technologies Program. METABO project. Chronic diseases related to metabolic disorders. ICT-26270. www.metabo-eu.org
19. America Diabetes Association (2008) Standards of medical care in diabetes-2008. Diabetes Care 31(1):S12–S54
20. Carey, M. P., Jorgensen, R. S., Weinstock, R. S., Sprafkin, R. P., Lantinga, L. J., Carnrike, C. L. M., Jr., Baker, M. T., & Meisler, A. W. (1991). Reliability and validity of the appraisal of diabetes scale. Journal of Behavioral Medicine, 14, 43–51.
21. Davis, F.D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly, September 1989, 319–339.
22. Hassenzahl, M., Burmester, M., Koller, F. (2003). AttrakDiff.: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In J. Ziegler, & G. Szwillus, (Hrsg.), Mensch & Computer 2003 (S. 187-196). Stuttgart: B. G. Teubner.
23. Boy, O., Ohmann, C., Aust, B., Eich, H.P., Koller, M., Knode, O., Nolte, U. (2000). Systematische Evaluierung der Anwenderzufriedenheit von Ärzten mit einem Krankenhausinformationssystem – Erste Ergebnisse. In: Hasman, A. et al. (Eds.). Medical Infobahn for Europe: proceedings of MIE2000 and GMD2000. IOS Press. Pp. 518-522

Chapter 14

Data Analysis in Cardiac Arrhythmias

Miguel Rodrigo, Jorge Pedrón-Torecilla, Ismael Hernández, Alejandro Liberos, Andreu M. Climent, and María S. Guillem

Abstract

Cardiac arrhythmias are an increasingly present in developed countries and represent a major health and economic burden. The occurrence of cardiac arrhythmias is closely linked to the electrical function of the heart. Consequently, the analysis of the electrical signal generated by the heart tissue, either recorded invasively or noninvasively, provides valuable information for the study of cardiac arrhythmias. In this chapter, novel cardiac signal analysis techniques that allow the study and diagnosis of cardiac arrhythmias are described, with emphasis on cardiac mapping which allows for spatiotemporal analysis of cardiac signals.

Cardiac mapping can serve as a diagnostic tool by recording cardiac signals either in close contact to the heart tissue or noninvasively from the body surface, and allows the identification of cardiac sites responsible of the development or maintenance of arrhythmias. Cardiac mapping can also be used for research in cardiac arrhythmias in order to understand their mechanisms. For this purpose, both synthetic signals generated by computer simulations and animal experimental models allow for more controlled physiological conditions and complete access to the organ.

Key words Data analysis, Cardiac arrhythmias, Signal analysis, Cardiac simulation, Cardiac alternans, Electrocardiogram, Noninvasive imaging, Inverse problem of electrocardiography, Body surface potential mapping

1 Introduction

1.1 The Electrical System of the Heart

The heart is muscular viscera located in the center of the thoracic cavity. Its main function is to pump blood throughout the body by rhythmic contractions allowing a continuous flow of blood. The contractions of the heart occur as a consequence of the electrical activation of myocardial cells, which experiment an increase in their intracellular potentials. This change of the potential, called action potential, is generated by a sequence of ion fluxes through specific ion channels located in the membrane of myocardial cells.

In a healthy heart, activation is initiated by a stimulus generated at the sinoatrial node. This small mass of cells with pacemaker properties is located in the wall of the right atrium and has the ability to auto generate an action potential. Electrical impulses are

Carlos Fernández-Llatas and Juan Miguel García-Gómez (eds.), *Data Mining in Clinical Medicine*, Methods in Molecular Biology, vol. 1246, DOI 10.1007/978-1-4939-1985-7_14, © Springer Science+Business Media New York 2015

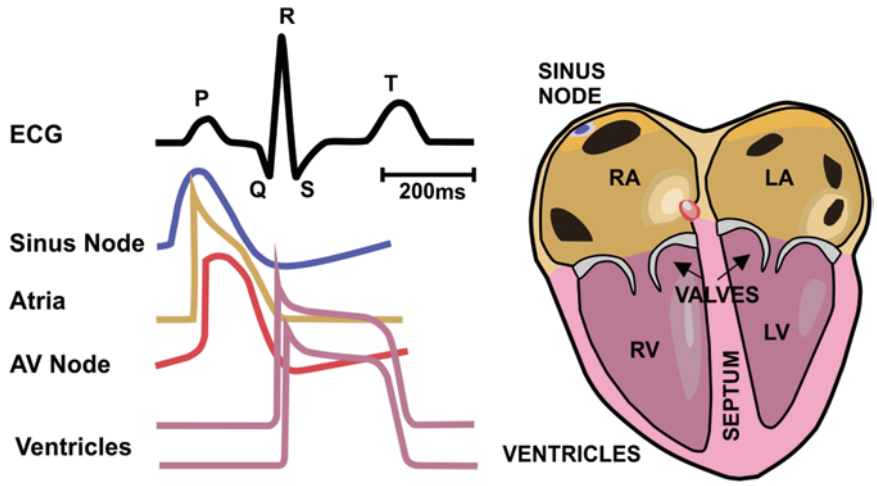

Fig. 1 Electrocardiogram (ECG) and action potential signals (*left*) from the different parts of the heart (*right*). The ECG signal (*black*) is formed by adding the action potential signals (*blue* to *purple*) from the different parts of the heart

then propagated through the atrial wall allowing the atria to contract and arrive to the atrioventricular node, delaying the electrical propagation before activating the ventricular muscle promoting its coordinate contraction.

The electrical activity of the heart can be measured indirectly on the surface of the torso by recording an electrocardiogram (ECG), as depicted in Fig. 1. An ECG tracing shows several deflections or waves that reflect the electrical activity of heart chambers. The so-called P wave corresponds to the electrical activation of the atria or depolarization, term used to refer that the cell leaves its resting or polarized state during the development of an action potential. After the P wave, the QRS complex reflects the ventricular depolarization, whereas the T wave corresponds to the ventricular repolarization or return to the resting state [1].

1.2 Heart Arrhythmias

As we have previously discussed, the mechanical activity of the heart is synchronized by its electrical activity. When the electrical activation of the heart does not follow a natural spontaneous activation of the sinus node at 60–120 beats per minute (the so-called sinus rhythm), the heart rhythm is referred as a cardiac arrhythmia. Cardiac arrhythmias can be classified as supraventricular arrhythmias, when they involve heart structures above the atrioventricular node or as ventricular arrhythmias, when they mainly involve the ventricles. While supraventricular arrhythmias mainly decrease the quality of life of patients, ventricular arrhythmias are lethal. There are several mechanisms that are involved in the development of cardiac arrhythmias. The most frequent cause for the development of arrhythmias is the modification of the myocardial substrate as a

consequence of a prolonged ischemia, or lack of oxygen that may result in necrosis of myocardial cells. This modified substrate favors the electrical impulse to reenter and if this reentry perpetuates it leads to an arrhythmia. Modification of the myocardial substrate can also be a consequence of aging and development of fibrosis, which can be favored by genetic factors. Finally, mutations of cardiac channels can also cause a modification of cardiac action potentials that may lead to an increased susceptibility to cardiac arrhythmias. We will introduce some cardiac arrhythmias whose diagnosis can be improved by applying signal processing to cardiac electrical signals.

Most atrial arrhythmias are caused by a macro or micro-reentry of the electrical wavefront in the atria. Atrial flutter is by a macro-reentry around an anatomical obstacle in the atria. Typically, this anatomical structure is a heart valve, most commonly the tricuspid valve, which is referred as typical atrial flutter [2]. This electrical behavior in the right atrium is reflected on the ECG by a continuous "sawtooth" pattern between QRST complexes that replaces the P wave which is present during sinus rhythm (*see* Fig. 2). Atrial fibrillation (AF) is the most common supraventricular arrhythmia and is characterized by a chaotic electrical activation of the atria,

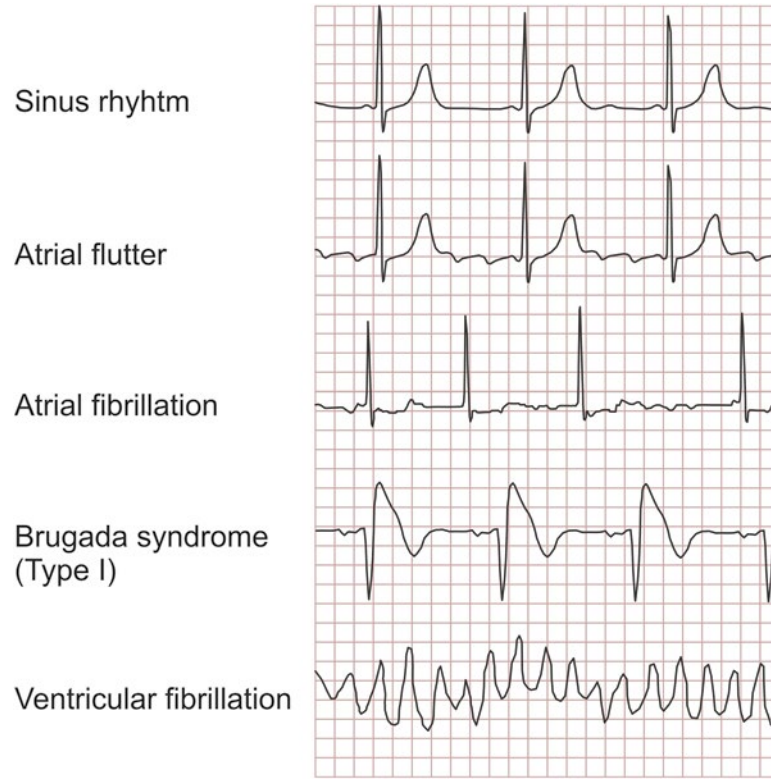

Fig. 2 ECGs (Lead-I) from different cardiac arrhythmias

which causes an inefficient blood pumping to the ventricles and increases the risk of embolism due to blood statis. The mechanisms underlying the initiation and maintenance of AF are not completely understood and there are several theories that partially explain its clinical manifestation: multiple random propagating wavelets, focal electrical discharges or functional reentrant activity with fibrillatory conduction [3]. During AF the P wave is absent and substituted by low amplitude and irregular atrial waves (Fig. 2). The shape of the QRS complex is not altered during atrial arrhythmias, since the ventricles are activated via the atrioventricular node, which blocks some of the activations and allows the ventricles to activate at a slower rate than the atria. However, the ventricular rate does not follow the activation of the sinus node and thus does not adapt to patient's activity, being usually too fast during atrial flutter and irregular during atrial fibrillation.

The most lethal ventricular arrhythmia is ventricular fibrillation (VF), which causes sudden death after a few minutes and is caused by a reentrant activation pattern on the ventricles. VF is typically caused by reentry at anatomical obstacles caused by ischemia or infarction. As opposed to AF, VF is a lethal arrhythmia, since results in an ineffective contraction of the ventricles. ECG tracings during VF present a chaotic signal resulting from the disorganized activity of the ventricles and PQRST complexes cannot be observed (Fig. 2). Among non-ischemic causes of VF, mutations in the genes encoding cardiac channels account for 5 % sudden cardiac deaths in infants and young adults. Brugada syndrome (BrS), which has been linked to mutations in the gene that encodes the sodium in, is a heritable arrhythmia syndrome that causes sudden death in young adults with structurally normal hearts [4]. BrS is diagnosed on the basis of the clinical and familiar history of the patient and sometimes they present a characteristic ECG pattern displaying a coved-type ST segment $\geq$0.2 mV in right precordial leads (referred as type I ECG) (Fig. 2). However, the clinical manifestation is often dynamic and shows variations over time. ECGs in BrS may also present a saddleback-type ECG or present any abnormality, which are not considered as diagnosis unless converted to a type I ECG after administration of ajmaline or flecainide [5].

2 Analysis of Invasive Cardiac Data

In clinical cardiology, electrophysiologists often need to explore the electrical activity of the heart by directly placing electrodes in close contact with the cardiac muscle in order to treat arrhythmias whose origin cannot be determined non-invasively. Moreover, invasive recordings, termed electrograms (EGM), allow greatly improve the diagnosis of cardiac arrhythmias in which it is very important to know the local electrical activity with anatomical

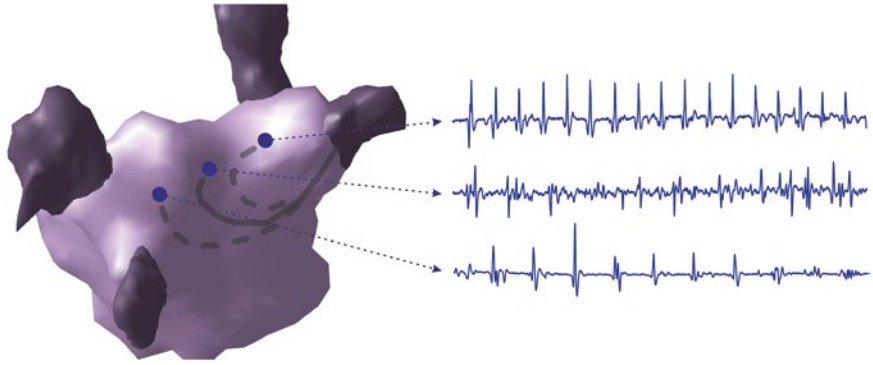

Fig. 3 Anatomical representation of the left atrium (*left*) and intracardiac recorded signals (*right*). The electroanatomic exploration of the heart chamber allows obtaining the local electrical activity of the heart wall

accuracy, as from this anatomical diagnosis will depend the subsequent treatment. In order to reach the heart, a catheter is inserted via a large vessel to the cardiac chamber which to be explored. The most frequently used technique makes use of a catheter with a few poles that allow obtaining the electrical activity at different locations of the wall chamber sequentially. The recording system can also allow determining the location of the recording catheter inside the cardiac chamber and reconstruct its electrical activity on top of the anatomy of the heart, technique which is referred to as electroanatomical mapping (*see* Fig. 3) [6–10]. There are also mapping systems that enable the recording of the electrical activity of the complete chamber simultaneously, by using catheters with multiple electrodes [11, 12]. However, these catheters are not commonly used in clinical practice because of the complications associated to their size.

Electroanatomical mapping can be performed in order to determine the re-entrant circuit responsible of the maintenance of atrial flutters and produce a lesion in the tissue by radiofrequency or crioablation [13–15] which can interrupting the reentrant activity. Electroanatomical mapping is especially useful as a guide to ablation strategies in atypical flutters [13] and is increasingly used for the determination of the mechanism responsible of the maintenance of atrial fibrillation [7, 10].

2.1 Analysis Methods for Electroanatomic Cardiac Data

Each commercial system for electroanatomical mapping includes different software packages for analysis of the recorded electrical signals, but they all share the same graphical display, plotting the measured parameter according to a color scale on top of the reconstructed anatomy. Measured parameters include activation times, dominant frequency, fractionation of the recorded signals, and their amplitude. Election of the parameter to measure depends on the clinical characteristics of the patient.

By computing activation times in recorded signals referenced to a common signal acquired at the same site during the mapping procedure, isochronal maps can be reconstructed. Activation times are determined as those at which the recorded bipolar EGM presents its maximum amplitude. Isochronal maps allow determining the activation pattern in the atria when it is stable over time and thus allow obtaining the location of the re-entrant circuit during atrial flutter [13].

More complex cardiac arrhythmias, such as atrial fibrillation, require different analysis methods since the activation pattern is unstable and activation times cannot be reproducibly determined. Analysis of mapping data during AF primarily focus on the identification of atrial sites involved in the maintenance of the arrhythmic activity [16, 17] since ablation of these sites often result in termination of the arrhythmia [11].

Dominant frequency (DF) mapping aims at determining the location of atrial sites activated at a higher rate, under the assumption that these sites are the primary sources that activate the rest of the atria [7]. The dominant frequency of each signal is obtained by using the Fast Fourier Transform or the Welch Periodogram on the recorded signals, which is inversely related to the activation rate. Ablation of highest DF sites has been proven to be an effective method for terminating AF in patients in which a gradient of DFs can be found [7].

Other analysis method used to guide ablation procedures in atrial fibrillation patients is the Complex Fractionation Atrial Rate mapping [8]. This methodology aims at quantifying the fractionation of the atrial signal based on the assumption that fractionated signals corresponds to areas with slow and irregular conduction which may be responsible of the maintenance of re-entrant circuits. There are different methodologies to obtain the fractionation level of EGMs, but the most widely used are based on the measure of the delay between consecutive deflections of the signal. However, although it has been proved that the primary generators cause fractional activity, there are other mechanisms that cause fractionation, and thus guidance of ablation procedures based on fractionation of EGMs is controversial [18–20]. Amplitude of EGMs can also serve also as guidance for ablation procedures since they have been proven to allow identifying scar and fibrotic areas in the atria, which favor the existence re-entrant patterns [21, 22].

Recently, Granger Causality, a statistical concept based on the prediction, has also been used to analyze intracardiac signals during AF. This concept has been applied to various analytical techniques of biomedical signals, both in the heart [12] and in the brain [23].

Causal theory has been presented as a useful tool in detecting the propagation patterns of cardiac electrical activity, since the tissue area that is acting as a source of the electrical activity can be identified. In this case the location of the primary mechanisms is

based on relations between adjacent atrial areas instead of single recordings, and thus a more global view of the electrical activity is achieved [12, 24–26]. Although this method has been proved effective in experimental studies, it has not been introduced in clinical practice yet.

3 Body Surface Potential Mapping

Diagnosis of cardiac arrhythmias can be approached noninvasively by using the surface electrocardiogram (ECG). ECGs acquired in clinical practice typically consist of 12 signals obtained from nine electrodes located at standard positions on the body surface and allow the diagnosis of most cardiac diseases. However, the recording of more ECG leads on the surface of the torso can improve the diagnosis of specific cardiac conditions. For years it was thought that the standard 12-lead ECGs recordings contained enough electrical information from the body surface which could be used for clinical diagnosis, and that the information obtained with more electrodes was redundant and was only a linear combination of the standard 12-leads, based on the assumption that the myocardial electrical activity can be modeled as a single dipole. However, under some clinical circumstances this assumption of dipolarity of the cardiac electric field is not accurate, such as during fibrillation in which multiple simultaneous propagation wavefronts travel with different directions.

