Health Informatics Journal

Volume 24 Number 4 December 2018

ISSN 1460-4582
journals.sagepub.com/home/jhi

Health Informatics Journal

Volume 24 Number 4 December 2018

Contents

Special Issue: UCAmI-IWAAL-AmIHEALTH conference

Special Issue: ISHIMR conference

Original Articles

Health Informatics Journal

Subscriptions and advertising

Health Informatics Journal (ISSN: 1460-4582 print, 1741-2811 online) is published in March, June, September and December by SAGE (Los Angeles, London, New Delhi, Singapore, Washington DC and Melbourne). Annual subscription (2018) including postage: Individual Rate (print only) £100/ US$185. Combined Institutional Rate (print and electronic) £492/US$911. Electronic only and print only subscriptions are available for institutions at a discounted rate. Note VAT is applicable at the appropriate local rate. Visit journals.sagepub.com/home/jhi for more details. To activate your subscription (institutions only) visit online.sagepub.com. Abstracts, tables of contents and contents alerts are available online free of charge for all. Student discounts, single issue rates and advertising details are available from SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP, UK, tel. +44 (0)20 7324 8500, fax +44 (0)20 7324 8600 and in North America, SAGE Publications Inc, PO Box 5096, Thousand Oaks, CA 91320, USA.

SAGE is a member of CrossRef.

Abstracting and indexing

Please visit journals.sagepub.com/home/jhi and click on the 'More' button (under about this journal), then click the Abstracting/Indexing tab to view a full list of databases in which this journal is indexed.

Disclaimer: The authors, editors, and publisher will not accept any legal responsibility for any errors or omissions that may be made in this publication. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed by Henry Ling Ltd, Dorchester, UK.

Special Issue: UCAmI-IWAAL-AmIHEALTH conference

Health Informatics Journal

Health Informatics Journal
2018, Vol. 24(4) 337–355
© The Author(s) 2016
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1460458216669555
journals.sagepub.com/home/jhi
SAGE

An eHealth information technology platform to help the treatment of mental disorders

Alexis Quesada-Arencibia, Enrique Pérez-Brito and Carmelo R García-Rodríguez
University of Las Palmas de Gran Canaria, Spain

Ana Pérez-Brito
Fundación Canaria Contra la Leucemia Alejandro da Silva, Spain

Abstract

For this project, we have used new technologies to create a new channel of communication between doctors and patients in the treatment of mental disorders. We have created a web application using an adaptable design accessible from any mobile device, which allows doctors to adapt their patients' therapy to real-time knowledge of their current condition. In turn, patients can express their mood state with respect to the component elements of their therapy.

Keywords

doctor–patient communication, emotional diary, mental disorders, psychological test, therapeutic adherence

Introduction

Mental illnesses rank among those that cause the greatest impact to the patient and their family. One of the main shortcomings in the treatment of mental disorders is the lack of real-time information on the status of the patient. Through this project, we aim to provide a new channel of communication between doctors and patients, using the Internet and new technologies, to facilitate the adaptation of patient therapy to their needs by receiving real-time updates on their condition. These real-time updates on the mood state of the patient in relation to the component elements of their therapy will enable the doctor to perform an immediate adjustment to the therapy to improve patient outcomes.

Corresponding author:
Alexis Quesada-Arencibia, University Institute of Cybernetic Science and Technology, University of Las Palmas de Gran Canaria, Campus de Tafira, 35017 Las Palmas, Spain.
Email: alexis.quesada@ulpgc.es

This project was developed in the context of the public health system of the Autonomous Community of the Canary Islands. No similar tool is currently used by this health service. Only one system, the 'drago system', is used to obtain information on scheduled visits and view clinical and medication history.

In the private sector, there are various psychological care websites (none of them psychiatric) offering online emotional care (entire treatment) or a first consultation method. Email is also used frequently as a tool for monitoring and/or virtual consultations between two appointments that are very far apart.

Currently, various techniques based on new technologies are being employed in the field of health care. These include, specifically, the field of cognitive software, specialised in stimulating and enhancing mental skills associated with the processes of learning, visual memory or linguistic stimuli. This type of software permits the tests to be configured to parameters defined by the patient's needs, taking into account aspects such as modality (visual or audible instructions), level of difficulty or response time. Some examples of cognitive software are telepsychology, virtual reality, augmented reality, video games, telecare or robots.

Related works

With the increasing use of Internet and information and communications technology (ICT) in recent years, it is natural to observe the union of health scientists with computer scientists in the research and development of technological solutions that advance the current state of global health.[1,2] Indeed, a great deal of significant literature has already been published on the application of ICT to help treat and monitor people with mental disorders. Health information technology (HIT) and its subset, ICT, are increasingly being applied to facilitate communication between health-care provider and caregiver.[3] Following this line of research, we can highlight the work of Richards,[4] who reports on the use of the online counselling service at Trinity College Dublin, including its uptake and usage, the issues and benefits of online counselling to students and whether clients are satisfied with their experience of online counselling. Colder-Carras[5] examined the extent to which patients at an inner-city community psychiatry clinic had access to ICT and how they used those resources. They concluded that a majority of patients in that community psychiatry clinic sample use ICT. Greater access to and use of the Internet by those with mental illness has important implications for the feasibility and impact of technology-based interventions. Timpano et al.[6] focused on the use of telehealth in neurological practice, highlighting the potential benefits of also applying information and communication technology to psychosocial and educational aspects of the treatment of neurological diseases. They concluded that one of the main advantages in the application of ICT solutions to neurology is the ability to build relationships across families and care systems. Indeed, not only are the client and the health-care professional linked together but also others within the client community, such as family and other specialists and physicians. This 'social network' allows a multiple perspective evaluation: all the parties meet together and work towards a treatment plan based on specialist recommendations. Meiland et al.[7] explain the results of the European Rosetta Project. In this project, a user participatory design was adopted to develop an integrated system, which combines three previously developed assistive technology systems and is, in close cooperation with the target groups, adjusted to their needs and wishes. The three previously developed systems are the COGKNOW Day Navigator (CDN),[8] the EMERGE system[9] and the Unattended Autonomous Surveillance (UAS) system.[10] The functionality most often mentioned as relevant and useful by persons with dementia was help in cases of emergencies (with movement sensors). The functionalities most often preferred by carers were support with navigation outdoors and the calendar function. Other studies, such as those carried out by

Lopez,[11] Mateu-Mateu and Navarro-Gómez[12] and Hu and Naseer,[13] show the advantages of ICT for social integration and clinical improvement of people with severe mental disorders (SMDs).

There are a huge number of ICT tools for specific mental disorders. Lozano et al.[14] and Charitaki[15] studied software tools for teaching emotions to students with autism spectrum disorder. Romdhane et al.[16] presented an automatic video monitoring system for assessment of Alzheimer's disease symptoms, and Robert et al.[17] gave some important recommendations for ICT use in Alzheimer's disease assessment. Välimäki et al.[18] evaluated the effects of ICT in patient education and support for people with schizophrenia, and Van der Krieke et al.[19] presented a usability evaluation of a web-based support system for people with a schizophrenia diagnosis. There are also a lot of game tools to help in the diagnosis or treatment of different mental disorders. Tárrega et al.[20] proposed a videogame as an additional therapy tool for training emotional regulation and impulsivity control in severe gambling disorder, and Bagga et al.[21] discussed a framework for designing games for cognitive assessment. Some studies deal with the use of mobile phones as medical devices in mental disorder treatment. Gravenhorst et al.[22] discussed how mobile phones can support the treatment of mental disorders by (1) implementing human–computer interfaces to support therapy and (2) collecting relevant data from patients' daily lives to monitor the current state and development of their mental disorders.

However, although there are many specific tools to aid the treatment of mental disorders, our article presents a complete, comprehensive tool without focusing on one specific pathology. The aim is to design a complementary, cross-cutting tool that makes it possible to

1. Establish continuity of care;
2. Address time constraints;
3. Bridge geographical barriers, a particularly critical aspect in the geographical context in which we propose to implement the system: the Canary Islands, where most of the specialised services are centralised on the two main islands (Gran Canaria and Tenerife), services which are, in many cases, non-existent on the smaller islands (Lanzarote, Fuerteventura, La Graciosa, La Gomera, El Hierro and La Palma).

Furthermore, in section 'Case of use', we describe the structure of the Canary Islands health system, where we aim to implement the proposed system and the metrics that we intend to use to evaluate its usefulness and impact once implemented.

eHealth

eHealth represents a change in our approach to health care. It is a technology at the service of everyone, so we can live a healthier life. We can monitor our vital signs and keep track of our treatments for better therapy compliance and even use remote medical consultation. This change is enabling data collection which, when analysed, will offer a more accurate and reliable diagnostic and therapeutic approach.

Mobile applications and new online communication tools are undergoing exponential growth as we seek immediate answers to all our health queries. The technologies that have been adopted to improve doctor–patient communication have brought some benefits, although they have also introduced new medical–legal and patient privacy risks.

One of the main advantages of eHealth is access to reliable quality information that helps resolve general queries about the health of the patient when they arise. The patient becomes the protagonist of the process and is able to store all information on his or her condition in his or her own medical history that he or she can share with the therapists.

Another advantage is the ability to stay in contact with therapists with waiting times that are shorter than under the current system. It reduces delays and unnecessary travel for all kinds of administrative tasks and consultations that can now be done online.

Finally, it facilitates learning so that the patients become increasingly autonomous in caring for their own health and for that of their dependants. Some of the advantages of eHealth are speed, low cost, asynchrony, accessibility, permanence, the absence of barriers and a reduction in unnecessary visits.

Design and development

In the analysis phase, we identified three different types of actors who interact with the web application. The most general user type is the 'unregistered user', all those who are not 'registered users'. Within the registered users, we have the roles 'Doctor' and 'Patient':

- *Unregistered user.* Users who access the application without identifying themselves. They may register if they want to see a doctor or login to identify themselves and access the corresponding features and be they doctor or patient.
- *Doctor.* Uses the features provided by the Doctor module; has a profile with personal, professional and login information; is responsible for creating patient user profiles; and organises patient information, managing their history, treatment and clinical information.
- *Patient.* Uses the features provided by the Patient module, can edit login information and some basic fields in their medical history through their profile and responds to events generated by his or her treating doctor and can generate others in turn through his or her emotional diary.

The design of the web application separates the system functions into three modules, each module covering the actions that can be performed by each user. The modules that have been developed are as follows:

- *Application*. Contains the public part of the web application; displays information on the project features and some mental illnesses and their effects on the family; and has access to login area and, for doctor users, user account creation.
- *Doctor.* Contains the private part intended for doctor users; permits administration of all information pertaining to patient records, clinical history and emotional diary; and contains psychological test editor.
- *Patient*. Contains the private part intended for patient users. In this module, the patient can manage elements of his or her emotional diary and send feedback on elements of his or her therapeutic adherence (Figure 1).

Information access control

Since this web application handles highly sensitive data, we chose to use various systems that guarantee the privacy of patients' personal data and their treatments. These systems include restricted access via personal login that asks the users to identify themselves when accessing the application's database and an access control list (ACL).

The ACL enables us to define a range of user roles and grant or remove permission to access certain parts of the application. This means that the list is a tool that controls information access for the corresponding user. To properly define the ACL, we first need to clarify two concepts:

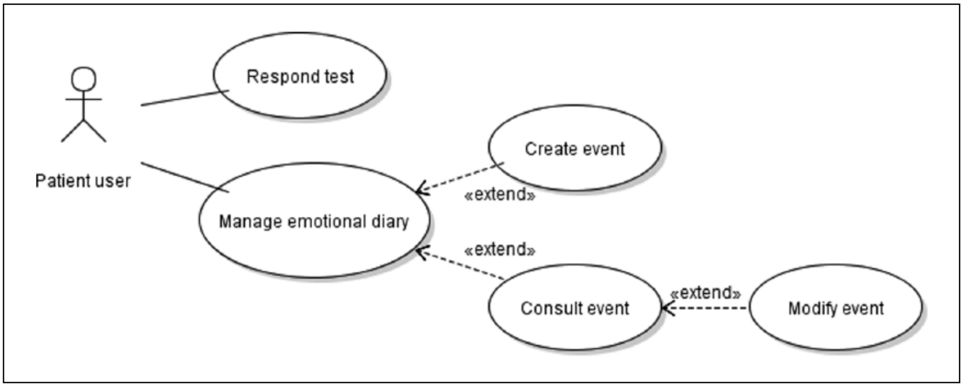


Figure 1. Patient user functions.

- *Resources*. Access-restricted objects;
- *Roles*. Subjects that will request access to a resource.

In simple terms, roles request access to resources. A well-defined ACL will enable an application to control which roles have access to each resource.

This web application has three types of users:

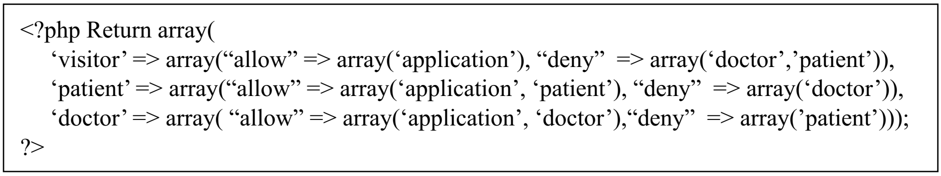
- *Visitor*. Users who have not identified themselves on the system and can only access public content;
- *Doctor*. Role assigned to doctors, who can access the functions available in the Doctor module;
- *Patient*. The role assigned to the patients, who can access the functions available in the Patient module (Figure 2).

Psychological test editor

One of the most salient features of this project is the psychological test editor that allows the doctor to create and modify tests adapted to the needs of each patient. With this editor, the doctor can create tests with various types of questions: simple, yes or no, true or false, relationship or cause or Likert scale. The editor has other features such as creating private entries for the doctor or a system to evaluate the test results, either as a whole or the individual answer to each question.

After the more general pre-diagnostic tests, and after assessment and diagnosis of the patient, this tool may be used to ascertain the patient's new needs that result from the changes brought about by adherence to a therapy. Being able to measure progress on a regular basis supplies us with information on these changes, and we can also, therefore, provide the patient with this same information (goals achieved and those yet to be attained). Furthermore, the language of the tests may be adapted to the mental capacity of the patient, adapting them to their reality.

To obtain the parameters for the online psychological test editor, we drew on the work by Bobes et al.,[23] which contains a compilation of all test types used in psychology as well as their objectives and types of questions used. By studying the questionnaires in this book, we gained an idea of the kinds of questions required to create the tests. Table 1 shows the questionnaires that were analysed and the objectives that they pursue.



```
<?php Return array(
    'visitor' => array("allow" => array('application'), "deny"  => array('doctor','patient')),
    'patient' => array("allow" => array('application', 'patient'), "deny"  => array('doctor')),
    'doctor' => array( "allow" => array('application', 'doctor'),"deny"  => array('patient')));
?>
```

Figure 2. ACL code.

Table 1. Types of questionnaire analysed.

Test	Objective
Assessment tools for organic mental disorders	Mini–Mental State Examination (MMSE) Alzheimer's Disease Assessment Scale (ADAS)
Assessment tools for disorders caused by the consumption of psychotropic substances: alcohol and other drugs	CAGE Questionnaire MALT AUDIT European Addiction Severity Index (EuropASI)
Assessment tools for schizophrenic disorders	Positive and Negative Syndrome Scale (PANSS) Brief Psychiatric Rating Scale (BPRS) Overt Aggression Scale (OAS) Scale to Assess Unawareness of Mental Disorder (SUMD) Drug Attitude Inventory (DAI)
Assessment tools for mood disorders	Udvalg für Kliniske Undersogelser (UKU) Hamilton Depression Rating Scale (HDRS) Montgomery–Åsberg Depression Rating Scale (MADRS) Geriatric Depression Scale (GDS) Calgary Depression Scale (CDS) Mood Disorder Questionnaire (MDQ)
Assessment tools for suicidal ideation and behaviour	Beck Hopelessness Scale (BHS) Scale for Suicidal Ideation (SSI) Beck's Suicide Intent Scale (SIS) Plutchik's Impulsivity Scale (IS) Plutchik's Violence Risk Scale (VR) Plutchik's Suicide Risk Scale (SR)
Assessment tools for neurotic and stress-related disorders	Hamilton Anxiety Rating Scale (HAM-A) Bandelow's Panic and Agoraphobia Scale (P&A) Liebowitz Social Anxiety Scale (LSAS) Watson and Friend's Social Avoidance and Distress Scale (SADS) Yale–Brown Obsessive Compulsive Scale (Y-BOCS) Davidson Trauma Scale (DTS) Trauma Questionnaire (TQ) 8-item Treatment-Outcome PTSD Scale (TOP-8) Duke Global Rating Scale for PTSD–Improvement (DGRP-I)
Assessment tools for disorders associated with physiological dysfunctions and somatic factors	Eating disorders Eating Disorder Inventory (EDI) Bulimic Investigatory Test, Edinburgh (BITE) Oviedo Sleep Questionnaire (OSQ) Changes in Sexual Functioning Questionnaire (CSFQ)
Assessment tools for global clinical status	Clinical Global Impression (CGI)

Table 1. (Continued)

Test	Objective
Personality disorder assessment tools	Eysenck Personality Questionnaire (EPQ-A) International Personality Disorder Examination (IPDE) Sensation Seeking Scale (SSS)
Level of functioning assessment tools	WHO Disability Assessment Schedule (WHODAS) Global Assessment of Functioning Scale (GAF) Sheehan Disability Inventory (SDI)
Health-related quality-of-life assessment tools	Short Form (36) Health Survey (SF-36) WHOQOL Quality-of-Life Assessment (WHOQOL-100) Seville Quality-of-Life Questionnaire (CSCV) Alzheimer's Disease-Related Quality of Life (ADRQL)

PTSD: post-traumatic stress disorder; MALT: Munich Alcoholism Test; AUDIT: Alcohol Use Disorders Identification Test.

Bibliotherapy

The therapeutic function resides in the healing, restorative and preventive effects of reading. Reading encourages a change of individual behaviour not only at the time of the crisis but also in individual routines. It assists the patient in developing the faculty of self-criticism and prompts a desire to make changes to help them adapt to their customary environment. Autonomous learning linked to experience is what, therefore, motivates them to constantly update and review their conduct and its repercussions beyond intentions without defined goals or personal statements and situations.

Workshop monitoring

Workshop monitoring is another tool for improving the behaviour of individuals engaging in activities to achieve the objectives proposed in therapy (socially proactive, search for alternative solutions, time management, information about toxic products, etc.).

Doctor–patient communication model – implementation

Doctor–patient communication conforms to several models that define how to handle conversations involving situations related to the patient's health and how to make them see the reality that surrounds them.

To carry out these communications, various models have been defined that determine how the doctor achieves this reality approach when establishing contact with the patient.

Following these models, we can define this relationship as a meeting between two people, one of them the patient who needs help to recover his or her health and the other the doctor who is trained to provide this help. This relationship depends on the cultural, scientific and technical circumstances of each time and place.

Of all the models of doctor–patient communication, we have selected the Veatch[24] model for the purposes of this project. This model considers that the contract to be established is a consensus or agreement based on the theme that motivates the meeting: the health of the patient.

The diagnosis is made by the doctor, but the responsibility for the therapy is shared. There is respect for the autonomy of the patient who is informed in order to be able to make an informed choice.

This model is the one that seems best suited to the nature of this project; when attending therapy, doctor and patient can agree on the steps to be taken, but the doctor cannot force the patient to follow them as directed. Another significant aspect of this model of communication is the importance that is given to the feedback that the patient gives to the doctor with respect to the therapy, an essential aspect in reinforcing the rationale of the project: real-time therapy adjustments according to the patient's needs.

Communication models follow several phases as the therapy progresses, according to renowned doctor Laín Entralgo.[25] These phases are summarised as follows:

- *Cognitive moment*. Stage at which the link between doctor and patient is established. The interest that binds both sides of this relationship is represented by the desire to recover health, but the person suffering from the condition is the patient, not the doctor.
  In this interaction, the doctor employs scientific knowledge to name, describe and set out what ails the patient; at the same time, the patient contributes with his or her ideas and emotions. The result is a medical diagnosis.
- *Operative moment*. Refers to the therapeutic activity of the doctor, from empathic listening at the start until the final send-off.
  Therapeutic action begins when the patient decides to seek medical advice, before the actual appointment, and does not end until final discharge. The moment of diagnosis is also therapeutic.
- *Affective moment*. The author argues that there are two forms of affectional bonding between the doctor and the patient.
- *Medical camaraderie*. Both the doctor and the patient seek to remedy the condition and achieve good health but with little personal commitment. The patient, if cured, is grateful and becomes emotionally attached to the doctor, albeit not very deeply, because of the service provided.
- *Medical friendship*. Characterised by trust whereby the patient can confide their innermost thoughts and emotions in the doctor.

The most important element for the doctor is the principle of bioethics: the intention to 'do good' for the patient, bearing in mind that the sought-after good is their health. The relationship in this case is both technical and affectional.

In this project, we may observe the following three stages described above:

- *Cognitive moment*. Applicable to all communication between doctor and patient. Recovery of the patient's health via the doctor's knowledge is enhanced whenever communication is established through the application, primarily through the doctor's messages and feedback from the patient as they follow the steps indicated by the doctor during therapy.
- *Operative moment*. Developed throughout the patient's therapy, from the first contact between doctor and patient, through the use of the application as a tool during therapy, to patient discharge and the end of the relationship with the doctor.
- *Affective moment*. Reflected in the use of the application itself, as the aim to improve communication between doctor and patient presupposes an intention between both parties to work on the affectional aspect of therapy. This improvement in communication between doctor and patient develops the spirit of both medical camaraderie and medical friendship, as described above.

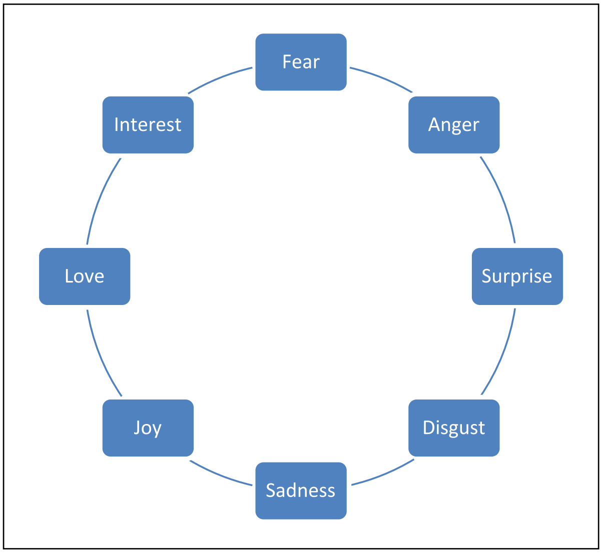


Figure 3. Primary emotions.

The following sections will describe aspects of the application that reflect and use the aforementioned doctor–patient communication model.

Emotional diary

This is another tool that is normally included in a wider package of measures to help patients express their emotions in their daily lives. The aim of this diary is to enable the patient, after a period of learning in which they relate emotions to physical stimuli and cognitions, to be able to recognise their own emotions, both primary and secondary. They will thus be able to, if need be, curb ill-adapted conduct and will learn to redirect their thoughts and put into practice learnt relaxation techniques to control and stimulate their own impulses in order to better adapt to the situation. The goal is to learn how to channel expressions of emotion and feeling in a healthier way by understanding, managing and using them to grow psychologically.

The emotional diary is a very useful tool to complement the therapy of a patient suffering from a mental disorder. It is an extra resource that helps to resolve certain problems that are emotional or have their origins in emotion. The purpose of this record is not to provide a solution to every feeling but to identify it and give it its exact name (Figure 3).

The emotional diary aims to elicit pure emotional expression, as this allows patients to better understand themselves in order to build self-assurance, reduce their fears and anxiety in new situations, learn to resolve problems and understand how to identify and self-regulate emotions.

Writing down these experiences will increase the patient's perception of different situations and the correct way to deal with them because it will enable them to recognise what they are feeling in certain situations and create behavioural patterns that they will then analyse so that they can decide how to act or react. It also provides the doctor, over time, with relevant information about the patient's evolving management of their emotions.

To reflect the emotional diary on the web application, the patient is provided with a range of features to express their mood state using three methods:

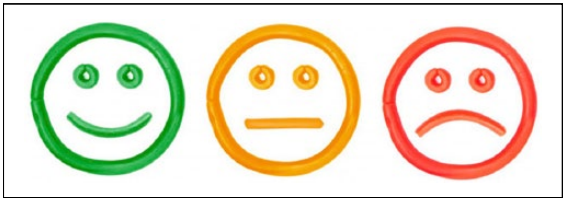


Figure 4. Mood state symbols.

- *Write a diary entry*. As if it were a physical diary, the patient has a section in which they can type up an entry in their emotional diary. As with a normal diary, the patient can add the date, title and development to the diary entry.
- *Create new mood state*. Another addition to the emotional diary is an indication of mood state. When the patient suffers any mood change and wants to communicate it to the doctor, they can do so in this section. Options are available here to indicate when the mood arose, give it a name, define what primary emotion most adequately describes their emotions, describe its intensity and explain the whole episode by reporting it.
- *Create new mood state associated with an image*. The patient is provided with this alternative way of describing a mood state. In this case, a mood is also described with the aforementioned features, but the explanation is reinforced by uploading an image. This caters to those patients who cannot write, or who are disorientated and have difficulties maintaining a coherent conversation due to the stage of their illness or their life situation, or for whom it is tedious and arduous to name or label their emotions. They are, however, capable of recognising their physical symptoms and the thoughts associated with the various mood states. They are provided with simple drawings that reflect them (Figure 4).

Guidelines to remember

There is no point in keeping an emotional diary if the doctor does not adjust the therapy after seeing the reactions of the patient. It is, therefore, important that the doctor develops guidelines to remember, which will enable the patient to recognise the emotions or situations that trigger an emotional crisis. Thus, they will be able to adequately channel events. Using these behaviour guidelines, the patient will be able to act correctly when these delicate moments arise and overcome their emotional deficiencies.

Therapeutic adherence

In addition to following medical advice, a number of changes in patient habits, lifestyle, thoughts and abilities are required to increase the efficacy of the treatment and obtain a better quality of life as the outcome. This combined approach is called ‘therapeutic adherence’.

The web application project seeks to group all elements of the patient’s therapy in one place. All elements of the patient’s therapy can be managed under the label of therapeutic adherence (consultations, medication, tests, workshops and bibliotherapy).

Feedback

In all the aforementioned sections, once the doctor has read the patient’s message, they can send a reply if they consider it appropriate. The patient, in turn, can send a message to his or her doctor

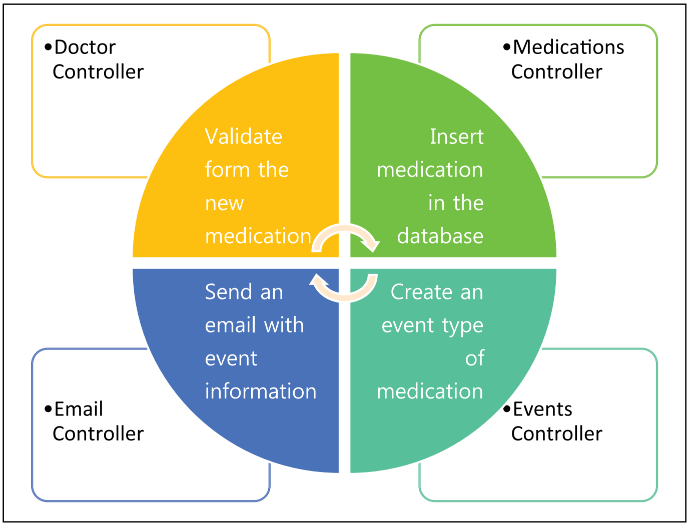

•Doctor Controller
•Medications Controller
Validate form the new medication
Insert medication in the database
Send an email with event information
Create an event type of medication
•Email Controller
•Events Controller

Figure 5. Controllers involved in the creation of an event.

about any element of his or her therapy whenever necessary. Both actors will be notified by the system of any entry or reply.