The recording of multiple ECG leads from the surface of the torso is termed Body Surface Potential Mapping (BSPM). There is no standard or number of electrodes—from 64 to 256 electrodes—or their location, but ECG leads are distributed around the torso surface in order to capture all the electrical information of the heart available on the body surface. Several studies have reported improved diagnosis by using BSPM. Bruns et al. [27], by analyzing the QRS integral and the gradient of the ST segment in BSPM recordings elucidated that upper right precordial leads allowed an improved detection of ECG markers of Brugada syndrome. Eckardt et al. [28] demonstrated that the presence of late potentials and ST segment elevation on BSPM recordings allowed the prediction of the inducibility of ventricular tachyarrhythmias in Brugada syndrome patients. Dubuc et al. [29] used isopotential maps from BSPM recordings for the localization of accessory pathways in Wolf–Parkinson–White patients to guide catheter ablation. In the field of diagnosis of atrial arrhythmias, SippensGroenewegen et al. showed that the origin of focal tachycardias [30] or the reentrant circuit during atrial flutter [31] could be determined by analysis of BSPM maps. Guillem et al. introduced the wavefront propagation maps as a technique to summarize the electrical propagation pattern into a single map during atrial flutter [32] and to characterize the organization degree during atrial fibrillation [33].

Guillem et al. also showed that spatial gradients of activation frequency in human AF can be detected on BSPM recordings by computing the power spectral density of multiple surface leads [34], which cannot be accomplished by using the standard 12-lead ECG .

In spite of the improved diagnostic accuracy of BSPM, its use is still restricted to research studies given the complexity of data interpretation. Techniques for computing and displaying isochronal maps developed for summarizing the propagation pattern the torso into a single image or the estimation of surface dominant frequencies may help in the interpretation of BSPM maps but still, validation of these results need to be extended to larger population cohorts before introduction of BSPM into the clinical practice.

4 Noninvasive Imaging of the Myocardial Electrical Activity

One approach that has been widely discussed in the literature that aims to bring BSPM data into the clinical practice is the noninvasive reconstruction of the electrical activity in the heart by using BSPM recordings. This computation can be undertaken by solving the so-called inverse problem of the electrocardiography [35, 36] that departs from the mathematical formalism employed for computing surface potentials from epicardial potentials, formalism known as forward problem of the electrocardiography (*see* Fig. 4).

4.1 The Forward and Inverse Problems of the Electrocardiography

The forward problem of the electrocardiography consists in the calculation of the potential distribution on the surface of the torso departing from the electrical activity of the heart and a geometrical model of the heart and torso (*see* Fig. 5), which can be obtained from segmentation of computer axial tomography or magnetic resonance images from patients.

Potentials on the surface of the torso can be computed from potentials on the heart surface by using the Boundary Element Method (BEM) formulation [37–41], as shown in Eqs. 1–3:

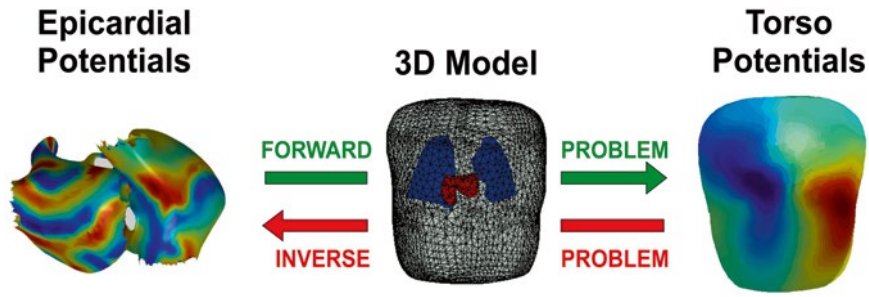

Fig. 4 Forward and inverse problem. The forward problem obtains the ECG signals by using the intracardiac electrical activity and the torso model. The inverse problem obtains the intracardiac electrical activity from the ECGs and the torso model

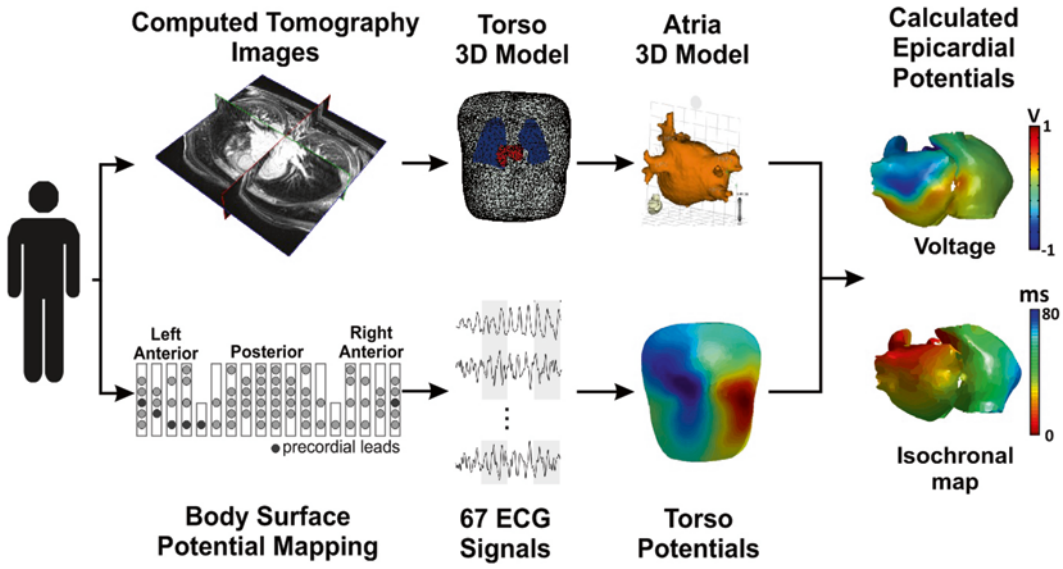

Fig. 5 Clinical application of the inverse problem of the electrocardiography for noninvasive atrial imaging. The models of the torso and the heart chambers are constructed from the Computed Tomography Images, and the ECG is recorded by using a multi-lead recording system. By applying the mathematical method of the inverse problem, the intracardiac electrical activity of the heart is obtained from the ECG signals and the torso model

$$A_1 x = b \tag{1}$$

$$A_1 = \begin{pmatrix} P_{HH(nxn)} & G_{HH(nxn)} \\ P_{BH(mxn)} & G_{BH(mxn)} \end{pmatrix}, \quad x = \begin{pmatrix} \Phi_H \\ \Gamma_H \end{pmatrix}, \quad b = \begin{pmatrix} -P_{HB(nxm)}\Phi_B \\ -P_{BB(mxm)}\Phi_B \end{pmatrix} \tag{2}$$

$$\Phi_B = M\Phi_H = \left(D_{BB} - G_{BH}G_{HH}{}^{-1}D_{HB}\right)^{-1} \cdot \left(G_{BH}G_{HH}{}^{-1}D_{HH} - D_{BH}\right)\Phi_H \tag{3}$$

where Φ_H is the potential on the surface of the heart, Φ_B is the potential on the surface of the torso, Γ_H is the potential gradient of the heart, P_{XY} is the potential transfer matrix from point Y to point X, G_{XY} is the potential gradient transfer matrix from point Y to point X, and M is the full-transfer matrix.

The forward problem is well-conditioned and well-posed and has a unique solution.

On the other hand, the inverse problem of the electrocardiography consists in the calculation of the electrical activity of the heart from surface recordings. The inverse problem is ill-posed, it is very unstable and has multiple solutions, requiring regularization methods for its resolution [36, 39]. One of the most used regularization methods is Tikhonov's regularization [42, 43], which is formulated as a minimization problem (Eq. 4):

$$\min\left\{\left\|M\Phi_H - \Phi_B{}^2\right\| + \left\|\lambda B\Phi_H{}^2\right\|\right\} \tag{4}$$

where λ is a regularization parameter that can be obtained with the L-curve method [39, 44] and B is the spatial regularization matrix which can be the identity matrix (zero-order), a gradient matrix (first order) or a Laplacian matrix (second order). Therefore, the inverse problem can be solved by using the expression (Eq. 5)

$$\Phi_H(\lambda) = \left(M^t M + \lambda B^t B\right)^{-1} M^t \Phi_B \tag{5}$$

4.2 Applications of the Noninvasive Cardiac Imaging

Noninvasive imaging of the electrical activity of the myocardium by solving inverse problem present advantages over invasive techniques. First of all, it avoids the risks, high expenses time required for the procedure. In addition in contact mapping systems, multiple recordings are obtained sequentially instead of simultaneously and thus they offer multiple local information of the electrical activity of the heart whereas the global activity in the chamber it is not recorded at every time instant. With the noninvasive method it is possible to obtain electrical information at any time and any location in the myocardial wall.

The noninvasively determination of the electrical activity of the myocardium is a promising tool in clinical diagnosis, since it may allow locating and identifying specific sites in the myocardium with a particular electrical behavior prior to a clinical intervention, helping in planning clinical procedures. Solution of the inverse problem showed to serve as guidance for locating single or multiple ectopic beats, origin of atrial arrhythmias, by identifying the earliest activated region in inverse-computed isochronal maps [45] or the reentrant circuit involved in atrial flutter [46, 47]. Moreover, it could be potentially used for diagnosis of Brugada syndrome in order to confirm the involvement of the right ventricular outflow tract in the clinical manifestation of the syndrome, which is not always univocal [48].

5 Ex Vivo Models of Cardiac Arrhythmia

As seen above, invasive and noninvasive techniques offer advantages in the diagnosis of cardiac arrhythmias. However, these methods do not allow a complete recording of the cardiac activity of the entire heart. The understanding of arrhythmia mechanisms and the design of effective antiarrhythmic therapies require experimentation models with access to more physiological variables and a more complete access to the heart, and also information regarding the coupling between excitation and contraction of the heart [49]. Optical mapping technique has revolutionized this field due to its ability to provide simultaneous recordings of electrical activity and molecular dynamics from multiple sites over the surface of the heart at high spatiotemporal resolution [50].

6 Optical Mapping

Optical mapping uses sensitive fluorescent dyes which react to alterations in membrane potential (V_m) and ion concentrations such as intracellular free calcium ($[Ca^{2+}]_i$), being possible to observe multiple parameters simultaneously in the same preparation [51]. These experiments require a light source with an excitation wavelength specific to each type of dye used. For this purpose, light-emitting diodes (LEDs), offer a more effective and efficient alternative to other lighting sources [52]. Recording of each parameter (i.e., V_m or $[Ca^{2+}]_i$) is performed by high-speed low-noise photodetectors combined with filters at the emission band of the different dyes (*see* Fig. 6).

By obtaining optical mapping recordings it is possible to study interactions of V_m and $[Ca^{2+}]_i$ in different preparations such as single cells, cardiac monolayers or a whole heart. Moreover, optical mapping may be combined with multitude of stimulation protocols by varying the pacing protocol spatially and temporally, or applying drugs to characterize their effects [53].

Analysis of optical mapping recordings, which are composed of a large number of pixels with different activities, requires of visual representations to characterize the behavior of the whole heart. First, it is necessary a preprocessing to remove signal drift and fluorescence noise. Pixels not belonging to the tissue are removed by a user defined mask. Two-dimensional maps are developed to explore the spatial distribution of parameter of interest over the surface of preparation. Depending on the measured parameter may be useful to calculate activation time maps [52], dominant frequency maps (previously described) or phase maps [54, 55], among others. These representations have managed to essential findings about such phenomena as alternans, bidirectional tachycardia, atrial fibrillation (AF) and ventricular fibrillation (VF) [56, 57].

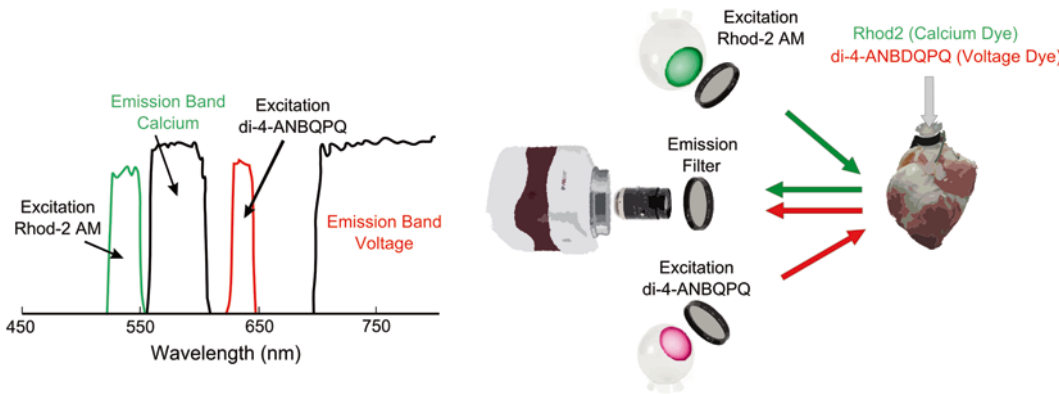

Fig. 6 Transmission spectrum of excitation and emission of voltage and calcium dye fluorescence (*left*). Schematic representation of the optical mapping system (*right*)

6.1 Phase Mapping

Phase mapping offers a representation of the time course of a propagating wave into a single map and offers an alternative to isochronal mapping. Phase can be computed in optical mapping signals by using Hilbert's phase transform [11], as shown in Eq. 6:

$$\text{Phase signal} = \angle\left(HT\left(\text{Electrogram}\right)\right) \tag{6}$$

where $\angle(\,)$ is the phase operator and $HT(\,)$ the Hilbert transform. By using this transform domain, functional reentries appear as sites with phase singularities, and thus can be easily identified [58]. Phase singularities are defined as the points on the map where all phases converge.

6.2 Analysis of Cardiac Alternans in Ex Vivo Models: Insights into Fibrillation Mechanisms

Arrhythmias, such as AF and VF, are due to a broken electrical wavefront causes a current reentry that stimulates new fronts of propagation [59]. However, the underlying mechanism of this process is not fully understood. Traditionally, it has been linked to the heterogeneity in tissue remodeling caused by heart disease, but the dynamic factors have recently gained importance as a precursor to this mechanism [60]. This is especially present in calcium dynamics, because it is the mediator of cellular contraction. During a normal contraction $[Ca^{2+}]_i$, is directly dependent on transmembrane potential, but this association is destabilized under pathological conditions.

The analysis of cardiac alternans, defined as to a beat-to-beat alternation in the action potential duration (APD) or intracellular calcium transient amplitude (CaA), focus on the role of calcium–voltage coupling in the initiation of arrhythmic processes.

Cardiac alternans can be classified as spatially concordant or discordant [61]. Alternans are spatially concordant if all regions of the tissue alternate in phase with each other, both APD and CaA, and are spatially discordant if some regions of tissue alternate in a long-short-long pattern, whereas other regions simultaneously alternate in a short-long-short pattern. Regions with different patterns are separated by a transition without alternans, termed nodal line [62].

In the other hand, V_m and $[Ca^{2+}]_i$ are bidirectional systems and variations in calcium can promote changes in voltage dynamics in cardiac tissue. In the V → Ca coupling, the L-type calcium current is the predominant factor of alternaces; so if APD alternates, the CaA will also alternate in response to the changes of the L-type calcium current. Conversely, in the Ca → V coupling, the CaA modulates APD due to its effect on Ca-sensitive currents during the action potential plateau. These couplings can be positive or negative because of Na–Ca exchange current. Positive coupling is present when a large CaA produces a long APD and negative coupling when a large CaA causes a short APD (*see* Fig. 7) [61, 62].

In order to quantify cardiac alternans on optical mapping recordings, each beat must be delineated by computing the activation time in all pixels. Alternating features are present in the action

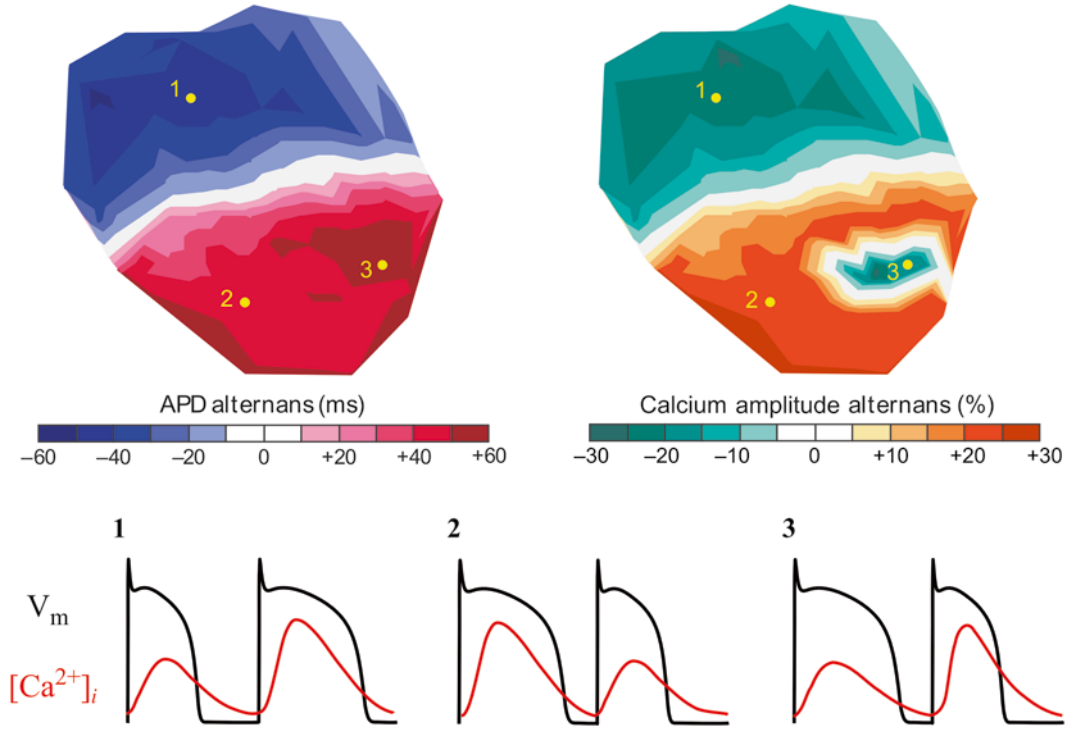

Fig. 7 Representation of dual optical mapping recording. Spatially discordants alternate for action potential duration (APD, *left*) and $[Ca^{2+}]_i$ (*right*) in a short-long pattern with positive coupling (*1*), in a long-short pattern with positive coupling (*2*) and negative coupling (*3*). Voltage and $[Ca^{2+}]_i$ signals (*down*) obtained with the optical mapping system are depicted for the locations marked in the maps

potential duration (APD) and intracellular calcium transient amplitude (CaA). APDs may be measured at 75 % repolarization (APD75) threshold using linear interpolation of samples [63], and CaAs is calculated as the difference between the local maxima and minima of a beat.

In order to characterize the distribution of APD alternans over the surface of the preparation, the difference between two consecutive APDs is defined in (Eq. 7):

$$\Delta \mathrm{APD}(x,y)_n = \mathrm{APD}(x,y)_{n+1} - \mathrm{APD}(x,y)_n \tag{7}$$

where n is the beat number and $APD(x,y)$ is the action potential duration at (x,y) coordinates in the 2D map.

The difference between two consecutive CaAs is calculated as the alternans ratio, as shown in Eq. 8:

$$\Delta \mathrm{CaA}(x,y)_n = \frac{\mathrm{CaA}(x,y)_{n+1} - \mathrm{CaA}(x,y)_n}{\max\left(\mathrm{CaA}(x,y)_{n+1}, \mathrm{CaA}(x,y)_n\right)} \tag{8}$$

Other 2D maps can be measured depending on signal quality and when the alternans pattern is masked by noise, average differences between pulses in the maximum continuous segment can be computed as estimation for the alternance degree [64]. In addition alternans analysis can be performed in other domains, such as the frequency or the phase domain [57].

The study of calcium dynamics in the heart is a new approach to know more about cardiac arrhythmias. This approach establishes that the arrhythmia is caused by the dynamic substrate that an ectopic beat encounters [65]. For this reason, antiarrhythmic strategies that attempt to inhibit premature contractions are not always effective, since it is necessary to take in consideration how these drugs affect the substrate.

In the future, the development of methods to detect alternans in the heart could serve as diagnostic tools to predict the vulnerability to lethal arrhythmias or to design more effective antiarrhythmic drugs.

7 Cardiac Simulation

The advantage of mathematical models applied to cardiac electrophysiology is that they allow controlling and monitoring all physiological parameters while avoiding technical, economical or ethical issues associated to in vivo, ex vivo, or in vitro experimentation.

The electrical activity of the entire heart can be computed by using an anatomically realistic structures modeled by finite elements, either volumetric [66–68] or surface [69] structures and a mathematical formulation for the cardiac electrophysiology. The structural elements in the geometrical model are nodes that model cells or groups of cells, and are connected into triangular [69] or square [70] faces for surface models or hexahedral or tetrahedral [68] in the case of volumetric models. Electrophysiology of cardiac cells is typically modeled by a set of differential equations based Hodgkin and Huxley formalism [71]. According to this formulation, the ionic currents, pumps and exchangers, together with the membrane capacity (C_m) and an externally applied current (I_{stim}) modulate the transmembrane potential as shown in Eq. 9:

$$\frac{dV_m}{dt} = -\frac{I_{ion} + I_{stim}}{C_m} \tag{9}$$

The whole tissue can be formulated in terms of a monodomain model, which assumes that cardiac tissue behaves as an excitable medium, with diffusion and local excitation of membrane voltage, according to Eq. 10:

$$\frac{\partial V_{\mathrm{m}}}{\partial t} = \nabla\left(D\nabla V_{\mathrm{m}}\right) - \frac{I_{\mathrm{ion}} + I_{\mathrm{stim}}}{C_{\mathrm{m}}} \tag{10}$$

where ∇ is the gradient operator, and D is a diffusion coefficient [72].

After discretization of the spatial derivatives for an isotropic medium, the voltage evolution of the transmembrane voltage of each cell (i.e., $V_{\mathrm{m},i}$ for the ith cell), is given by the following first-order, time-dependent ordinary differential equation:

$$\frac{\mathrm{d}V_{\mathrm{m},i}}{\mathrm{d}t} = \frac{I_{\mathrm{ion}} + I_{\mathrm{stim}}}{C_{\mathrm{m}}} - D\sum_{j}\frac{V_{\mathrm{m},i} - V_{\mathrm{m},j}}{d_{i,j}^2} \tag{11}$$

where $d_{i,j}$ is the distance between neighbor cells i and j.

Typically, each cell is modeled with 15–40 differential equations describing the behavior of the potential and other gating variables [65, 73, 74]. In other cardiac models [75–77], each cell is represented as a one-dimensional chain of sarcomeres, with associated Ca-cycling dynamics and Ca concentrations for each sarcomere. In this model, membrane voltage is determined by the average states of all sarcomeres, computed from all ionic membrane currents, and is thus synchronized over the entire cell [77]. Sarcomeres are coupled to one another by Ca diffusion and through bidirectional coupling with membrane voltage. By using this model, it is possible to simulate the existence of calcium and voltage alternans and reproduce experimental observations.

An accurate 3D simulation requires thousands to millions of cells or sarcomeres; hence, the required computing power is huge and simulation of a few seconds of cardiac activity requires hours or even days of computation in standard computers. Simulation time can be shortened by performing the resolution of the system in a Graphic Processing Unit (GPU). Although GPUs were designed for managing the Graphic User Interface, today the General Purpose GPU (GP-GPU) are mainly used for high-performance computations [78, 79].

7.1 Cardiac Simulation to Understand Surface Patterns During Arrhythmias

Cardiac models can be used to simulate cardiac arrhythmias and compute surface potentials associated to the simulated conditions. Surface potentials are computed by using simulated electrograms together with the forward problem of the electrocardiography [69, 80]. By using this approach, the effects on the ECG of alteration in ion channels caused by either genetic mutations or drugs, aging or ambient factors can be modeled and compared to observed ECGs in patients (*see* Fig. 8).

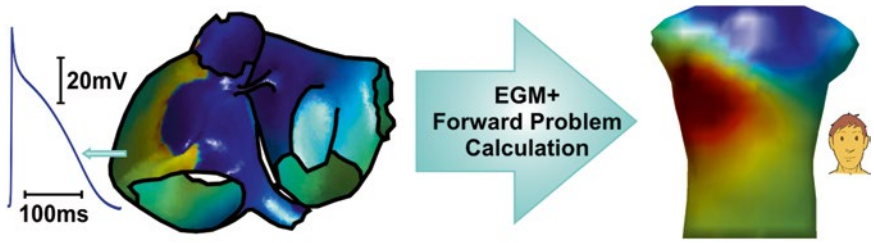

Fig. 8 Simulated action potentials in the atria (*left*) and torso electric potential (*right*) during atrial flutter. The forward problem is used to obtain the torso electric potential from the atrial electrical activity

8 Final Remarks

Computer analysis of cardiac signals helps in diagnosis, procedure planning and understanding of arrhythmia mechanisms. Given the large amount of data recorded, computer analysis is essential for interpretation of mapping data at all levels, from endocardial or epicardial mapping data to body surface mapping. Going one step further, computational models can be used to reproduce a wide range of pathologic conditions and validate mechanistic hypothesis of the observed signals in patients.

References

1. Malmivuo J, Plonsey R (1995) The heart. In: Malmivuo J, Plonsey R (eds) Bioelectromagnetism. Oxford University Press, New York, pp 119–132
2. Olgin JE, Kalman JM, Fitzpatrick AP et al (1995) Role of right atrial endocardial structures as barriers to conduction during human type I atrial flutter. Activation and entrainment mapping guided by intracardiac echocardiography. Circulation 92:1839–1848
3. Calkins H, Brugada J, Packer DL et al (2007) HRS/EHRA/ECAS expert consensus statement on catheter and surgical ablation of atrial fibrillation: recommendations for personnel, policy, procedures and follow-up. Europace 9:335–379
4. Antzelevitch C, Brugada P, Brugada J et al (2005) Brugada syndrome: from cell to bedside. Curr Probl Cardiol 30:9–54
5. Antzelevitch C, Brugada P, Borggrefe M et al (2005) Brugada syndrome: report of the second consensus conference: endorsed by the Heart Rhythm Society and the European Heart Rhythm Association. Circulation 111:659–670
6. Atienza F, Almendral J, Moreno J et al (2006) Activation of inward rectifier potassium channels accelerates atrial fibrillation in humans: evidence for a reentrant mechanism. Circulation 114:2434–2442
7. Atienza F, Almendral J, Jalife J et al (2009) Real-time dominant frequency mapping and ablation of dominant frequency sites in atrial fibrillation with left-to-right frequency gradients predicts long-term maintenance of sinus rhythm. Heart Rhythm 6:33–40
8. Nademanee K, McKenzie J, Kosar E et al (2004) A new approach for catheter ablation of atrial fibrillation: mapping of the electrophysiologic substrate. J Am Coll Cardiol 43:2044–2053
9. Sanders P, Berenfeld O, Hocini MZ et al (2005) Spectral analysis identifies sites of high-frequency activity maintaining atrial fibrillation in humans. Circulation 112:789–797
10. Atienza F, Calvo D, Almendral J et al (2011) Mechanisms of fractionated electrograms formation in the posterior left atrium during paroxysmal atrial fibrillation in humans. J Am Coll Cardiol 57:1081–1092
11. Narayan SM, Krummen DE, Shivkumar K et al (2012) Treatment of atrial fibrillation by the ablation of localized sources CONFIRM (conventional ablation for atrial fibrillation with or without focal impulse and rotor modulation) trial. J Am Coll Cardiol 60:628–636
12. Richter U, Faes L, Cristoforetti A et al (2011) A novel approach to propagation pattern analysis in intracardiac atrial fibrillation signals. Ann Biomed Eng 39:310–323

13. Morady F (1999) Radio-frequency ablation as treatment for cardiac arrhythmias. N Engl J Med 340:534–544
14. Isobe N, Taniguchi K, Oshima S et al (2004) Factors predicting success in cryoablation of the pulmonary veins in patients with chronic atrial fibrillation. Circ J 68:999–1003
15. Mack CA, Milla F, Ko W et al (2005) Surgical treatment of atrial fibrillation using argon-based cryoablation during concomitant cardiac procedures. Circulation 112:I1–I6
16. Haissaguerre M, Jais P, Shah DC et al (1998) Spontaneous initiation of atrial fibrillation by ectopic beats originating in the pulmonary veins. N Engl J Med 339:659–666
17. Jalife J, Berenfeld O, Mansour M (2002) Mother rotors and fibrillatory conduction: a mechanism of atrial fibrillation. Cardiovasc Res 54:204–216
18. Porter M, Spear W, Akar JG et al (2008) Prospective study of atrial fibrillation termination during ablation guided by automated detection of fractionated electrograms. J Cardiovasc Electrophysiol 19:613–620
19. Stiles MK, Brooks AG, John B et al (2008) The effect of electrogram duration on quantification of complex fractionated atrial electrograms and dominant frequency. J Cardiovasc Electrophysiol 19:252–258
20. Di Biase L, Elayi CS, Fahmy TS et al (2009) Atrial fibrillation ablation strategies for paroxysmal patients randomized comparison between different techniques. Circ Arrhythm Electrophysiol 2:113–119
21. Badger TJ, Daccarett M, Akoum NW et al (2010) Evaluation of left atrial lesions after initial and repeat atrial fibrillation ablation lessons learned from delayed-enhancement MRI in repeat ablation procedures. Circ Arrhythm Electrophysiol 3:249–259
22. Jadidi AS, Cochet H, Shah AJ et al (2013) Inverse relationship between fractionated electrograms and atrial fibrosis in persistent atrial fibrillation combined magnetic resonance imaging and high-density mapping. J Am Coll Cardiol 62:802–812
23. Baccalá LA, Sameshima K, Ballester G et al (1998) Studying the interaction between brain structures via directed coherence and granger causality. Appl Signal Process 1:40–48
24. Rodrigo M, Guillem MS, Liberos A et al (2012) Identification of fibrillatory sources by measuring causal relationships. CinC 2012 39:705–708
25. Rodrigo M, Liberos A, Guillem MS et al (2011) Causality relation map: a novel methodology for the identification of hierarchical fibrillatory processes. CinC 2011 38:176–179
26. Richter U, Faes L, Ravelli F et al (2012) Propagation pattern analysis during atrial fibrillation based on sparse modeling. IEEE Trans Biomed Eng 59:1319–1328
27. Bruns HJ, Eckardt L, Vahlhaus C et al (2002) Body surface potential mapping in patients with Brugada syndrome: right precordial ST segment variations and reverse changes in left precordial leads. Cardiovasc Res 54:58–66
28. Eckardt L, Bruns HJ, Paul M et al (2002) Body surface area of ST elevation and the presence of late potentials correlate to the inducibility of ventricular tachyarrhythmias in Brugada syndrome. J Cardiovasc Electrophysiol 13:742–749
29. Dubuc M, Nadeau R, Tremblay G et al (1993) Pace mapping using body-surface potential maps to guide catheter ablation of accessory pathways in patients with Wolff–Parkinson–White syndrome. Circulation 87:135–143
30. SippensGroenewegen A, Roithinger FX, Peeters HAP et al (1998) Body surface mapping of atrial arrhythmias—atlas of paced P wave integral maps to localize the focal origin of right atrial tachycardia. J Electrocardiol 31:85–91
31. SippensGroenewegen A, Lesh MD, Roithinger FX et al (2000) Body surface mapping of counterclockwise and clockwise typical atrial flutter: a comparative analysis with endocardial activation sequence mapping. J Am Coll Cardiol 35:1276–1287
32. Guillem MS, Quesada A, Donis V et al (2009) Surface wavefront propagation maps: noninvasive characterization of atrial flutter circuit. Int J Bioelectromagn 11:22–26
33. Guillem MS, Climent AM, Castells F et al (2009) Noninvasive mapping of human atrial fibrillation. J Cardiovasc Electrophysiol 20: 507–513
34. Guillem MS, Climent AM, Millet J et al (2013) Noninvasive localization of maximal frequency sites of atrial fibrillation by body surface potential mapping. Circ Arrhythm Electrophysiol 6:294–301
35. MacLeod RS, Brooks DH (1998) Recent progress in inverse problems in electrocardiology. IEEE Eng Med Biol Mag 17:73–83
36. Rudy Y, Messingerrapport B (1988) The inverse problem in electrocardiography: solutions in terms of epicardial potentials. Crit Rev Biomed Eng 16:1047–1058
37. Geselowitz DB (1967) On bioelectric potentials in an inhomogeneous volume conductor. Biophys J 7:1–11