Feedback on therapies, visits and medication improves ongoing doctor–patient communication between appointments, which are often too far apart. It gives continuity to the information and consequently allows the treatment to be adapted to the changing needs of the patient. Moreover, a closer bond is forged with the family social unit (primary caregivers) which generates information that is useful when making timely changes to the therapy.

Event system

All this therapy monitoring would be impossible without a system that supports real-time communication between doctor and patient.[26,27] To this end, we have developed an event system that notifies users if they have any pending action to attend to. These notifications are displayed on the user's main screen. When a user creates a new action, an event is automatically generated together with a notifying email (Figure 5).

When the main controller – either the Doctor or Patient module – creates a new activity, it invokes the associated controller, and this second controller is responsible for creating the activity in the database, invoking the controller that manages the events and receives the emails.

As shown in Figure 5, if the doctor creates a new medication for the patient, the action is inserted into the medication table of the database and a medication type event is generated, and if the patient has notifications enabled in his or her profile, an email with information about the event is sent to him or her. Once the process is finished, we return to DoctorController, which will display the patient's list of medications with this latest medication already recorded.

The status of an event may vary over time, following a 'now the ball is in your court' approach, that is, when a user generates an event with an action, for them the status of this event is 'terminated', while the other user now has a 'new event' status. The aim is that when a doctor or patient generates an activity, the event informs the other so that they know they should attend to it (Figure 6).

This method has been developed with two status fields on the table that stores the events in the database: one reflecting event status for the doctor and one for the patient. An illustration of the

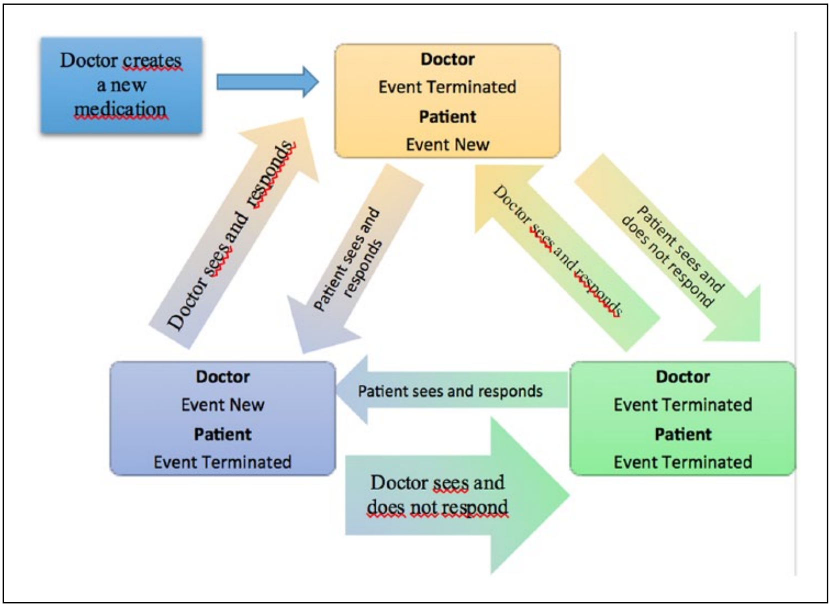


Figure 6. Event system flow.

changes that an event may undergo during use of the web application is shown in Figure 6; in this figure, it is the doctor who creates a new activity.

Interface

The interface is designed to be simple and user-friendly. The main actions that the user can perform are located in a horizontal bar at the top of the screen. Overloading the screen with elements will be avoided, while maintaining the maximum amount of information in the horizontal space. To avoid the user having to scroll vertically whenever possible, several systems have been designed:

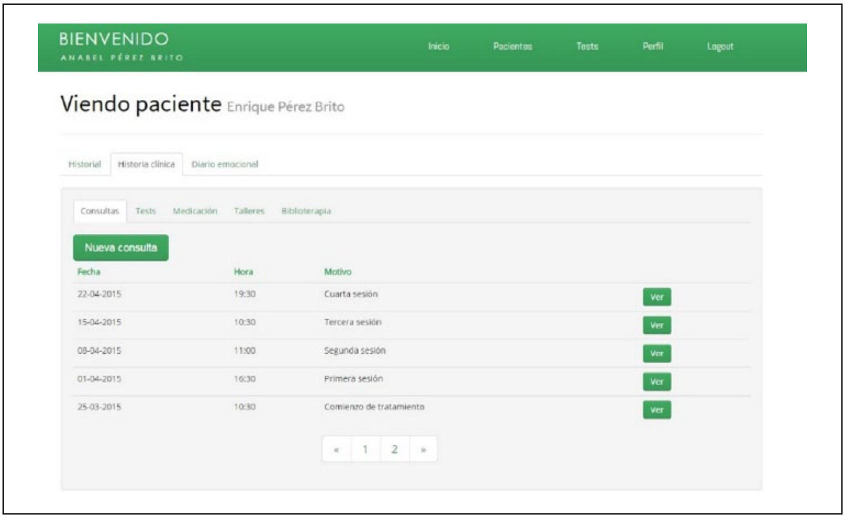
Paginated tables. Non-detailed information is displayed in paginated tables with up to five rows. If the table has more entries, these will be moved to a new page (Figure 7).

Search filters. The user can filter search results through a system of filters. All they need to do is select the desired filter and apply it to their search (Figure 8).

Tab system. The information is divided by a tabbed browsing system. Each tab represents an information category, hiding the other categories until the user wishes to consult them. Each time the user returns from a detailed view of an event, the system remembers the last selected tab (Figure 9).

Case of use

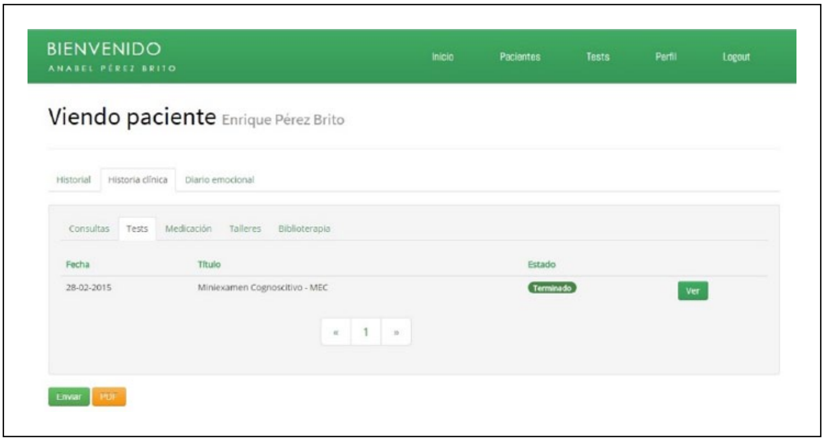
The aim of the tool that we have developed is to improve mental health care in the Canary Islands. More specifically, the tool is intended to obtain more and better patient information through improved communication between the patient and the doctor. Although the tool may be used with



Viendo paciente Enrique Pérez Brito

Fecha	Hora	Motivo	
22-04-2015	19:30	Cuarta sesión	Ver
15-04-2015	10:30	Tercera sesión	Ver
08-04-2015	11:00	Segunda sesión	Ver
01-04-2015	16:30	Primera sesión	Ver
25-03-2015	10:30	Comienzo de tratamiento	Ver

Figure 7. Paginated tables.



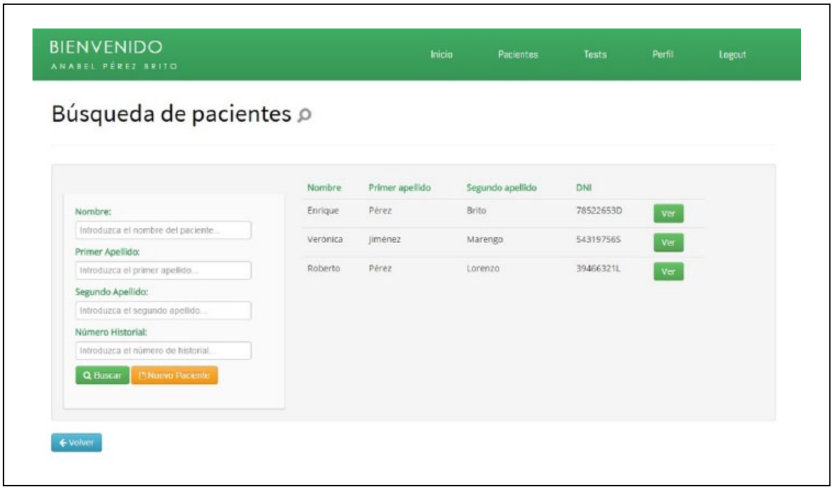
Viendo paciente Enrique Pérez Brito

Fecha	Título	Estado	
28-02-2015	Miniexamen Cognoscitivo - MEC	Terminado	Ver

Figure 8. Search filters.

patients capable of managing ICT resources, it could be a particularly effective resource for young patients because of the interest that this sector of the population shows in ICT tools. To contextualise the impact of the tool, we have provided an overview of the organisation, resources and activities of health-care provision for mental disorders in the public health network of the Canary Islands. We then propose a case of use for the tool and the metrics to be used to assess the impact of the tool in improving patient care, and we consider a proposal for a pilot project to assess its impact.

Mental health care in the Canary Islands is provided by a network of centres, of which there are two types: outpatient units and inpatient units. The use of our proposed system makes the most sense in outpatient units; hence, the following brief description of their structure and information relating to relevant activities.



Búsqueda de pacientes

Nombre: Introduzca el nombre del paciente...

Primer Apellido: Introduzca el primer apellido...

Segundo Apellido: Introduzca el segundo apellido...

Número Historial: Introduzca el número de historial...

Nombre	Primer apellido	Segundo apellido	DNI	
Enrique	Pérez	Brito	78522653D	Ver
Verónica	Jiménez	Marengo	54319756S	Ver
Roberto	Pérez	Lorenzo	39466321L	Ver

Figure 9. Tab system.

Mental health care in the public health system of the Canary Islands: outpatient units

Outpatient units are distributed throughout the Canary Islands, and there are currently 31 such units. This network of centres covers 100 per cent of the population, once they have been seen by the corresponding primary care centre. These units are, in turn, divided into three types:

- *Community Mental Health Units (CMHU)*. These units serve the entire population, and there are 24 such units.
- *Community Child Mental Health Units*. These are specific units to treat patients under 18 years. They have specific teams of physicians, and there are five units of this type.
- *Child Day Hospitals*. Currently, there are two units of this type, consisted of psychiatrists, psychologists, paediatricians specialised in neuropsychiatry, nurses, nursing assistants, occupational therapists, psychomotor specialists, social workers, special education teachers, porters and administrative assistants.

With regard to human resources in mental health outpatient units, these are organised into multidisciplinary teams. Table 2 shows the number of professionals assigned.

With regard to the provision of care in these units, Table 3 shows activity for the Community Mental Health Units during 2011–2015. The most prevalent diagnoses in these units in 2015 are as follows:

- Schizophrenia and other psychoses;
- Affective disorders;
- Anxiety and somatic symptom disorders;
- Behavioural and emotional disorders with onset usually occurring in childhood and adolescence.

Mental health-care provision from 2011 onwards for the population aged under 18 years is shown in Table 4.

Table 2. Human resources assigned to mental health outpatient units of the public health system in the Canary Islands.

Human resources CMHU Canary Islands 2015	Number of staff
Psychiatrists	70
Psychologists	68
Nurses	48
Clinical assistants	33
Administrative assistants	28
Social workers	20
Porters	3

Table 3. Provision of care related to mental health in outpatient units of the public health network of the Canary Islands, since 2011.

Care provision	2011	2012	2013	2014	2015
First consultation	4631	4522	4455	4434	3967
Follow-up consultations	33,306	37,016	41,116	40,226	37,683
Total consultations	37,937	41,538	45,571	44,660	41,650
Follow-up/first consultation	7.19	8.19	9.23	9.07	9.50
Patients seen	8333	8931	9598	10178	9541
Incidence rate	12	12	12	12	11
Prevalence rate	13	16	18	20	20
Attendance rate	84	94	105	103	97

Table 4. Provision of care related to mental health for the population aged under 18years in the public health network of the Canary Islands, since 2011.

Care provision	2011	2012	2013	2014	2015
First consultation	20,478	19,920	20,039	18,997	17,538
Follow-up consultations	324,088	336,669	352,940	341,410	330,478
Total consultations	344,566	356,589	372,979	360,407	348,016
Follow-up/first consultation	15.83	16.90	17.61	17.97	18.84
Patients seen	49,926	52,341	54,855	56,324	55,119
Incidence rate	10	10	10	9	8
Prevalence rate	18	20	22	23	23
Attendance rate	144	154	160	156	151

For patients under 18, the most frequent diagnosis in 2015 was by far for behavioural and emotional disorders with onset usually occurring in childhood and adolescence, with 4684 cases diagnosed.

Planned case of use: metrics to evaluate usefulness

The tool that we developed is a resource that enables practitioners to gather information about the mood state and emotions of their patients, complemented with the information gathered during consultations in health centres in order to monitor patient outcomes. Another feature of the tool is that it helps the patient recognise and express their emotions, facilitating the development of emotional intelligence in young patients with pathologies that affect their social skills. The usefulness of the tool will be evaluated by considering two sources of information that it provides:

- *The patient's diary entries*. This source of information will reveal to what extent the tool provides relevant information on the patient's history: information related to the patient's ability to represent and control their mood and the effects of the therapies.
- *Communications between the patient and the doctor*. This source of information will measure the impact of the tool in improving communication with the patient, insofar as it prevents unnecessary travel by the patient to the health-care facilities.

With regard to the indicators used to measure the relevant information provided by the tool, these are obtained by analysing the entries made by patients in their diaries. Specifically, these indicators are as follows:

- Number of entries created by the patient: This indicator measures whether patient trust and fluency with their doctor improves. In addition, sudden changes in the number of entries may be symptoms of patient status changes due to stages of worsening or crisis.
- Number of mood changes noted by the patient: This indicator reports the effects of treatments and also indicates worsening or crisis patient situations.
- Number of new mood states created by the patient: This indicator is used to detect improvements in patient status since it is able to identify their mood states and monitor them when they manifest new mood states.
- Effect of the therapies, analysing the correlation between the number and type of entries made by the patient and the prescribed therapies.

To assess how the tool enables improvement in the quality of life of patients, those entries in their diary that indicate better emotional control or represent their mood states are analysed. Indicators associated with improved patient quality of life are as follows:

- Entries that indicate the acquisition of new social skills or development thereof. These are related to social proactivity or the ability to find solutions to everyday situations.
- Diary entries that indicate changes in habits and lifestyle.

To measure the impact on patient care, patient–doctor communication records will be used to temporarily analyse said communications and entries to assess the effect of these communications.

To evaluate the tool, we propose developing a pilot project with a sample of young patients aged between 12 and 16 years with diagnosed disorders associated with behavioural and emotional disorders with onset usually occurring in childhood and adolescence. The reason for choosing this group of patients is that young people are a sector of the population that normally show interest in ICT and more easily acquire the ICT skills necessary to manage the application. In addition to the intervention of patients and doctors, this trial will require the collaboration of their teachers and tutors to check the findings of the indicators related to the development of new skills (social proactivity, finding solutions to everyday situations and changes in habit and lifestyle). Moreover, as initial data for each participating patient, their mental age will be obtained by performing a standardised test for this purpose. Based on this initial information, there will be a learning phase on how to use the tool, taking into account the characteristics of each patient. Once patients have mastered the use of the tool, it will be used during the school year in order to check the evolution of the patient with their tutors and parents.

Conclusion

The web application we have developed provides a number of essential tools for functional improvement of communication between doctors and patients. Some of the most salient features are psychological test editor, doctor–patient communication system, emotional diary and combining medical and clinical history.

The tool represents an advance in patient proximity and provides more continuous monitoring outside doctor visits. It offers the possibility of adjusting or changing medical and psychological treatment in a simple, fast way that reinforces patient autonomy.

Despite this being a tool that will enhance therapy, there are certain types of patients with whom it should not be used. Patients who may potentially suffer from using this tool are those who suffer from certain obsessive compulsive disorders, who are exhibiting signs of severe depression or patients with symptoms of paranoia.

The proposed tool can be used by all patients receiving any psychological therapy including cognitive behavioural counselling, systemic therapy, humanistic therapy … all apart from psychodynamic therapy. Furthermore, it could also be used in psychiatry by doctors and liaison nurses who are mental health specialists in home care programmes and day care units.

The web application could be used by patients with severe psychopathologies who remain under the care of their families; in these cases, the latter would have access to this channel of communication with health professionals in order to provide information on changes and/or new requirements that arise between scheduled visits. This would improve one of the major shortcomings of the Canary Islands health system, in particular, but one that also affects health care nationally: overlong waiting times between doctor–patient visits as a result of an overloaded health system.

In short, the development of all the tools described above facilitates, on one hand, the work of the professionals (more and better access to the experiences of the patient through the information that they and their family have reported); on the other hand, it promotes adherence to therapy through closer monitoring by the health professional and, finally, brings relatives closer to the therapeutic exchange that their loved one is undergoing.

Moreover, although it is not an essential requirement, the system provides benefits to the patient if used with mobile devices since there are several features of the web application that benefit from the use of this technology.

The patient can use the camera on their mobile device (mobile phone or tablet) to take a picture that describes a feeling better than any text in their emotional diary.

Another useful option that using the application on mobile devices offers is the ability to communicate with their doctor in real time in case they want to express some feeling they think is appropriate or indicate the trigger for a crisis once it has been identified, thanks to the guidelines to remember that were defined in therapy. The way in which the patient perceives many of these feelings may be affected over time if the patient has to wait for the next appointment to discuss them with their doctor. The immediacy of the platform on a mobile device is, therefore, vital.

Future work

After the current phase of system development, the first step would be to implement it in the Canary Islands health system, so that we may obtain the metrics described in section ‘Case of use’ and evaluate the real impact of the system.

The web application developed for this project has many features that can be expanded upon and improved. Instead of opting to use an adaptable design, a native mobile application could be developed for both Android and iOS. This would enhance the functionality of the system. For example, the representation of a mood in the emotional diary via an image would benefit from mobile device geolocation. Currently, patient information and treatment is stored in a database developed in SQL. A way could be found to export patient information by creating a format that presents the information in such a way that it may easily be transferred to another professional. Similarly, it would be useful to add a feature that allows the sharing of information on therapies or to create a method by which a psychologist can make an online inquiry to another colleague if they have any doubts about a patient's therapy. Finally, new stakeholders can be added to the system. As noted above, in some therapies, it is convenient to receive the active support of relatives or social workers, and they could be introduced to the system under new roles and functions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

References

1. Broderick M and Smaltz D. *HIMSS E-health white paper: e-health defined*. Chicago, IL: Healthcare Information and Management Systems Society, 2008.
2. Anacleto J, Silvestre R, Filho CS, et al. Therapist-centred design of NUI based therapies in a neurological care hospital. In: *Proceedings of the IEEE international conference on systems, man, and cybernetics*, Seoul, South Korea, 14–17 October 2012, vol. 12, pp. 2318–2323. New York: IEEE.
3. Gentles SJ, Lokker C and McKibbon KA. Health information technology to facilitate communication involving health care providers, caregivers and pediatric patients: a scoping review. *J Med Internet Res* 2010; 12(2): e22.
4. Richards D. Features and benefits of online counselling: Trinity College online mental health community. *Brit J Guid Couns* 2009; 37(3): 231–242.
5. Colder-Carras M, Mojtabai R, Furr-Holden CD, et al. Use of mobile phones, computers and internet among clients of an inner-city community psychiatric clinic. *J Psychiatr Pract* 2014; 20: 94–103.
6. Timpano F, Bonanno L, Bramanti A, et al. Tele-Health and neurology: what is possible? *Neurol Sci* 2013; 34: 2263–2270.
7. Meiland FJ, Hattink BJ, Overmars-Marx T, et al. Participation of end users in the design of assistive technology for people with mild to severe cognitive problems. *Int Psychogeriatr* 2014; 26(5): 769–779.
8. Meiland FJM, Reinersmann A, Sävenstedt S, et al. User-participatory development of assistive technology for people with dementia – from needs to functional requirements. First results of the COGKNOW project. *Non-Pharmacol Therap Dementia* 2010; 1: 71–91.
9. Storf H, Kleinberger T, Becker M, et al. An event-driven approach to activity recognition. In: *Proceedings of the AmI '09* (ed M Tscheligi, B de Ruyter, P Markopoulus, et al.) (LNCS, vol. 5859), Salzburg, 18–21 November 2009, pp. 123–132. Berlin, Heidelberg: Springer.
10. Jans A, Overmars-Marx T, Van Hoof J, et al. *Evaluatieonderzoek van het UAS-project van Zorgpalet Baarn-Soest, Zorg aan huis* [Evaluation study of UAS project Zorgpalet Baarn-Soest, Care at home]. Utrecht: Vilans, 2009.
11. Lopez A. An investigation of the use of internet based resources in support of the therapeutic alliance. *Clin Soc Work J* 2015; 43:189–200.
12. Mateu-Mateu JM and Navarro-Gómez N. Keys and evidences of the use of ICT in severe mental disorder: psychology. *Society Educ* 2015; 7: 85–95.

13. Hu B and Naseer A. Human centric ICT support to young persons with mental disorders. In: *Proceedings of the IEEE 28th international symposium on computer-based medical systems*, Sao Carlos, Brazil, 22–25 July 2015, vol. 15, pp. 354–355. New York: IEEE.
14. Lozano J, Ballesta J and Alcaraz S. Software for teaching emotions to students with autism spectrum disorder. *Comunicar* 2011; XVIII(36): 139–147
15. Charitaki G. The effect of ICT on emotional education and development of young children with autism spectrum disorder. (International Conference on Communication, Management and Information Technology (ICCMIT 2015)) *Proced Comput Sci* 2015; 65: 285–293
16. Romdhane R, Mulin E, Derreumeaux A, et al. Automatic video monitoring system for assessment of Alzheimer's disease symptoms. *J Nutr Health Aging* 2012; 16: 213–218.
17. Robert PH, Konig A, Andrieu S, et al. Recommendations for ICT use in Alzheimer's disease assessment: Monaco CTAD Expert Meeting. *J Nutr Health Aging* 2013; 17: 653–660.
18. Välimäki M, Hätönen H, Lahti M, et al. Information and communication technology in patient education and support for people with schizophrenia. *Schizophr Bull* 2013; 39(3): 496–498.
19. Van der Krieke L, Emerencia AC, Aiello M, et al. Usability evaluation of a web-based support system for people with a schizophrenia diagnosis. *J Med Internet Res* 2012; 14(1): e24.
20. Tárrega S, Castro-Carreras L, Fernández-Aranda F, et al. A serious videogame as an additional therapy tool for training emotional regulation and impulsivity control in severe gambling disorder. *Front Psychol* 2015; 6: 1721 (12 pp.).
21. Bagga V, Kahol K and Chandra S. Game design for pre-screening patients with mental health complications using ICT tools. In: *Proceedings of the AMBI-SYS 2013* (ed CT Angelis, D Fotiadis and AT Tzallas) (LNICST, vol. 118), Athens, 25 March 2013, pp. 16–22. Cham: Springer.
22. Gravenhorst F, Muaremi A, Bardram J, et al. Mobile phones as medical devices in mental disorder treatment: an overview. *Pers Ubiquit Comput* 2015; 19: 335–353.
23. Bobes J, González MP, Sáiz PA, et al. *Instrumentos básicos para la práctica de la psiquiatría clínica*. Asturias: Área de Psiquiatría, Universidad de Oviedo, 2002.
24. Veatch RM. Models for ethical medicine in a revolutionary age. What physician-patient roles foster the most ethical relationship? *Hastings Cent Rep* 1972; 2(3): 5–7.
25. Laín Entralgo P. The friendship between physician and patient in hippocratic medicine. *J Art* 1962; 47: 1–18.
26. Miranda M, Jadresic E, Chomali M, et al. The use of e-mail in the communication between physicians and their patients. *Chile Med J* 2013; 141(6): 814–815.
27. Garvi Soler P, Villanueva Rodríguez C and Andrés Martínez E. Launching a consultation by email, to provide solutions, not to create problems. *Pediatr Aten Prim* 2014; 16(64): 311–316.

Health Informatics Journal

Special Issue: UCAmI-IWAAL-AmIHEALTH conference

Health Informatics Journal
2018, Vol. 24(4) 356–367
© The Author(s) 2016
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1460458216671560
journals.sagepub.com/home/jhi
SAGE

Ubiquitous and ambient-assisted living eHealth platforms for Down's syndrome and palliative care in the Republic of Panama: A systematic review

Juan Jose Saldaña Barrios, Luis Mendoza, Edgardo Pitti and Miguel Vargas
Technological University of Panama, Panama

Abstract

In this work, the authors present two eHealth platforms that are examples of how health systems are migrating from client-server architecture to the web-based and ubiquitous paradigm. These two platforms were modeled, designed, developed and implemented with positive results. First, using ambient-assisted living and ubiquitous computing, the authors enhance how palliative care is being provided to the elderly patients and patients with terminal illness, making the work of doctors, nurses and other health actors easier. Second, applying machine learning methods and a data-centered, ubiquitous, patient's results' repository, the authors intent to improve the Down's syndrome risk estimation process with more accurate predictions based on local woman patients' parameters. These two eHealth platforms can improve the quality of life, not only physically but also psychologically, of the patients and their families in the country of Panama.

Keywords

ambient-assisted living, Down's syndrome, eHealth, palliative care, ubiquitous computing

Introduction

As mentioned in Saldaña and Vargas-Lombardo[1] and Tran et al.,[2] in Panama, a lot of medical information are still being record in paper. The information systems related to health are not developed with standards that help to manage the patient's clinical information. Down's syndrome and palliative care (PC) are some examples of these affected areas.

The information provided by the finance and economic ministry[3] shows that 1 in every 100 births presents Down's syndrome, and around 15,000 cases were registered by 2012 in Panama.

Corresponding author:
Juan Jose Saldaña, Software Engineering Department, Technological University of Panama, Chiriqui, Panama.
Email: juan.saldana@utp.ac.pa

Senior population aged 60 years in Panama is increasing because advances in medicine have achieved an increase in life expectancy, but most people also arise likely to have advanced disease, involving prioritization attentions at the end of life. That is why providing control pain in patients with terminal disease takes more importance everyday in the health sector as a humanitarian necessity is not a medical obligation.

Structuring a platform capable of bringing the management of patients receiving PC in Panama is necessary to help achieve the main goal of PC, providing a better quality of life for patients and their families.

To research and improve the Down's syndrome risk estimation process and how PCs are provided in Panama, it is necessary to collect, organize and share information using artificial intelligent methods and ubiquitous computing. In this article, we present two case studies of eHealth platforms designed and developed based on ubiquity, machine learning techniques and interoperability.

The rest of the document is organized as follows: section "Ubiquity" describes what ubiquitous is. Section 3 explains the origins of PC. Section "NB" explains briefly what naive Bayes (NB) method is. Section "PC in Panama" provides information of PC in the country of Panama. Section "Down's syndrome" resumes what is Down's syndrome and the impact in the country of Panama. Section "Ubiquitous eHealth platform design, development and implementation" explains the main point of the analysis, design, development and implementation of both platforms. Section "Conclusion" presents the conclusion.

Ubiquity

Ubiquity is the quality of ubiquitous and it refers to the ability to have presence everywhere. Initially, this term was used as a reference to God who is capable of being everywhere.

Mark Weiser describes in his work "The Computer for the Twenty-First Century"[4] the impact that the communication and information technologies would have on the everyday life of the human being. He developed a program in the late 1980s that he called Ubicomp (Ubiquitous Computing). In this model, the communications' capacity was beyond what was expected at the time, so it opened the next generation of computing with information technology accessible wherever and whenever.

Weiser thought that Ubicomp was opposite to virtual reality because virtual reality puts people on a computer-generated world, while Ubicomp places computers at the service of people in the real world. Based on that, Weiser expected to create an environment where devices regardless of the size and functionality could interconnect and manage information, making it more accessible and consistent with the people's daily activities.

Ubicomp has many areas of research and application, with healthcare being one of them,[5] which gives rise to the term Ubicomp in the area of health or pervasive healthcare. It aims to provide technology services to the health sector of Ubicomp allowing access to information inside and outside the medical facilities.

Ubicomp has become notorious in recent years with several projects. The telemonitoring service offered by telemedicine is the result of one such project, which allows specialists to perform remote and real-time monitoring on older patients or PC patients.[6]

Ubiquitous System Patient Medical Records or SUHPC is another example of ubiquitous projects created based on Ubicomp, which allows to manage the patient record remotely. The information can be accessed in real-time in diverse institutions based on the health information requirements.[7]

NB

NB[8–10] is a probabilistic classifier and a machine learning technique that uses the Bayes' theorem, but at the same time assumes a "naive" strong independence between the variables which are independent of each other. NB is a technique that requires first to learn using a training set of classified data. After one introduces the selected data as a representative sample of the population, a model is created. This model will receive the non-classified data to be analyzed and classify it based on the rules of the model.

The advantages of using NB are as follows:

- It is not complicated to implement.
- It provides accurate results although the sample of data for training is small.

Some of the disadvantages are as follows:

- If the variables to analyze present some dependencies, it would reduce considerably the results of the test.

Using NB, we intent to provide a method to add an extra evaluation layer to the prediction process already presented in Saldaña and Vargas-Lombardo.[1]

PC in Panama

Studies conducted in 2012 indicated that 8 percent of the world population are more than 65 years, and it is estimated that within 20 years this percentage will increase to 20 percent.[11] This increase in older people is due to the great strides we have today in medicine, as it provides improvements in the treatment of various infectious diseases and other innovations. However, this increase in life expectancy involves chronic degenerative diseases in the patient, which also affects the families of the patient.

According to the 2010 census of Panama, adult population aged 60 years and older is about 9.7 percent, and it is estimated that by 2020 this will be around 12.4 percent,[12] indicating that this increase will involve a great impact on the health sector in the country, bringing with it the need to ensure greater emphasis on these people.

These home care services were given the emergence of HOSPES Association for Palliative Care in 1992, this being the first in the country to offer care in home mode. Three years later in 1995, the Program for Palliative Care and Pain Relief was created within the premises of the National Cancer Institute (ION), allowing it to provide the care in outpatient and inpatient modes. In 2003, law 68 arose, which required all health facilities in the country to provide the PC with professionals within their facilities. In the years 2006–2007, the PC was provided inside the country, covering every type of care. Finally, on 21 June 2010 under Resolution 499, the National Palliative Care Program of Panama was created.

The hospice has been providing in Panama for over 20 years, which has evolved over time but still the information is not electronically saved and many times the PC is not applied to the patient when it is necessary. In some cases, the PC arrived the patient's home after the patient had died.

Down's syndrome

Trisomy 21 also known as Down's syndrome is an aneuploidy where the fetus shows a genetic alteration having three chromosome 21.[13] This trisomy is one of the major causes of deficiencies or physical disabilities in children and premature deaths that take place before birth, situation that many mothers are unaware of. This chromosomal disorder causes various physical deformities, hearth defects, organ malformations, mental retardation, thyroid disorder and diseases such as Alzheimer's. Trisomy 21 responsible for Down's syndrome is the most frequent aneuploidy. How is the Down's syndrome currently calculated? As mentioned in Saldaña et al.,[14] screening is a probabilistic technique applied to a population to calculate the risk or probability that the fetus suffers a particular disease.

In the screening, first serum and biochemical markers are established and compared with historical median reference values of the population. When the test results of the patient and the multiple of median (MoM) of the markers have different values, the test is considered positive.

The screening methods for Down's syndrome are performed in the second and first trimesters, the first trimester being the most difficult to execute. One of the main barriers of this test is the lack of sampling data to perform the test.

For the first trimester trisomy 21 screening, it is necessary to perform a more effective detection taking into consideration some of the characteristics of the mother such as her weight, her ethnicity, whether she has diabetes or whether she smokes. These factors could affect the result of the test so it needs to be corrected.

Ubiquitous eHealth platform design, development and implementation

Ubiquitous palliative healthcare platform

The first step was to make a state of the art of PC in Panama. All the requirements, resources, actors, process and current issue were gathered from the specialist, current documents, final user interviews and patients. All this information was analyzed to develop a ubiquitous platform to provide an improved PC to the patients in the country. Figure 1 shows the general use cases and actors of the system that help us understand the context of the ambulatory care.

In order to cover all the steps in which the patients need to receive PC, the treatment has been divided into three types: home care, ambulatory care and hospital care.

- *Home care*. This attention mode is very important because the patient will not be in the health institutions. It lets the patient to share with his family at home in his last stage.
- *Ambulatory care*. This mode has two ways to perform. In the first case, the patient has the ability to attend the institution for care. In the second case, the presence of a person (family or friend) is necessary to ensure the care of the patient. This person is known as the primary caregiver and he needs to receive the ongoing training on how should give care to the sick.
- *Hospital care*. This last method is applied when the patient cannot remain at home or the caregiver no longer has the professional skills to care for the sick. The patient has increased suffering caused by the disease, thus requiring more treatments onerous for each symptom and pain relieving suffering of the patient.

Evolution notes. In PC, it is really important to record the current status of the patient in each stage to evaluate the evolution of the illness with each applied treatment. The evolutions notes are divided into four sections, called SOAP or Subjective, Objective, Assessment Plan:

- *S (Subjective)*. This section records all the information provided by the patient, such as symptoms and pains. The subjective impressions of the specialist are also included.

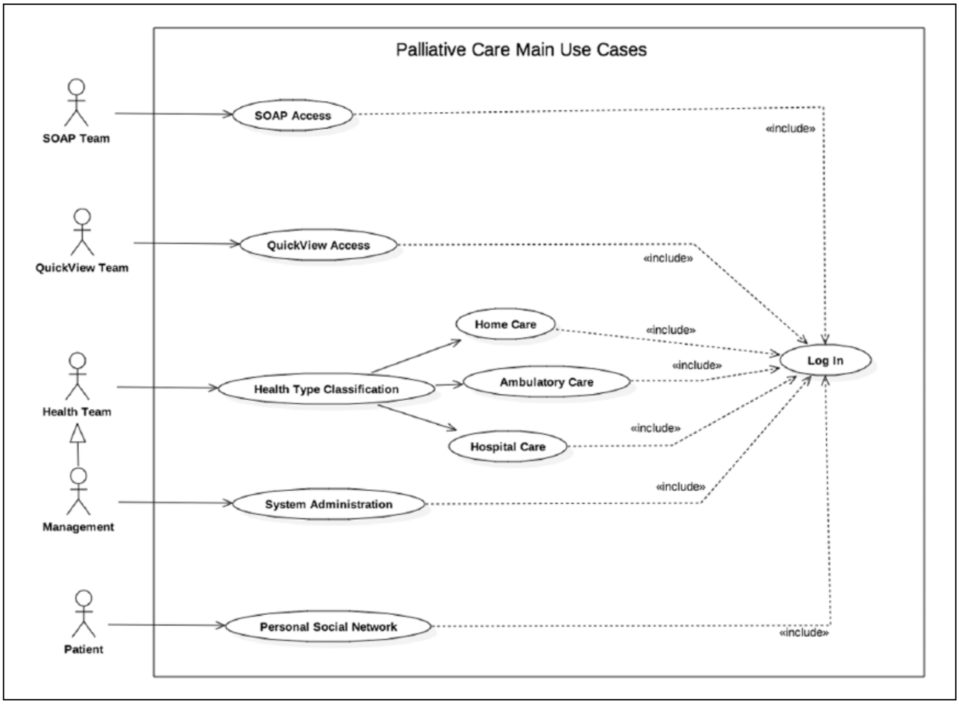


Figure 1. Palliative care use case diagram.

- *O (Objective)*. In this section, the vital signs, physical and complementary examinations of the patients are recorded.
- *A (Assessment)*. In this section, the specialist evaluates the status and its evolution.
- *P (Plan)*. This section modifies the plan applied previously according to the patient's new tests and evaluation.

The main classes of the platforms are presented in Figure 2. The usability was a very important factor in the design of the platform and it was based on goal-oriented design by Allan Cooper.[15,16]

The interaction with the platform was designed based on the usability and specific goal that each type of user will have with the platform.

The system includes more classes, but the intention of the diagram is to show only the classes that are related to the attribute, operations and functional requirements that are specific to the context of PC.

The general architecture of the system can be appreciated in Figure 3 as was proposed in Saldaña and Vargas-Lombardo.[1]

Implementation and testing. For the development, Laravel was used as the development framework, connecting a relational MySQL database to Eloquent ORM and working with Bootstrap to provide a rich user interface.

The platform was tested using white box testing method by two requirement engineers. After all the functionalities were working, the platform was deployed in the cloud of the university and was enabled for the hospital specialist for about 2 months. They used the platform by recording data of

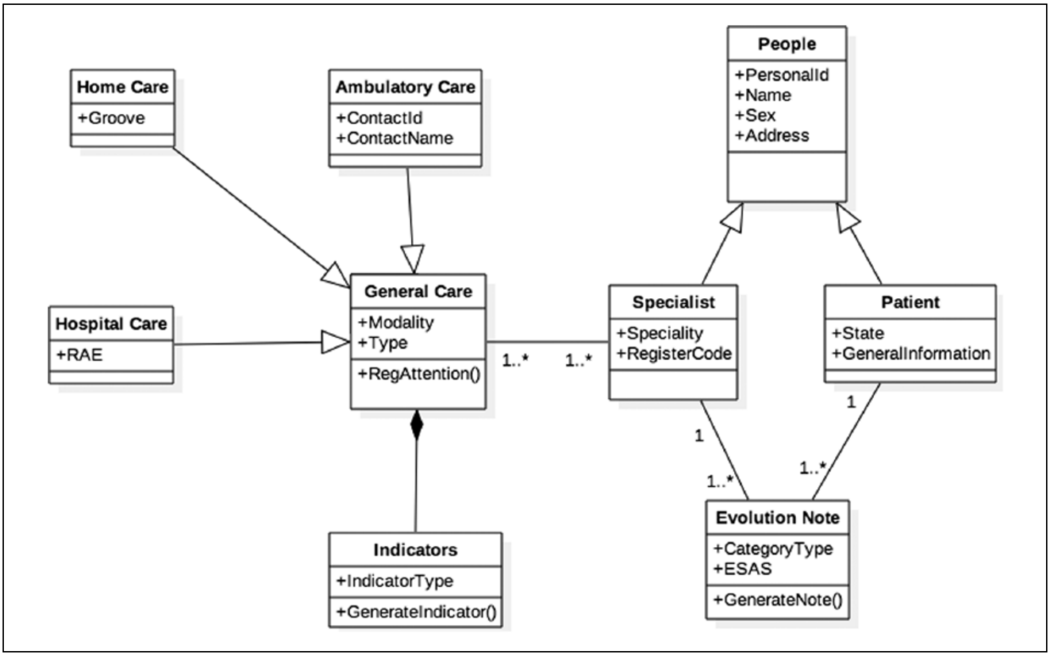


Figure 2. Palliative healthcare class diagram.

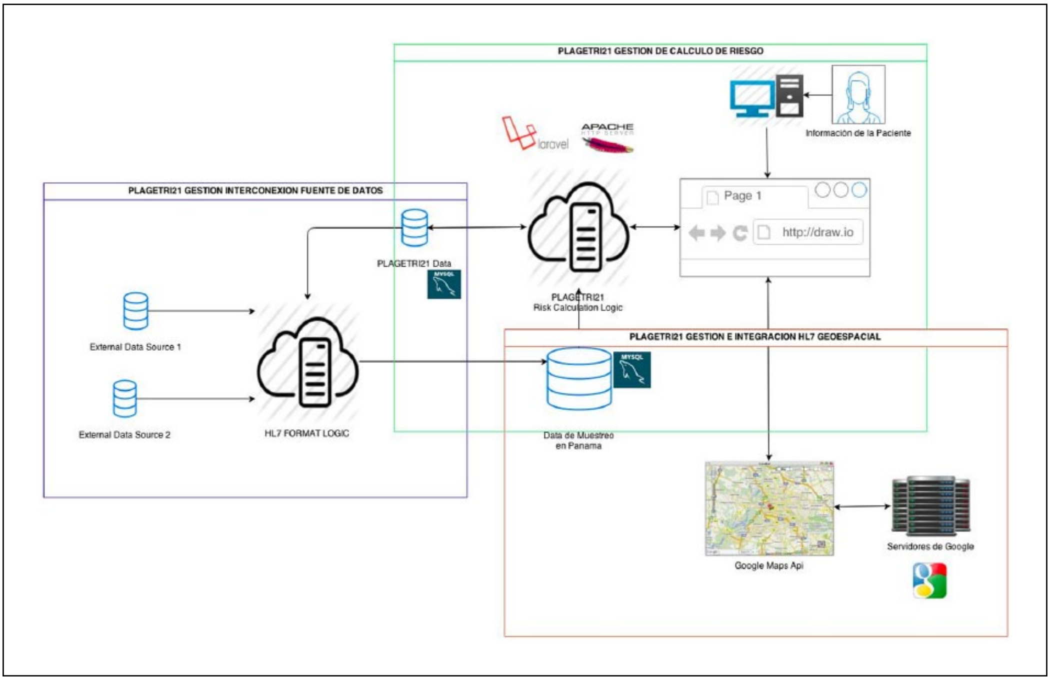


Figure 3. PLAGETRI21 eHealth management platform architecture.

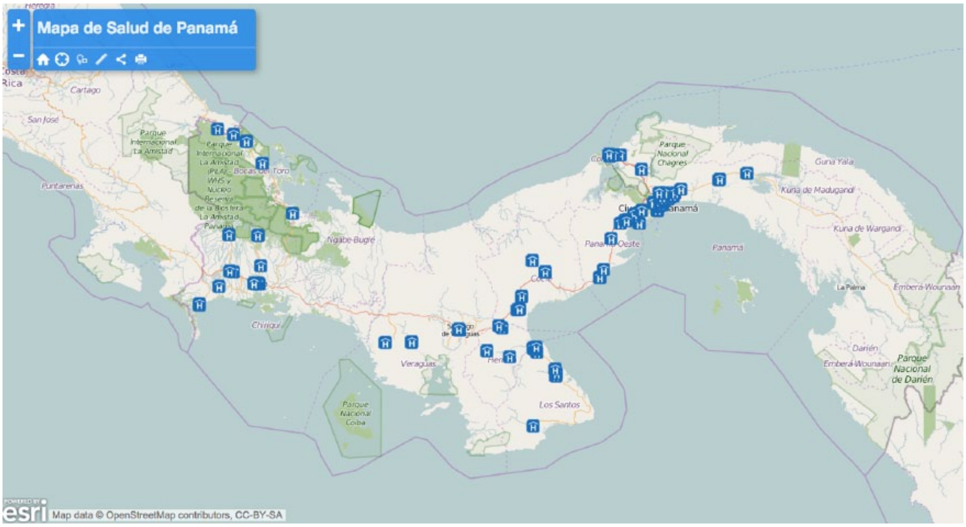


Figure 4. Health institutions that will utilize the eHealth platform for palliative care in the country of Panama.

patients from 1 year ago allowing them to test all the functionalities and provide their feedback. The next step was to sign the term and conditions agreement and deploy the platform to the medical institutions. Figure 4 shows the medical institutions and hospitals of the country that will have access to the platform.

Smart platform for Down's syndrome risk estimation process

The eHealth Management Platform PLAGETRI21[1] is based on the calculation risk method of likelihood, published in Benn,[17] combining the a priori risk for maternal age obtained from the meta-analysis with the likelihood obtained from combining the MoM of the different markers used in each profile. The MoM was calculated using the multivariate normal distribution. This calculation follows a mathematical and statistical process:

- Risk estimation based on the maternal age;
- Markers' standardization;
- MoM calculation;
- Weight and correction factors' adjustments;
- Maternal weight corrections;
- Covariates' corrections;
- Likelihood ratio estimation;
- Risk estimation.

Additional to the normal process, the platform provides two new functionalities. First, it adds an extra layer of analysis applying NB techniques. Using a training set previously selected, the platform generates a model that posteriori receives the not classified data and separates the normal from the abnormal values, predicting whether the test is positive or negative. The following steps are performed to implement the NB classification:

1. Calculating the average $\mu_{F_iC_j}$ of each feature for the classes;
2. Calculation of the variance $\sigma^2_{F_iC_j}$ of the classes;
3. Estimate the probability for each class C_j;
4. Estimate the probability $\sigma^2_{F_iC_j}$ of each feature F_i due to class C_j;
5. Calculate the evidence of the value that it is really the probability of occurrence of all the features $p(F_1, F_2, \ldots, F_i)$;
6. Evaluate the class that presents the higher probability $\hat{p} = \underset{j \in \{1,\ldots,J\}}{argmax}\, p(C_j) \prod_{k=1}^{i} p(F_k \mid C_j)$.

Table 1 shows the implementation of these processes.

Second, the architecture's platform is based on the architecture presented in Barbarito et al.,[18] Esri[19] and Feldmann et al.,[20] which allows the interoperability between many hospital information systems. PLAGETRI21 is able to interact using the HL7 standard with many data source of clinical information. It uses the Clinical Document Architecture (CDA) to save domain-sampling data from diverse sources around the country without the necessity of installing any software at the client side and using laptop and mobile devices. The CDA standard allows the interoperability with other platforms that also implement this standard in their architecture. The message body structured HL7 CDA consists of two parts that are the header and message body as shown in Figure 5.

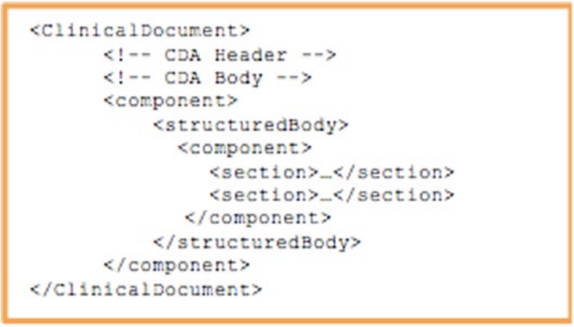
As an example, to send data about the height and weight of the patient, the tag <entry> is used and is structured as follows:

```
<entry>
<observation classCode="OBS" moodCode="EVN">
<code code="363808001" codeSystem="2.16.840.1.113883.6.96" codeSys-
temName="SNOMED CT" displayName="Peso Corporal"/>
<effectiveTime value="201504071430"/>
<value xsi:type="PQ" value="71.6" unit="kg"/>
</observation>
</entry>
<entry>
<observation classCode="OBS" moodCode="EVN">
<code code="384627007" codeSystem="2.16.840.1.113883.6.96" codeSys-
temName="SNOMED CT" displayName="Estatura"/>
<effectiveTime value="201504071430"/>
<value xsi:type="PQ" value="1.65" unit="m">
</observation>
</entry>
```

The interconnection with other data sources involves security and privacy of the information as presented by Geissbuhler. The data that are extracted from the different data sources include general information of the patient such as age, sex, ethnic, blood type, place of birth and residence. Specific fields such as name, last names and personal ID are not included in the sampling data source. The platform also implements the geospatial interoperability standard proposed in Granell et al.[21] and Ahern.[22] The inclusion of geographical information to the platform enables the clinical information being georeferenced by birth place, residence place and location where the patient receives the medical care, allowing us to research how the location data affect the process.

Table 1. Down's syndrome estimation applying naive Bayes to the data.

	p(Normal| Criterion 1, Criterion 2, Criterion 3)	*p*(Abnormal| Criterion 1, Criterion 2, Criterion 3)	*p*(Criterion 1|Normal)	*p*(Criterion 1|Abnormal)	*p*(Criterion 2|Normal)	*p*(Criterion 2|Abnormal)	*p*(Criterion 3|Normal)	p(Criterion 3|Abnormal)	Evidence
Normal	9.7946E−01	2.0545E−02	1.5860E−01	2.2124E−01	5.2578E−01	1.5711E−01	7.7873E−01	3.7226E−01	5.9983E−02
Abnormal	3.6384E−01	6.3616E−01	2.8865E−01	2.2939E−01	7.1001E−01	1.1676E+00	1.5217E−02	1.9341E−01	7.7554E−03
Normal	7.6056E−01	2.3944E−01	3.7669E−01	2.3326E−01	6.1834E−01	1.0993E+00	1.0290E−01	2.7955E−01	2.8512E−02
Abnormal	8.8520E−08	1.0000E+00	2.8432E−01	2.5526E−01	1.0589E+00	1.2358E+00	2.5672E−11	2.6295E−03	7.8997E−05
Normal	9.5582E−01	4.4182E−02	8.7175E−01	2.4733E−01	1.0241E+00	1.2479E+00	1.7386E−01	2.2083E−01	1.4692E−01
Abnormal	2.2593E−01	7.7407E−01	8.8236E−01	2.6085E−01	9.6253E−01	1.2556E+00	1.4177E−03	1.1965E−01	4.8216E−03
Normal	8.0611E−01	1.9389E−01	1.2441E+00	2.5726E−01	7.8952E−01	3.2329E−01	5.9359E−03	1.6019E−01	6.5442E−03
Normal	8.1843E−01	1.8157E−01	1.7509E−01	2.2254E−01	4.2356E−01	8.9411E−01	3.4393E−01	2.7017E−01	2.8197E−02
Abnormal	1.4100E−01	8.5900E−01	1.1726E−01	2.4911E−01	1.0531E+00	1.2383E+00	7.1843E−03	1.6647E−01	5.6931E−03
Abnormal	4.6286E−01	5.3714E−01	1.1795E+00	2.5465E−01	6.7275E−01	1.1420E+00	5.1809E−03	1.5585E−01	8.0359E−03
Normal	9.1255E−01	8.7448E−02	2.0307E−01	2.2452E−01	1.1375E+00	1.1731E+00	3.3593E−01	2.6823E−01	7.6941E−02
Abnormal	4.2107E−01	5.7893E−01	1.3384E−01	2.1905E−01	8.4016E−02	2.7885E−01	7.1563E−02	1.7207E−01	1.7290E−03
Abnormal	3.2302E−01	6.7698E−01	9.8361E−01	2.4986E−01	1.0582E+00	6.0721E−01	7.7312E−04	1.0560E−01	2.2539E−03
Normal	7.1689E−01	2.8311E−01	9.0362E−01	2.4806E−01	6.7241E−01	1.1418E+00	2.6880E−02	2.1635E−01	2.0613E−02
Abnormal	6.4765E−04	9.9935E−01	2.4634E−05	1.8604E−01	1.1011E+00	6.7738E−01	8.5628E−02	2.7016E−01	3.2445E−03
Abnormal	1.7207E−28	1.0000E+00	8.7490E−28	2.3910E−02	5.0820E−01	9.9355E−01	5.9926E−05	6.1926E−02	1.4010E−04
Normal	9.7666E−01	2.3342E−02	7.0629E−01	2.4339E−01	9.0319E−01	1.2501E+00	7.8031E−01	3.7145E−01	4.6113E−01
Normal	9.5241E−01	4.7592E−02	3.1621E−01	2.3070E−01	1.0963E+00	1.2139E+00	5.2713E−01	3.0977E−01	1.7359E−01
Normal	9.2440E−01	7.5604E−02	3.1600E−01	2.3069E−01	1.0060E+00	1.2518E+00	4.1522E−01	3.5512E−01	1.2919E−01
Normal	9.6669E−01	3.3314E−02	3.3280E−01	2.3144E−01	1.1539E+00	1.1485E+00	7.8171E−01	3.6972E−01	2.8095E−01
Normal	8.2330E−01	1.7670E−01	8.4631E−02	2.1332E−01	1.1584E+00	8.0449E−01	1.9638E−01	2.2872E−01	2.1156E−02
Normal	6.7343E−01	3.2657E−01	2.6284E−01	2.2807E−01	4.1702E−02	1.5791E−01	1.5138E−01	2.1225E−01	2.2292E−03



```
<ClinicalDocument>
      <!-- CDA Header -->
      <!-- CDA Body -->
      <component>
          <structuredBody>
            <component>
               <section>…</section>
               <section>…</section>
             </component>
          </structuredBody>
      </component>
</ClinicalDocument>
```

Figure 5. HL7 CDA structure.

Table 2. Results of the screening test.

Data	Values
Total samples	100
Valid test	98
Test with double verification	6
Valid and approved result tests	92
Doubtful test	4
Acceptance rate for doubtful test	50%–60%
Acceptance test wit 90% of accuracy rate	100%
Acceptance test wit 95% of accuracy rate	92.58%
General acceptance rate	94.3877551

Implementation and testing. The platform was tested initially by the development team and three medicine specialists from the hospital. After all the bugs were fixed, the screening group of the laboratory took a sample data from 100 patients. The chemical tests were introduced and analyzed by the platform and the results can be found in Table 2.

The platform is running on a private cloud inside the Technological University of Panama and the Electronics Health and Supercomputing Research's Group developed it.

Conclusion

The Down's syndrome risk estimation platform enhances the accuracy of the result because it adds an extra layer of data analysis applying machine learning methods to establish smart classifiers extracted from the population sampling and includes valuable geographical information to the procedures, not taken into consideration before.

Thanks to the information that is being captured, organized and shared from diverse sources in the country, the first trimester screening test will be applied allowing to detect any illness earlier in order to provide the treatment in a timely manner.

The PC platform allows, thanks to its ubiquitous properties, to record personal and medical information in real-time even from the patient's home. It also lets specialist of the field to have access to this information and apply adequate medicines and treatment. The platform is a tool that allows the patient's family being in touch with the medical specialist.

Ubicomp, machine learning techniques and ambient-assisted living open to us a broad source of resources that can be applied to improve many areas of healthcare.

Acknowledgements

The Research Group Electronics Health and Supercomputing (GISES) of the Technological University of Panama supported this research as well as Dr Miguel Vargas-Lombardo National Research Investigator award (SNI) from the SENACYT. Both projects are being presented as main topics looking for founds with SENACYT.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Saldaña J and Vargas-Lombardo M. eHealth management platform for screening and prediction of Down's syndrome in the Republic of Panama. *E-Health Telecommun Syst Netw* (Scientific Research Publishing) 2014; 3(3): 33–42, http://www.scirp.org/journal/PaperInformation.aspx?PaperID=49724&#abstract (accessed 27 February 2015).
2. Tran V-A, Johnson N, Redline S, et al. OnWARD: ontology-driven web-based framework for multi-center clinical studies. *J Biomed Inform* 2011; 44 (Suppl. 1): S48–S53, http://www.sciencedirect.com/science/article/pii/S1532046411001468 (accessed 20 February 2015).
3. Domínguez Y. Sindrome de Down Batalla social. *Día a Día*, 2015, http://www.diaadia.com.pa/primerplano/s%C3%ADndrome-de-down-batalla-social-268675 (accessed 1 January 2015).
4. Weiser M. The computer for the 21st century. *IEEE Pervas Comput* 2002; 1(1): 19–25.
5. Pomares ES. *Computación Ubicua: un gran desafío*. Cholula, Mexico: National Institute of Astrophysics, Optics and Electronicsp, 2010, p. 2.
6. Escayola J, Martínez I, Serrano L, et al. Propuesta de una Nueva Arquitectura de Software para uso del Estándar ISO/IEEE 11073 en Dispositivos Médicos de Limitada Capacidad de Procesado y Memoria, 2008, pp. 2–5, http://diec.unizar.es/intranet/articulos/uploads/Propuesta%20de%20una%20Nueva%20Arquitectura%20de%20Software%20para%20uso%20del%20Estandar%20ISO-IEEE%2011073%20en%20Dispositivos%20Medicos%20de%20Limitada%20Capacidad%20de%20Procesado%20y%20Memoria.pdf
7. Rodríguez Robledo G. Sistema Ubicuo de Historia Clínica del Paciente. *Instituto Politécnico Nacional*, 2006, http://www.saber.cic.ipn.mx/cake/SABERsvn/trunk/Repositorios/webVerArchivo/366/1
8. Soria D, Garibaldi JM, Ambrogi F, et al. A "non-parametric" version of the naive Bayes classifier. *Knowledge-Based Syst* 2011; 24(6): 775–84, http://www.sciencedirect.com/science/article/pii/S0950705111000414 (accessed 27 May 2015).
9. Naive Bayes text classification. http://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html (2009, accessed 8 October 2015).
10. Marucci-Wellman HR, Lehto MR and Corns HL. A practical tool for public health surveillance: semi-automated coding of short injury narratives from large administrative databases using naïve Bayes algorithms. *Accident Anal Prev* 2015; 84: 165–76, http://www.sciencedirect.com/science/article/pii/S0001457515300099 (accessed 5 October 2015).
11. Valencia B and Inés M. Revista Colombiana de Anestesiología Envejecimiento de la población. *un reto para la salud pública* 2012; 40(69): 192–194.