38. Sarvas J (1987) Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. Phys Med Biol 32:11–22
39. Horacek BM, Clements JC (1997) The inverse problem of electrocardiography: a solution in terms of single- and double-layer sources on the epicardial surface. Math Biosci 144:119–154
40. De Munck JC (1992) A linear discretization of the volume conductor boundary integral-equation using analytically integrated elements. IEEE Trans Biomed Eng 39:986–990
41. Cowper GR (1972) Gaussian quadrature formulas for triangles. Int J Numer Meth Eng 7(3):405–408
42. Tikhonov A (1963) On the solution of incorrectly posed problems and the method of regularization. Sov Math Dokl 4:1035–1038
43. Tikhonov A, Arsenin V (1977) Solutions of ill-posed problems. Wiley, New York
44. Hansen PC, Oleary DP (1993) The use of the L-curve in the regularization of discrete ill-posed problems. SIAM J Sci Comput 14:1487–1503
45. Pedrón-Torrecilla J, Climent AM, Liberos A et al (2012) Non-invasive estimation of the activation sequence in the atria during sinus rhythm and atrial tachyarrhythmia. CinC 2012 39:901–904
46. Ramanathan C, Ghanem RN, Jia P et al (2004) Noninvasive electrocardiographic imaging for cardiac electrophysiology and arrhythmia. Nat Med 10:422–428
47. Roten L, Pedersen M, Pascale P et al (2012) Noninvasive electrocardiographic mapping for prediction of tachycardia mechanism and origin of atrial tachycardia following bilateral pulmonary transplantation. J Cardiovasc Electrophysiol 23:553–555
48. Pedron-Torrecilla J, Climent AM, Millet J et al (2011) Characteristics of inverse-computed epicardial electrograms of Brugada syndrome patients. Conf Proc IEEE Eng Med Biol Soc 2011:235–238
49. Trayanova NA (2011) Whole-heart modeling applications to cardiac electrophysiology and electromechanics. Circ Res 108:113–195
50. Herron TJ (2012) Optical imaging of voltage and calcium in cardiac cells & tissues. Circ Res 110:E49
51. Lee P, Yan P, Ewart P et al (2012) Simultaneous measurement and modulation of multiple physiological parameters in the isolated heart using optical techniques. Pflugers Arch 464:403–414
52. Lee P, Bollensdorff C, Quinn TA et al (2011) Single-sensor system for spatially resolved, continuous, and multiparametric optical mapping of cardiac tissue. Heart Rhythm 8:1482–1491
53. Chang P, Hsieh Y, Hsueh C et al (2013) Apamin induces early afterdepolarizations and torsades de pointes ventricular arrhythmia from failing rabbit ventricles exhibiting secondary rises in intracellular calcium. Heart Rhythm 10:1516–1524
54. Auerbach DS, Grzeda KR, Furspan PB et al (2011) Structural heterogeneity promotes triggered activity, reflection and arrhythmogenesis in cardiomyocyte monolayers. J Physiol 589: 2363–2381
55. Pandit SV, Jalife J (2013) Rotors and the dynamics of cardiac fibrillation. Circ Res 112:849–862
56. Yamazaki M, Vaquero LM, Hou L et al (2009) Mechanisms of stretch-induced atrial fibrillation in the presence and the absence of adrenocholinergic stimulation: interplay between rotors and focal discharges. Heart Rhythm 6:1009–1017
57. Girouard SD, Pastore JM, Laurita KR et al (1996) Optical mapping in a new guinea pig model of ventricular tachycardia reveals mechanisms for multiple wavelengths in a single reentrant circuit. Circulation 93:603–613
58. Gray RA, Pertsov AM, Jalife J (1998) Spatial and temporal organization during cardiac fibrillation. Nature 392:75–78
59. Weiss JN, Qu ZL, Chen PS et al (2005) The dynamics of cardiac fibrillation. Circulation 112:1232–1240
60. Karma A (2013) Physics of cardiac arrhythmogenesis. Annu Rev Condens Matter Phys 4:313–337
61. Weiss JN, Karma A, Shiferaw Y et al (2006) From pulsus to pulseless: the saga of cardiac alternans. Circ Res 98:1244–1253
62. Sato D, Shiferaw Y, Garfinkel A et al (2006) Spatially discordant alternans in cardiac tissue: role of calcium cycling. Circ Res 99:520–527
63. Gizzi A, Cherry EM, Gilmour RFJ et al (2013) Effects of pacing site and stimulation history on alternans dynamics and the development of complex spatiotemporal patterns in cardiac tissue. Front Physiol 4:71
64. Jia Z, Bien H, Entcheva E (2008) A sensitive algorithm for automatic detection of space-time alternating signals in cardiac tissue. Conf Proc IEEE Eng Med Biol Soc 2008:153–156
65. Shiferaw Y, Aistrup GL, Wasserstrom JA (2012) Intracellular Ca^{2+} waves, afterdepolarizations, and triggered arrhythmias. Cardiovasc Res 95:265–268
66. Harrild DM, Henriquez CS (2000) A computer model of normal conduction in the human atria. Circ Res 87:E25–E36
67. Seemann G, Hoper C, Sachse FB et al (2006) Heterogeneous three-dimensional anatomical and electrophysiological model of human atria. Philos Trans A Math Phys Eng Sci 364:1465–1481

68. Gong Y, Xie F, Stein KM et al (2007) Mechanism underlying initiation of paroxysmal atrial flutter/atrial fibrillation by ectopic foci: a simulation study. Circulation 115:2094–2102
69. van Dam PM, van Oosterom A (2003) Atrial excitation assuming uniform propagation. J Cardiovasc Electrophysiol 14:S166–S171
70. Shajahan TK, Nayak AR, Pandit R (2009) Spiral-wave turbulence and its control in the presence of inhomogeneities in four mathematical models of cardiac tissue. PLoS One 4:e4738
71. Hodgkin AL, Huxley AF (1990) A quantitative description of membrane current and its application to conduction and excitation in nerve (reprinted from Journal of Physiology, vol 117, pp 500–544, 1952). Bull Math Biol 52:25–71
72. Clayton RH, Panfilov AV (2008) A guide to modelling cardiac electrical activity in anatomically detailed ventricles. Prog Biophys Mol Biol 96:19–43
73. Courtemanche M, Ramirez RJ, Nattel S (1998) Ionic mechanisms underlying human atrial action potential properties: insights from a mathematical model. Am J Physiol Heart Circ Physiol 275:H301–H321
74. Nygren A, Fiset C, Firek L et al (1998) Mathematical model of an adult human atrial cell: the role of K^+ currents in repolarization. Circ Res 82:63–81
75. Shiferaw Y, Watanabe MA, Garfinkel A et al (2003) Model of intracellular calcium cycling in ventricular myocytes. Biophys J 85: 3666–3686
76. Fox JJ, McHarg JL, Gilmour RF (2002) Ionic mechanism of electrical alternans. Am J Physiol Heart Circ Physiol 282:H516–H530
77. Shiferaw Y, Karma A (2006) Turing instability mediated by voltage diffusion in paced cardiac cells. Proc Natl Acad Sci U S A 103: 5670–5675
78. Garcia VM, Liberos A, Vidal AM et al (2014) Adaptive step ODE algorithms for the 3D simulation of electric heart activity with graphics processing units. Comput Biol Med 44:15–26
79. Sato D, Xie Y, Weiss JN et al (2009) Acceleration of cardiac tissue simulation with graphic processing units. Med Biol Eng Comput 47:1011–1015
80. van Oosterom A, Oostendorp TF, van Dam PM (2011) Potential applications of the new ECGSIM. J Electrocardiol 44:577–583

Chapter 15

Knowledge-Based Personal Health System to Empower Outpatients of Diabetes Mellitus by Means of P4 Medicine

Adrián Bresó, Carlos Sáez, Javier Vicente, Félix Larrinaga, Montserrat Robles, and Juan Miguel García-Gómez

Abstract

Diabetes Mellitus (DM) affects hundreds of millions of people worldwide and it imposes a large economic burden on healthcare systems. We present a web patient empowering system (PHSP4) that ensures continuous monitoring and assessment of the health state of patients with DM (type I and II). PHSP4 is a Knowledge-Based Personal Health System (PHS) which follows the trend of P4 Medicine (Personalized, Predictive, Preventive, and Participative). It provides messages to outpatients and clinicians about the achievement of objectives, follow-up, and treatments adjusted to the patient condition. Additionally, it calculates a four-component risk vector of the associated pathologies with DM: Nephropathy, Diabetic retinopathy, Diabetic foot, and Cardiovascular event. The core of the system is a Rule-Based System which Knowledge Base is composed by a set of rules implementing the recommendations of the American Diabetes Association (ADA) (American Diabetes Association: http://www.diabetes.org/) clinical guideline. The PHSP4 is designed to be standardized and to facilitate its interoperability by means of terminologies (SNOMED-CT [The International Health Terminology Standards Development Organization: http://www.ihtsdo.org/snomed-ct/] and UCUM [The Unified Code for Units of Measure: http://unitsofmeasure.org/]), standardized clinical documents (HL7 CDA R2 [Health Level Seven International: http://www.hl7.org/index.cfm]) for managing Electronic Health Record (EHR). We have evaluated the functionality of the system and its users' acceptance of the system using simulated and real data, and a questionnaire based in the Technology Acceptance Model methodology (TAM). Finally results show the reliability of the system and the high acceptance of clinicians.

Key words Diabetes mellitus, P4 Medicine, Personal Health System, Empower, Rule-based system, Decision support system

1 Introduction

Chronic diseases are the greatest cause of death in the world. One of these diseases is diabetes mellitus (DM), which has become one of the main health problems in the world due to its prevalence,

Carlos Fernández-Llatas and Juan Miguel García-Gómez (eds.), *Data Mining in Clinical Medicine*, Methods in Molecular Biology, vol. 1246, DOI 10.1007/978-1-4939-1985-7_15,

morbidity, and mortality. The World Health Organization (WHO[1]) estimates that by 2030, DM will be the seventh leading cause of death worldwide mainly due to the high mortality associated to several complications: retinopathy, neuropathy (diabetic foot), nephropathy, and heart disease or stroke [1]. Additionally, DM imposes a large economic burden on national healthcare systems [2]. In 2010, healthcare expenditures for diabetes accounted for 12 % of the total healthcare budget in the world. This cost is expected to increase at 2030 in 30–34 %, so more prevent efforts are needed to reduce this burden [3].

In addition to the important health and economic issues, DM changes the life of patients and the people around them for their lifetime. Patients must learn many routines that are different from their former habits (such as a controlled diet, self-monitoring, and aerobic exercise) in order to manage their disease [4]. Most scientific studies report a loss in quality of life (QoL) for people with diabetes with respect to the general population, being more pronounced in type II diabetes than in type I [5].

The system described in this study has been developed to match certain medical, functional, and technological requirements defined by the clinical and technical staff. In order to meet the medical requirements, the system is based on medical evidences; daily disease progression is monitored; the risks of pathologies associated with diabetes (heart disease, diabetic foot, nephropathy, and retinopathy) are calculated; and recommendations or alerts are inferred. With regard to the functional requirements, the system empowers the patients to be active participants in their healthcare (self-care) in order to improve the treatment adherence, the health-related lifestyle changes, and the outcomes [6–8]. The technological requirements are met by facilitating accessibility and interoperability with the use of international standards (i.e., HL7-CDA[2] and SNOMED-CT[3]).

In this context, we present a web Knowledge-Based Personal Health System to provide P4 Medicine (PHSP4) to diabetic patients, so it is (1) predictive, able to determine the risk of each individual to develop comorbidities associated with diabetes; (2) preventive, able to suggest appropriate prophylactic measures based on potential health problems; (3) personalized, able to tailor the treatment suggestions to each individual's condition; and (4) participatory, able to empower patients, by means of information and education in the continuous care of their disease.

[1] World Health Organization (WHO/OMS): http://www.who.int/diabetes/en/

[2] Health Level Seven Clinical Document Architecture standard: http://www.hl7.org/implement/standards/cda.cfm

[3] The International Health Terminology Standards Development Organization: http://www.ihtsdo.org/snomed-ct/

This chapter is organized as follows: Subheading 2 describes the state of the art of the techniques and technologies related to our work. Subheading 3 describes the functional design and the implementation of the system. Next, the results obtained in the evaluations are given in Subheading 4. Finally, Subheading 5 discusses and concludes the chapter and introduces the future work.

2 Background

The healthcare paradigm is changing from the reactive approach (based on managing a person's disease) to preventive approach (based on managing a person's health). In 2003, this medical trend was called by L. Hood as P4 Medicine (Personalized, Predictive, Preventive, and Participative). As defined by the P4 Medical Institute,[4] P4 Medicine is (1) personalized because it is based on the genetic, biomedical, or environmental information of each individual; (2) predictive because it is able to determine the risk for certain diseases in each individual; (3) preventive because, given the prediction of risk, prophylactic measures (lifestyle or therapeutic) can be taken to decrease risk; and (4) participative because many of these prophylactic interventions will undeniably require the participation of the patient. P4 Medicine approach decreases the number of patients, the number of deaths, the economic costs generated, and can improve the patient's QoL [9]. The philosophy of P4 Medicine is especially recommended for chronic diseases [10] such as DM [9], where the use of Personal Health System (PHS) is of great importance.

The main objective of a PHS is to assist in providing continuous and personalized health services to individuals regardless of location. As a result, it is able to improve lifestyle management and prevention, early diagnosis, treatment, and disease management [11]. Moreover, PHSs reduce costs in the medium-long term. Several studies show the benefits of these systems in chronic diseases like diabetes [12, 13]. Additionally, PHSs permit to easily empower patients, allowing them to be actively involved in their health, by sharing decisions about their health and taking the responsibility for managing their illnesses. Specifically, chronic conditions require continuous monitoring, which would be more feasible with an automatic system. In the case of DM type I (and often in type II), a blood sample must be taken several times a day to monitor glucose or to measure blood pressure. The most feasible strategy to get a real monitoring is to involve the patients as part of the control of their disease. This engagement is a clear example of patient empowerment. In diabetes, patient empowerment includes:

[4] P4 Medicine Institute: http://p4mi.org/

(1) the patient education; (2) the collaborative use of behavioral-change techniques to foster lifestyle changes; (3) the adoption of health-promoting behaviors; (4) easy contact with health care providers; and (5) skill development across a range of chronic conditions [7, 14]. The WHO pointed out that the optimal care for chronic conditions such as diabetes is achieved when healthcare providers interact with informed patients [15]. We extend this idea to empowered patients: when patients are more informed and involved, they interact more effectively with the healthcare providers and strive to take actions that will promote healthy outcomes. In addition, the empowerment of patients allows the development of individualized care plans to keep the patients informed with specific knowledge and provides behavioral suggestions that they might need on a daily basis. The empowerment of patients reduces the risk of disease and the cost of treatment, improves QoL [16, 17] and clinical outcomes [18, 19], and also increases the efficiency of medical system [20–22]. There is a strong link between the concept of empowerment and the participative concept defined at the P4 Medicine paradigm. Hence, the combination of PHSs and P4 Medicine is a powerful strategy for developing an intelligent medical system that adapts to the current and near future needs of chronic diseases like diabetes.

PHSs require an intelligent system in order to infer new results from data and clinical knowledge (e.g., vital guidelines, biomedical markers, or activity), such as Rule-Based Systems (RBSs). RBSs are a branch of applied Artificial Intelligence (AI) which have proven to be effective in improving health care, decreasing the cost of treatments, and enhancing QoL [23]. Generally, RBSs represent knowledge by rules in the form of "If…Then" rules. For medical domains, RBSs implement medical evidence as suggestions that are linked to specific conditions of the patients given their clinical data.

Several RBSs have been developed in the diabetes domain. Most of them aim to give advice during diagnosis [24–28], to educate patients [6, 29], to monitor the treatment (such as the insulin dose adjustment) [29–33], and to assist in disease management [34, 35]. Ma et al. [6] proposed an architecture to help patients through empowerment and participative by providing tailored and prioritized information about diabetes. Başçiftçi et al. [27] developed a small RBS with ten variables (such as age, blood glucose, or blood pressure) to diagnose Type I, II, or gestational diabetes. Yu et al. [36] carried out a review and evaluation of web-accessible tools for the management of diabetes and related cardiovascular risk factors by patients and healthcare professionals. This review [36] showed the large number of tools that are available and demonstrates that these tools have the potential to improve health outcomes and complement healthcare delivery. Dixon et al. [37] implements a pilot study of a multidisciplinary web-based CDS with only 11 preventive care rules extracted from clinical repository, not from clinical guideline like

PHSP4. Other differences between the work of Dixon [37] and ours is no calculated risk pathologies, and only shows preventive care reminders rather than recommendations related with the treatment, achievement of objectives, general, follow-up for both clinicians and patients. Yung-Hsiu et al. [38] presented a web-accessible diabetes care support system for diabetic patient and care provider. The major differences regarding our work are that in [38] no clinical guidelines, standardized protocols, or Electronic Health Records (EHR) were used, neither the calculation of the pathological diabetes risks factors. Similarly to [38], we carried out a users' acceptance evaluation based in the Technology Acceptance Model (TAM) methodology [39]. TAM permits measuring the acceptance of a new technology based in the users' perspective on its usefulness and ease of use. Other development similar to ours is the system proposed by Lähteenmäki et al. [35]. However, it also does not calculate the pathological diabetes risks, neither does not implemented the American Diabetes Association (ADA) clinical guideline.

3 Material and Methods

In this work we are describing a DM (type I and II) continuous monitoring use-case. Nevertheless, the PHSP4 system has been designed to be generic and configurable hence, permitting its adaptation to other pathologies (such as hypertension) modifying some configurations such as the rule file. The PHSP4 design is based on a three-tier architecture (*see* Fig. 1) to provide scalability, flexibility, and maintainability: the Presentation layer, where interaction with the patients and doctors are managed; the Logic layer,

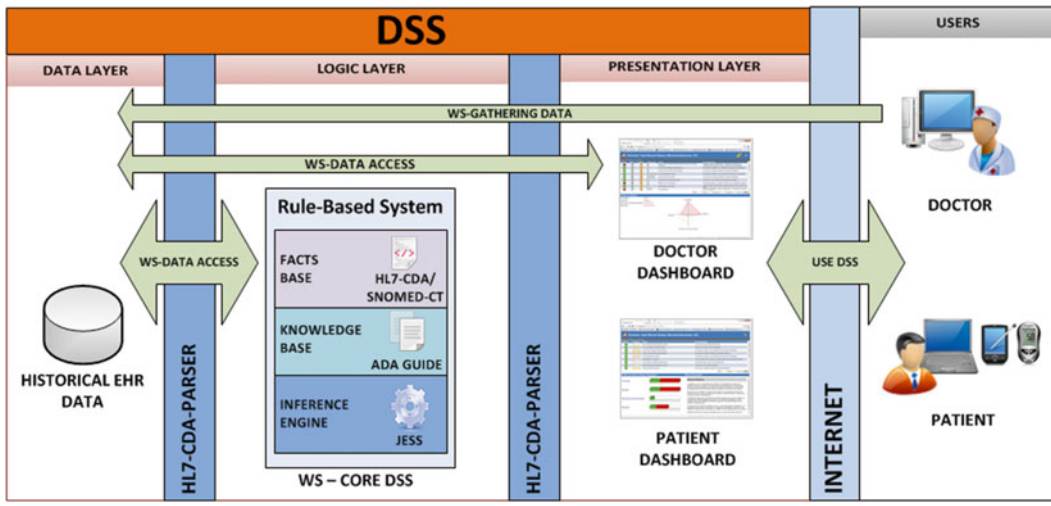

Fig. 1 Diagram of the Global System divided in a three-tier architecture: Data, Logic, and Presentation

where reasoning is performed; and the Data layer, where communications and data storing is managed.

We distributed each layer of the system in a different server. With this layout, the interoperation among layers became explicit and could be easily tested. Third party applications can join the system by using the provided interface of interoperability. The division also eased the maintenance of the system by means of the detection and relation of errors to specific layers or the provision of new functionality in a layer without affecting others.