12. INEC Contraloría General de la República de Panamá. *Estimaciones y proyecciones de la población en la republica, provincia, comarca indígena por distrito, según sexo y edad; 2010–20*, https://www.contraloria.gob.pa/inec/Publicaciones/Publicaciones.aspx?ID_SUBCATEGORIA=10&ID_PUBLICACION=499&ID_IDIOMA=1&ID_CATEGORIA=3
13. Gyselaers W, Vereecken A, Van Herck E, et al. Screening for trisomy 21 in Flanders: a 10 years review of 40.490 pregnancies screened by maternal serum. *Eur J Obstet Gyn R B* 2004; 115(2): 185–9, http://www.sciencedirect.com/science/article/pii/S0301211503006456 (accessed 26 February 2014).
14. Saldaña BJJ, Rovetto C, Pitti E, et al. Modelado formal de la metodología para la predicción de pacientes con Síndrome de Down en Panamá, 2015, http://www.laccei.org/LACCEI2015-SantoDomingo/RefereedPapers/RP082.pdf
15. Guersenzvaig A. El usuario arquetípico: creación y uso de personajes en el diseño de productos interactivos (Human-computer interact), http://www.alzado.org/imgconts/autor_id3/personajes_alzado2.pdf
16. Honduvilla M, Bernabé Poveda MA and Manrique Sancho MT. La usabilidad de los geoportales: aplicación del Diseño Orientado a Metas (DOM). In: *IV Jornadas Técnicas de las Infraestructuras de Datos Espaciales de España*, Santiago de Compostela, España, 17–19 October 2007.
17. Benn PA. Advances in prenatal screening for Down syndrome: I. general principles and second trimester testing. *Clin Chim Acta* 2002; 323(1–2): 1–16, http://www.sciencedirect.com/science/article/pii/S0009898102001869 (accessed 26 February 2014).
18. Barbarito F, Pinciroli F, Mason J, et al. Implementing standards for the interoperability among healthcare providers in the public regionalized Healthcare Information System of the Lombardy Region. *J Biomed Inform* 2012; 45(4): 736–45, http://www.sciencedirect.com/science/article/pii/S153204641200007X (accessed 20 February 2014).
19. Esri A. HL7 and Spatial interoperability standards for public health and health care delivery, 2011, https://www.esri.com/library/whitepapers/pdfs/hl7-spatial-interoperability.pdf
20. Feldmann RL, Shull F, Denger C, et al. A survey of software engineering techniques in medical device development 2007. In: Joint workshop on high confidence medical devices, software, and systems and medical device plug-and-play interoperability (HCMDSS-MDPnP 2007), Cambridge, MA, 25–27 June, pp. 46–54. New York: IEEE, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4438163 (accessed 19 July 2013).
21. Granell C, Fernández ÓB and Díaz L. Geospatial information infrastructures to address spatial needs in health: collaboration, challenges and opportunities. *Future Gener Comp Sy* 2014; 31: 213–222, http://www.sciencedirect.com/science/article/pii/S0167739X13000629 (accessed 5 February 2014).
22. Ahern DK. Challenges and opportunities of eHealth research. *Am J Prev Med* 2007; 32(Suppl. 5): S75–S82, http://www.sciencedirect.com/science/article/pii/S0749379707000451 (accessed 26 February 2014).

Special Issue: ISHIMR conference

Health Informatics Journal

Health Informatics Journal
2018, Vol. 24(4) 368–378
© The Author(s) 2016
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1460458216675497
journals.sagepub.com/home/jhi
SAGE

Understanding Internet access and use to facilitate patient portal adoption

Prudence W Dalrymple and Michelle Rogers
Drexel University, USA

Lisl Zach
Information Insights, LLC, USA

Anthony Luberti
The Children's Hospital of Philadelphia, USA

Abstract

Understanding the information-seeking preferences and Internet access habits of the target audiences for a patient portal is essential for successful uptake. The resource must deliver culturally and educationally appropriate information via technology that is accessible to the intended users and be designed to meet their needs and preferences. Providers must consider multiple perspectives when launching a portal and make any needed adjustments once the launch is underway. We report results of a study of 270 parents and caregivers of paediatric patients in a major health system during the process of implementing a patient portal. Through a 26-question paper-and-pencil survey, data were collected on participant demographics, Internet access and use, health information–seeking behaviours, health literacy, and potential use of a patient portal. Results indicate a positive attitude towards portal use but also suggest that low health literacy may be a key issue to portal adoption.

Keywords

health literacy, information behaviour, information seeking, Internet access, patient portals

Introduction

One of the goals of Healthy People 2020 is the use of health information technology (HIT) to improve health outcomes, quality, and equity. Among the several specific objectives are the increase in patient–provider communication, the delivery of health information that is relevant and

Corresponding author:
Prudence W Dalrymple, College of Computing & Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA.
Email: pdalrymple@drexel.edu

understandable to the target audiences, and the improvement of health literacy skills.[1] There are numerous HIT mechanisms designed to achieve these objectives, most of which require access to the Internet – especially broadband, and increasingly, mobile devices. While health-related websites have become ubiquitous, of growing importance are those extensions of electronic health records, often enhanced with knowledge-based resources, known as patient portals. The recent push to implement electronic health records in both inpatient and outpatient environments, as well as the recognition of the importance of patients' involvement in their own healthcare, has stimulated interest in portals through which patients and designated caregivers can view test results, request medication refills, communicate with clinicians, and make/change appointments.

The advent of these portals presents additional opportunities for technology to play a role in health information and communication activities such as health information seeking and health literacy. A portal enables providers to tailor information and communication resources to the specific needs of the intended recipients, thus enabling providers to deliver culturally and educationally appropriate information. However, for a portal to reach its full potential, the healthcare providers and organizations must take some preliminary measures to ensure that both the materials and the interface are optimized. Designing and implementing an effective portal require clarifying its purpose, identifying the intended users, becoming familiar with their information-seeking patterns and preferences, and assessing their capabilities and access habits.[2] In this article, we report on one phase of the process that a major paediatric hospital and associated clinical practices used to implement its patient portal; we present our story so that others may benefit as they embark on selecting, refining, and implementing a patient portal.

Background

In the portal implementation discussed here, the provider organization is a major paediatric hospital that has a significant investment in making educational materials available to the patients' families and caregivers as well as to the patients themselves, although (since the patients are children) the materials for each group are of necessity quite different from one another. Prior to portal design and development, the organization had already established procedures and practices for communication between providers and parents or caregivers and had taken steps to assure that their written materials met basic health literacy guidelines. Nonetheless, the organization lacked knowledge about the specific health literacy and information-seeking behaviours of the families they hoped to engage, particularly since the practices were located in diverse socioeconomic areas. Furthermore, the organization was concerned about the rate of uptake for the portal. This situation provided an opportunity for a team of researchers to investigate the phenomenon and to provide its insights to the organization. The research team, comprised of information scientists and systems designers, had experience in assessing the information-seeking habits and preferences of diverse populations and was able to contribute its expertise in these areas to the experience of the healthcare providers. Involving an interdisciplinary team enabled a comprehensive or 'systems approach' to assessing readiness for the implementation.

Adopting a systems approach when introducing any technological innovation helps to ensure that both the user and the designer perspectives are considered. That is, because the patient/family unit and the provider/designer unit are related to one another, the characteristics of each part may affect the overall system; changing one component invites an impact on the others. Understanding the information-seeking preferences and Internet access habits of the desired audiences for a patient portal – or any other digital tool – is essential for successful uptake. The resource must also be attractive enough to be used, employ technology that is accessible to the intended users, and be designed so that it meets their needs and preferences. Thus, provider organizations must consider multiple perspectives when determining how and whether to launch a portal and make any needed adjustments once the launch is underway. Doing an initial assessment of the intended audience can be a critical success factor in ensuring adoption.[3,4]

Information-seeking behaviour is the topic of a large body of research, and numerous resources are available that provide important insights, particularly in health information seeking.[5,6,7] While much Internet health information–seeking behaviour is conducted by adults seeking answers to their own personal questions, a significant amount of health information seeking is done by family members or caregivers on behalf of others. What follows here is a summary of what is known about the information seeking of parents and caregivers on behalf of their children.

Early studies investigating parental use of the Internet have shown that parents use Internet resources to learn more about their child's condition and that they are willing to use them for managing their children's care.[8,9] On the other hand, misunderstandings and confusion can occur, and there is a need to educate both parents and the general population on how to evaluate the quality of online health information.[10,11,12,13] Since parents appear to demand a very high standard of credibility when it comes to information that may affect their children, trustworthiness is a key factor in determining whether parents will seek and accept information that will be used to make decisions about their children's care. At the same time, patients and families often share information with each other and often report that family and friends are an important source of health-related information.[14]

Health literacy is defined by the Centers for Disease Control and Prevention (CDC) as the degree to which an individual has the capacity to obtain, communicate, process, and understand basic health information and services to make appropriate health decisions.[15] The National Assessment of Adult Literacy (NAAL) measures the health literacy of adults living in the United States. At the time of its most recent assessment, approximately 36 per cent of adults in the United States have limited health literacy. An additional 5 per cent of the population is not literate in English. Only 12 per cent of the population has a proficient health literacy level.[16] As noted earlier, improving health literacy is one of the goals of Healthy People 2020; helping providers to understand and appreciate the health literacy 'profile' of their patient population can be an important first step in addressing this goal.

Building on previous research and our own experience, we focused our attention on learning more about the patients and families served by a specific healthcare organization to enable it to prepare for implementing its patient portal more effectively. The specific aims of this research were to explore the current technologies parents of this health system use to access the Internet, to ascertain how parents find health information for themselves and their family, and to assess the level of awareness and use of digital technologies to meet the information needs parents may have. To address these aims, we surveyed parents/caregivers at five clinics affiliated with the health system and located throughout a large metropolitan area to determine their health information–seeking practices and preferences. We also asked screening questions to assess their level of health literacy and their interest in accessing and using a patient portal. Because we were conducting the research in conjunction with the operations of a specific health system, our sample reflects the goals of that system and does not necessarily represent the population as a whole.

Methods

Setting and sample

The study was conducted in five clinical practices associated with a major paediatric hospital in a large metropolitan area in the eastern United States. The clinic locations were selected to capture the variety of patients and practices present in this metropolitan area. We included rural, urban, and suburban practices to encompass social and demographic as well economic differences. Using a convenience sample, we conducted a paper-and-pencil survey of parents and other responsible adult caregivers of children and adolescents. All respondents were 21 years or older and were able to read and write English. The study was approved by the Institutional Review Boards (IRBs) of both the health system and the researchers' university. Data were collected from August to December 2012. A total of 270 usable surveys were collected with a minimum of 50 from each clinical practice.

Survey instrument

The paper-and-pencil survey instrument used for this study was adapted from an interview protocol designed for a previous study conducted in a medically underserved area in the same metropolitan area.[17] The survey investigated the various forms of access to the Internet in use by patients. The conversion to a paper-and-pencil format was done to be minimally intrusive in the clinical setting, and the new format was field tested and refined before implementation. The research coordinator for the health system and members of the research team met with the practice managers and clerical staff to discuss the study protocol and to finalize the procedure for administering the survey instrument.

To establish the reading level of the survey, the text was pasted into a Microsoft Word (Microsoft Corporation, Redmond, Washington) document and then evaluated for reading ease and grade level using the Flesch Reading Ease formula, which has been widely used in evaluating medical literature. For this survey, the Reading Ease score was 66.1 (scale 0–100) and the grade level was 7.1, which approximates the reading level recommended by the National Library of Medicine's[18] consumer health resource MedlinePlus. The final version of the survey contained 26 questions and took 10–15 min to complete by checking appropriate boxes. Survey questions included demographics (age, race, and education) and questions about Internet use, mobile device access, health information–seeking behaviours, health literacy, and potential use of a patient portal. Access was assessed using the following questions: 'Do you currently have Internet access through a computer/laptop/cell phone/smartphone or other device?' and 'How do you get to the Internet most often?' Healthcare information–seeking patterns were elicited by questions about looking for health information for themselves or others in their family. Health literacy was assessed using the three screening questions proposed by Chew[19] in 2004 for use in identifying patients with inadequate or marginal health literacy.

Data collection and analysis

The 26-item survey was administered to participants while they were waiting to be seen by clinicians at the individual practices. A student researcher trained in the procedure for conducting the survey recruited a convenience sample of participants by approaching parents/caregivers in the practice waiting rooms after they had checked in for their child's appointment. Parents/caregivers received a brief introduction to the study and instructions for completing the survey; they were encouraged to take the survey with them to the examination rooms to complete the survey while waiting if necessary. The surveys were collected as participants checked out. Participants' time was compensated by a US$15 gift card for a local store.

Data from the 270 usable surveys were entered manually into SurveyMonkey for subsequent analysis. The data were collected anonymously, and only descriptive statistics are presented here. Relationships between the responses to several different pairs of questions were correlated using SPSS but nothing of statistical significance was found.

Table 1. Demographics of sample (n = 270).

Variable	%
Age (years)	
21–29	28.6
30–39	32.3
40–49	23.0
50–64	11.2
65+	2.2
Gender	
Female	80.4
Male	19.6
Race/ethnicity	
American Indian/Alaska Native	1.5
Asian	1.9
Black/African American	38.1
Hispanic/Latino	13.7
Native Hawaiian/Pacific Islander	0.4
White	40.7
More than one race/ethnicity	4.4
Other	1.5
Education	
Some high school	6.7
High school grad/GED	29.4
Some college	22.3
College grad (AS/BS/tech)	30.5
Graduate school	9.3
Household income	
>US$24,999	24.8
US$25,000 to US$49,999	28.9
US$50,000 to US$100,000	27.0
<US$100,000	15.6

GED: General Educational Development.

Since some respondents did not answer all questions or selected more than one choice, not all variables total 100.

Results

Study participants

Demographic data of the participants are summarized in Table 1. As shown, the majority of the participants were women of child-bearing age. This is consistent with the nature of the population, that is, parents and caregivers of paediatric patients. The number of White and African-Americans was nearly equal, at 41 per cent White and 38 per cent African American. Close to 25 per cent of participants had income levels of less than US$24,999, and 36 per cent had completed no more than a high school education or its equivalent (GED). This suggested that a number of study participants might have health literacy issues. A profile of the study participants appears in Table 1.

Table 2. Access to the Internet.

Method of Internet access	%
Computer or laptop	57.8
Cell phone/other mobile device	36.9
Smartphone	21.3
iPad or tablet	11.4
Other/error/no access	10.6

Internet access

Because successful implementation of patient portals depends on users having Internet access, several questions were designed to capture various aspects of Internet access and use. In total, 94 per cent of respondents reported having access to the Internet. Methods of access are displayed in Table 2, indicating that while computers (including laptops) were used by over half of those surveyed, mobile devices, including tablets, and mobile phones accounted for more than a third of access. Since some respondents checked more than one mode of access, totals sum to more than 100 per cent.

Over 80 per cent of respondents reported using the Internet several times a day, and fewer than 3 per cent reported using the Internet less than once a week. Of the respondents, 83 per cent reported that they were either 'comfortable' or 'very comfortable' using the Internet. These findings indicate a well-connected population of parents and caregivers.

Health information seeking

Of the respondents, 85 per cent reported using the Internet to look for general information, although only 42 per cent reported using the Internet specifically to look for health-related information as we noted in our previous study.[17] Only about 11.5 per cent of respondents use the Internet daily to look for some type of health-related information. Table 3 displays the frequency of health information–seeking activity.

Nearly 75 per cent of respondents reported that they 'usually' or 'always' found the health-related information that they were seeking; more than 52 per cent of respondents were confident that the information they found was accurate. Respondents had various ways of determining whether to trust information that they found on the Internet. Table 4 displays the reasons for trusting health information on the Internet.

In addition to looking for health information on the Internet, respondents indicated that they used other sources of information. Table 5 displays the most frequently used sources of health information.

Health literacy of participants

The survey instrument used the three screening questions proposed by Chew et al.[19] to identify patients with marginal or inadequate health literacy in the population served at the five clinical practices. Table 6 indicates that a substantial portion of the respondents may have marginal or inadequate health literacy. More than a quarter of the participants indicated that they 'sometimes' had difficulty reading health materials, and fully one-third indicated that they 'sometimes' needed help reading health materials. These results are consistent with those found in the population at large, according to the NAAL, cited above.

Table 3. Frequency of health information–seeking activity.

How often do you use the Internet to look for information about you or your family's healthcare?	%
Several times a day	6.0
About once a day	5.6
3–5 times a week	5.6
1–2 times a week	15.1
Every few weeks	29.4
Less often	30.5
Never	2.2
Don't know or N/A	8.9

Table 4. Reasons for trusting health information.

I trust/believe health information I find on the Internet if it (please check all that apply)	%
Comes from my healthcare system	61.6
Comes from a government website	46.3
Has a doctor's or nurse's name on it	19.0
The website comes up on the first page of Google	16.9
Looks like it was written for people who have questions similar to mine	16.1
The website is recommended to me by a friend	14.5
Don't know or N/A	7.9

Table 5. Sources of health information.

Where do you go most often for health information?	%
I ask a nurse or doctor	63.2
I look things up on the Internet	35.7
I ask somebody in my family	20.4
Other (books/magazines, TV, friend, library)	<6.0

Attitudes towards portal adoption

The final set of questions in the survey queried participants about their interest in using the Internet for administrative tasks such as making appointments, or for viewing personal clinical data, and accessing their medical records. Table 7 shows that a majority of respondents would use a portal if it were available from their healthcare provider.

Discussion

This study was conducted during the early implementation stage of a patient portal designed by a leading paediatric health organization. As part of the investigation, our research team, guided by a hospital clinician, explored how the parents and caregivers of paediatric patients access and use

Table 6. Health literacy questions.

How often do you have someone help you read health-related materials?	
Never	55.6%
Sometimes	33.0%
Usually	8.2%
Always	3.0%
How often do you have problems learning about medical conditions because of difficulty reading health-related materials?	
Never	68.9%
Sometimes	26.3%
Usually	2.2%
Always	1.8%
How confident are you in filling out medical forms by yourself?	
Very comfortable	59.6%
Comfortable	25.2%
Ok	10.4%
Not very comfortable	2.6%
I don't fill out medical forms by myself	0.7%

Table 7. Portal-related questions.

If your healthcare provider (had a portal that) allowed you to do administrative tasks online, (how often) would you use it?	
Never	9.7%
Sometimes	29.4%
Usually	15.3%
Mostly	21.0%
Always	23.8%
If your healthcare provider (had a portal that) allowed you to view personal clinical things online, (how often) would you use it?	
Never	11.3%
Sometimes	28.6%
Usually	11.7%
Mostly	16.1%
Always	31.5%
If your healthcare provider put your medical records on a secure Internet website that only you could access, (how often) would you use it?	
Never	10.5%
Sometimes	25.0%
Usually	10.5%
Mostly	19.0%
Always	33.5%

the Internet. We were interested in learning how parents/caregivers look for health information and in understanding more about their information-seeking patterns and preferences. The research yielded pertinent descriptive data about the ways in which parents and caregivers seek and use information. The results showed that the overwhelming majority of parents/caregivers surveyed were connected to the Internet through a variety of devices. This finding suggests that Internet access is not a barrier to portal implementation, at least among the population being served at the clinics where the study took place. The study also confirms findings from our previous research indicating that mobile devices such as smartphones were often the means by which patients connect to the Internet.[17]

While 85 per cent of the respondents to the survey reported that they use the Internet to look for general information, only 42 per cent indicated that looking for health information was among their most frequent information-seeking activities. The picture that emerged from our investigation suggests that although parents/caregivers may often use the Internet to look for health information, their preferred source of health information remains a nurse or doctor. This finding suggests that as health organizations begin developing patient portals, care should be taken to involve clinicians in selecting health resources as users put a high value on materials associated with their doctor and healthcare team. Furthermore, when asked whether they would go online to do various administrative tasks and/or to view their medical records on a secure website (i.e. a portal), the response was positive. When combined with the findings from previous research indicating the positive effect of having health information tailored to specific users, the portal idea seems powerful indeed.[20]

As clinics and hospitals look into deploying patient portals, they will need to include efforts to understand their users and the factors that are likely to influence their use of the Internet to meet their healthcare information needs. As patients and caregivers depend more on mobile devices for Internet access, the ways in which health information is presented will evolve to include resources that are optimized for mobile use. Clinicians and designers of health information resources are advised to take this into account to ensure that health information is provided in the most appropriate ways. Digital literacy and health literacy are also key considerations in establishing effective ways to increase health information access and use. (After the completion of this study, the organization implemented a mobile version of the patient portal.)

Future studies may wish to focus on patient and family education, specifically directed at increasing portal use in the ways that their patients access the portal. Respondents' positive responses to using the Internet for administrative tasks and for viewing clinical items bode well for clinics and hospitals in the United States because they must meet the 'Meaningful Use' requirements mandated by the Federal government. These regulations require providers who received Federal funding to implement HIT and to demonstrate that they are using this technology 'meaningfully'. That is, they must work towards achieving specific goals such as installing a patient portal that will inform and engage patients. The next step will be to explore different use patterns to examine whether users will limit their use of mobile devices to accessing their personal health information or whether they will engage in more complicated tasks such as searching for health information.[21]

Our findings indicate that providers must also consider the health literacy of those expected to use their portal. Almost all of the respondents to our survey reported having a high school education or above, but a quarter to a third admitted having difficulty with reading health materials. This finding aligns with the perception that low health literacy is widespread across the US population. Calls for making health information more understandable and accessible are especially relevant to the designers of patient portals. As patient portals are enhanced with educational and informational resources, care must be exercised to create materials at appropriate levels and are tailored to the needs, preferences, and abilities of the intended users so that they can be read, understood, and followed.

Limitations

Our sample was a convenience sample of parents and caregivers in five paediatric clinics in a single metropolitan area; respondents may not have had the same characteristics as non-responders. Further study is needed to assess the generalizability of these findings to other types of clinics with adult populations. In addition, participants' education and income were generally higher than the city average, which may indicate some response bias to the survey. If patients had limited health literacy, they may have been less likely to respond to the survey, in which case our findings may overestimate patient interest in using the Internet for health information seeking and possible portal usage. As a result, the lessons learned from these experiences may not be generalizable to all communities.

Conclusion

This study looked at a population of parents and caregivers in five urban and suburban/rural paediatric practices to understand their health information–seeking patterns and preferences and to assess their capabilities and habits in order to gauge the potential for portal use. We found that a majority of patients cared for by these practices had access to the Internet and were interested in using a patient portal as a way to manage their care. Their preference for nurse- or doctor-provided information may be a factor that influences portal adoption among this population. Our findings lead us to recommend that organizations such as hospital and clinics that are contemplating the implementation of a patient portal should conduct a similar survey to understand more about their potential users. Digital technology is rapidly expanding in the healthcare arena, along with calls for greater involvement of patients. Our data suggest that these are welcome changes that could result in more informed and engaged patients and families. At the same time, it is critically important to identify communities where access to resources is limited and therefore may not be able to access use the resources. Low health literacy remains a serious problem among the general population, and it is a potential barrier to portal adoption within target populations. With simple survey techniques such as those described here, providers can be confident that their investment in a patient portal will have the intended results. Only by taking the time to survey potential users and assess their health literacy and their patterns and preferences for digital technology will portal designers and implementers have the confidence that they are truly advancing the goals of Healthy People 2020.

Acknowledgements

The authors gratefully acknowledge the contributions of Jordan Jobs and Debra Abrams in the collection and preparation of the data and Marcy Polansky for statistical assistance. All authors contributed equally to this study. A copy of the survey instrument may be requested from the corresponding author Dr Prudence Dalrymple at pdalrymple@drexel.edu

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Office of Disease Prevention and Health Promotion, https://www.healthypeople.gov/2020/topics-objectives/topic/health-communication-and-health-information-technology (accessed 29 November 2015).
2. Kreps GL and Neuhauser L. New directions in eHealth communication. *Patient Educ Couns* 2010; 78: 329–336.
3. Sanders MR, Winters P, Fortuna RJ, et al. Internet access and patient portal readiness among patients in a group of inner-city safety-net practices. *J Ambul Care Manage* 2013; 36(3): 251–259.
4. Reinhardt ER. Technical paradigms for realizing ubiquitous care. *Stud Health Technol Inform* 2008; 134: 129–134.
5. Johnson JD and Case DO. *Health information seeking*. New York: Peter Lang, 2012.
6. Neuhauser L and Kreps GL. Rethinking communication in the e-health era. *J Health Psych* 2003; 8: 7–22.
7. Lustria ML, Smith SA and Hinnant CC. Exploring digital divides: an examination of eHealth technology use in health information seeking, communication and personal health information management in the USA. *Health Informatics J* 2011; 17(3): 224–243.
8. Khoo K, Bolt P, Babl FE, et al. Health information seeking by parents in the Internet age. *J Paediatr Child Health* 2008; 44(7–8): 419–423.
9. Britto MT, Jimison HB, Munafo JK, et al. Usability testing finds problems for novice users of pediatric portals. *J Amer Med Inform Assoc* 2009; 16(5): 660–669.
10. Wainstein B, Sterling-Levis K, Baker S, et al. Use of the Internet by parents of paediatric patients. *J Paediatr Child Health* 2006; 42(9): 528–532.
11. Britto MT, Hesse EA, Kamdar OJ, et al. Parents' perceptions of a patient portal for managing their child's chronic illness. *J Pediatr* 2013; 163: 280–281.
12. Walsh AM, Hyde MK, Hamilton K, et al. Predictive modelling: parents' decision making to use online child health information to increase their understanding and/or diagnose or treat their child's health. *BMC Med Inform Decis Mak* 2012; 12: 144.
13. Ancker JS, Barrón Y, Rockoff ML, et al. Use of an electronic patient portal among disadvantaged populations. *J Gen Intern Med* 2011; 2: 1117–1123.
14. Pecchioni LL and Sparks L. Health information sources of individuals with cancer and their family members. *Health Commun* 2007; 21: 143–151.
15. Centers for Disease Control and Prevention. What is health literacy? http://www.cdc.gov/healthliteracy/learn/index.html (accessed 29 November 2015).
16. National Center for Education Statistics. Health literacy of America's adults: results from the 2003 National Assessment of Adult Literacy, 2006, http://nces.ed.gov/pubs2006/2006483.pdf (accessed 29 November 2015).
17. Rogers M, Zach L, Multak N, et al. The digital divide in Philadelphia: drawing a portrait of technology use in an urban area. In: *5th annual healthcare informatics symposium*, Center for Biomedical Informatics (CBMi), Philadelphia, PA, April 27, 2012.
18. National Library of Medicine. How to write easy-to-read health materials. *MedlinePlus*, http://www.nlm.nih.gov/medlineplus/etr.html
19. Chew LD, Bradley KA and Boyko EJ. Brief questions to identify patients with inadequate health literacy. *Fam Med* 2004; 36: 588–594.
20. Hawkins RP, Kreuter M, Resnicow K, et al. Understanding tailoring in communicating about health. *Health Educ Res* 2008; 23: 454–466.
21. Using patient portals to achieve meaningful use, www.himss.org/using-patient-portals-achieve-meaningful-use-ep-edition (accessed 22 October 2016).

Original Article

Health Informatics Journal

Health Informatics Journal
2018, Vol. 24(4) 379–393
© The Author(s) 2016
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1460458216675500
journals.sagepub.com/home/jhi
SAGE

A soft computing approach for diabetes disease classification

Mehrbakhsh Nilashi, Othman Bin Ibrahim, Abbas Mardani, Ali Ahani and Ahmad Jusoh
Universiti Teknologi Malaysia, Malaysia

Abstract

As a chronic disease, diabetes mellitus has emerged as a worldwide epidemic. The aim of this study is to classify diabetes disease by developing an intelligence system using machine learning techniques. Our method is developed through clustering, noise removal and classification approaches. Accordingly, we use expectation maximization, principal component analysis and support vector machine for clustering, noise removal and classification tasks, respectively. We also develop the proposed method for incremental situation by applying the incremental principal component analysis and incremental support vector machine for incremental learning of data. Experimental results on Pima Indian Diabetes dataset show that proposed method remarkably improves the accuracy of prediction and reduces computation time in relation to the non-incremental approaches. The hybrid intelligent system can assist medical practitioners in the healthcare practice as a decision support system.

Keywords

clustering, diabetes disease diagnosis, incremental principal component analysis, incremental support vector machine, machine learning

Introduction

Diabetes has been one of the leading health problems in the United States.[1] It has attained the dubious distinction of becoming the fifth leading cause of disease-related death.[2] Diabetes is a chronic endocrine disorder affecting the body's metabolism and resulting in structural changes affecting the organs of the vascular system.[3,4] Generally, diabetes is characterized as existing in two major forms: (a) insulin-dependent (Type I)[5] and (b) noninsulin-dependent (Type II).[6] The latter appears to be the more common, accounting for 80 percent of all cases.[2] The Pima is one of the most studied populations regarding diabetes, not only among American Indians, but in the world.[7] The most studied populations regarding diabetes is Pima, not only among American Indians but also in the world.[7] The samples of studied populations regarding diabetes refer to discrete Type-2 positive and negative instances.

Corresponding author:
Mehrbakhsh Nilashi, Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia.
Email: nilashidotnet@hotmail.com

The only way for the diabetes patient to live with this disease is to keep the blood sugar as normal as possible without serious high or low blood sugars, and this is achieved when the patient uses a correct management (therapy) which may include diet and exercising, taking oral diabetes medication or using some form of insulin.[2] However, treating the diabetes disease is also a difficult, an expensive and a complex task for the medical staff.[8] There are a number of important things to record about the patient and disease that help the doctors to make an optimal decision about the patient to make his or her life better.

Machine learning deals with the development of technologies which allow machines to learn. The challenge is to create algorithms that can take a group of patterns (on a broader range, the existing knowledge) and automatically make new inferences from the initial information, with or without human intervention.

From the machine learning perspective, classification is the problem of identifying a set of observations into several categories, based on the training result of a subset of observations whose belonging category is known. The unsupervised learning is defined as cluster analysis. It is also called clustering. Clustering is a process of putting a set of observations into several reasonable groups according to certain measures of similarity within each group. The clustering problem has been addressed in many disease diagnosis systems.[9–11] This reflects its broad appeal and usefulness as one of the steps in exploratory health data analysis.

There is a vast sea of different techniques and algorithms used in data mining, especially for supervised machine learning techniques; therefore, selecting the appropriate technique has been a challenge among researchers in developing the diabetes disease diagnosis systems.[12,13] In addition, although these data mining methods can be used to classify the diabetes disease through a set of real-world datasets, most of the methods developed by supervised methods in the previous researches do not support the incremental approaches for diabetes disease prediction. Furthermore, standard supervised methods usually cannot be performed in incremental situation and therefore they require to recompute all the training data to construct the classification model. Hence, in order to improve predictive accuracy and computation time of diabetes disease classification, a new method is proposed by applying noise removal, classification and clustering techniques. To the best of the authors' knowledge, there is no implementation of classification method (support vector machine (SVM)), clustering method (expectation maximization (EM)) and noise removal method (principal component analysis (PCA)) for diabetes disease diagnosis from the real-world dataset. In addition, since in medical datasets constantly new information is available, it is desirable to incrementally update the once trained models to reduce computation time in classifying the data. The proposed method in the study at hand supports incremental updates and re-learning of data and is more efficient in memory requirement.

Our study at hand is organized as follows. In section "Related work," we present the related work. In section "Methodology of research," the research methodology and all techniques incorporated to the proposed method are explained. In section "Results of methods," the evaluations of methods are presented. Finally, we conclude our work in section "Conclusion and future work."

Related work

Polat et al.[14] used discriminant analysis and SVM for diabetes classification. Using 10-fold cross-validation, they achieved 82.05 percent of accuracy on Pima diabetes dataset. Kayaer and Yıldırım[15] developed a method using general regression neural network (GRNN) for diabetes classification. The method was tested on Pima Indian Diabetes (PID) and achieved 80.21 percent accuracy for classification. Aslam et al.[16] proposed a method using genetic programming (GP) for diabetes classification. The method includes three stages: features selection, features generation and testing. Two classifiers, the *k*-nearest neighbor (*k*-NN) and SVM, were used for evaluating the selected features. The authors tested the performance of method using Pima Indians diabetes dataset. A hybrid intelligent system was developed by Kahramanli and Allahverdi[12] using fuzzy neural network (FNN) and artificial neural network (ANN). They evaluated the method on two public medical datasets, Pima Indians diabetes and Cleveland heart disease. Using *k*-fold cross-validation, the method obtained classification accuracies of 84.24 and 86.8 percent for Pima Indians diabetes dataset and Cleveland heart disease dataset, respectively. An intelligent system was proposed by Erkaymaz and Ozer[13] for diagnosis of diabetes. The method was based on the small-world feed forward artificial neural network (SW-FFANN). The accuracy of the method was 91.66 percent. Ganji and Abadeh[17] developed a method, FCS-ANTMINER, by ant colony optimization (ACO). They extracted a set of fuzzy rules to classify the diabetes disease. The obtained classification accuracy was 84.24 percent. An intelligent diagnosis system, linear discriminant analysis–adaptive neuro-fuzzy inference system (LDA-ANFIS), was developed by Dogantekin et al.[18] for diabetes using LDA classification method and neuro-fuzzy (ANFIS) system. The classification accuracy of LDA-ANFIS was about 84.61 percent. A comparative study of diabetes disease on Pima Indian diabetes disease was conducted by Temurtas et al.[19] They used multilayer NN which was trained by Levenberg–Marquardt (LM) algorithm and probabilistic NN. An automatic diagnosis system, linear discriminant analysis–Morlet wavelet support vector machine (LDA–MWSVM), was developed for diabetes by Çalişir and Doğantekin.[20] They used Morlet wavelet support vector machine (MWSVM) classifier and LDA. Their method classification accuracy was about 89.74 percent.

From the literature on diabetes disease diagnosis from experiments with Long Beach and Cleveland Clinic Foundation, we found that at the moment there are no implementations of PCA, Gaussian mixture model with EM and SVM method for distinguishing between presence and absence of diabetes disease in patients. This research accordingly tries to develop a diabetes disease diagnosis intelligent system based on these methods. Overall, in comparison with research efforts found in the literature, in this research

- EM is used for data clustering. The clustering problem has been addressed in many disease diagnosis systems.[9–11] This reflects its broad appeal and usefulness as one of the steps in exploratory health data analysis. In this study, EM clustering is used as an unsupervised classification method to cluster the data of experimental dataset into similar groups.
- SVM is used for data classification. SVM is widely employed in diagnosis of diseases for their efficiency and robustness. It is a promising classification approach which has been used in many researches on diseases classification.[21–24]
- PCA is used for dimensionality reduction and dealing with the multi-collinearity problem in the experimental data. This technique has been used in developing in many disease diagnosis systems to eliminate the redundant information in the original health data.[25]
- Incremental techniques, incremental support vector machine (ISVM) and incremental principal component analysis (IPCA), are used for incremental learning. Incremental techniques have been used in many disease diagnosis systems[23,26,27] to enhance the predictive accuracy and decrease the computation time of classification.

By combination of EM, PCA and SVM, a hybrid intelligent system is proposed to increase the predictive accuracy and decrease the computation time of diabetes disease.

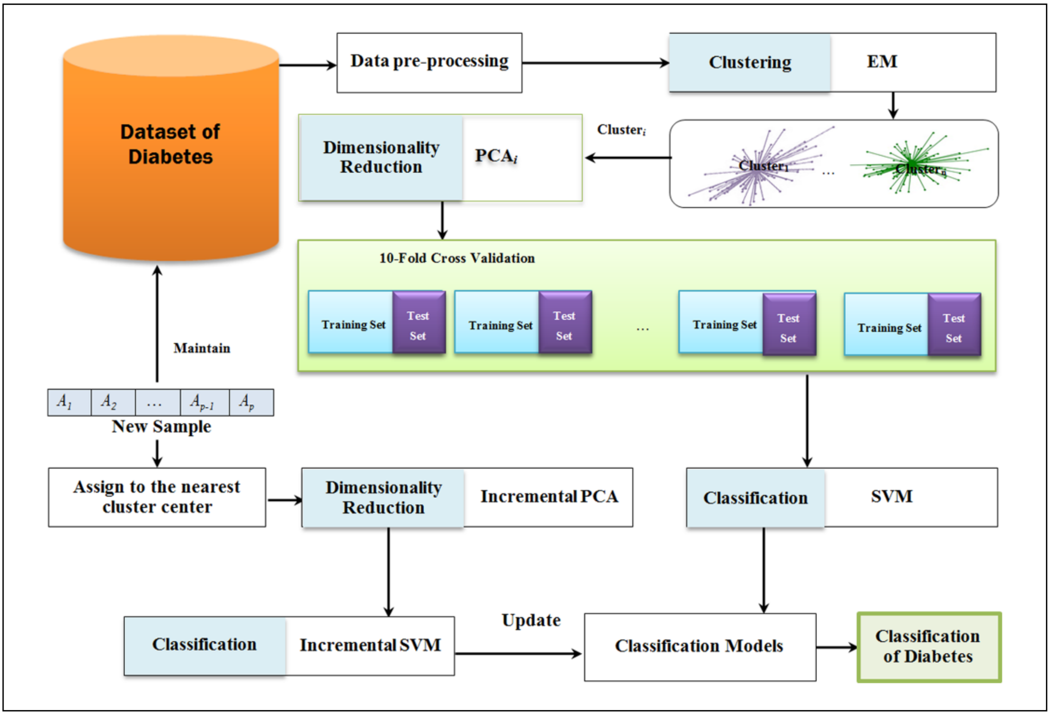


Figure 1. Proposed method for the diabetes diseases diagnosis.

Methodology of research

Focusing on the prediction and classification of diseases, this study uses PCA, EM and classification (SVM) methods. We also develop the method for incremental situation using incremental noise removal method (IPCA) and incremental classification (ISVM) method. The general framework of proposed model is shown in Figure 1. We propose to rely on classification methods to learn the classification functions. Additionally, PCA is employed for dimensionality reduction and to overcome the multi-collinearity problem of the datasets. In addition, since in medical datasets the data are constantly collected from the new observations, it is beneficial to incrementally update previous model of classification by considering only new arrived data to reduce the computation time in classification tasks. The proposed method therefore supports incremental updates using IPCA and ISVM to re-learn the medical data which can be more efficient in memory requirement. These methodologies are addressed in the following sections.

Dataset for the experiments

The Pima aboriginals diabetes dataset is provided at the courtesy of National Institute of Diabetes and Digestive and Kidney Diseases and Vincent Sigillito of the Applied Physics Laboratory of the Johns Hopkins University who was the original donor of the dataset. The actual data itself are obtained by the author of this research from the website of the UCI (University of California, Irvine).[28] These data have been used in the past by the researchers to investigate possible vital signs that may be used to indicate the presence of diabetes within patients according to World Health Organization (WHO) standards. There are a total of 768 training instances included in this dataset.

Table 1. Description of the features of Pima Indian Diabetes dataset.

Feature label	Variable type	Range
Number of times pregnant	Integer	0–17
Plasma glucose concentration in a 2 h oral glucose tolerance test	Real	0–199
Diastolic blood pressure	Real	0–122
Triceps skin fold thickness	Real	0–99
2 h serum insulin	Real	0–846
Body mass index	Real	0–67.1
Diabetes pedigree function	Real	0.078–2.42
Age	Integer	21–81
Class	Binary	Tested positive for diabetes = 1

Each training instance has eight features and a class variable that provides the label for that training instance (see Table 1). The features are number of times pregnant, plasma glucose concentration, diabetes pedigree function, triceps skin fold thickness (mm), diastolic blood pressure (mmHg), 2-h serum insulin (mU/mL), body mass index (kg/m^2) and years of age. The class variable takes on the binary value of 0 or 1, with 0 indicating a healthy person and 1 indicating a patient with diabetes.

EM clustering

One of the commonly used model-based clustering approaches is mixture-approach EM algorithm, which was first officially proposed by Dempster et al.[29] Later, Wu[30] has corrected a flawed convergence analysis in the method. The EM algorithm is widely used because of its simplicity, easy implementation and its efficient iterative procedure in computing the maximum likelihood (ML).[31–34]

Since it is not easy to maximize the log-likelihood directly, EM algorithm maximizes the expectation of complete log-likelihood instead. The complete data in EM algorithm are considered to be (x, z). z is the missing data indicating the mixture component origin label of each observation. $z = (z_1, ..., z_n)$ where $z_i = k$ when x_i belongs to the component k. The complete log-likelihood takes the form

$$CL(\Phi, Z \mid X) = \sum_{k=1}^{K} \sum_{k=1}^{K} z_{ik} \log\,(\pi_k f_k(x;\theta_k)) \tag{1}$$

EM algorithm starts from the initial parameter θ^0, then computes the expectation step (E step) and the maximization step (M step) iteratively:

E step. In this step, the expected value of the complete log-likelihood function is calculated. The calculation is with respect to the conditional distribution of z given x under the current estimate of the parameters Φ

$$Q(\Phi, \Phi^{(q)}) = E[P] \tag{2}$$

$$P = \log(CL(\Phi, Z \mid X) \tag{3}$$

that is, calculate the posterior probabilities $t_{ik}^{(q)}$ of x_i belonging to the kth component as

$$t_{ik}^{(q)} = \frac{\pi_k^{(q)} f_k(x;\theta_k^{(q)})}{\sum_l \pi_l^{(q)} f_l(x;\theta_l^{(q)})} \tag{4}$$

M step. In this step, the parameter $\Phi^{(q+1)}$ is found that maximizes the expectation

$$\Phi^{(q+1)} = \arg\max_{\Phi} Q(\Phi \mid \Phi^{(q)}) \tag{5}$$

PCA

PCA is a statistical technique for multivariate analysis and is used as a dimensionality reduction technique in data compression to retain the essential information and is easy to display.[35] The method identifies patterns in data and represents the data in a way that highlights similarities and differences. The central idea is to reduce the number of dimensions of the data while preserving as much as possible the variations in the original dataset.[36] PCA has four goals. The first goal is to extract the most information from the data. The second goal is to compress the data by only keeping the most characterizing information. The third goal is to simplify the description of the data and the fourth goal is to enable analysis of the structure of the observations. The analysis enables conclusions to be drawn regarding the used variables and their relations. The analysis is performed through transforming the data to a new set of variables, called the principal components (PCs).[37] The PCs are uncorrelated and ordered so that the first few PCs retain most of the variations of the total dataset.[38,39] The first PC describes the dimension in which the data have the biggest variation (variance) and the second component describes the dimension in which it has the second largest variation (variance).

PCA is chosen for this study because the method exemplifies a category of analysis methods. If the data have linear relations and are correlated, as data often are in medical datasets, the method will give a compression that maintains a high amount of the information in the original dataset. The described solution saves a compact summary of the data, which is derived by applying ideas from statistics to enable an analysis while preserving its characteristics. In this study, we use an algorithm for IPCA proposed by Hall et al.[40] that updates eigenvalues and eigenvectors incrementally.

ISVM

SVMs are large-margin classifiers which have found successful applications in many scientific fields such as engineering[41] and disease classification,[21] information retrieval,[38] finance and business[42] among many others. An important and crucial point in the SVM formulation is that it can provide a good generalization independent of the training set's distribution by making use of the principle of structural risk minimization. This principle provides a trade-off between the complexity of the classifier (accuracy in the training set) and the quality of fitting the training data (generalization-empirical error). Therefore, the SVMs belong to a class of algorithms which are known as maximum-margin classifiers. The size of the gap is decided upon the training samples which are between the margins. These samples are the so-called support vectors (SVs).

Classical SVMs method has been originally developed as offline classification algorithms that are trained with a pre-determined dataset before they can be used for classification problems. Cauwenberghs and Poggio[43] proposed ISVM by analyzing the changes of the Karush–Kuhn–Tucker (KKT) conditions for online learning when a new (incremental) sample was added into the

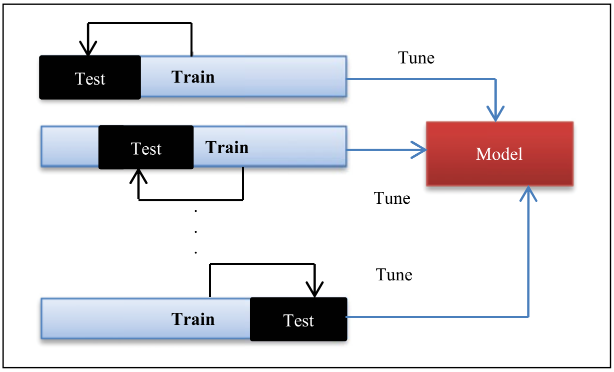


Figure 2. *k*-fold cross-validation.

old samples. Employing a partition of the dataset, ISVM trains an SVM which reserves only the SVs at each step of training the samples and creates the training set for the next step using these SVs. Hence, the key of ISVM is to preserve the KKT conditions on all existing training data while adiabatically adding a new vector. In this project, the MATLAB scripts of incremental learning are created based on Cauwenberghs and Poggio's[43] work.

Suppose the current working set is X and the incremental set is I. First, X is clustered by EM; thus, X is clustered to $\{X_1, X_2,\dots, X_b,\dots, X_M\}$ $(b = 1,\dots, M;$ M is the number of clusters). Then, each X_b is trained by SVM, respectively, and its corresponding training functions $f(x)$ can be obtained. For each sample (x_c, y_c) in I (new sample), its distance to each cluster is first calculated (Euclidean distance between the observation and the cluster center), and after performing IPCA, incremental learning is carried out using ISVM.

Cross-validation

Cross-validation is a statistical method that, in this research, is used for the performance evaluation of learning algorithms and performance of a predictive model on an unknown dataset. For this reason, using cross-validation, the datasets used in the research are divided into several equally sized subsets (see Figure 2). The learning model is then trained on some subsets known as training sets. After training process, the model is tested on the remaining subsets, known as test sets. According to the number of subsets partitioned, researcher tests *k*-fold cross-validation. For 10-fold cross-validation, researchers use 10 result of 10-fold cross-validation. In the experiments of this research, for the training of models, it is considered to test different 10 for 10-fold cross-validation, so that researchers can make sure that there are enough training instances to learn the models.[44]

Results of methods

The experimental results of the proposed method on real-world datasets are explained in this section. Here, the results of applying all incorporated methods in the proposed system are discussed.

Clustering with EM algorithm

In this research, EM algorithm is applied on experimental dataset. As far we know, in any clustering algorithm, the right number selection of the clusters is an important task. The selection of

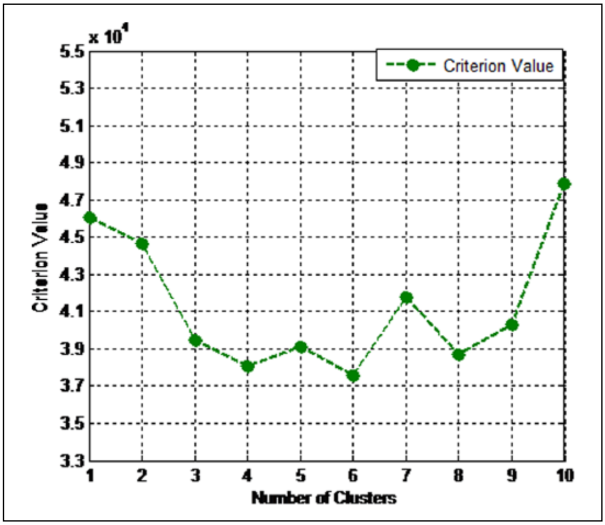


Figure 3. Best cluster using EM algorithm for PID dataset.

number needs to be performed to provide the best quality for clustering. In EM algorithm, the maximization of likelihood is important for the Gaussian mixture model. Akaike information criterion (AIC), as a model selection approach, can be used for the maximization of likelihood.[45] Accordingly, for the dataset used in this study, we have applied resubstitution AIC to select the value optimal number of clusters in EM algorithm. Additionally, 10-fold cross-validation was applied in the clustering procedure to obtain unbiased results. Hence, as we used resubstitution AIC estimate to choose the value optimal number of clusters, we need to test the number of clusters from $n = 1$ to $n = m$, in which for $n > m$, the criterion value be always increased. From the results, we found the minimum criterion value for $n < 10$ and, accordingly, we decided $m = 10$ for obtaining optimal criterion value. The results of clustering by EM is presented in Figure 3 where based on chosen criterion, the various numbers of clusters are shown to select the best cluster for the datasets.

In addition, from Table 2, it can be seen that the best criterion value (37577.854250) is obtained when EM generates six clusters. For visualizing clusters of EM for each dataset in scatter plot, we use two PCs of PCA in order to obtain a two-dimensional (2D) representation. In Figure 4, the clusters (six clusters) generated by EM are visualized. As can be seen, we project the observations in the first two dimensions generated by PCA.

PCA evaluation

As PCA generates PCs instead of original factors, choosing the right number selection of these PCA is an important task. If we select too many factors, we include noise from the sampling fluctuations in the analysis. If we choose too few factors, we lose relevant information, and the analysis is incomplete. As we know that the eigenvalue associated to a factor corresponds to its variance, the eigenvalue indicates the importance of the factor. The higher the value, the higher the importance of the factor. The eigenvalues for each factor can be indicators for its importance. In this study, we have applied the rule proposed by Cattell.[46] Accordingly, we create "scree" plots that show the eigenvalues of the factors. In the "scree" plots, we can simply detect "elbows" to decide the number of PCAs to be used in the classification process.

Table 2. Best cluster using EM algorithm for PID dataset.

Number of clusters	Criterion
1	46092.154901
2	44644.787691
3	39507.303442
4	38080.146803
5	39121.694768
6	**37577.854250**
7	41783.431618
8	38675.813840
9	40309.951994
10	47894.584548

EM: expectation maximization; PID: Pima Indian Diabetes.
The boldface in the table indicates the best criterion value.

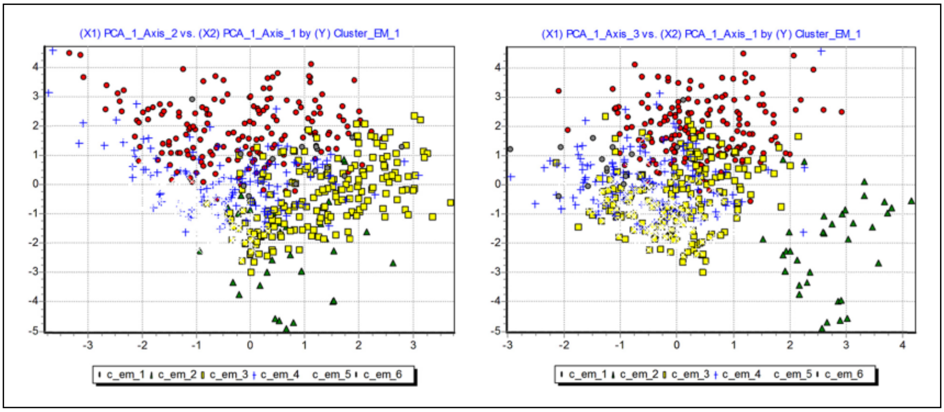


Figure 4. Clusters visualization of PID dataset.

We employed the PCA technique for the clusters of experimental dataset obtained by EM algorithm. Based on the rule proposed by Cattell,[46] in PID, for Cluster 1, we included the elbow into the selection, that is, we selected $k = 2$ factors. Indeed, the eigenvalues associated with the second factor was high. In addition, three PCs for Clusters 2 and 4 and four PCs for Clusters 3, 5 and 6 were chosen.

Performance evaluation of ISVM

This section provides the experimental results of diabetes disease classification with non-incremental and incremental SVM classifiers based on PID. In addition, comparison experiments with other methods in the literature are performed using non-incremental and incremental SVM based on the same dataset.

As far we know the classical SVM, it can be used as offline classification and prediction methods which are trained with a pre-determined dataset before they can be used for the disease classification and prediction. In addition, the capability of classical SVM is limited by fixed number of training samples. Therefore, there was a need for a classifier that be able to augment itself with new data constantly. Accordingly, we have implemented ISVM to overcome this issue by taking their ability in learning incrementally.

The models of classification were trained under a 4 GHz processor PC and Microsoft Windows 7 running MATLAB 7.10 (R2010a). We applied ISVM with radial basis function (RBF) kernel on experimental dataset clustered by EM algorithm. To show the predictive accuracy of the proposed method, we use area under the curve (AUC) of receiver operating characteristic (ROC) chart. ROC is a graphical display that provides the measure of classification accuracy of the model using sensitivity and specificity.[24] For predicting events, sensitivity in ROC can be used as a measure of accuracy which can be calculated by dividing the true positive over total actual positive. For predicting nonevents, specificity can be used as a measure of accuracy which can be calculated by dividing true negative over the total actual negative of a classifier for a range of cutoffs.

As we have selected RBF kernel for SVM classifier, there are two parameters, C and γ, which are unknown and we need to set a best value for them. Hence, some kind of model selection methods are required to find an optimal value for C and γ. The aim of this task is to find good parameters for RBF kernel so that SVM classifier can provide good classification models and accurately predict the unknown classes in testing data. To do so, we used k-fold cross-validation (k = 10) as a statistical model selection method. Using 10-fold cross-validation, the data used in the research were divided into 10 equally sized subsets. Accordingly, a single subsample was retained as the test data and the remaining nine subsamples were used as the training data. The learning models were then trained on nine subsamples. After training process, the model was tested on the single subset and the 10 results from each of the folds could be averaged to produce a single generalization estimation. By trying several values for the parameters C and γ ,[47] we then set the value of penalty parameter C and γ in RBF kernel equal to the optimal one determined via 10-fold cross-validation.

In order to experimentally demonstrate the effectiveness of EM clustering, IPCA and ISVM, we divide the data in the clusters into two categories. The first category is considered as initial clustering and the second one is considered for incremental phase that is incrementally added to the initial clusters data. The aim is to calculate the classification prediction time method after adding the second category data incrementally. We perform this procedure on all clusters and present the average computation time. The general procedure of this evaluation is demonstrated in Figure 5.