DSS's layers are linked by Web Services (WS) implemented with SOAP[5] against REST because it offers security services in data transfers. All data embedded in the communications are parsed into Health Level Seven (HL7) standards which provide a standardized framework for the exchange, integration, sharing, and retrieval of the I. Specifically, we use HL7 Clinical Document Architecture on its release 2 (HL7 CDA R2) which is a document markup standard that specifies the structure and semantics of clinical documents for the purpose of exchange between healthcare providers and patients. It defines a clinical document as having the following six characteristics: (1) Persistence, (2) Stewardship, (3) Potential for authentication, (4) Context, (5) Wholeness, and (6) Human readability. CDA is approved by ANSI and widely accepted as clinical document standard.

The PHSP4 development was carried out following a Spiral software development methodology. This allows us to build more and more complete prototypes. Each of these prototypes was tested and discussed by technicians and clinicians in order to refine them for use in the following spiral cycle.

3.1 The Presentation Layer

We implemented the GUI using Java Server Pages[6] (JSP). Two different user roles were defined: patients (Patient User Interface) and clinicians (Clinical User Interface). In both cases, we followed The Microsoft Health Common User Interface (MSCUI) guides.[7] Even though MSCUI cannot be considered a standardized methodology, it was defined under an exhaustive use by physicians, proving to increase clinical effectiveness and improve patient safety. During our interface development, continuous feedback from clinicians was taken into account to ensure the acceptability of the system by the clinical community.

Users (patients and clinicians) can access these interfaces using the internet web browser over different devices (e.g., Laptop, PDA). The GUI shows information generated according to the gathering of biomedical data and adapted to each user role: (1) The

[5] http://www.w3.org/TR/soap/
[6] http://www.oracle.com/technetwork/java/javaee/jsp/index.html
[7] Microsoft Health Common User Interface guidance overview: http://www.mscui.net/DesignGuide/DesignGuide.aspx

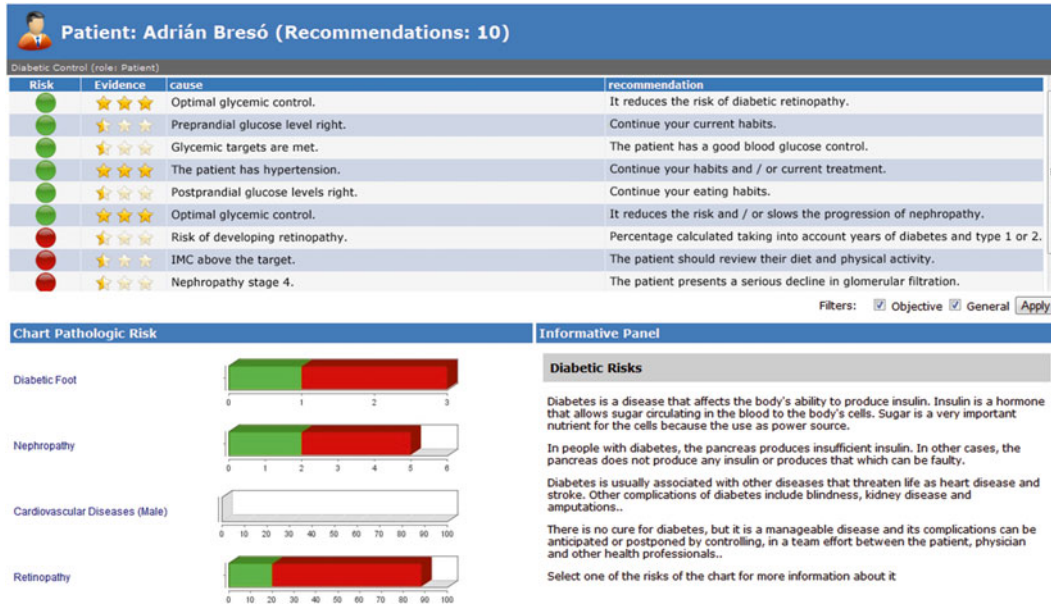

Fig. 2 Web patient user interface screenshot in which patients can view their personalized recommendations (general and objective types), information about DM, and its pathologies, and the calculated risk levels for each of the four conditions associated with DM: Nephropathy, Diabetic retinopathy, Diabetic foot, and Cardiovascular event

Patient User Interface (*see* Fig. 2) shows an easy-to-use information layout to give patients clear information about their state. The screen was divided into three areas. The first area shows a table with personalized alerts (general recommendations and alerts related to their personal objectives with regard to DM self-management). Every alert contains information about the cause (what values triggered the alert), the recommendation (what actions should be followed to revert the alert), and the related risks (what health complications can occur if the recommendation is ignored). The second area shows graphs about the risk of having one of the four pathologies associated with DM. The third area shows a brief information of these comorbidities for educational purposes. (2) The Clinical User Interface shows a more detailed patient status. It contains more types of alerts (general, and objective alerts and alerts related to the follow-up of the patient and treatment recommendations). In addition, each alert is supported with bibliographic information. It is also possible to review the path that generated the alerts in order to study the patient's situation. Hence, the clinician could make a thorough analysis about the situation of each patient based on the triggered domain rules. Additionally, the current pathology risks are shown on a spider plot: four normalized axes make up the spider plot, one for each pathology risk. The calculated risk levels are showed with a red-shaded area.

Additionally, clinicians have access to other two screens. The former is a simple web-form in which doctors could register data collected during checkups. The later contains four historical plots (*see* Fig. 3) of the monitored signals. Three of these plots were predefined by clinical experts and they showed information about three more important variable groups in DM (blood glucose, blood pressure, and lipid values). Last one was a customizable plot in which the user can add any valuable variable or treatment milestone (which indicates changes or important moments in the treatment).

3.2 The Logic Layer

In this layer we implement the RBS of the PHSP4. The RBS consists of three components: a Working Memory (WM) where the facts are stored, a Knowledge Base (KB) where the (if-then) rules are stored, and an Inference Engine (IE) that carries out the reasoning to generate patient recommendations or alerts.

In our work, the WM consisted of a medical data set received from patient sensors and forms filled out by the doctor. This set of data (received in HL7-CDA format) was managed by the database, and parsed and serialized by the parser [40].

Knowledge acquisition is the starting point in the construction of KBs in Rule-Based Clinical Decision Support Systems (CDSS). Our knowledge source was the ADA clinical guideline [41] and the adaptations to the Spanish population, which were interpreted by professionals working at the Health Center of Gandia (Spain). Knowledge acquisition was performed by manually parsing the ADA clinical guideline overseen by the clinical experts. Our KB was defined by more than 100 rules extracted from the ADA recommendations and stored in the rules file using the JESS 7 language. Our rules were divided into two main groups: calculation rules and recommendation rules. Recommendation rules (*see* Fig. 4) were additionally divided into four groups: general rules, objective rules, follow-up rules, and treatment rules. General rules refer to information about the consequences of having an optimal control of the different values. In the category of objective rules, the recommendations refer to the objectives that DM patients have to reach regarding several biomedical levels (glucose levels, blood pressure, lipids, BMI, among others). Follow-up rules are recommendations that are related to reviews and planning of the next patient's visit. Finally, treatment rules inform about the treatment to be applied to each patient. Calculation rules were used to make internal calculations (such as updates, and increments) and to infer the risk of pathologies.

We chose JESS, which is based on Rete algorithm [42], as IE because it is fast, stable, and tightly integrated with the Sun's Java platform. Hence, rule consistency, rule prioritization, and conflict resolution are managed by JESS.

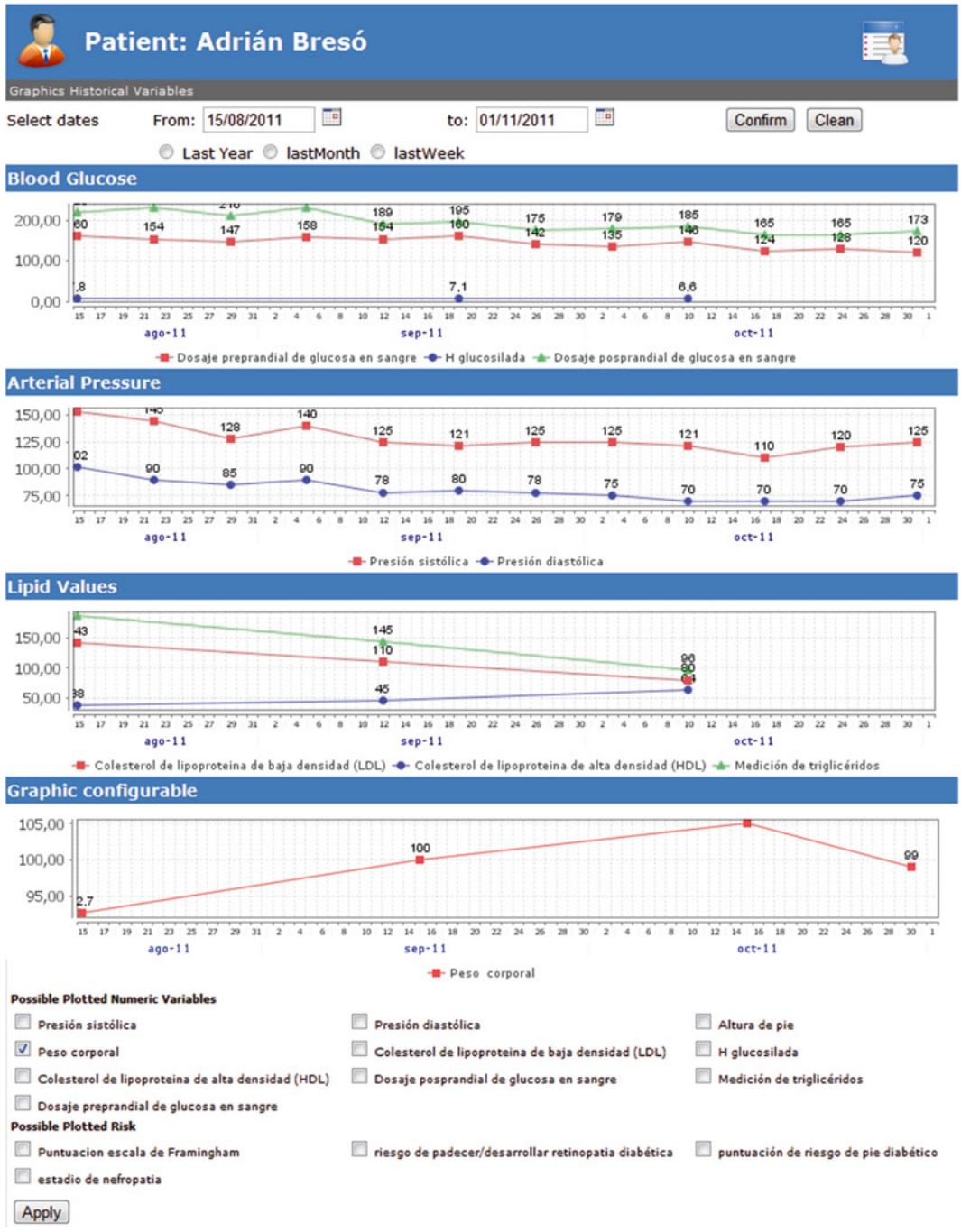

Fig. 3 Historical clinical interface screenshot where the clinician can view three specific plots (blood glucose, blood pressure, and lipid values) and one customizable plot, in which the clinician can add the available variables in the database system. All plots show the data between the date ranges selected by the clinician

```
;; POSTPRANDIAL BLOOD SUGAR LEVELS RULE
(defrule OBJ_X_GPostprandialOK "Correct postprandial glucose level"
?patient <- (Data (name "GPostprandial")
{value != nil && (value < ?*MAX_GPostprandial*)} (variable ?myVar))
 (User (id ?myID))
 (Case (id ?myCase))
 =>
(bind ?result (new FinalResult ))
(?result setText "Correct postprandial glucose level.")
(?result setRecommendation "Correct postprandial glucose level.")
(?result setRisk FALSE)
(?result setNameRule "OBJ_X_GPostprandialOK")
(?result setType "objective")
(?result setIdUser ?myID)
(?result setIdCase ?myCase)
(?result addValuesOUT ?patient.value)
(?result addVariablesOUT (?myVar getName))
(add ?result)
)
```

Fig. 4 Objective rule example for checking whether the postprandial glucose level of the patient is under a defined threshold. If the value is under the threshold, the rule is triggered, making a new recommendation in which the system informs the patient about his/her current state of health

We developed the RBS to be customizable, using six external configuration files. These files were organized in two main groups: general-purpose files and DSS-specific files.

The general-purpose files allowed the internationalization of the system multilingual. The file variables.xml is an example of general-purpose file, which gathered the 95 variables used in our system and were defined by name; maximum and minimum threshold values; text description; unit and type (UCUM codification); text unit description; and variable code (SNOMED or internal codification).

The DSS specific files included (1) the DSS Fact Base or data files; (2) the KB or rules file; and (3) input and output files. The input document (inputCDA.xml) contained the patient personal information and patient data measures. The output document (outputCDA.xml) saved the inferred recommendations and assessment of risks, using HL7 CDA structure. Sáez et al. [40] described in detail the implemented framework for wrapping this layer facilitating the semantic interoperability.

4 The Data Layer

4.1 Data Storage

In our system, the data storage tier stored the patient's EHR, containing two types of data: (1) historical clinical and demographical data; and (2) HL7 CDA documents, as inputs and outputs to/from

the IE. Additionally, this layer was responsible for keeping the data neutral and independent from other tiers. To store EHR, we used a relational database management system, the Oracle DBMS Enterprise Edition. Data collected by the different applications were stored in tables following a relational model using the Structured Query Language (SQL) as the programming language.

4.2 Codification of Data

In our development, clinical variables were coded using SNOMED-CT terminology. SNOMED-CT is a systematically organized computer collection of medical terminology that includes the most important areas of clinical information such as diseases, findings, procedures, microorganisms, and pharmaceuticals. Additionally, units of measure were coded using UCUM, which is a code system intended to include all units of measures currently being used in international science, engineering, and business. The purpose of using UCUM was to facilitate unambiguous electronic communication of quantities together with their units.

4.3 Exchange of Data

We allocated HL7-CDA parsers between layers. These parsers build HL7-CDA documents from data stored in databases (data and logic layer communication, *see* Fig. 1), and extract recommendations and data from HL7-CDA documents generated by the inference system (logic and presentation layers communication). Parsers produced HL7 CDA input documents to feed the inference system, which in time produced HL7 CDA output documents as a result of the input record. Each pair of documents was stored in the database system.

We used Java Architecture for XML Binding (JAXB) in the parser to automatically build the HL7 CDA documents from Java objects (marshall) and vice versa (unmarshall). JAXB is a technology that is used to map Java classes into XML documents. JAXB is particularly useful when the XML specification is complex as is the case with HL7 CDA. In such a case, JAXB uses XML Schema definitions to create Java classes automatically. Those classes can later be instantiated and populated. Finally, the objects can be serialized to create XML documents. We used JAXB to create the HL7 CDA objects, map data in each field, and serialize or marshall the object into an XML document. This document was used as the fact base input in the inference system. The inverse process was also performed: CDA documents obtained as the result of the inference system were unmarshalled from XML to object, and data was retrieved and stored in the database (*see* Fig. 5).

4.4 Gathering of Data

In our system, the patient data were collected by means of telemetry devices such as digital blood pressure monitors, scales, and glucometers using a web service. Clinicians also registered data gathered during medical visits by means of forms provided at the presentation layer. This data gathering process used Procedural

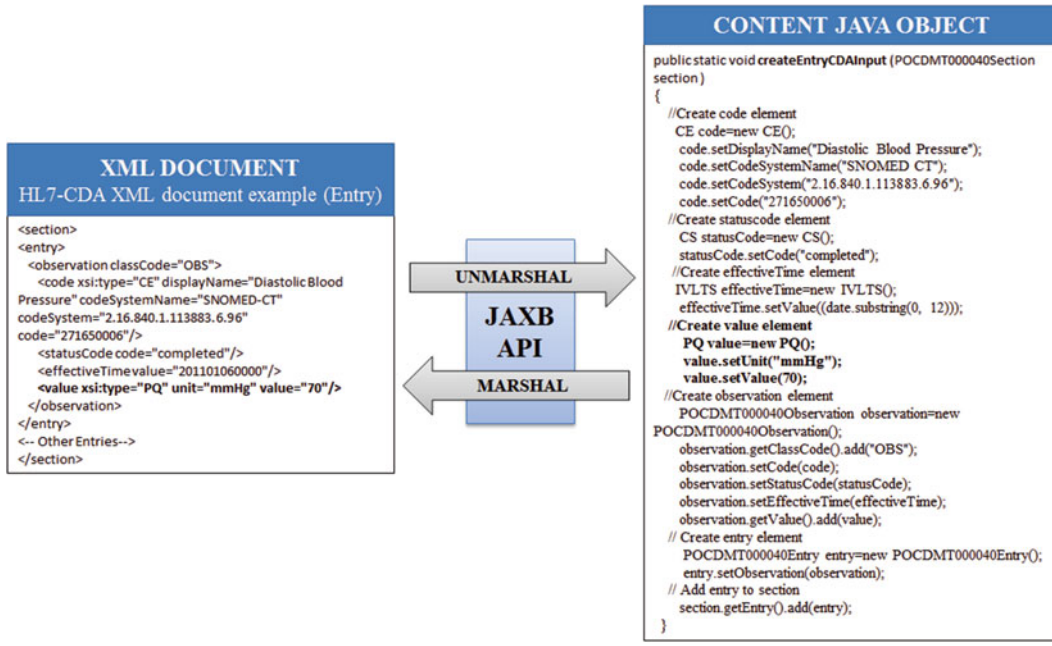

Fig. 5 This figure illustrates the parser of a CDA document and Java objects. In this example, the entry holds a diastolic blood pressure measurement. In addition to the parameter measurement, the entry presents information about the measurement unit employed, the time it was collected, the code system used, and the status

Language/Structured Query Language (PL/SQL) functions and procedures offered at the data layer to maintain both types of information in the database. These functions were encapsulated in WS. Figure 6 shows a WS method and a PL/SQL function that was used to insert health parameter values in the database.