For evaluating the ISVM, we initially considered 20 percent of data in any cluster for test set, 20 percent for initial data clusters and 60 percent for incremental set which is incrementally added to the clusters, initial data.

The increment ratio is considered 10 percent of incremental set and added to training set and calculated computation time. Specifically, we consider six measurement points add the 10 percent of data to the initial clusters in each measurement point. In that direction, for different measurement points, the average computation time and accuracy were calculated for all clusters.

To experimentally show the effectiveness of EM and incremental approach (ISVM), we conduct the experiments on the public PID dataset and compare with the methods of the non-incremental learning for computation time. It should be noted that the kernel parameters and penalty parameter C have been determined by 10-fold cross-validation.

In Figure 6(a), the classification accuracy of ISVM measured by ROC in each cluster for PID is presented. From all plots in Figure 6(a), we can see the influence of using ISVM on accuracy is significant and the incremental update has provided a good classification accuracy measured by ROC in each cluster. The average accuracy obtained by the proposed method is about 97.95 percent for all clusters. It should be noted that the increment ratio for ISVM is considered 10 percent of incremental set and added to training set, and we calculated accuracy in each fold of 10-fold cross-validation.

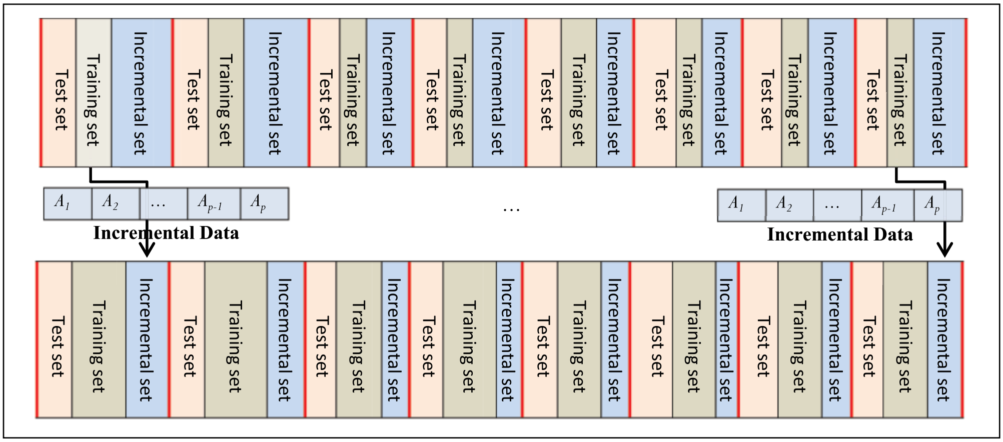


Figure 5. ISVM evaluation procedure.

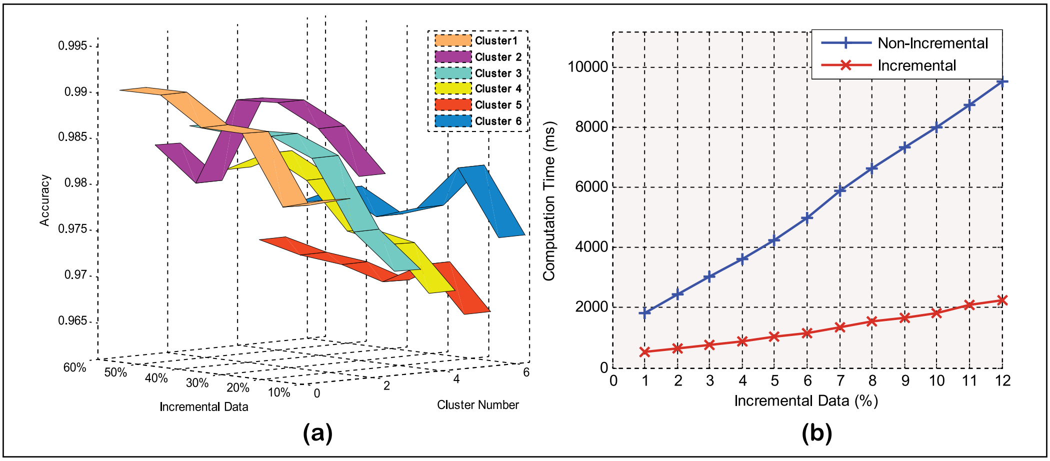


Figure 6. Incremental SVM evaluation for (a) accuracy and (b) computation time.

Figure 6(b) presents the computation time results of our experiments for the proposed method in the incremental situation. The computation time is plotted as a function of the incremental data percentage. Note that in the comparative experiments, the non-incremental SVM and ISVM were tested with a 10-fold cross-validation. From all curves in Figure 6(b), we can see that the incremental method has significantly reduced the computation time in relation to the non-incremental one. In addition, as the figure shows, non-incremental methods perform poor with respect to time for PID dataset. From the curves as shown in the figures, it can be also observed that by increasing the number of incremental data, the computation time is slightly raised. A possible explanation could be that, since the non-incremental method cannot learn in the incremental situation, it requires to recompute all the training data to build the classification and prediction models. In addition, the non-incremental SVM method can be used as an offline method and is trained with a predetermined dataset before it can be used for the disease classification. Thus, the capability of non-incremental SVM is limited by fixed number of training samples in each cluster, and it is not be

Table 3. Comparison of proposed method with other classifiers for PID.

Method	Reference	Accuracy
General regression neural network	Kayaer and Yıldırım[15]	80.21%
GDA-LSSVM	Polat et al.[14]	79.16%
MWSVM	Çalişir and Doğantekin[20]	89.74%
SW-FFANN	Erkaymaz and Ozer[13]	91.66%
IPCA-EM-ISVM	This study	97.95%

PID: Pima Indian Diabetes; MWSVM: Morlet wavelet support vector machine; SW-FFANN: small-world feed forward artificial neural network; IPCA-EM-ISVM: incremental principal component analysis–expectation–maximization–incremental support vector machine.

able to augment itself with new data constantly. In other words, those medical records in the experimental dataset, which have been incrementally added, need to be retained along with the previous data in each cluster through non-incremental SVM. However, the methods that use ISVM reduce computation time results as it needs to train only the data which have been added incrementally. Finally, the method that combines clustering, IPCA and ISVM lead to the computation time reduction results. Overall, the results showed that the main practical advantage of using ISVM as a training method is a great saving in computation time.

We compare the accuracy of our proposed method with the classification accuracy of the methods GRNN,[15] General Discriminant Analysis and Least Square Support Vector Machine (GDA-LSSVM),[14] MWSVM[20] and SW-FFANN[13] for PID. The performance of the classifiers that were compared with our method is shown in Table 3. From the results shown in this table, our proposed method proves to have a better accuracy (0.9795) in relation to the other classification systems. Compared to GRNN (80.21%), GDA-LSSVM (79.16%), MWSVM (89.74%) and SW-FFANN (91.66%), our classification, clustering and noise removal techniques help to improve the classification accuracy of diabetes disease by more than 17, 18, 8 and 6 percent, respectively. This shows the effectiveness of incorporating the clustering and PCA techniques for the classification accuracy of diabetes disease.

Conclusion and future work

In this article, we propose a new hybrid intelligent system for diabetes disease classification using machine learning techniques. We applied EM clustering algorithm to cluster the experimental diabetes disease dataset and SVM for classification of disease types. In addition, PCA was used for dimensionality reduction and to address multi-collinearity in the dataset. Furthermore, since new information is constantly available in medical datasets, it is desirable to incrementally update the trained models to reduce the computation time. The proposed method in this study at hand then supports incremental updates that were more efficient in memory requirement. In order to analyze the effectiveness of the proposed method and validate the system, several experiments were conducted on PID. The dataset was taken from Data Mining Repository of the UCI. The results indicated that the method which combines clustering, IPCA and ISVM obtains good classification accuracy and significantly reduces the computation time in relation to the non-incremental methods. All of the approaches used in this study may also be applicable to other classification problems within the medical domain. However, there is still plenty of work in conducting researches on incremental algorithms for disease diagnosis in order to exploit all their potential and usefulness. In the future work, more attention should be paid to the datasets for disease classification and prediction using the incremental machine learning approaches. Hence, in our future study, we plan to evaluate the proposed method on additional datasets and in particular on large datasets to show the effectiveness of the incremental methods on computation time of large data in relation to the non-incremental ones.

Data availability

The actual data itself are obtained by the author of this research from the website of the UCI (University of California, Irvine).

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors would like to thank the Research Management Center (RMC) at Universiti Teknologi Malaysia (UTM) for supporting and funding this research under the Post-Doctoral Fellowship Scheme Grant (Vote no. Q.J130000.21A2.03E26) and the UTM-GUP Research Grant (Vote no. Q.J130000.2506.13H49).

References

1. Onitilo AA, Stankowski RV, Berg RL, et al. A novel method for studying the temporal relationship between type 2 diabetes mellitus and cancer using the electronic medical record. *BMC Med Inform Decis Mak* 2014; 14(1): 1.
2. Hamburg BA and Inoff GE. Relationships between behavioral factors and diabetic control in children and adolescents: a camp study. *Psychosom Med* 1982; 44(4): 321–339.
3. Court S, Sein E, McCowen C, et al. Children with diabetes mellitus: perception of their behavioural problems by parents and teachers. *Early Hum Dev* 1988; 16(2): 245–252.
4. Egede LE. Diabetes, major depression, and functional disability among U.S. adults. *Diabetes Care* 2014; 27(2): 421–428.
5. Frandsen CS, Dejgaard TF and Madsbad S. Non-insulin drugs to treat hyperglycaemia in type 1 diabetes mellitus. *Lancet Diabetes Endocrinol* 2016; 4(9): 766–780.
6. Kramer CK, Zinman B and Retnakaran R. Short-term intensive insulin therapy in type 2 diabetes mellitus: a systematic review and meta-analysis. *Lancet Diabetes Endocrinol* 2013; 1(1): 28–34.
7. Knowler WC, Pettitt DJ, Bennett PH, et al. Diabetes mellitus in the Pima Indians: genetic and evolutionary considerations. *Am J Phys Anthropol* 1983; 62(1): 107–114.
8. Egede LE and Miohel Y. Perceived difficulty of diabetes treatment in primary care: does it differ by patient ethnicity? *Diabetes Educ* 2001; 27(5): 678–684.
9. Hruschka ER and Ebecken NF. Extracting rules from multilayer perceptrons in classification problems: a clustering-based approach. *Neurocomputing* 2006; 70(1): 384–397.
10. Chen CH. A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection. *Appl Soft Comput* 2014; 20: 4–14.
11. Polat K. Classification of Parkinson's disease using feature weighting method on the basis of fuzzy C-means clustering. *Int J Syst Sci* 2012; 43: 597–609.
12. Kahramanli H and Allahverdi N. Design of a hybrid system for the diabetes and heart diseases. *Expert Syst Appl* 2008; 35(1): 82–89.
13. Erkaymaz O and Ozer M. Impact of small-world network topology on the conventional artificial neural network for the diagnosis of diabetes. *Chaos Soliton Fract* 2016; 83: 178–185.
14. Polat K, Günes S and Arslan A. A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine. *Expert Syst Appl* 2008; 34(1): 482–487.

15. Kayaer K and Yıldırım T. Medical diagnosis on Pima Indian diabetes using general regression neural networks. In: *Proceedings of the international conference on artificial neural networks and neural information processing*, 26–29 June 2003, pp. 181–184. Istanbul: Springer.
16. Aslam MW, Zhu Z and Nandi AK. Feature generation using genetic programming with comparative partner selection for diabetes classification. *Expert Syst Appl* 2013; 40(13): 5402–5412.
17. Ganji MF and Abadeh MS. A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis. *Expert Syst Appl* 2011; 38(12): 14650–14659.
18. Dogantekin E, Dogantekin A, Avci D, et al. An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA-ANFIS. *Digit Signal Process* 2010; 20(4): 1248–1255.
19. Temurtas H, Yumusak N and Temurtas F. A comparative study on diabetes disease diagnosis using neural networks. *Expert Syst Appl* 2009; 36(4): 8610–8615.
20. Çalişir D and Doğantekin E. An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. *Expert Syst Appl* 2011; 38(7): 8311–8315.
21. Long NC, Meesad P and Unger H. A highly accurate firefly based algorithm for heart disease prediction. *Expert Syst Appl* 2015; 42: 8221–8231.
22. Awad M, Motai Y, Näppi J, et al. A clinical decision support framework for incremental polyps classification in virtual colonoscopy. *Algorithms* 2010; 3: 1–20.
23. Molina JFG, Zheng L, Sertdemir M, et al. Incremental learning with SVM for multimodal classification of prostatic adenocarcinoma. *PLoS ONE* 2014; 9: e93600.
24. Yu W, Liu T, Valdez R, et al. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Med Inform Decis Mak* 2010; 10(1): 16.
25. Çali-ir D and Dogantekin E. A new intelligent hepatitis diagnosis system: PCA–LSSVM. *Expert Syst Appl* 2011; 38: 10705–10708.
26. Gerlá V, Lhotska L, Murgas M, et al. An incrementalapproach to clinical EEG data classification. In: *Proceedings of the 6th European conference of the international federation for medical and biological engineering*, Dubrovnik, 7–14 September 2014, pp. 489–492. Switzerland: Springer International Publishing. Available at: http://link.springer.com/chapter/10.1007/978-3-319-11128-5_122
27. Tortajada S, Robles M and García-Gómez JM. Incremental logistic regression for customizing automatic diagnostic models. In: Fernández-Llatas C and García-Gómez JM (eds) *Data mining in clinical medicine*. New York: Springer Science+Business Media, 2015, pp. 57–78.
28. Bache K and Lichman M. *UCI Machine Learning Repository*. University of California, School of Information and Computer Science, 2013. Available at: http://archive.ics.uci.edu/ml
29. Dempster AP, Laird NM and Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc B Met* 1977; 39: 1–38.
30. Wu CJ. On the convergence properties of the EM algorithm. *Ann Stat* 1983; 11: 95–103.
31. Ordonez C and Omiecinsk E. FREM: fast and robust EM clustering for large data sets. In: *Proceedings of the eleventh international conference on information and knowledge management*, McLean, VA, 4–9 November 2002, pp. 590–599. New York: ACM.
32. Jung YG, Kang MS and Heo J. Clustering performance comparison using K-means and expectation maximization algorithms. *Biotechnol Biotechnol Equip* 2014; 28(1): S44–S48.
33. Nathiya G, Punitha SC and Punithavalli M. An analytical study on behavior of clusters using K means, EM and K-means algorithm. *Int J Comput Sci Inform Secur* 2010; 7(3): 185–190.
34. Mitra P, Pal SK and Siddiqi MA. Non-convex clustering using expectation maximization algorithm with rough set initialization. *Pattern Recogn Lett* 2003; 24(6): 863–873.
35. Moore BC. Principal component analysis in linear systems: controllability, observability, and model reduction. *IEEE T Automat Contr* 1981; 26(1): 17–32.
36. Nilashi M, Esfahani MD, Roudbaraki MZ, et al. A multi-criteria collaborative filtering recommender system using clustering and regression techniques. *J Soft Comput Decis Support Syst* 2016; 3(5): 24–30.
37. Nilashi M, Ibrahim O, Ithnin N, et al. A multi-criteria collaborative filtering recommender system for the tourism domain using Expectation Maximization (EM) and PCA–ANFIS. *Electron Commer R A* 2015; 14(6): 542–562.

38. Nilashi M, Ibrahim OB, Ithnin N, et al. A multi-criteria recommendation system using dimensionality reduction and Neuro-Fuzzy techniques. *Soft Comput* 2015; 19: 3173–3207.
39. Nilashi M, Jannach D, Ibrahim O, et al. Clustering-and regression-based multi-criteria collaborative filtering with incremental updates. *Inform Sciences* 2015; 293: 235–250.
40. Hall PM, Marshall AD and Martin RR. Incremental eigenanalysis for classification. *BMVC* 1998; 98: 286–295.
41. Farahmand M, Desa MI and Nilashi M. A comparative study of CCR-(e-SVR) and CCR-(e-SVR) models for efficiency prediction of large decision making units. *J Soft Comput Decis Support Syst* 2015; 2(1): 8–17.
42. Wu WW. Beyond business failure prediction. *Expert Syst Appl* 2010; 37(3): 2371–2376.
43. Cauwenberghs G and Poggio T. Incremental and decremental support vector machine learning. *Adv Neur In* 2001; 409–415.
44. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI* 1995; 14(2): 1137–1145.
45. Pelleg D and Moore AW. X-means: extending K-means with efficient estimation of the number of clusters. In: *Proceedings of the seventeenth international conference on machine learning (ICML)*, Stanford, CA, 29 June–2 July 2000, pp. 727–734. San Francisco, CA: Morgan Kaufmann Publishers.
46. Cattell RB. The scree test for the number of factors. *Multivar Behav Res* 1966; 1(2): 245–276.
47. Chang CC and Lin CJ. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2011; 2(3): 27.

Original Article

Health Informatics Journal

Health Informatics Journal
2018, Vol. 24(4) 394–409
© The Author(s) 2016
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1460458216675661
journals.sagepub.com/home/jhi
SAGE

Chronic obstructive pulmonary disease phenotypes using cluster analysis of electronic medical records

Rodrigo Vazquez Guillamet, Oleg Ursu, Gary Iwamoto, Pope L Moseley and Tudor Oprea
University of New Mexico School of Medicine, USA

Abstract

Chronic obstructive pulmonary disease is a heterogeneous disease. In this retrospective study, we hypothesize that it is possible to identify clinically relevant phenotypes by applying clustering methods to electronic medical records. We included all the patients >40 years with a diagnosis of chronic obstructive pulmonary disease admitted to the University of New Mexico Hospital between 1 January 2011 and 1 May 2014. We collected admissions, demographics, comorbidities, severity markers and treatments. A total of 3144 patients met the inclusion criteria: 46 percent were >65 years and 52 percent were males. The median Charlson score was 2 (interquartile range: 1–4) and the most frequent comorbidities were depression (36%), congestive heart failure (25%), obesity (19%), cancer (19%) and mild liver disease (18%). Using the sphere exclusion method, nine clusters were obtained: depression–chronic obstructive pulmonary disease, coronary artery disease–chronic obstructive pulmonary disease, cerebrovascular disease–chronic obstructive pulmonary disease, malignancy–chronic obstructive pulmonary disease, advanced malignancy–chronic obstructive pulmonary disease, diabetes mellitus–chronic kidney disease–chronic obstructive pulmonary disease, young age–few comorbidities–high readmission rates–chronic obstructive pulmonary disease, atopy–chronic obstructive pulmonary disease, and advanced disease–chronic obstructive pulmonary disease. These clusters will need to be validated prospectively.

Keywords

asthma, comorbidity, chronic obstructive pulmonary disease, epidemiology, factor analysis, phenotype

Introduction

Chronic obstructive pulmonary disease (COPD) is a heterogeneous disease characterized by persistent airflow limitation. It is caused by the inhalation of cigarette smoke and other noxious particles and gases. Mortality rates are 40.8 per 100,000 United States inhabitants every year, and as of 2010, chronic respiratory diseases were the fourth leading cause of death in the United States and are projected to be the third by 2020.[1,2]

Corresponding author:
Rodrigo Vazquez Guillamet, Division of Pulmonary, Critical Care and Sleep Medicine, University of New Mexico School of Medicine, Albuquerque, NM 87131, USA.
Email: rvazquezguillamet@salud.unm.edu

Until recently, international guidelines were basing specific treatment recommendations solely on airway obstruction, quality of life, number of exacerbations and exercise capacity oversimplifying in practice a very heterogeneous group of patients. This approach has resulted in improved symptoms and decreased number of COPD exacerbations, but its impact on survival has been disappointing.[1] Main causes of death in patients with respiratory conditions are related to cardiovascular disease and cancer[3,4] with other comorbidities also playing an important role.[5] Therefore, in its latest revision, the Global Initiative for Obstructive Lung Disease started providing guidance for the management of common comorbidities. However, their recommendations are limited by the lack of disease-specific outcome studies.[1]

Several studies have attempted to capture the heterogeneity of COPD patients using clustering techniques in order to describe phenotypes, provide more personalized therapies and pencil possible pathophysiology links.[6–13] However, most of them have had restrictive inclusion criteria, small sample sizes, have relied on highly specialized measurements and have rarely included United States subjects. The motivation of our study is to fill the gap left by previous similar studies by adding clinically relevant COPD phenotype categories[14] using cluster analyses on readily available electronic medical record data. This can constitute the first step toward stratified treatment in patients with COPD.[15]

Materials and methods

Study location and patient population

This retrospective analysis included all the patients older than 40 years, admitted to the University of New Mexico Hospital, a 580-bed University tertiary hospital, between 1 January 2011 and 1 May 2014 and carrying a diagnosis of COPD (ICD9 codes: 490, 491, 492 or 496), regardless of their primary admission diagnosis.[16]

This study was conducted in accordance with the amended Declaration of Helsinki. The University of New Mexico Health Science Center review board approved the protocol and waived the need for informed consent, protocol number: 14-312.

Study design and data collection

We used i2b2, a de-identified replica of our hospital medical records system that includes data on diagnoses, procedures, prescriptions, hospital admissions and laboratory results.[17]

Our data collection included the following: demographics, comorbidities included in Charlson's comorbidity index, presence of atopy, obesity, number of admissions, prescriptions for inhalers grouped as short acting beta-agonist, long-acting beta-agonist, anticholinergics, steroids and their combinations, prescriptions for oral steroids, beta-blockers and statins. We collected comorbidities according to previously validated methods.[18] I2b2 does not include pulmonary function tests. To capture the severity of disease, we included weight loss[19] and elevated plasma bicarbonate[20] among the variables collected. All the variables, including age (40–65 years and >65 years) and number of admissions (one admission and ⩾two admissions), were coded as binary for the analysis. The denominator for the number of admissions was the duration of the study.

Data analysis

Cluster analysis is a set of methodologies that group objects (e.g. patients) based on their characteristics. We used the sphere exclusion method with applications in cheminformatics, bioinformatics and pattern recognition.[21,22] It is a disjoint, similarity-based method; that is, a patient can belong

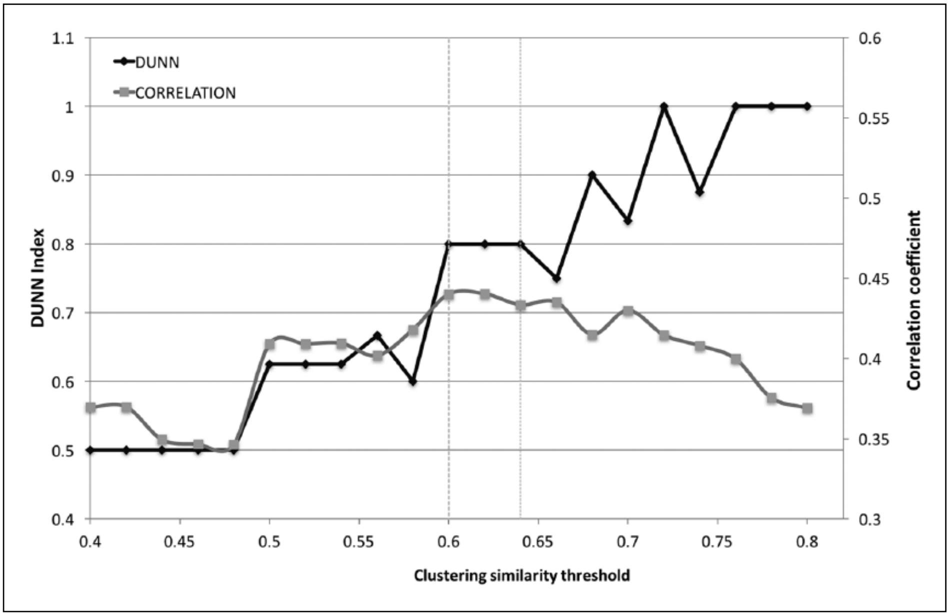


Figure 1. Dunn index and clustering coefficient against similarity.

to only one cluster, and the measure used for grouping is similarity. In a multidimensional space with as many dimensions as variables, a distance metric between individuals is dissimilarity, which is complementary to similarity (1: similarity). By definition, similarity can have a value between 0 if all the variables are different and 1 if they are equal.

In sphere exclusion, the only input needed from the analyst is a similarity threshold. The algorithm first computes the similarity between all individuals. It then chooses the individual with the most "neighbors" within the specified similarity cut-off. These individuals form the first cluster are excluded from further analysis. The process is then repeated iteratively until the only individuals left are singletons—individuals without neighbors.[21–23]

To choose the optimal similarity threshold, the clustering algorithm is run over a range of similarity thresholds without excluding individuals at any step. In this case, each subject can belong to more than one cluster. For this data set, we found the optimal balance between number of clusters and clustering overlap to be at similarity threshold of 0.62.[23] Plotting the Dunn index[24] and correlation coefficient[25] for different similarity thresholds provided consistent results (Figure 1).

All the collected variables were candidates for the clustering algorithm. Number of admissions was also included given the clinical relevance of the frequent exacerbation phenotype.[1]

After applying factor analysis to exclude inter-correlated variables, 24 common variables (out of 40 candidate variables) were selected with 10 latent variables, $p=0.54$ (H_0 model with 10 latent variables). The relevant variables were as follows: age, ICD9-CM codes 496 and 490, congestive heart failure, cerebrovascular disease, myocardial infarction, diabetes mellitus (DM) with complications, chronic kidney disease (CKD), obesity, depression, dementia, severe liver disease, plegias, rheumatologic disease, atopy, diagnosis of cancer, prescription for anticholinergic bronchodilators, prescription for fluticasone–salmeterol, prescription for albuterol–ipratropium, prescription for non-cardio selective beta-blocker, prescription for salmeterol, bicarbonate level >30 mEq/L, weight loss and number of admissions $\geqslant 2$ (Appendix 1).

Table 1. Demographics and general descriptors.

	All subjects (n = 3144)
Age less than 65 years	1436 (45.7%)
Male (%)	1636 (52%)
COPD ICD9 code	
ICD9 = 490	421 (13.4%)
ICD9 = 491	671 (21.3%)
ICD9 = 492	307 (9.8%)
ICD9 = 496	1745 (55.5%)
Number of comorbidities	2.4 ± 1.7
Charlson, median (IQR)	2 (1–4)
Admission ⩾2	1603 (51%)
Advanced disease	
Weight loss	413 (13.1%)
Bicarbonate > 30 mEq/L	300 (9.5%)

IQR: interquartile range; COPD: chronic obstructive pulmonary disease.

Analyses were computed using STATA/SE 13.1 (StataCorp LP, College Station, TX, USA) and MATLAB with the statistical toolbox installed (MATLAB version 8.3.0.532 (R2014a), Natick, MA, USA: The MathWorks Inc., 2014).

Results

A total of 3144 patients met the inclusion criteria. Of these, 1436 (46%) were older than 65 years, 1636 (52%) were males and 1745 (55.5%) had the ICD9 code 496. The median Charlson score was 2 (interquartile range (IQR): 1–4) With the most frequent comorbidities being depression (36%), congestive heart failure (25%), obesity (19%), cancer (19%) and mild liver disease (18%) (Table 1).

We obtained nine clusters with 189 patients remaining as outliers (Figure 2). The characteristics of each one of the clusters as compared to the rest are detailed below.

The largest cluster, cluster 1 contains 1748 patients and is characterized by a large proportion of patients older than 65 years and by depression. The patients have relatively few comorbidities, without a clear pattern and a Charlson score of 2 (IQR: 1–3).

We found two “malignancy” clusters: cluster 2 with 312 patients, few comorbidities and low number of readmissions and cluster 5 with 144 patients, signs of advanced disease and frequent readmissions.

We also identified two “cardiovascular” clusters: cluster 3 with 291 patients with a significant proportion of patients older than 65 years and predominantly coronary artery disease and congestive heart failure and Cluster 6, respectively, with 120 patients, higher proportion of patients younger than 65 years and predominantly cerebrovascular disease.