5 Results

We performed three different evaluations of our system: (1) *functional tests* during implementation in order to get a reliable and robust system; (2) *simulations of clinical profiles* defined by the clinicians; and (3) *acceptability tests by clinicians* after the use of the system in real cases.

5.1 Functional Evaluation

We performed tests of each of the prototypes generated at each iteration of the spiral development. We evaluated the KB and the IE in order to test the correctness and efficiency of operation. To carry out this test, we employed a unit tester. A unit tester is a method by which individual units of source code are tested to determine if they are fitting for use. A unit is the smallest testable part of an application. We implemented a tester tool for this first

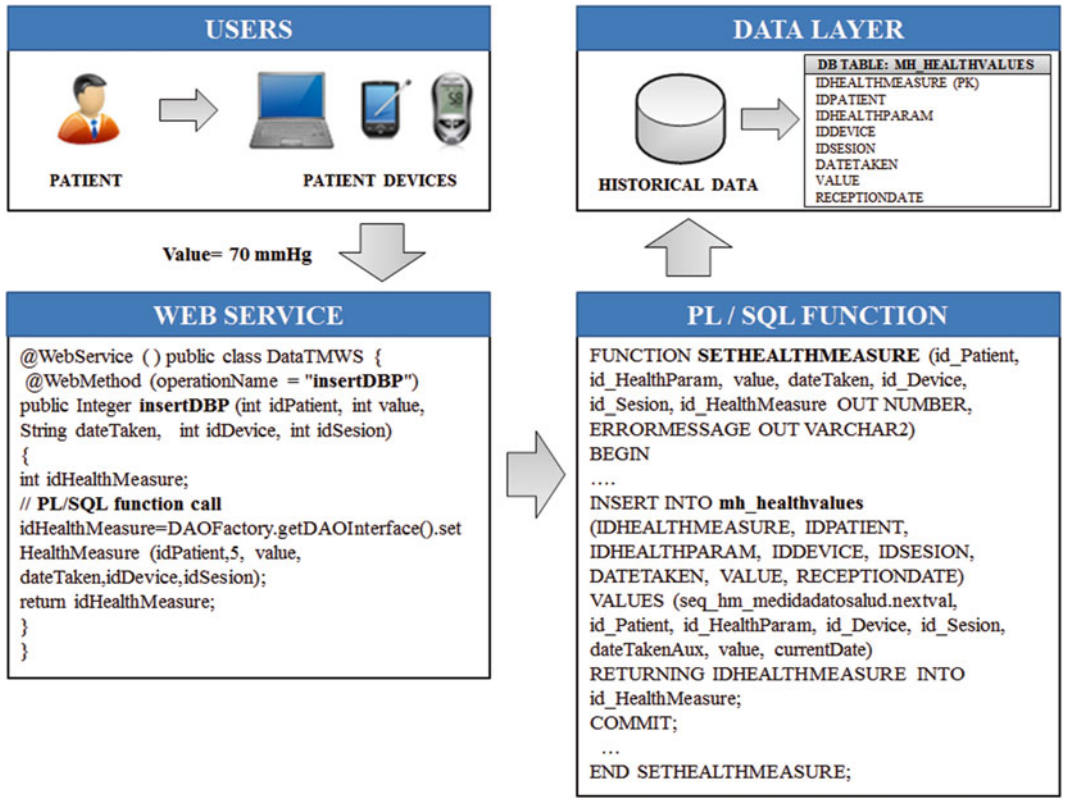

Fig. 6 Diagram of the WS-Gathering data. Patient user uses the telemetry devices to invoke the WS with the collected data (e.g., Diastolic Blood Pressure). WS calls PL/SQL function (e.g., setHealthMeasure function). Finally, PL/SQL inserts the new value into historical data (e.g., mh_HEALTHVALUES table)

evaluation, which allowed us to generate massive simulated data or to enter profile data (fact base) in the KB. The tester tool executed the rules for every patient's situation and showed the results generated in different tailored views for the review of the experts who conducted the testing.

First, we evaluated whether or not the results inferred by the rules met the clinical guideline specification. We obtain 100 % success rate. Second, we evaluated whether the inference process was efficient in terms of computational time. The system efficiency was tested by launching a massive analysis to check the load capacity of the system. The feedback of this simulation was used for fixed technical problems like integration, response speed, or rule execution met the targets. Finally, we fine-tuned the parsing and inference process and we obtained the expected results. Table 1 shows the time spent results on parse some CDA files and infer the alerts in a Pentium i5 2.40 GHz computer with 4GB RAM. Each CDA file contained 95 variables and the RBS's KB was defined by 106 rules.

Table 1
Summary of time spent (s) on parse the CDA data and infer the alerts in the inference engine

CDAs	Inferred alerts	Time spent (s)
1	2	0.074
2	5	0.093
3	8	0.107
4	9	0.152
5	14	0.178
50	80	0.757
100	262	3.445

5.2 Simulated Clinical Cases Evaluation

The expert clinicians involved in this project defined 10 clinical profiles (see Appendix 1). Each clinical profile was composed by 58 variables such as gender, weight, height, age, postprandial glucose, triglycerides, kidney failure, or family history. Clinicians defined several stages for each clinical profile, simulating a treatment of one year. The results for these profiles were inferred manually by the clinicians. After this, the profiles were run in the system. An internal audit checked rule by rule whether the manual results matched the inferred ones. As a result, we confirmed that all the rules were working properly and that the system allowed us to follow the patient evolution over the different stages.

5.3 User Acceptance Evaluation with Real Data

Finally, we performed a retrospective evaluation with real patient data to measure the utility and usability of PHSP4 with TAM methodology [39]. TAM provides the acceptance (ease of use and usefulness) measurement of new technologies by end-users. We added three additional quantitative questions (Q13, Q14 and Q15) about the predictive, preventive and personalized aspect of the PHS; and two qualitative questions (Q16 and Q17) in order to collect the opinions of the users on how to improve the tool (see Appendix 2). The evaluation of quantitative questions was carried out using Likert scales (ranging from 1 to 5).

Hence we collected daily data from 10 real diabetic patients (aged 50–60) for 3 months from the same center. We added these data to the PHSP4 and seven clinicians (over six different clinical centers) tested the response of the system to those patients. To avoid possible biases we recruit independent clinicians for the evaluation, and we set a secure web access (user/password) to the system to ensure anonymity of the clinicians who were involved in the

Table 2
Summary of TAM's quantitative questions outcomes

Question	Mean ± standard deviation
Usefulness (Q1–Q6 struct)	4.48 ± 0.36
Ease of use (Q7–Q12 struct)	4.36 ± 0.44
Predictive, Preventive and Personalized (Q13–Q15 struct)	4.60 ± 0.48

evaluation. In addition, to avoid a central tendency bias we reduced the original seven items to five, as shown in [43].

After testing the system, the clinicians filled the TAM questionnaire. The summary of quantitative questions outcomes are shown in Table 2.

The results of qualitative questions showed that clinicians agreed that the system offered an understandable design and got adequate information to motivate and involve the patient in self-management of their disease. The clinical information was personalized, it was presented clearly and it was complemented by an overview of the disease. Some experts recommended the integration of the PHSP4 into the regional health information system.

6 Discussion

The system described in this study is classified by the Australian standard defined by Brand [44] as a CDSS of level III. A CDSS of level III uses deductive inference engines to operate on a specific knowledge base to automatically generate diagnostic or intervention recommendations for the changing patient clinical condition [44].

Our developed PHSP4 web system implements a generic and configurable system to monitor and treat DM (type I and II) based on rules, clinical guideline (ADA), and international standards (SNOMED-CT, UCUM and HL7 CDA R2) that allows a connection with different EHR infrastructures. Our system provides continuous, quality controlled, and personalized health service to the individual. It provides positive and negative recommendations that are classified into four groups: general, treatment, objective, and follow up. Additionally, our system calculates numerical risk assessment for the most common comorbidities associated with DM (Nephropathy, Diabetic retinopathy, Diabetic foot, and Cardiovascular event).

As far as we are concern, there are several systems based on medical evidences that use a rules engine to implement guidelines, several of which focus on DM, but our PHSP4 system is the only that is focus on P4 Medicine. It collects, analyzes and stores individual data in order to provide an effective and empowered

feedback that helps in lifestyle management, and preventive, and treatment process of patient care. The PHS4P empowers DM outpatients to be more informed and motivated, provides a patient-provider communication, and facilitates the achievement of objectives, follow-up, and suggested treatment in a way that both patients and doctors can understand. Our PHSP4 is based on self-management and improves the QoL of chronic patients by helping them become more participative, and preventing and predicting possible future complications of the disease. The PHSP4 is geared towards personalized results (such as treatment recommendations or messages about the patient's current state) improving acceptance and efficiency.

The results obtained in this work have demonstrated that this system is ready to be integrated and used as a core system in a clinical information system for chronic patients. We plan to deploy the system in a multinational company and this integration will provide us a valuable feedback from patient users and health professionals that will help us to improve our system in the future. The feedback obtained from this integration will be analyzed in order to measure the frequency of the recommendation triggered, or the usefulness of each clinical variable. Additionally, this deployment will enable us to access diabetic patients who want to participate in a future acceptance study.

We are aware that the number of clinicians involved in the evaluation of acceptability is low ($n=7$). Despite this limitation, the responses of the clinicians who participated in the evaluation of acceptability have been positive. Additionally, the clinicians made interesting recommendations that could improve the system. The evaluation revealed that several clinicians considered important a more exhaustive monitoring of the blood pressure.

We are considering adding in a near future some of these improvements such as the update of the rules to guide ADA to its most recent version (2013) and adding new rules to the system to expand coverage of the domain from DM to pre-diabetes. Since the PHSP4 has been designed as a generic and configurable system, other future improvement will contemplate the adaptation of the system to diseases such as hypertension, where a continuous monitoring and empowerment of the patient might benefit his QoL.

Acknowledgements

We thank Dr. Tomás Fuster and his team from Centro de Salud Gandia—Beniopa for their collaboration and clinical support during the development of the system. We thank the collaboration from Universidad de Mondragon, specially, Urtzi Markiegi. We

also thank Fagor Electrodomésticos S.Coop[8] for their support and collaboration in the development of this work, especially to Juan Ramón Inurria and Jorge de Antonio Prieto. Regarding the clinical evaluation, we thank the collaboration of Dr. Pablo Díaz-Munio from Medicalquatro (Madrid); Dr. Alejandro Rodríguez from Hospital Virgen del Castillo (Yecla); Dr. Francisca Moreno from Hospital La Fe (Valencia); Dr. Llucia Palacios from Verge dels Lliris (Alcoy); Dr. Tomás Fuster and Dr. Pilar Alonso from Centro de Salud Gandia—Beniopa; and Dr. Inmaculada Ibáñez from Centro de Salud Laboral UPV (Valencia). This work was funded by the Ministerio de Ciencia e Innovación of Spain (INNPACTO 2011 ref. IPT-2011-1087-900000).

Appendix 1: Clinical Profiles Defined by Expert Clinicians

Profile	Patient	Stages
1	**Gender**: Male **Age**: 56 **DM**: Type II (1 year)	**Stage 1**: Good health. **Stage 2**: He gains weight and the BMI rises to 25. **Stage 3**: The lipid profile gets worse. **Stage 4**: The lipid profile gets better.
2	**Gender**: Male **Age**: 56 **DM**: Type II (1 year)	**Stage 1**: Smoking patient. **Stage 2**: He starts a program to quit smoking. **Stage 3**: The diabetic foot risk is 2. **Stage 4**: He finally quits smoking.
3	**Gender**: Female **Age**: 40 **DM**: Type II (1 year)	**Stage 1**: Good health. **Stage 2**: The lipid profile gets worse. **Stage 3**: The lipid profile gets worse. **Stage 4**: The lipid profile improves, but not to desirable levels.
4	**Gender**: Female **Age**: 40 **DM**: Type II (1 year)	**Stage** 1: Good health. **Stage 2**: The blood pressure (BP) gets worse. **Stage 3**: The BP gets worse until reaching hypertension.
5	**Gender**: Female **Age**: 60 **DM**: Type II (10 years)	**Stage 1**: Good health. **Stage 2**: The BP get worse. **Stage 3**: The BP gets worse until reaching hypertension and albuminuria.
6	**Gender**: Female **Age**: 60 **DM**: Type II (1 year)	**Stage 1**: Good health. **Stage 2**: The glycemic targets gets worse. **Stage 3**: The glycemic targets improve.
7	**Gender**: Male **Age**: 31 **DM**: Type I (3 years)	**Stage 1**: Good health. He is smoker and he is in lifestyle therapy. **Stage 2**: After 3 months he has hypoglycemia and high BP. **Stage 3**: Hypoglycemia continues. Hypertension was resolved after administering medication treatment. He keeps smoking.

(continued)

[8] http://www.fagor.com/web/es/home

(continued)

Profile	Patient	Stages
8	**Gender**: Male **Age**: 45 **DM**: Type I (10 years)	**Stage 1**: He is in treatment. He suffers hypertension and he has risk of coronary artery disease and retinopathy. **Stage 2**: After six months, he suffers myocardial infarction (MI). **Stage 3**: After six months, the patient changes his current treatment (angiotensin II receptor antagonist treatment) by enzyme inhibitor of angiotensin converting treatment. Additionally, he is treated with metformin and beta blockers. The hypertension improves.
9	**Gender**: Female **Age**: 55 **DM**: Type II (22 years)	**Stage 1**: She is in treatment. She is obese and she has hypertension. She has altered sensitivity on her feet, kidney damage and glomerular count below 90. **Stage 2**: After six months, the glomerular count (nephropathy) gets worse. She develops calluses and deformities in one foot. Her vision also starts to be affected, which confirms that she has a diabetic retinopathy. **Stage 3**: A year later, the lipids are not controlled. The patient has proliferative diabetic retinopathy. The glomerular count significantly dropped and finally requires dialysis. She develops foot ulcers.
10	**Gender**: Female **Age**: 28 **DM**: Type I (5 years)	**Stage 1**: She is in treatment. She suffers microalbuminuria, hyperglycemia and hypertension. **Stage 2**: The patient changes her current treatment (angiotensin II receptor antagonist treatment) by enzyme inhibitor of angiotensis converting treatment and hypertension improves. The microalbuminuria becomes macroalbuminuria with renal failure. **Stage 3**: Treatment keeps desired lipid levels and BP. The macroalbuminuria becomes microalbuminuria.

Appendix 2: TAM Questionnaire

Q1. The new tool makes my work of integral monitoring of diabetic patients easier.

Q2. The new tool allows me to be productive.

Q3. The new tool allows me to be effective in the integral monitoring of diabetic patients.

Q4. The new tool allows me to accomplish my tasks quickly.

Q5. The new tool allows me to provide a quality service of integral monitoring of diabetic patients.

Q6. I consider useful the new tool in my work in order to monitor diabetic patients.

Q7. Learning to use the tool was easy for me.

Q8. I think that with the new tool is easy to get what I propose to do.

Q9. My interaction with the new tool is clear and understand its operation.

Q10. The interaction with the new tool is flexible.

Q11. Currently I am skillful using the new tool.

Q12. I believe that the new tool is easy to use.

Q13. The new tool provides me access to clinical documentation of patient alerts.

Q14. The new tool allows me to observe the causes and recommendations regarding the current status of the disease.

Q15. The new tool allows me to quickly observe pathological risks associated with the patient's situation.

Q16. Which improvements would you include in order to make the new tool more useful in your work of integral monitoring of diabetic patients?

Q17. Which improvements would you include in order to make the new tool easier to use?