The remaining clusters were cluster 4 with 152 patients, most of them younger than 65 years, with few comorbidities, the highest number of prescriptions for bronchodilators and also with frequent readmissions. Cluster 7 includes 81 patients, the majority younger than 65 years, with asthma/atopy and many readmissions. Cluster 8 has 64 patients, younger than 65 years, who suffer from CKD or diabetes and with few readmissions. Cluster 9 with 41 patients includes patients with a high prevalence of signs of advanced disease and frequent readmissions. A classification rule

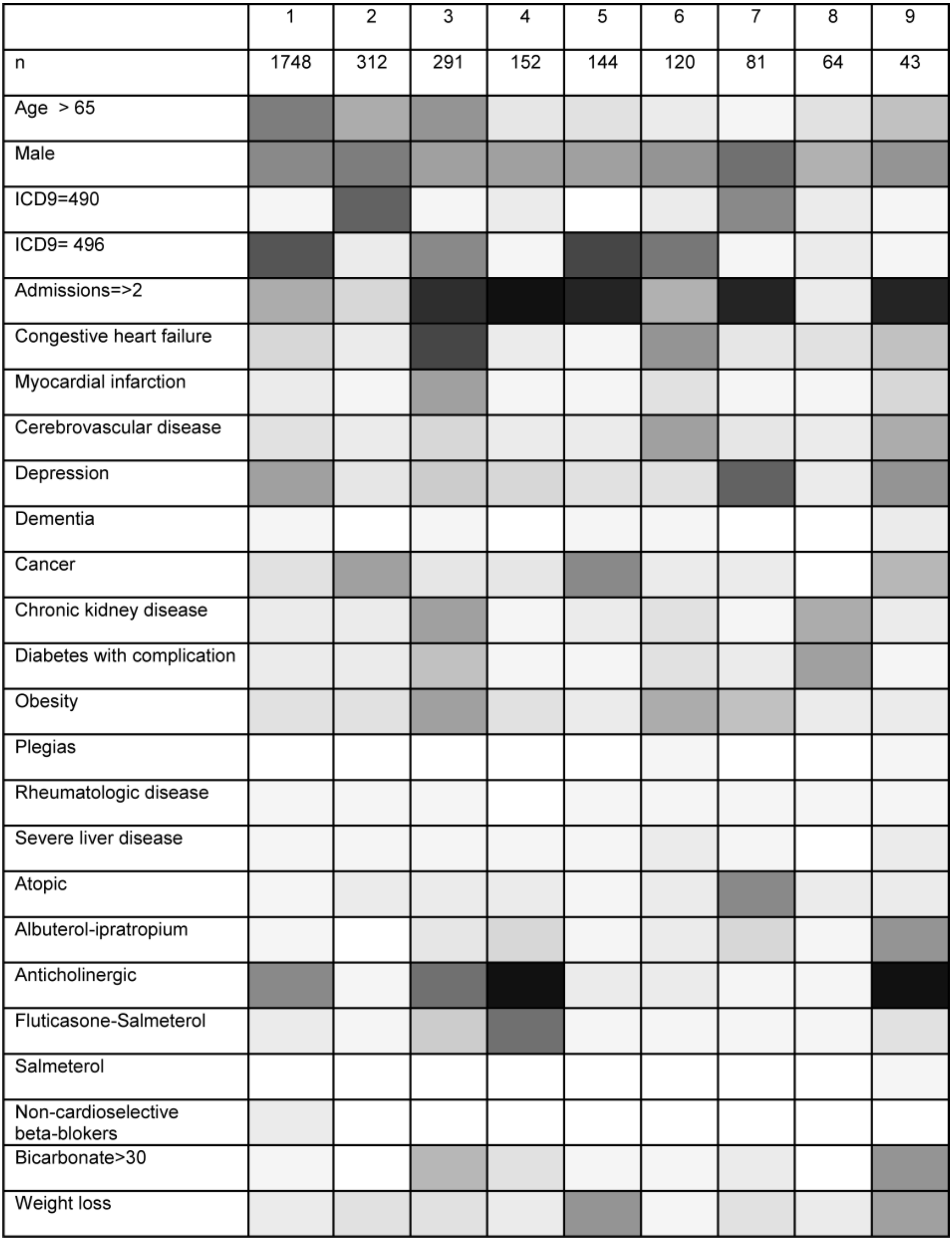


Figure 2. Grayscale heat map with the results of cluster analysis (white = 0%, black = 100%). Clusters are represented in the horizontal axis: 1: depression–COPD, 2: malignancy–COPD, 3: coronary artery disease–COPD, 4: young age–low comorbidity–high readmission–COPD, 5: advanced malignancy–COPD, 6: cerebrovascular disease–COPD, 7: atopy–COPD, 8: DM–CKD–COPD and 9: advanced disease–COPD.

consisting of a series of decisions that replicates the clustering algorithm can be found in Appendix 1 and Tables 3 and 4.

Discussion

In this study, we showed that patients with a diagnosis of COPD admitted to our hospital can be divided into clinically relevant phenotypes. Based on readily available data and using cluster analysis methodology, we obtained nine phenotypes, with only 6 percent of the patients as outliers. We also derived a classification rule for use in future validation and clinical practice. Most of our phenotypes confirm few previously known phenotypes obtained using different methodologies (Figure 2, Table 5). Given our sample size and more inclusive criteria, we observed new phenotype categories not previously described that may further stratify the COPD patient population.

The development of this classification scheme for patients with COPD can be used to generate new phenotype-specific outcomes and interventions.

The first and largest cluster describes the COPD-depression phenotype. Due to similar symptoms, regular screening tools that differentiate depression in COPD have limited validity.[26] In pulmonary practices, the prevalence of depression in COPD patients is estimated at 40 percent, and this increases with the severity of ventilatory obstruction as measured by the forced expiratory volume in first second (FEV1).[27] The relationship between depression and COPD is complex, as depression can further impact the social isolation, mobility impairment and quality of life in COPD patients. There are data suggesting that COPD precedes depression. In a prospective cohort study, the relative risk for developing depression 2 years after a new diagnosis of COPD was estimated at 2.21 (95% confidence interval: 1.64–2.97).[28] It is also known that there is a primary association between depressive symptoms and smoking and that depression severely limits the effectiveness of any smoking cessation intervention.[28] Understanding the COPD, depression phenotype could help develop COPD-specific depression screening tools and evaluate the effectiveness of preventive and therapeutic strategies.

Very relevant from a clinical perspective are the COPD—cardiovascular disease phenotypes. One is dominated by cerebrovascular disease, while the other by coronary artery disease. Both phenotypes have different secondary prevention strategies and therapeutic needs making the distinction clinically relevant.[29] The strong association between COPD and cardiovascular disease has been observed using different methodologies. In the Lung Health Study, a prospective cohort, more than 30 percent of the deaths were related to cardiovascular disease.[4] In terms of chronology, members of our study group described trajectories of disease in a population wide data registry with 6.2 million individuals. They found that all the trajectories starting with a diagnosis of atherosclerosis were followed by the diagnosis of COPD supporting a pathophysiological link and a temporal relationship.[30] Five other studies using either only comorbidities or more complex data sets have each identified at least one cluster characterized by the presence of cardiovascular disease.[6–8,10,11] This pattern has generated a great interest in discovering the underlying pathophysiology that leads to COPD after onset of atherosclerosis, and so far, the common link is attributed to inflammation.

Another phenotype previously described in the literature and confirmed in our cohort is the COPD–asthma overlap. Using a different set of variables, at least three studies that employed cluster analysis identified this phenotype.[6,12,31] COPD–asthma phenotype is of special interest as it highlights a subpopulation of patients usually excluded from therapeutic clinical trials, which have a poor quality of life and consume a disproportionate amount of healthcare resources.[1]

Our analysis also revealed five phenotypes, not previously described using cluster analysis, and we feel these phenotypes are very common and relevant in clinical practice. All of them highlight the disconnection between most COPD studies and the real-life COPD spectrum seen in hospital wards and outpatient clinics. One of the malignancy clusters incorporates patients with signs of more advanced disease (number 5), whereas cluster number 2 includes patients who rarely necessitate hospital readmission despite a diagnosis of cancer. The link between malignancy and COPD has been well established, expanding over lung cancers but also extra-pulmonary cancers.[3,4] In the Danish health registry, the overall risk for cancer was elevated in COPD, independent of comorbidities suggesting a different pathogenesis than, for example, cardiovascular disease.[32] Regarding the biological basis for these malignancy phenotypes, it has been noted that smokers overexpress repair genes and oncogenes and underexpress tumor suppression genes.[33] Although the expression of repair genes returns to normal several years after smoking cessation, oncogenes and tumor suppression genes continue to be altered decades later. To what extent the differential expression of these genes is deranged may determine which individuals develop cancer versus other comorbidities.[34] From a therapeutic perspective, COPD patients with malignancies tend to receive limited treatment. Grouping them into a more homogeneous phenotype and access to care in integrated clinics may lead to better outcomes. This model of healthcare delivery has been successfully used in other pulmonary diseases such as cystic fibrosis.[35]

A third phenotype of patients not previously reported in the literature is the chronic kidney disease–diabetes mellitus (CKD-DM)–COPD cluster. We note the low readmission rates of this phenotype. Interestingly, in previous studies, both CKD and DM have been associated with hospital readmissions in patients admitted with any diagnosis, but this does not hold true for COPD patients.[36] To which extent this characteristic is attributable to their model of care or to common underlying mechanism of disease is unknown. For example, DM targets the lung with reduction in carbon monoxide diffusion capacity, FEV1 and forced vital capacity (FVC) that show a dose response effect to fasting plasma glucose.[37] Furthermore, recently, a study reported improved asthma control in patients treated with thiazolidinediones.[38]

Another new phenotype described in our cohort is the "advanced COPD phenotype," represented by cluster 9. These patients who have more advanced disease based on the frequency of weight loss and elevated serum bicarbonate are also readmitted frequently.[19] One potential intervention would be initiation of end-of-life discussions especially if higher mortality is associated with this phenotype.

Finally, cluster number 4 can intuitively be labeled as "COPD Resistant to Treatment" or "COPD Non-Compliant."[39,40] Although these patients receive the highest number of prescriptions, they accrue very frequent readmissions. Since information on compliance was not available in our database, we were unable to differentiate between the two possible explanations.

Eight other studies have used similar clustering techniques to group COPD patients into phenotypes. They differ in the number and selection of individuals, the choice and number of variables and in the clustering techniques utilized. Choices at each of these levels can be expected to influence the results. This is reflected in Table 2 with each study describing phenotypes according to the variables selected for analysis.

Clustering techniques are most valuable when the phenotypes obtained describe the etiology, pathogenesis or clinical characteristics of the patients. Studies clustering detailed physiologic data and biomarkers with clinical characteristics may be better suited to investigate the pathogenesis and etiology of the different phenotypes, for example, Fens et al.,[9] chronic bronchitis and inflammatory eNose profile. Studies based on readily available data like ours[6,10] can also generate hypotheses about pathogenesis as detailed above, but more importantly, they can describe phenotypes outside of the research setting. They can help organize care and generate clinical research with phenotype-specific interventions.[41] For example, psycho-social interventions for the "Non-Compliant" phenotype or palliative care options for the advanced disease phenotype.

Table 2. Previous articles with >100 patients using cluster analysis and phenotypes as described by the authors.

	Number of patients	Population, country	Variables used for clustering	Clusters obtained
Baty et al.[6]	340,948	Population based Sweden	Comorbidities	1. Asthma/COPD 2. Anxiety/depression 3. Malignancy 4. Heart failure 5. Coronary artery disease
Burgel et al.[7]	322	Pulmonary units France	Age, smoking, airflow obstruction, exacerbations, BMI, QoL, anxiety, depression, MMRC	1. Young severe respiratory disease 2. Older, mild respiratory disease, mild comorbidities 3. Young, severe disease, mildly symptomatic 4. Older, severe respiratory disease, severe comorbidities
Burgel et al.[8]	527	COPD clinics + NELSON[a] Belgium, Netherlands	Comorbidities, COPD physiologic data	1. Low mortality, low comorbidities 2. Young, severe emphysema, low BMI, low comorbidities 3. Older, less severe disease, high BMI, comorbidities
Fens et al.[9]	157	NELSON[a] Netherlands	pbFEV1, FEV1 reversibility, chronic bronchitis, coronary artery disease, BMI, dyspnea at rest, packs year, LABA, eNose signature, emphysema score	1. Mild COPD, minimal symptoms, good QoL 2. Poor lung function, emphysema or chronic bronchitis and eNOSE 3. Emphysema, preserved lung function 4. Severe symptoms, mild lung disease
Vanfleteren et al.[10]	213	Pulmonary rehabilitation Netherlands	Comorbidities	1. Less comorbidities 2. Cardiovascular 3. Cachectic 4. Metabolic 5. Psychological
Garcia Aymerich et al.[11]	342	COPD admissions Spain	Symptoms, sputum microbiology, QoL, radiological emphysema score, PFT, nutrition, inflammation biomarkers total of 224 variables	1. Severe airflow limitation and poor respiratory performance 2. Mild airflow limitation 3. Mild airflow limitation and high BMI
Weatherall et al.[12]	175	Population based New Zealand	pbFEV1/FVC %, pbFEV1%, FEV1 change %, FRC%, DLCO%, IgE serum, mean FeNO, sputum production and smoking history	1. Severe airflow obstruction, low QoL, overlap of asthma, emphysema and chronic bronchitis 2. Emphysema 3. Asthma with eosinophilic airway inflammation 4. Mild airflow obstruction 5. Chronic bronchitis non-smokers
Renard et al.[13]	2164	Pulmonary clinics North America and Europe	Demographic Symptoms Biochemical Clinical/functional	1. Moderate—quasi stable 2. Functional emphysema 3. Mixed 4. Exacerbator emphysema 5. Inflamed comorbid

BMI: body mass index; QoL: quality of life; MMRC: Modified Medical Research Council dyspnea scale; COPD: chronic obstructive pulmonary disease; pb: pre-bronchodilator; FEV1: forced expiratory volume in first second; FVC: forced vital capacity; LABA: long-acting beta-agonist inhaler; FRC: functional residual capacity; DLCO: carbon monoxide lung diffusion capacity; FeNO: fraction of exhaled nitric oxide; PFT: pulmonary function test.

[a]NELSON: population-based cancer-screening trial in heavy smokers and ex-smokers.

Our results are similar to the studies that used only comorbidities for clustering.[6,10] The larger number of clusters found in our study can be explained by its sample size, our inclusion criteria and the variable selection which also included severity of disease markers and medications. Baty et al.[6] described a malignancy cluster; we were able to detail further, describing two malignancy clusters, one with signs of advanced disease and frequent readmissions and one without.

Another pre-requisite to make phenotyping useful is providing the classification rules for use in future validation and clinical practice[42] (Appendix 1, Tables 3 and 4).

Previous cohort studies and industry-sponsored randomized controlled trials have had stringent enrollment criteria excluding many clinically relevant and common COPD subpopulations. One strength of our study was including all adult patients (>40 years) with a diagnosis of COPD. As stated above, these lax criteria have helped us to recognize a larger number of phenotypes than previous studies. This is also the largest US COPD population is analyzed with this methodology.[13]

Another strength of our study is the use of the sphere exclusion method for clustering. It does not require a priori determination of the number of clusters, and using homogeneity measures to determine the optimal similarity threshold automatically (Figure 1) makes the clustering algorithm autonomous and minimizes the risk of bias.

Our study has several limitations. First, the proposed clusters rely on the available variables.[21] We lacked data on spirometry, 6-min walk test, mortality and quality of life questionnaires. In their absence, our ability to detect some clinically relevant phenotypes may have been limited; for example, the previously described upper lobe–predominant emphysema with poor exercise performance was not observed.[43] We also had to use alternative measures of severity of disease (sodium bicarbonate and weight loss), which may not apply to all the phenotypes. However, under normal clinical settings, such information is rarely available. In a recent report, only 32 percent of newly diagnosed COPD patients had undergone spirometry testing,[44] and thus, relying on these data to clinically classify patients beyond the pulmonary office is, as of today's practice, of limited value. Furthermore, even in the absence of these variables, we were able to detect clinically relevant phenotypes with plausibly different underlying pathogenesis. Our methodology can only describe co-occurrence of diseases and not cause effect relationships.

Another limitation of the study is single center study design. People from New Mexico may have different distribution of risk factors than patients in other areas of the United States. They have a higher exposure to biomass fuels, and the Hispanic population has also been found to be particularly resistant to the development of this disease.[45,46]

The main impact of our study is the development of a classification scheme for patients with COPD that can be used to generate new phenotype-specific outcomes and interventions. While other studies have mainly focused on mortality and readmission rates, it is debatable whether these are the most or only relevant outcomes. For instance, two phenotypes can have the same mortality and hospitalization rate for very different reasons thus requiring different therapeutic approaches. Phenotype-specific outcomes such as time to myocardial infarction, depression recurrence or improved compliance along with mortality and readmission rates may offer an extra dimension to better differentiate between and within subpopulations of COPD patients.

Conclusion

Moving from generalized to stratified medicine, clustering studies are needed to both discover pathophysiology links and group patients with clinically similar phenotypes allowing for more personalized and integrated care. Our study confirms previous COPD phenotypes and adds new ones to better understand the interplay between COPD and comorbidities. The use of readily available data in defining our clusters makes this methodology appealing for validation and implementation at other centers.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study is supported by The University of New Mexico School of Medicine.

References

1. GOLD—the Global Initiative for Chronic Obstructive Lung Disease, http://goldcopd.org/gold-reports/ (accessed 4 November 2016).
2. CDC. Chronic Obstructive Pulmonary Disease (COPD)—data and statistics, http://www.cdc.gov/copd/data.htm (accessed 28 March 2015).
3. McGarvey LP, John M, Anderson JA, et al. Ascertainment of cause-specific mortality in COPD: operations of the TORCH Clinical Endpoint Committee. *Thorax* 2007; 62: 411–415.
4. Anthonisen NR, Connett JE, Enright PL, et al. Hospitalizations and mortality in the Lung Health Study. *Am J Respir Crit Care Med* 2002; 166: 333–339.
5. Cully JA, Graham DP, Stanley MA, et al. Quality of life in patients with chronic obstructive pulmonary disease and comorbid anxiety or depression. *Psychosomatics* 2006; 47: 312–319.
6. Baty F, Putora PM, Isenring B, et al. Comorbidities and burden of COPD: a population based case-control study. *PLoS ONE* 2013; 8: e63285.
7. Burgel P-R, Paillasseur J-L, Peene B, et al. Two distinct chronic obstructive pulmonary disease (COPD) phenotypes are associated with high risk of mortality. *PLoS ONE* 2012; 7: e51048.
8. Burgel P-R, Paillasseur J-L, Caillaud D, et al. Clinical COPD phenotypes: a novel approach using principal component and cluster analyses. *Eur Respir J* 2010; 36: 531–539.
9. Fens N, van Rossum AGJ, Zanen P, et al. Subphenotypes of mild-to-moderate COPD by factor and cluster analysis of pulmonary function, CT imaging and breathomics in a population-based survey. *COPD* 2013; 10: 277–285.
10. Vanfleteren LE, Spruit MA, Groenen M, et al. Clusters of comorbidities based on validated objective measurements and systemic inflammation in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2013; 187: 728–735.
11. Garcia-Aymerich J, Gómez FP, Benet M, et al. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. *Thorax* 2011; 66: 430–437.
12. Weatherall M, Travers J, Shirtcliffe PM, et al. Distinct clinical phenotypes of airways disease defined by cluster analysis. *Eur Respir J* 2009; 34: 812–818.
13. Rennard SI, Locantore N, Delafont B, et al. Identification of five chronic obstructive pulmonary disease subgroups with different prognoses in the ECLIPSE cohort using cluster analysis. *Ann Am Thorac Soc* 2015; 12: 303–312.
14. Han MK, Agusti A, Calverley PM, et al. Chronic obstructive pulmonary disease phenotypes: the future of COPD. *Am J Respir Crit Care Med* 2010; 182: 598–604.
15. Hingorani AD, Windt DA, Riley RD, et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. *BMJ* 2013; 346: e5793.
16. Prieto-Centurion V, Rolle AJ, Au DH, et al. Multicenter study comparing case definitions used to identify patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2014; 190: 989–995.
17. CTSC Biomedical Informatics, http://hsc.unm.edu/research/ctsc/BMI/i2b2.shtml (accessed 29 March 2015).
18. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005; 43: 1130–1139.

19. Prescott E, Almdal T, Mikkelsen KL, et al. Prognostic value of weight change in chronic obstructive pulmonary disease: results from the Copenhagen City Heart Study. *Eur Respir J* 2002; 20: 539–544.
20. Groenewegen KH, Schols AM and Wouters EFM. Mortality and mortality-related factors after hospitalization for acute exacerbation of COPD. *Chest* 2003; 124: 459–467.
21. Everitt BS, Landau S, Leese M, et al. Index. In: *Cluster analysis*. New York: John Wiley & Sons, Inc., 2011, pp. 321–330.
22. Taylor R. Simulation analysis of experimental design strategies for screening random compounds as potential new drugs and agrochemicals. *J Chem Inf Comput Sci* 1995; 35: 59–67.
23. MacCuish J, Nicolaou C and MacCuish NE. Ties in proximity and clustering compounds. *J Chem Inf Comput Sci* 2001; 41: 134–146.
24. Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J Cybern* 1973; 3: 32–57.
25. Farris JS. On the cophenetic correlation coefficient. *Syst Biol* 1969; 18: 279–285.
26. Wilson I. Depression in the patient with COPD. *Int J Chron Obstruct Pulmon Dis* 2006; 1: 61–64.
27. Polsky D, Doshi JA, Marcus S, et al. Long-term risk for depressive symptoms after a medical diagnosis. *Arch Intern Med* 2005; 165: 1260–1266.
28. Cinciripini PM, Wetter DW, Fouladi RT, et al. The effects of depressed mood on smoking cessation: mediation by postcessation self-efficacy. *J Consult Clin Psychol* 2003; 71: 292–301.
29. Kernan WN, Ovbiagele B, Black HR, et al. Guidelines for the prevention of stroke in patients with stroke and transient ischemic attack: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 2014; 45: 2160–2236.
30. Jensen AB, Moseley PL, Oprea TI, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat Commun* 2014; 5: 4022.
31. Wardlaw AJ, Silverman M, Siva R, et al. Multi-dimensional phenotyping: towards a new taxonomy for airway disease. *Clin Exp Allergy* 2005; 35: 1254–1262.
32. Kornum JB, Sværke C, Thomsen RW, et al. Chronic obstructive pulmonary disease and cancer risk: a Danish nationwide cohort study. *Respir Med* 2012; 106: 845–852.
33. Rutgers SR, Postma DS, ten Hacken NH, et al. Ongoing airway inflammation in patients with COPD who do not currently smoke. *Chest* 2000; 117: 262S.
34. Brody JS and Spira A. State of the art. Chronic obstructive pulmonary disease, inflammation, and lung cancer. *Proc Am Thorac Soc* 2006; 3: 535–537.
35. Lobo J, Rojas-Balcazar JM and Noone PG. Recent advances in cystic fibrosis. *Clin Chest Med* 2012; 33: 307–328.
36. Bahadori K and FitzGerald JM. Risk factors of hospitalization and readmission of patients with COPD exacerbation—systematic review. *Int J Chron Obstruct Pulmon Dis* 2007; 2: 241–251.
37. Walter RE, Beiser A, Givelber RJ, et al. Association between glycemic state and lung function: the Framingham Heart Study. *Am J Respir Crit Care Med* 2003; 167: 911–916.
38. Rinne ST, Feemster LC, Collins BF, et al. Thiazolidinediones and the risk of asthma exacerbation among patients with diabetes: a cohort study. *Allergy Asthma Clin Immunol Off J Can Soc Allergy Clin Immunol* 2014; 10: 34.
39. Vestbo J, Anderson JA, Calverley PMA, et al. Adherence to inhaled therapy, mortality and hospital admission in COPD. *Thorax* 2009; 64: 939–943.
40. Chung KF. New treatments for severe treatment-resistant asthma: targeting the right patient. *Lancet Respir Med* 2013; 1: 639–652.
41. Schizophrenia IPS of Organization WHO. Report of the International Pilot Study of Schizophrenia, http://apps.who.int//iris/handle/10665/39405 (1973, accessed 28 March 2015).
42. Weatherall M, Shirtcliffe P, Travers J, et al. Use of cluster analysis to define COPD phenotypes. *Eur Respir J* 2010; 36: 472–474.
43. Fishman A, Martinez F, Naunheim K, et al. A randomized trial comparing lung-volume-reduction surgery with medical therapy for severe emphysema. *N Engl J Med* 2003; 348: 2059–2073.
44. Han MK, Kim MG, Mardon R, et al. Spirometry utilization for COPD: how do we measure up? *Chest* 2007; 132: 403–409.

45. Sood A, Petersen H, Blanchette CM, et al. Wood smoke exposure and gene promoter methylation are associated with increased risk for COPD in smokers. *Am J Respir Crit Care Med* 2010; 182: 1098–1104.

46. Ramírez-Venegas A, Sansores RH, Quintana-Carrillo RH, et al. FEV1 decline in patients with chronic obstructive pulmonary disease associated with biomass exposure. *Am J Respir Crit Care Med* 2014; 190: 996–1002.

Appendix 1

Factor analysis

Factor analysis was carried out using factoran function in MATLAB version 2014a.

H_0 (null hypothesis) was tested against alternative hypothesis that more than N latent variables are required to explain variance of the whole data set.

Number of latent variables (common factors) was selected by iteratively setting the number of latent variables parameters to factoran until H_0 was not rejected, 10 latent variables are required to explain the variance of the data set in our analysis. Observed variables from initial data set were eliminated based on observed specific variance, and observed variables with specific variance close to 0 were eliminated based on the fact that these would be almost entirely determined by latent variables which are linear combinations of the observed variables, see list below.