References

1. World Health Organization (2008) World Health Organization 2008. WHO Press, Geneva
2. Busse R, Blümel M, Scheller-Kreinsen D, Zentner A (2010) Tackling chronic disease in Europe: strategies, interventions and challenges. European Observatory on Health Systems and Policies
3. Zhang P, Zhang X, Brown J, Vistisen D, Sicree R, Shaw J, Nichols G (2010) Global healthcare expenditure on diabetes for 2010 and 2030. Diabetes Res Clin Pract 87(3):293–301
4. Rogers IH (1959) Social problems of the diabetic. Postgrad Med J 35(403):281–284, 286
5. Rubin RR, Peyrot M (1999) Diabetes and quality of life. Diabetes Metab Res Rev 15(3):205–218
6. Ma C, Warren J, Phillips P, Stanek J (2006) Empowering patients with essential information and communication support in the context of diabetes. Int J Med Inform 75(8):577–596
7. Funnell M, Anderson R (2004) Empowerment and self-management education. Clin Diabetes 22:123–127
8. Hernandez-Tejada M, Campbell J, Walker R, Smalls B, Davis K, Egede L (2012) Diabetes empowerment, medication adherence and self-care behaviors in adults with type 2 diabetes. Diabetes Technol Ther 14(7):630–634
9. Golubnitschaja O (2010) Time for new guidelines in advanced diabetes care: paradigm change from delayed interventional approach to predictive, preventive & personalized medicine. EPMA J 1(1):3–12
10. Bousquet J, Anto JM, Sterk PJ, Adcock IM, Chung KF, Roca J, Agustí A, Brightling C, Cambon-Thomsen A, Cesario A (2011) Systems medicine and integrated care to combat chronic noncommunicable diseases. Genome Med 3(7):43
11. PHS2020 Project Personal Health Systems (2007) FP7-ICT-2007-215291. European Commission
12. Klug C, Bonin K, Bultemeiner N, Rozenfeld Y, Vasquez R, Johnson M, Cherry J (2011) Integrating telehealth technology into a clinical pharmacy telephonic diabetes management program. J Diabetes Sci Technol 5(5):1238–1245
13. Fico G, Fioravanti A, Arredondo M, Ardigó D, Guillén A (2010) A healthy lifestyle coaching-persuasive application for patients with type 2 diabetes. Conf Proc IEEE Eng Med Biol Soc 2010:2221–2224
14. Holman H, Loring K (2004) Patient self-management: a key to effectiveness and efficiency in care of chronic disease. Public Health Rep 119(3):239–243
15. World Health Organization (2002) Innovative care for chronic conditions: building blocks for action. World Health Organization, Geneve, Switzerland

16. Steed L, Cooke D, Newman S (2003) A systematic review of psychosocial outcomes following education, self-management and psychological interventions in diabetes mellitus. Patient Educ Couns 51(1):5–15
17. Cochran J, Conn VS (2008) Meta-analysis of quality of life outcomes following diabetes self-management training. Diabetes Educ 34(5):815–823
18. Ellis S, Speroff T, Dittus R, Brown A, Pichert J, Elasy T (2004) Diabetes patient education: a meta-analysis and meta-regression. Patient Educ Couns 52(1):97–105
19. Sarasohn-Kahn J (2013) A role for patients: the argument for self-care. Am J Prev Med 44(1):S16–S18
20. Farrell K, Wicks M, Martin J (2004) Chronic disease self-management improved with enhanced self-efficacy. Clin Nurs Res 13(4): 289–308
21. National Association of Chronic Disease Directors (2005) Strategies for chronic disease management. Marketing and Planning Leadership Council. www.chronicdisease.org/files/public/Managing_Chronic_Diseases.pdf
22. Norris SL, Engelgau M, Narayan K (2001) Effectiveness of self-management training in type 2 diabetes: a systematic review of randomized controlled trials. Diabetes Care 24(3): 561–587
23. Patel V, Shortliffe E, Stefanelli M, Szolovits P, Berthold M, Bellazzi R, Abu-Hanna A (2009) The coming of age of artificial intelligence in medicine. Artif Intell Med 46(1):5–17
24. Abu-Naser S, El-Hissi H, Abu-Rass M, El-Khozondar N (2010) An expert system for endocrine diagnosis and treatments using JESS. J Artif Intell 3:239–251
25. Sangi M, Than Win K, Fulcher J (2010) A knowledge-based risk advisor model for chronic complications of diabetes. In: PACIS proceedings, Australia
26. Lee C-S, Wang M-H (2011) A fuzzy expert system for diabetes decision support application. IEEE Trans Syst Man Cybern B Cybern 41(1):139–153
27. Başçiftçi F, Hatay O (2011) Reduced-rule based expert system by the simplification of logic functions for the diagnosis of diabetes. Comput Biol Med 41(6):350–356
28. Amine Chikh M, Saidi M, Settouti N (2012) Diagnosis of diabetes diseases using an Artificial Immune Recognition System2 (AIRIS2) with fuzzy K-nearest neighbor. J Med Syst 36:2721–2729
29. Renard E (2010) Online prediction of forthcoming glucose profile and outcomes for advice on insulin therapy: the DIAdvisor concept. Third international conference on advanced technologies and treatments for diabetes, Basel, 10 Feb 2010
30. Rudi R, Celler B (2005) Diabetes management in home telecare. In: Proceedings of the 15th ACRA annual scientific meeting, Australia
31. Tudor R, Hovorka R, Cavan D, Meeking D, Hejlesen O, Andreassen S (1998) DIAS-NIDDM—a model-based decision support system for insulin dose adjustment in insulin-treated subjects with NIDDM. Comput Methods Programs Biomed 56(2):175–191
32. AIDA: freeware diabetic software simulator program of blood glucose–insulin interaction [En línea]. http://www.2aida.net/welcome/.
33. García-Jaramillo M, Calm R, Bondia J, Tarín C, Vehí J (2010) Insulin dosage optimization based on prediction of postprandial glucose excursions under uncertain parameters and food intake. Comput Methods Programs Biomed 105(1):61–69
34. Georga E, Protopasppas V, Guillen A, Fico G, Ardigo D, Arredondo M, Exarchos T, Polyzos D, Fotiadis D (2009) Data mining for blood glucose prediction and knowledge discovery in diabetic patients: the METABO diabetes modeling and management system. Conf Proc IEEE Eng Med Biol Soc 2009:5633–5636
35. Lähteenmäki J, Leppänen J, Orsama A-L, Salaspuro V, Pirinen J, Sormunen M, Kaijanranta H, Ermes M (2011) Remote patient monitoring system with decision support. In: IASTED/BIOMED, Innsbruck, Austria
36. Yu CH, Bahniwal R, Laupacis A, Leung E, Orr MS, Straus SE (2012) Systematic review and evaluation of web-accessible tools for management of diabetes and related cardiovascular risk factors by patients and healthcare providers. J Am Med Inform Assoc 19(4):514
37. Dixon B, Simonatis L, Goldberg H, Paterno M, Schaeffer M, Hongsermeier T, Wright A, Middlenton B (2013) A pilot study of distributed knowledge management and clinical decision support in the cloud. Artif Intell Med 59:45–53
38. Yung-Hsiu L, Rong-Rong C, Sophie Huey-Ming G, Hui-Yu C, Her-Kun C (2012) Developing a web 2.0 diabetes care support system with evaluation from care provider perspectives. J Med Syst 36:2085–2095
39. Davis F (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 13(3):319–339
40. Sáez C, Bresó A, Vicente J, Robles M, García-Gómez JM (2013) An HL7-CDA wrapper for facilitating semantic interoperability to rule-based Clinical Decision Support Systems.

Comput Methods Programs Biomed 109(3): 239–249

41. American Diabetes Association (2010) Standards of medical care in diabetes—2010. Diabetes Care 3(Suppl 1):11–61
42. Forgy CL (1974) A network match routine for production systems. Working Paper
43. Sáez C, Martí-Bonmatí L, Alberich-Bayarrib A, Robles M, García-Gómez J (2014) Randomized pilot study and qualitative evaluation of a clinical decision support system for brain tumour diagnosis based on SV 1H MRS: evaluation as an additional information procedure for novice radiologists. Comput Biol Med 45(1):26–33
44. Brand D (2002) Electronic decision support for Australia's health sector. Report to Health Ministers by the national electronic decision support taskforce

Chapter 16

Serious Games for Elderly Continuous Monitoring

Lenin-G. Lemus-Zúñiga, Esperanza Navarro-Pardo, Carmen Moret-Tatay, and Ricardo Pocinho

Abstract

Information technology (IT) and serious games allow older population to remain independent for longer. Hence, when designing technology for this population, developmental changes, such as attention and/or perception, should be considered. For instance, a crucial developmental change has been related to cognitive speed in terms of reaction time (RT). However, this variable presents a skewed distribution that difficult data analysis. An alternative strategy is to characterize the data to an ex-Gaussian function. Furthermore, this procedure provides different parameters that have been related to underlying cognitive processes in the literature. Another issue to be considered is the optimal data recording, storing and processing. For that purpose mobile devices (smart phones and tablets) are a good option for targeting serious games where valuable information can be stored (time spent in the application, reaction time, frequency of use, and a long etcetera). The data stored inside the smartphones and tablets can be sent to a central computer (cloud storage) in order to store the data collected to not only fill the distribution of reaction times to mathematical functions, but also to estimate parameters which may reflect cognitive processes underlying language, aging, and decisional process.

Key words Aging, ICT, Serious games, Distribution components, Reaction times, Mobile devices

1 Introduction

In the last third of the last century, population has undergone a change of unusual social structure, growing the interest in examining the underlying processes of aging. This is the consequence of the general improvement in economic conditions and subsequent advances in technology. In general, population over 60 is increasing all around the world. Figure 1 shows the projection of population in the USA. And Fig. 2 shows the projection of population in the EU.

Aging is a natural and gradual process, with changes at several levels: biological, psychological, social and familiar. Usually, this term tends to be defined as a natural process that depends on our genetic structure and on environmental variables (how and where we live). But aging process can also be defined by people longevity, i.e., life expectancy. One example is the case of Cuba or Canada.

Carlos Fernández-Llatas and Juan Miguel García-Gómez (eds.), *Data Mining in Clinical Medicine*, Methods in Molecular Biology, vol. 1246, DOI 10.1007/978-1-4939-1985-7_16,

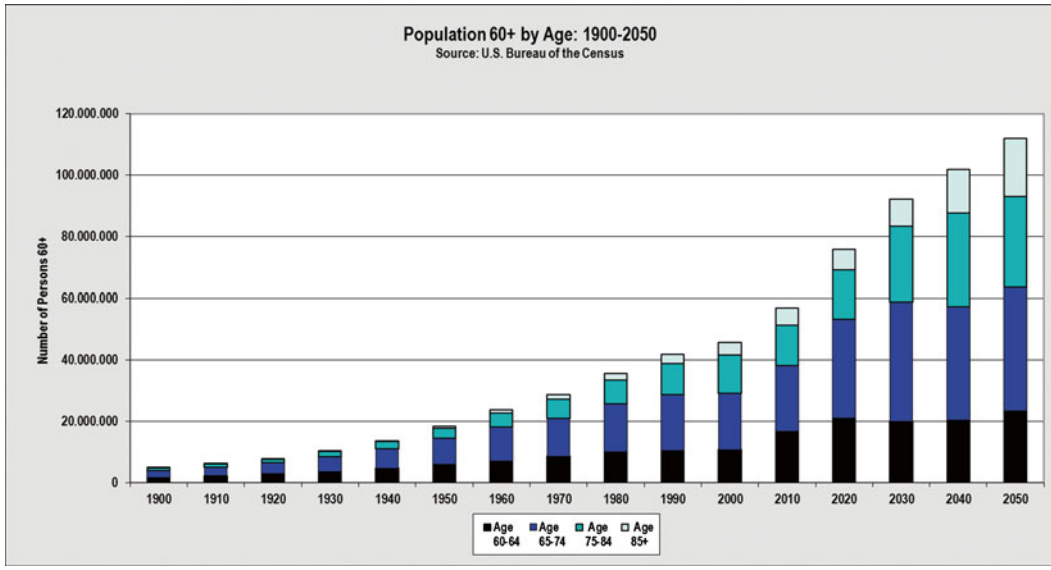

Fig. 1 Projections of the population by age and sex for the USA: 2010–2050

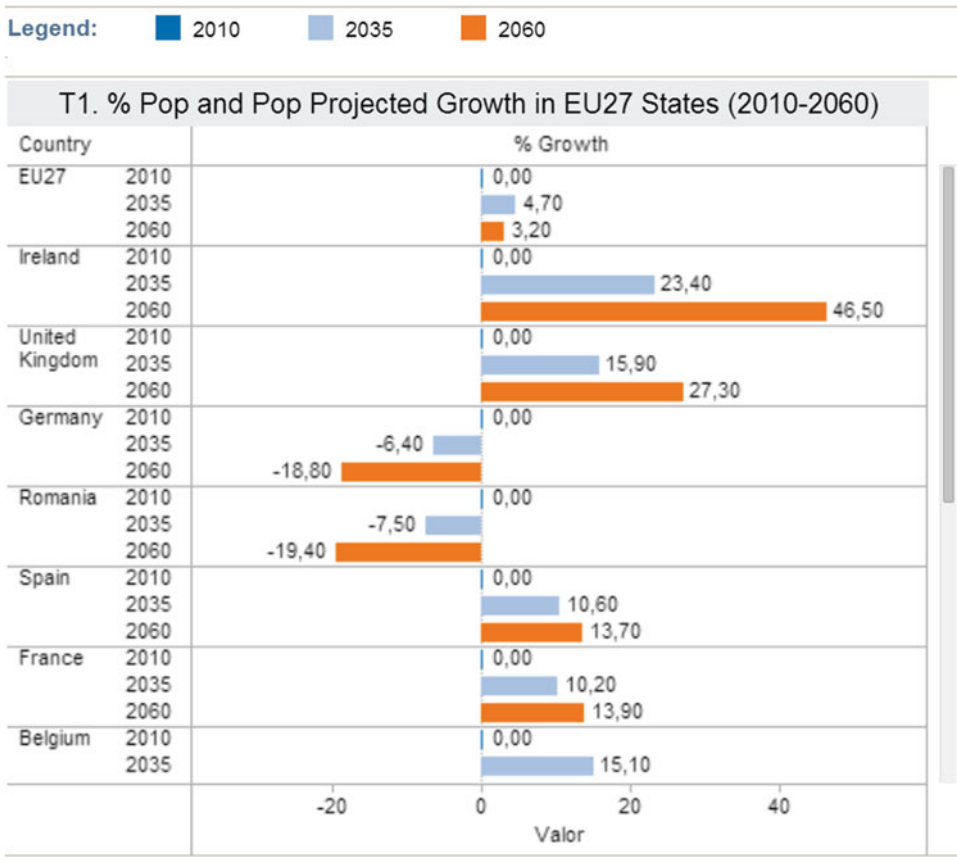

Fig. 2 Projections of the population for the EU: 2010–2060

According to the CIA world factbook report [1], life expectancy in Canada was 81.38 and 77.7 years in Cuba in 2011, much higher that some other European countries, such as the Republic of Moldova where life expectancy was 62.3 years.

This is also closely related to scientific and technological development, becoming an indicator of social and scientific development. Consequently, aging is the most important phenomenon of the twenty-first century in developed societies. It is important not only because it reflects individual changes, but also because that phenomenon has several *socioeconomical* implications. Therefore, this phenomenon transcend to both industrialized and developing countries. On the other hand, Osorio [2] indicated that the aging phenomenon transcends individual and collective interests of this social group because of its implications in the field of family, social, economic, and political.

Not surprisingly, the studies on cognition and behavior with older adults have grown. However, there remain many issues underlying the cognitive development of materials for the purpose of preventing situations of dependency degree of stimulation and improvement of cognitive functioning after use.

Regarding this last point, a whole new industry has sprung up around the possibility of keeping the brain young and healthy. We have recently experienced how companies have expanded the use of computer software under the banner of Dr. Kawashima (Brain Training for Nintendo DS) "keep your brain young." No wonder how quickly its use has been popularized in children and adolescents. Neither it is surprising that this process has been, without any doubt, much slower for the elderly. Bear in mind that this group along with other digital immigrants, have been forced, into a relatively short space of time, to migrate from the analog world to a new propositional/digital world. Regarding this point, it is remarkable the emerging growth of serious games designed for entertainment in the fields of education, scientific exploration, health care. Even if it is possible to find multiple definitions, traditionally, this kind of technology is defined as a game, where, during its activity, the participant has to deal with two or more independent decision-makers seeking to reach different goals [3]. As it was mentioned, serious games were designed for entertainment; however, studies have showed evidence on its effects and benefits, not only in education but also in therapy and diagnosis improving cognitive abilities [4]. Moreover, a vast number of studies on behavioral plasticity have been developed over these lines [5, 6].

The human brain has been described as a great processor information, constantly engaged in managing environmental information that allows processing. In order to do that, our brain makes use of basic cognitive functions such as attention. Current approaches have addressed that specific cognitive function, which is the basis for other basic and higher cognitive processes [7]. Attention is the process where certain information is selected or rejected; therefore, it is

a crucial process for daily life. It is said that older adults tend to be more conservative and might have difficulty in ignoring irrelevant or distracting information. However, practice can improve attention demands. Regarding this point, serious games have been proposed as a training tool. The innovative aspect of this work is it also to propose the use of senior for continuous monitoring employing RT.

2 The Reaction Time Variable for Continuous Monitoring

A relevant variable regarding attention is reaction time (RT). This variable can be used for continuous monitoring for cognitive impairment, because of its high sensitivity to cognitive processes such as attention [8, 9]. As expectable, the RT has turned into a star dependent variable on most of cognitive assessment tests. However, it is positively skewed data distribution usually difficult data analysis. An option to avoid trimming or other similar techniques is to perform a distributional analysis of the data. In the case of positively skewed data, an appealing possibility for this distribution is the ex-Gaussian distribution function [10]. This function is the convolution of two processes; a Gaussian (normal) and an exponential distribution. *See* Figs. 3 and 4.

Luce [11] describes this function as a model for the decision-making inside the temporal space (and therefore, a model which might describe different cognitive processes). The ex-Gaussian distribution is specified through three parameters: μ, τ, and σ. The first and second parameters (μ and σ), correspond to the average and standard diversion of the Gaussian component, while the third parameter (τ) is the decay rate of the exponential component.

When analyzing the results from an ex-Gaussian fit, one must be careful because μ and σ should not be interpreted as the distribution's average and standard deviation. The average of the ex-Gaussian distribution in terms of its components' parameters is $M = \mu + \tau$ and

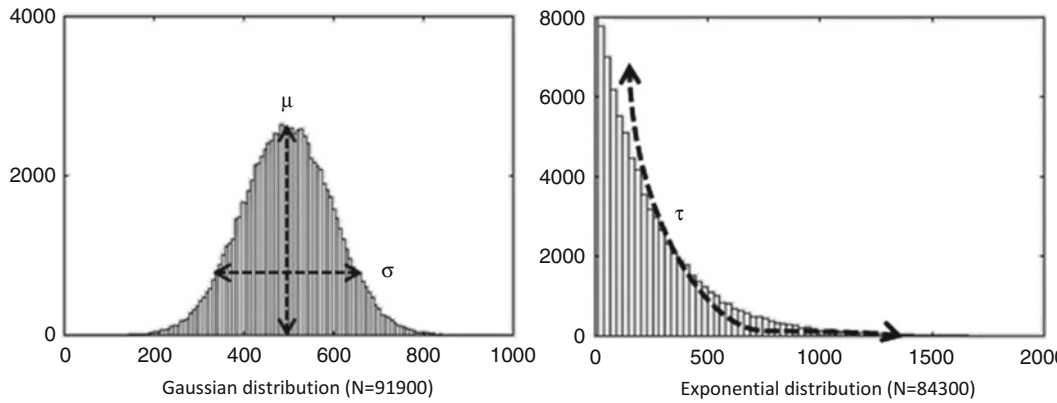

Fig. 3 A Gaussian and an exponential distribution characterized by its respectively parameters: μ, τ and σ

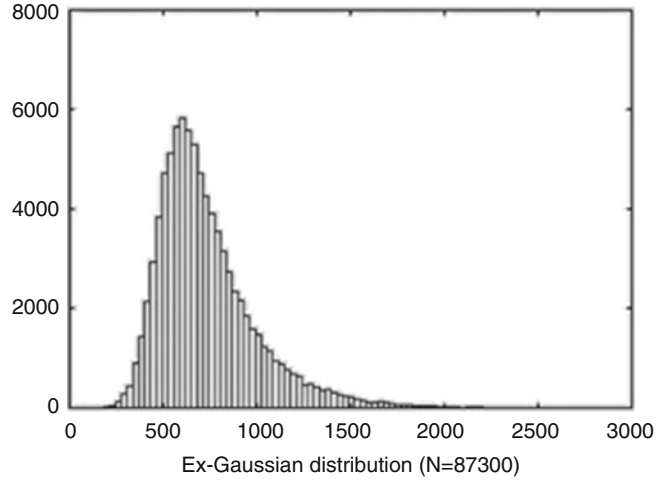

Fig. 4 An ex-Gaussian distribution: a convolution of two processes; a Gaussian (normal) and an exponential distribution

its variance is $S^2 = \sigma^2 + \tau^2$. Luce [11] has argued that the ex-Gaussian function provides a good fit to multiple empirical response time distributions. In addition, many researchers have related these parameters to underlying cognitive processes. and Wagenmakers [12] provide a review on the interpretation on the ex-Gaussian parameters in terms of underlying cognitive processes, although the functional interpretation of those parameters is still debated in the literature. One of the most relevant works in the subject is the research performed by Leth-Steensen et al. [13]. These researchers compared groups of children with ADHD to two control groups and found different tailed distributions, slower response times and, what is more important to the aim of our study, differences on τ parameter for those with ADHD. The findings provide evidence about the role of τ parameter on attention and, this was supported by other literature [14, 15].

Old participants tend to be slower than the young while they are involved in serious games [16, 17]. Nevertheless, the effects of age on a task and how reaction times are affected, is the subject of much discussion in the literature. Many authors have shown that reaction time distributions of old students have longer tails than young students (e.g., Fozard et al. [18]), which means an enhanced asymmetry in the RT distribution and in other terms, poor attentional performance.

In sum, the main objective of this project is to deepen in the development of serious games that include cognitive task (specifically on attentional demands) to examine old participants performance. The innovative aspect of this work is to promote the use of reaction time as dependent variable and it is fitted, in terms of processing components (particularly in terms of processing efficiency) for continuous monitoring.

3 Guidelines for Developing Serious Games Using Mobile Devices

The RT can be recorded using experimental software such as DMDX [19], or even using an application such as Science XL [20]. The DMDX is a traditional software for experiments in a laboratory. However, it is said that this kind of research might suffer from problems on ecological validity. For those who are not familiar with the term "ecological validity," it is referred as how close the results are to the real world and daily life. Obviously, the RT collected in a laboratory is restricted to special situations, such as a quiet room, far away in many cases from daily life. However, Dufau et al. [20] compared the RT recorded in a laboratory versus the RT collected from participants who voluntarily sent them after using an app on lexical decision tasks. The authors not only were able to replicate one of the most robust effects in word recognition (the word frequency effect), but they also found a direct relation between the reaction times reported in both tasks. This shows us the benefits of new technologies, such as the smartphones and tablets. Furthermore, these emergent designs of serious game have showed several benefits for mild dementia [21]. Because of this, it is very interesting to use mobile devices tablets or smartphones for implementing smart games. Figure 5 shows the worldwide smartphones sales.

One arising problem is that smart games should be implemented on the major operating systems used on mobile devices. The trend changes very quickly. Right now Android and iOS are

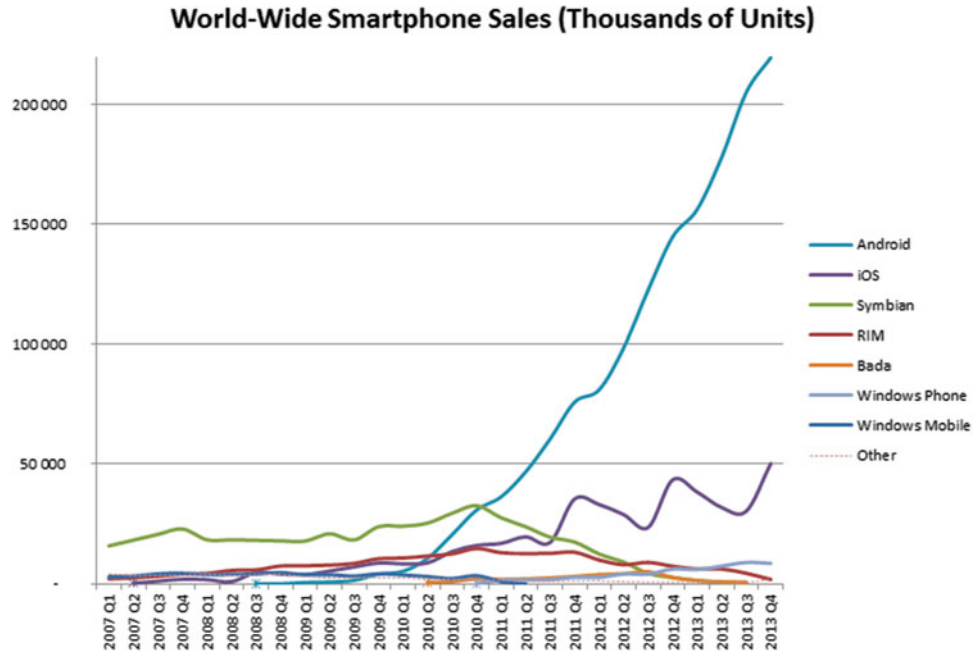

Fig. 5 Worldwide smartphone sales (thousands of units)

the most used operating systems, but at the beginning of 2010 Symbian, RIM and IOS were the most used.

In any case, it is very important to emphasize the fact that the software development process must take into account that the reaction time is the cornerstone to analyze data. Because neuroscience studies made in the past decade has suggested a problem with obtaining millisecond-accurate timing in some computer-based studies. Timing inaccuracies can affect not only response time measurements, but also stimulus presentation and the synchronization between equipment. However, as some researchers indicated, even, if it is desirable to examine this point, generally it is not required [22].

Because of such fact care must be taken to grant that experiments are repeatable, taking into account:

1. The presentation time is the same independently of the operating system (Android, iOS, Windows phone) used in the platform or the device hardware.
2. The reaction times are recorded with the minimal disturbance due to the device hardware.
3. The application is executed without disturbance caused by other processes running inside the mobile device.

In other words, the application must show the test at the same speed and must obtain the same values when running on different operating systems and even if it executed on the same platforms but with different hardware configuration (processor, graphics processing unit, etcetera.)

Finally, it is recommended that the mobile devices store the recorded data locally an in a cloud server, in order that the data could be analyzed using cloud computing and/or data mining tools offline.

There is a growing body of research on cognitive activities and games for the elderly as we have already mentioned, the older group tends to be slower than the young while they are involved in serious games. Nevertheless, if the older people are slower than the young, it is expected that a distribution shift occurs, but the shape would be similar to the young (in terms of distribution components). Yet, if a deficit on attentional demands is produced, differences on parameter τ are expectable, and therefore, in the distribution shape change. Analysis of Navarro et al. [9] claimed that changes in the τ parameter were found with word frequency but not with the load of the demand. However, more research in this issue is necessary.

The methodology process in this area is to test in a laboratory a developed battery for the seniors. Then, it is possible to test the serious games in an app format. Over these lines, a linear relationship between lab responses and iPhones and iPads has been found [20].

Furthermore, an optimization of these resources was suggested in the present work. It will allow us, even employing "noisy" (or contaminated) to fit data distribution, and the most interesting, to reflect underlying cognitive processes.

The study of new technologies, particularly for seniors, is a relatively new field that has been approached from many different disciplines. Sciences like biology or medicine have been commissioned to study the physical changes associated with aging. This type of multidisciplinary approaches offers new perspectives on the given topic. In this case, we are interested, on the one hand, in a mathematical approach through adjustment of probability functions. On the other hand, to design and implement games which makes data registration in cloud storage and then the data are analyzed using cloud computing and/or data mining.

References

1. Central Intelligence Agency (2001) The CIAWorld Factbook. http://www.cia.gov/cia/publications/factbook/
2. Osorio AR (2007) Os idosos na sociedade atual. In: Osório AR, Pinto FC (eds) As pessoas idosas: contexto social e intervenção educativa. Instituto PIAGET, Lisboa
3. Abt C (1987) Serious games. University Press of America, Washington, DC, USA
4. Van Muijden J, Band GP, Hommel B (2012) Online games training aging brains: limited transfer to cognitive control functions. Front Hum Neurosci 6:221
5. Ball K, Edwards JD, Ross LA (2007) The impact of speed of processing training on cognitive and everyday functions. J Gerontol B Psychol Sci Soc Sci. doi:10.1093/geronb/62.special_issue_1.19
6. Noack H, Lövdén M, Schmiedek F, Lindenberger U (2009) Cognitive plasticity in adulthood and old age: gauging the generality of cognitive intervention effects. Restor Neurol Neurosci. doi:10.3233/RNN-2009-0496
7. Moret-Tatay C (2013) Analysis of developmental changes in lexical decision tasks: differences between well elderly and university students. Doctoral Dissertation, Universidad Politécnica de Valencia
8. Moret-Tatay C, Moreno-Cid A, Argimon IIL et al (2014) The effects of age and emotional valence on recognition memory: an ex-Gaussian components analysis. Scand J Psychol. doi:10.1111/sjop.12136
9. Navarro-Pardo E, Navarro-Prados AB, Gamermann D et al (2013) Differences between young and old university students on a lexical decision task: evidence through an ex-Gaussian approach. J Gen Psychol. doi:10.1080/00221309.2013.817964
10. Lacouture Y, Cousineau D (2008) How to use MATLAB to fit the ex-Gaussian and other probability functions to a distribution of response times. Tutor Quant Methods Psychol 4:35–45
11. Luce RD (1986) Response times: their role in inferring elementary mental organization. Oxford University Press, New York
12. Matzke D, Wagenmakers EJ (2009) Psychological interpretation of the ex-Gaussian and shifted Wald parameters: a diffusion model analysis. Psychon Bull Rev. doi:10.3758/PBR.16.5.798
13. Leth-Steensen C, Elbaz ZK, Douglas VI (2000) Mean response times, variability, and skew in the responding of ADHD children: a response time distributional approach. Acta Psychol. doi:10.1016/S0001-6918(00)00019-6
14. Spieler DH, Balota DA, Faust ME (1996) Stroop performance in healthy younger and older adults and in individuals with dementia of the Alzheimer's type. J Exp Psychol Hum Percept Perform 22(2):461
15. West R, Murphy KJ, Armilio ML et al (2002) Lapses of intention and performance variability reveal age-related increases in fluctuations of executive control. Brain Cogn. doi:10.1006/brcg.2001.1507
16. Ijsselsteijn W, Nap HH, de Kort Y et al (2007) Digital game design for elderly users. In: Proceedings of the 2007 conference on future play. ACM, pp 17–22
17. Rogers WA, Fisk AD (2000) Human factors, applied cognition, and aging. In: Craik FIM, Salthhouse TA (eds) The handbook of aging and cognition, 2nd edn. Lawrence Erlbaum Associates, NJ, pp 559–592

18. Fozard JL, Thomas JC, Waugh NC (1976) Effects of age and frequency of stimulus repetitions on two-choice reaction time. J Gerontol 31(5):556–563
19. Forster KI, Forster JC (2003) DMDX: a windows display program with millisecond accuracy. Behav Res Methods Instrum Comput. doi:10.3758/BF03195503
20. Dufau S, Duñabeitia JA, Moret-Tatay C et al (2011) Smart phone, smart science: how the use of smartphones can revolutionize research in cognitive science. PLoS One. doi:10.1371/journal.pone.0024974
21. Cappeliez P, O'Rourke N, Chadbury H (2005) Functions in reminiscence and mental health in later life. Aging Ment Health 9:295–301
22. Damian MF (2010) Does variability in human performance outweigh imprecision in response devices such as computer keyboards? Behav Res Methods. doi:10.3758/BRM.42.1.205

Index

A

Actigraphy
- device 3–7
- quantification 4–8, 11, 16
- signal modelling 4–16
- signal pre-processing 5–8, 11
- simulation 4, 13, 15, 16

Aging 159, 219, 231–232, 259, 261, 266
Algorithms
- cost-sensitive 20
- incremental 53, 58–61, 63–65, 69, 70, 72, 76
- machine 11, 20, 125
- reinforcement 52
- statistical 58
- supervised 76, 126
- unsupervised 126

Alzheimer 159–173
Atrial fibrillation (AF), 219–224, 227, 228
Audit 39–55, 250

B

Bayesian inference 59–62
Brain tumours 30–32, 34, 41–43, 46, 48, 54, 55, 59, 66–69, 71–73, 76
Breast cancer 66, 71, 72, 102, 175–188

C

Cardiac
- arrhythmia 217–232
- simulation 230–232

Chronic disease 191–215, 237, 239, 240
Circadian rhythm 3, 4, 7
Classification 11, 21, 22, 24, 26, 35, 36, 40, 42, 47, 51–54, 58–61, 63–69, 71, 73, 76, 124–128, 137, 143, 163–166, 168–171, 176, 178, 181, 183–185, 188, 199–202
Classifier comparison 33, 40, 51, 52, 69
Clinical pathways 79–87, 99, 199
Cloud computing 112, 147–154, 265, 266
Context-aware 147–154
Cost-effective 3

D

Data
- complexity of 224
- heterogeneous 90, 93, 102, 103, 110, 121, 152
- imbalanced 19–36
- shift 53–55, 93
- temporal 89–103
- warehouse 100, 101, 103, 110, 112, 120, 121, 176

Data mining, temporal 89–103
Decision
- Bayesian 20, 21
- Support System (DSS) 5, 36, 39–55, 58, 76, 192, 193, 242, 244, 246, 251
- tree 7, 137, 160, 164, 165, 176, 177
- workflow 41

Dementia 4, 264
Diabetes mellitus 193, 237–255
Distribution in times 60

E

Electrocardiogram (ECG) 90, 218–220, 223–226, 231–232
Electronic Health Record (EHR) 91, 98, 109–113, 117–118, 120, 139, 148–151, 241, 246–247, 251
Empowerment 131, 141, 239, 240, 252
Evaluation, metric 19–36, 54

F

Feature extraction 4, 11, 12, 47, 67, 161
Functional data analysis 9

G

Generalization 22, 54, 59, 65
Genetic algorithm (GA) 177–181, 183, 186–188

H

Health recommender system 131–144
Hospital information system (HIS) 89, 102

Carlos Fernández-Llatas and Juan Miguel García-Gómez (eds.), *Data Mining in Clinical Medicine*, Methods in Molecular Biology, vol. 1246, DOI 10.1007/978-1-4939-1985-7, © Springer Science+Business Media New York 2015

I

Insomnia....3

K

Kernel density estimation....13
Knowledge Discovery in databases (KDD), 177

L

Learning....11, 19–36, 47, 53, 54, 58–61, 63, 66, 71, 73, 74, 76, 81, 82, 126, 135, 141, 176, 177, 179, 185, 196, 213, 215, 238, 254
Lexicon....124, 126, 127
Linear discriminant analysis (LDA)....160, 164, 165
Logistic regression, incremental....53, 57–76
Loss function....19–36

M

Machine learning....11, 20, 30, 65, 124–126, 128, 137, 143, 177, 197, 200
Magnetic resonance spectroscopy (MRS)....41, 46, 47, 54–55, 66, 68
Major depression....4–7, 36
m-Health....147–154
Mobile devices....147, 153, 264–266
Monitoring
  non-intrusive....4, 5
  non-stigmatizing....4, 5
  outpatients....3–16, 239, 241, 252

N

Natural language processing (NLP)....80, 102, 123–124
Neural networks....11, 137, 176, 177, 185–186
N-grams....124, 126

O

Odds
  dynamic....46, 47, 49, 53–55
  posterior....40, 41, 45, 47, 49–50, 53
  prior....40, 41, 47
  static....44, 46–47, 49, 53–55
Ontology....40, 42, 52, 55, 80, 93, 101, 102, 139, 143
Outcome....3, 4, 100, 109, 138, 141, 178, 181, 238, 240, 251

P

Part-of-speech (POS)....126
Personal
  Health Record (PHR),....131, 141–142, 148–151, 195
  Health Systems (PHS)....191, 237–255
  information....246
P4 medicine....57, 76, 237–255
Prediction....20–23, 25, 30–32, 39–41, 45–51, 53, 55, 57, 58, 61, 63–65, 98, 102, 119, 124, 126–128, 134, 137, 165, 176–179, 183–185, 187, 192, 194, 196–199, 222, 223, 230, 238, 239, 250–252
Probability, prior....27, 33, 40, 41, 48, 60–63, 72, 73
Process mining....79–87, 99, 102

R

Regularization....35, 52, 225–226
Rule
  Based Systems (RBSs)....80, 240, 244, 246, 249
  discovery....180, 204, 206

S

Sentiment analysis....124–129, 143–144
Serious games....259–266
Simulation....4, 13, 15, 16, 46, 47, 63, 64, 68, 69, 84, 198, 210, 230–232, 248–250
Sleep disorders....3, 4, 7
Smartphone....4–6, 147, 148, 195, 264
Social web....123, 128
Speech analysis....159–173
Support vector machines (SVMs)....11, 35, 125, 160, 164–166, 168–173, 198

T

Tagging....11, 124, 126
Temporal abstraction....92–96

V

Ventricular fibrillation (VF)....220, 227, 228

W

Web 2.0....128
Wireless connectivity....6
Workflows....40, 41, 80–82, 86, 87, 99