Observed variables (40) list is as follows:

Age > 65
Gender
History of cancer
Congestive heart failure
Chronic kidney disease
Cerebrovascular disease
Dementia
Depression
Diabetes with complications
Drug user
Human immunodeficiency virus infection
Mild liver disease
Metastatic cancer
Obesity
Alcoholism
Plegias
Peptic ulcer disease
Rheumatologic disease
Cirrhosis
Gout
History of myocardial infarction
Weight loss

Atopy
Treatment with albuterol–ipratropium
Treatment with albuterol
Treatment with long-acting antimuscarinic antagonists
Treatment with budesonide–formoterol
Treatment with fluticasone–salmeterol
Treatment with salmeterol
Treatment with statins
Treatment with steroids
Treatment with non-cardio selective beta-blockers
Treatment with cardio-selective beta-blockers
Plasma bicarbonate greater than 30
Number of admissions
ICD9 490
ICD9 491
ICD9 492
ICD9 496
COPD other than emphysema

Observed variables selected with specific variance > 0 (24), these variables form the following 10 latent variables:

Atopy
Age > 65 years
Treatment with albuterol–ipratropium
Treatment with long-acting anticholinergics
Plasma bicarbonate > 30 mEq/mL
History of cancer
Congestive heart failure
Chronic kidney disease
ICD9 490
ICD9 496
Cerebrovascular disease
Dementia
Depression
Diabetes mellitus with complications
Treatment with fluticasone–salmeterol
History of myocardial infarction
Treatment with non-cardio-selective beta-blockers

Obesity
Plegias
History of rheumatological disease
Treatment with salmeterol
Cirrhosis
Number of admissions
Weight loss

Sensitivity analysis

We used factor analysis to eliminative variables not contributing to latent variables (0 specific variance); this means that any perturbation of any remaining variables will lead to different clustering results. Scoring rules (1–9) are derived from sensitivity analysis where we analyzed systematically how observed variables change across clusters. Each of the equations has to be calculated sequentially; the variables are scored as 1 if present/true and 0 if not present/false. The procedure stops once the result of one of the equations is ⩾0.32, and the patient is allocated to that cluster (e.g. if score ⩾0.32, then patient is allocated to cluster 3):

Equation 1 = (Age > 65 + depression + anticholinergic bronchodilator + admission ⩾2 + ICD9: 496)/[(Σ all variables) + 5 − (age > 65 + depression + anticholinergic bronchodilator + admission ⩾2 + ICD9 = 496)]

Equation 2 = (Age > 65 + malignancy + admission ⩾2 + ICD9: 490)/[(Σ all variables) + 4 − (age > 65 + malignancy + admission ⩾2 + ICD9 = 490)]

Equation 3 = (Age > 65 + CHF + CKD + obesity + anticholinergic bronchodilator + bicarbonate > 30 + myocardial infarction + ICD9: 496)/[(Σ all variables) + 8 − (age > 65 + CHF + CKD + obesity + anticholinergic bronchodilators + bicarbonate > 30 + myocardial infarction + ICD9: 496)]

Equation 4 = (Anticholinergic bronchodilator + fluticasone–salmeterol)/[(Σ all variables) + 2 − (anticholinergic bronchodilators + fluticasone–salmeterol)]

Equation 5 = (Cancer + weight loss + ICD9:496)/[(Σ all variables) + 3 − (cancer + weight loss + ICD9:496)]

Equation 6 = (CH + CVD + obesity + non-cardioselective beta blockers + admissions ⩾2 + ICD9: 496)/[(Σ all variables) + 6 − (CH + CVD + obesity + non-cardioselective beta blockers + admissions ⩾2 + ICD9: 496)

Equation 7 = (Depression + anticholinergic bronchodilators + ICD9:490 + atopy)/[(Σ all variables) + 4 − (depression + anticholinergic bronchodilators + ICD9: 496 + atopy)]

Equation 8 = (CKD + DM + admissions ⩾2)/[(Σ all variables) + 3 − (CKD + DM + admissions ⩾2)

Equation 9 = (Malignancy + CVD + depression + weight loss + albuterol–ipratropium + anticholinergic bronchodilators + bicarbonate > 30)/[(Σ all variables) + 7 − (malignancy + CVD + depression + weight loss + albuterol–ipratropium + anticholinergic bronchodilators + bicarbonate > 30)

Table 3. Comparison of rule assignments with clusters from cluster analysis values are the number of patients from each cluster where score ⩾0.32.

Total	Cluster ID	1	2	3	4	5	6	7	8	9
1748	1	99.66%	0%	0%	0%	0%	0%	0%	0%	0%
312	2	0%	100%	0%	0%	0%	0%	0%	0%	0%
291	3	0%	0%	100%	0%	0%	0%	0%	0%	0%
152	4	0%	0%	0%	100%	0%	0%	0%	0%	0%
144	5	0%	0%	0%	0%	100%	0%	0%	0%	0%
120	6	0%	0%	0%	0%	0%	97.50%	0%	0%	0%
81	7	0%	0%	0%	0%	0%	0%	100%	0%	0%
64	8	0%	0%	0%	0%	0%	0%	0%	100%	0%
43	9	0%	0%	0%	0%	0%	0%	0%	0%	100%
189	Outlier	0%	0%	0%	0%	0%	0%	0%	0%	0%

The first column contains the number of patients assigned to each cluster using the sphere exclusion method. To the right a matrix with the percent of correct patients assignments using the derived classification rule sequentially.

Table 4. Sensitivity analysis.

Total	Cluster ID	1	2	3	4	5	6	7	8	9
1748	1	99.66%	22.43%	27.69%	8.81%	9.44%	23.28%	4.81%	3.49%	7.95%
312	2	0%	100%	5.45%	0.64%	15.38%	4.49%	21.79%	9.29%	2.24%
291	3	0%	0%	100%	7.56%	4.47%	24.05%	1.72%	5.84%	11.68%
152	4	0%	0%	0%	100%	2.63%	0%	9.21%	1.32%	14.47%
144	5	0%	0%	0%	0%	100%	6.94%	0%	1.39%	17.36%
120	6	0%	0%	0%	0%	0%	97.50%	2.50%	11.67%	3.33%
81	7	0%	0%	0%	0%	0%	0%	58.02%	0%	28.40%
64	8	0%	0%	0%	0%	0%	0%	4.69%	100%	0%
43	9	0%	0%	0%	0%	0%	0%	4.65%	0%	100%
189	Outlier	0%	0%	0%	0%	0%	0%	8.99%	0%	0%

The first column contains the number of patients assigned to each cluster using the sphere exclusion method. To the right a matrix with percentage of patients assigned to each cluster when the derived classification rule is not applied sequentially.

Table 5. Cluster analysis results.

	1	2	3	4	5	6	7	8	9
n	1748	312	291	152	144	120	81	64	43
Age > 65 years	1041 (59.6%)	118 (37.8%)	152 (52.2%)	19 (12.5%)	21 (14.6%)	13 (10.83%)	3 (3.7%)	9 (14.1%)	13 (30.2%)
Male (%)	927 (53%)	187 (59.9%)	131 (45%)	67 (44.1%)	68 (47.2%)	62 (51.7%)	54 (66.7%)	24 (37.5%)	21 (48.8%)
ICD9 = 490	98 (5.6%)	208 (66.7%)	11 (3.8%)	12 (7.9%)	0	10 (8.3%)	43 (53.1%)	7 (10.9%)	2 (4.7%)
ICD9 = 496	1288 (73.7%)	32 (10.3%)	154 (52.9%)	10 (6.6%)	118 (81.9%)	75 (62.5%)	3 (3.7%)	7 (10.9%)	2 (4.7%)
Admissions ⩾2	681 (39%)	70 (22.4%)	255 (87.6%)	145 (95.4%)	128 (88.9%)	45 (37.5%)	73 (90.1%)	5 (7.8%)	38 (88.4%)
Charlson, median (IQR)	2 (1–3)	3 (1–5)	3 (4–6)	2 (1–3)	3 (2–6)	3 (2–4)	2 (1–3)	2 (1–4.5)	3 (2–5)
Congestive heart failure (%)	376 (21.5%)	35 (11.2%)	232 (79.7%)	16 (10.5%)	9 (6.3%)	57 (47.5%)	10 (12.4%)	9 (14.1%)	12 (27.9%)
Cancer (%)	259 (14.8%)	138 (44.23%)	38 (13.1%)	19 (12.5%)	79 (54.9%)	8 (6.7%)	8 (9.9%)	1 (1.6%)	14 (32.6%)
Chronic kidney disease (%)	185 (10.6%)	30 (9.6%)	135 (46.4%)	5 (3.3%)	11 (7.6%)	16 (13.3%)	3 (3.7%)	26 (40.6%)	4 (9.3%)
Cerebrovascular disease (%)	258 (14.8%)	31 (9.9%)	64 (22%)	12 (7.9%)	15 (10.4%)	56 (46.7%)	10 (12.4%)	6 (9.4%)	18 (41.9%)
Depression (%)	821 (47%)	38 (12.2%)	71 (24.4)	30 (19.4%)	22 (15.3%)	19 (15.8%)	58 (71.6%)	7 (10.9%)	21 (48.8%)
Dementia (%)	57 (3.3%)	3 (1%)	15 (5.2%)	2 (1.3%)	4 (2.8%)	5 (4.2%)	1 (1.2%)	1 (1.6%)	4 (9.3%)
Diabetes with complication (%)	140 (8%)	26 (8.3%)	84 (28.9%)	8 (5.3%)	7 (4.9%)	19 (15.8%)	10 (12.4%)	28 (43.8%)	2 (4.7%)
Obesity (%)	245 (14%)	54 (17.3%)	126 (43.3%)	23 (15.1%)	11 (7.6%)	50 (41.7%)	26 (32.1%)	6 (9.4%)	4 (9.3%)
Plegias (%)	31 (1.8%)	3 (1%)	4 (1.4%)	2 (1.3%)	3 (2%)	5 (4.2%)	1 (1.2%)	0	2 (4.7%)
Rheumatologic disease (%)	86 (4.9%)	19 (6.1%)	20 (6.9%)	2 (1.3%)	10 (6.9%)	7 (5.8%)	2 (2.5%)	2 (3.1%)	2 (4.7%)
Severe liver disease (%)	51 (2.9%)	10 (3.2%)	18 (6.2%)	8 (5.3%)	9 (6.3%)	13 (10.8%)	2 (2.5%)	0	4 (9.3%)
Myocardial infarction (%)	198 (11.3%)	19 (6.1%)	136 (46.7%)	7 (4.65)	5 (3.5%)	18 (15%)	2 (2.5%)	3 (4.7%)	8 (18.6%)
Atopic (%)	108 (6.2%)	37 (11.9%)	27 (9.3%)	16 (10.5%)	7 (4.9%)	11 (9.2%)	40 (49.4%)	7 (10.9%)	4 (9.3%)
Prescriptions									
Albuterol–ipratropium (%)	120 (6.9%)	1 (0.3%)	39 (13.4%)	32 (21.1%)	5 (3.5%)	11 (9.2%)	15 (18.5%)	3 (4.7%)	22 (51.2%)
Anticholinergic bronchodilator (%)	918 (52.5%)	23 (7.4%)	186 (63.9%)	147 (96.7%)	14 (9.7%)	11 (9.2%)	57 (7.4%)	4 (6.3%)	40 (93%)
Fluticasone–salmeterol (%)	204 (11.7%)	8 (2.6%)	72 (24.7%)	98 (64.5%)	6 (4.2%)	7 (5.8%)	5 (6.2%)	2 (3.1%)	7 (16.3%)
Salmeterol (%)	6 (0.3%)	0	1 (0.3%)	0	1 (0.7%)	2 (1.7%)	0	0	2 (4.7%)
Non-cardioselective beta-blockers	169 (9.7%)	16 (5.1%)	54 (18.6%)	6 (4%)	10 (6.9%)	55 (45.8%)	11 (13.6%)	5 (7.8%)	9 (20.9%)
Advanced disease									
Bicarbonate > 30 (%)	127 (7.3%)	5 (1.6%)	100 (34.4%)	22 (14.5%)	5 (3.5%)	4 (3.3%)	8 (9.9%)	1 (1.6%)	22 (51.2%)
Weight loss (%)	171 (9.8%)	40 (12.8%)	40 (13.8%)	16 (10.5%)	71 (49.3%)	6 (5%)	11 (13.6%)	5 (7.8%)	20 (46.5%)

Health Informatics Journal

Original Article

Health Informatics Journal
2018, Vol. 24(4) 410–418
© The Author(s) 2016
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1460458216677841
journals.sagepub.com/home/jhi
SAGE

Elderly fall risk prediction based on a physiological profile approach using artificial neural networks

Jafar Razmara
University of Tabriz, Iran

Mohammad Hassan Zaboli
University of Tabriz, Iran

Hadi Hassankhani
Tabriz University of Medical Sciences, Iran

Abstract

Falls play a critical role in older people's life as it is an important source of morbidity and mortality in elders. In this article, elders fall risk is predicted based on a physiological profile approach using a multilayer neural network with back-propagation learning algorithm. The personal physiological profile of 200 elders was collected through a questionnaire and used as the experimental data for learning and testing the neural network. The profile contains a series of simple factors putting elders at risk for falls such as vision abilities, muscle forces, and some other daily activities and grouped into two sets: psychological factors and public factors. The experimental data were investigated to select factors with high impact using principal component analysis. The experimental results show an accuracy of ≈90 percent and ≈87.5 percent for fall prediction among the psychological and public factors, respectively. Furthermore, combining these two datasets yield an accuracy of ≈91 percent that is better than the accuracy of single datasets. The proposed method suggests a set of valid and reliable measurements that can be employed in a range of health care systems and physical therapy to distinguish people who are at risk for falls.

Keywords

elderly, fall risk prediction, neural networks, physiological profile

Introduction

Due to rapidly growing elderly population, today, societies are threatened to provide their life expectancy such as public health, medical, and social services. Based on World Health Organization estimation, by 2050, more than 2 billion people will be over 60 years, 80 percent of which will be

Corresponding author:
Jafar Razmara, Department of Computer Science, University of Tabriz, Tabriz 51666, Iran.
Email: razmara@tabrizu.ac.ir

from developing countries.[1] This growth leads governments to urgently provide technical facilities for health care requirements of elderly.

Falls among older people remain as a major problem in public health care issues. Among elderly living in the community, one in three people likely to fall at least one time in a year, and this falling rate is even increasing in older adults living in nursing home cares.[2,3] Falling is important in older people life due to making several mental, economical, and social issues. It is the most frequent reason for accidental death[4,5] and the third reason for inability[6] among elders. Fear of falling for the second time is caused by falling and makes limitations in the elder's daily activities.[7] Therefore, researchers are highly interested to find technical solutions to prevent falling events in elders.

The primary step to prevent falling events is determining elders potentially subjected to this risk. Many researches have been done to propose efficient tools for identifying people with high probability of falling risk.[8–10] The methods work based on different schemes varying from simple clinical tests to complex ones.[11] The attempts were to identify factors that highly affect the fall risk. Among these factors, older people (more than 65 years old) having multiple chronic illness has the highest risk of falls, and stroke, Parkinson disease, a history of falls, the presence of impaired gait, muscle weakness, arthritis, foot problems, impaired cognition, abnormal neurological signs, and the taking of psychoactive medications and multiple medications are the critical factors in fall prediction.[2] However, it is important to measure the degree of risk among these factors. A comprehensive review on the proposed tools for fall prediction has been presented by Oliver et al.[12] and Hassankhani et al.[13]

In recent years, machine learning-based methods have been widely used in diagnosis and prediction applications. These methods are also used in fall risk prediction where the system needs to learn from previous experiences in the prediction phase. In 2012, Marschollek et al.[14] proposed a classification tree model using the C4.5 algorithm as well as a logistic regression model and evaluated their predictive performance. In 2013, Rose[15] proposed an ensemble machine learning approach that combines multiple algorithms into a single algorithm and returns a prediction function with the best cross-validated mean squared error. Artificial neural networks (ANN) have received an increasing interest over the recent years, and its superior performance in different machine learning applications as diverse as engineering, medicine, finance, and many other areas have established it as an accepted model for a wide variety of scientific problems. They constitute a set of models inspired by biological neural systems that we call brain. The structure of ANNs generally is constructed from a set of neurons which exchange signals with others via an interconnected network. Each connection has a numeric weight that can be adjusted during training of the network, making the network adaptive to input patterns and capable of learning. ANNs have remarkable capabilities to solve a wide variety of tasks that are difficult to solve by ordinary rule-based algorithms.

In this article, we describe the development of a neural network to predict elderly fall risk based on their physiological profile. In section "ANNs," a summary on the structure of neural networks and their learning procedure is described. In section "PCA," the principal component analysis (PCA) approach for dimensionality reduction is introduced. In section "Materials and methods," methods and materials of the study are explained. In section "Results and discussion," results of the experimental study are discussed. In section "Conclusion," the conclusion of this study is drawn.

ANNs

ANNs propose a methodology for extracting knowledge from raw data.[16] Feedforward neural network is a typical type of ANNs constructed from at least three layers of artificial neurons (Figure 1) including input layer, output layer, and one or more hidden layers. The neurons in input

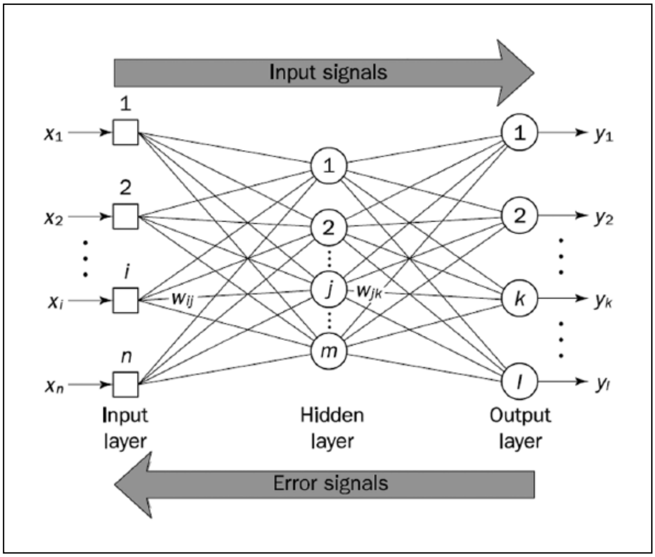


Figure 1. A three-layer feedforward neural network.

and output layers represent variables of predictor and predictand. The hidden layer is arranged between input and output layers including a number of neurons that is determined empirically to obtain an optimized performance for a particular problem.[16] The neurons in the hidden and output layers get their input from their previous layer via weighted connections and feed their output to the next layer. A weighted sum is calculated in each neuron of hidden layer via the following activation formula

$$a_i = \sum_{j=1}^{M} w_{ij} x_j + \theta_i$$

where θ is a certain threshold to fire a neuron. The output of each neuron is computed as a function of the activation value by

$$y = f\left(a_i\right)$$

that is a predefined function such as binary or linear threshold, sigmoid, hyperbolic tan, and Gaussian. A similar computation is done in the output layer neurons to produce output of network.

In the training phase of the above neural network, back-propagation learning algorithm is mostly used to adjust weights of synaptic connections. For each training sample, input signals are propagated through the network, and output of the network is produced. The error signals are calculated at the output layer and propagated backward to the hidden layer at iteration p via the formula

$$e_k\left(p\right) = y_{d,k}\left(p\right) - y_k\left(p\right)$$

where $y_{d,k}(p)$ is the desired output of the neuron k. The synaptic weights are updated using the formula

$$w_{jk}\left(p+1\right)=w_{jk}\left(p\right)+\Delta w_{jk}\left(p\right)$$

where $\Delta w_{jk}(p)$ is the correction value and calculated by

$$\Delta w_{jk}\left(p\right)=\alpha\times y_j\left(p\right)\times\delta_k\left(p\right)$$

where α is the learning rate and $\delta_k(p)$ is the error at neuron k. The network training phase runs iteratively until the defined error criterion is satisfied.

The standard method to develop a prediction system based on ANNs is training the network with a set of known samples according to the above procedure. After training, the network can be used to predict output for an unknown sample.

PCA

In machine learning problems, when input patterns are multidimensional, the use of all features may increase computational complexity and decrease the generalization ability of the model. To overcome these problems, one should look for the subset of features which decrease the learning ability of the model and increase the complexity of the algorithm and then remove this subset from feature set. Generally, these solutions fall in the dimensionality reduction methods and are widely used in pattern recognition and classification algorithms.

PCA[17,18] is a well-known statistical technique used for dimension reduction. The technique transforms a set of correlated variables into a set of uncorrelated variables by mapping data into its eigenspace. PCA chooses top K eigenvectors to reflect directions with maximum variability. The major advantage of PCA is the ability to determine degree of similarities among data.

To conduct PCA for a given set of d-dimensional input samples X, m principal axes T_1, T_2, …, T_m $(1 \leqslant m \leqslant d)$ are defined as orthonormal with a maximum preserved variance in the new space. In general, matrix T is created by the m prominent eigenvectors from the covariance matrix of samples

$$S=\frac{1}{N}\sum_{i=1}^{N}\left(x_i-\mu\right)^T\left(x_i-\mu\right)$$

where $x_i \in X$ and μ is the mean of samples. Therefore, we have

$$ST_i=l_iT_i$$

for $i=1, 2, \ldots, m$, where l_i is the ith largest eigenvalue of S. For a given input vector x, the m principal components are calculated by

$$y=\left[y_1,y_2,\ldots,y_m\right]=\left[T_1^T,T_2^T,\ldots,T_m^T\right]=T^Tx$$

The m principal components of x is used in the new space instead of the input vector.

Materials and methods

Dataset

Data used in this study were collected through a questionnaire designed by a group of experts in old adults. The questionnaires were filled out by 200 old members of the Tabriz pension center

Table 1. Physiological profile used in questionnaire.

	Psychological factors	Public factors
1	Cleaning the house	Age
2	Create a simple meal	Sex
3	Go to the store	Education
4	Go up or down the stairs	Employment
5	Getting hands above head	Living conditions
6	Walking on slippery surfaces	Life independency
7	Walking in a crowded places	Fear of falling
8	Go up or down the slope	Use of drug
9	Dressing or undressing	Needs to help for movement
10	Bathing or showering	Financial condition
11	Sit down or get up from the chair	Chronic diseases
12	Walking	Cardiovascular diseases
13	Answering phone	Muscular diseases
14	Go to visit friends or family	Respiratory diseases
15	Walking on an uneven surface	Metabolic diseases
16	Going out to gathering	Neurological diseases

having age more than 60 years. The questionnaire contains a series of simple factors putting elders at risk for falls and grouped into two sets: psychological factors and public factors. Table 1 shows the items in these two sets. For each item, four choices were considered indicating the degree for solicitous where score 1 shows the lowest degree and score 4 shows the highest degree.

The neural network for fall prediction

A feedforward neural network with back-propagation learning algorithm was used to predict falling risk based on the personal profile of older adults. The number of neurons in the input layer is decided by the number of input features for each sample. To determine the number of neurons in the hidden layer, a set of experiments was arranged using different numbers of neurons ranging from 10 to 50 as represented in Figure 2(a) for psychological factors and Figure 2(b) for public factors. As can be seen from this figure, the network obtains the highest accuracy with 20 neurons in the hidden layer. The output layer contains one neuron representing the falling risk with a value between 0 and 1 where a higher value indicates a higher risk. The input data were normalized such that the features are zero-mean and unit variance. For adaption of the network weights in the training phase, training samples were given to network expecting an output of 0 (no fall containing) and 1 (fall containing), and back-propagation learning algorithm was used to adjust network weights.

Dimensionality reduction by PCA

Dimension reduction is considered as an important step in classification and pattern recognition problems. The aim in this step is to filter out redundant information and choose the essential features from high-dimensional data. In this study, we note that the input data have relatively high dimensionality and generally contains some redundancy. Accordingly, the PCA was applied to remove redundant information and extract intrinsic features from the original data. Based on the resulting covariance matrix, the components with eigenvalue higher than 1 were chosen as the effective features for training the network.

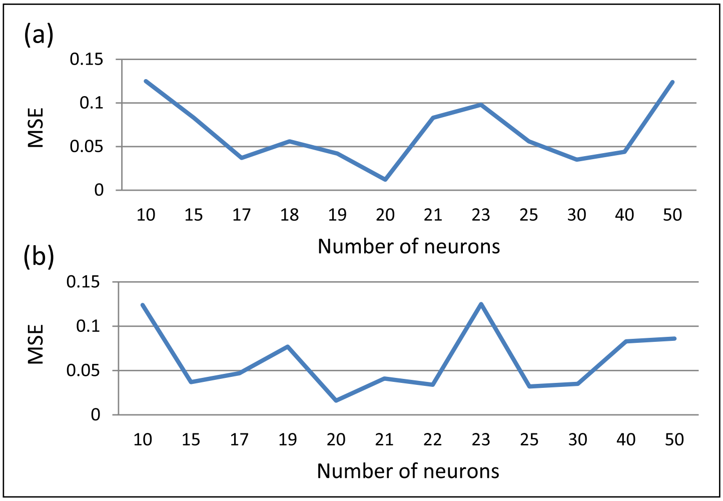


Figure 2. Mean square error (MSE) obtained by the neural network with different numbers of neurons in the hidden layer for (a) psychological and (b) public factors.

Results and discussion

Experimental setup

The above-proposed scheme was implemented using MATLAB software package on a Pentium desktop computer with 2.99-GHz central processing unit (CPU) and 4-GB main memory. In the first experiment, we trained the ANN using whole features of the dataset. The collected dataset through questionnaires was first normalized to be in the same range for all features. The dataset was divided into two training (75% or 150 samples) and testing samples (25% or 50 samples). For both psychological and public factors, training of the network was run independently using the training set and then benchmarked using testing set.

In the second experiment, we used PCA to reduce dimensionality among data before training the network. The analysis on the covariance matrix of both psychological and public factors shows that there are five components having eigenvalue greater than 1. The validity of the principal components was examined by scree graph of eigenvalues over the training samples as shown in Figure 3. The graph in this figure represents eigenvalues in descending order versus the number of the components. It can be visually observed from the graph that five components have the most variability within the feature set.

Results of the ANN models

A various number of architectures were examined by varying the activation function, number of hidden layers and their neurons, and the learning algorithms. Finally, we chose multilayer perceptron (MLP) with one hidden layer having 20 neurons, logistic activation function, and back-propagation learning algorithm. The chosen architecture produced the best accuracy rate on the validation set and used in our experiments.

To illustrate the performance of our prediction system, we plot the receiver operating characteristic (ROC) curve. The curve plots the true-positive rate (TPR) known as sensitivity against the false-positive rate (FPR) that can be calculated as 1−specificity. Generally, ROC curve analysis facilitates to compare and select the optimal model with a rational performance. The optimal

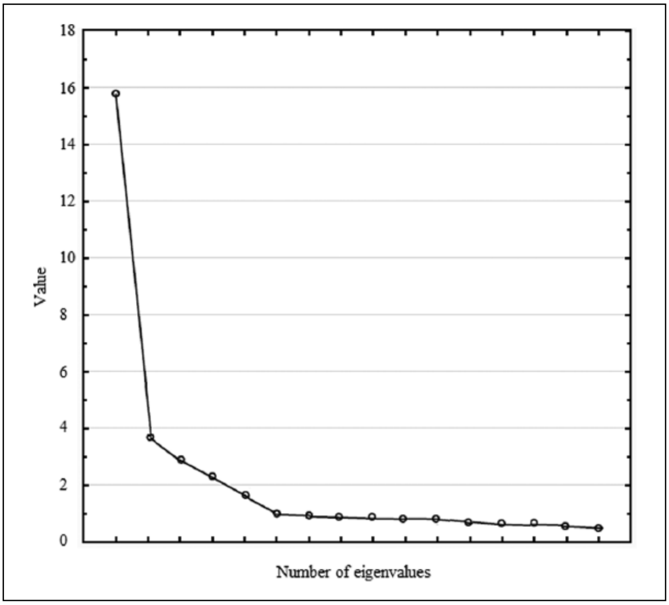


Figure 3. Scree plot of the principal component eigenvalues over the training samples.

prediction model in this curve obtains a point in the upper left corner of the plot where the value of the TPR reaches 1 or very near value to 1 for a very small value of FPR.

Figure 4 shows the ROC plots of the trained ANN models using different datasets including psychological and public factors before and after dimension reduction. As can be observed from this figure, the best result belongs to psychological factors where its curve quickly climbs to high TPR values at low FPR values. Based on the ROC curves, the performance of the ANN model over the psychological factors is expectedly decreased after dimension reduction. Furthermore, similar observations can be seen from this figure for the public factors before and after dimension reduction. However, the overall performance of the ANN models over public factors is generally lower than that of psychological factors. We arranged another experiment to train the ANN model using both psychological and public factors after dimension reduction. From this figure, the ROC curve of this model shows a better performance in comparison with the results produced by the trained models using single datasets.

Comparing the results obtained from the ANN models over different datasets (Table 2) reveals that psychological factors provide the best evidence for fall prediction among the physiological profile of elderly adults (accuracy≈90%). Performance of the models has been decreased when dimensionality of the datasets was decreased by PCA. However, combining two psychological and public factors produces the highest accuracy among the proposed models (accuracy≈91%). Furthermore, this combined dataset after dimension reduction yields a performance relatively near to the best results produced by psychological factors. Considering the higher efficiency of the combined dataset with reduced dimensions in terms of training time of the ANN, the minor decrease in the accuracy of this model can be ignored.

Extraction of the key features for fall risk prediction

We performed another study to extract features that are the most important among the features to predict falling risk in elderly adults based on the PCA. According to the PCA on psychological

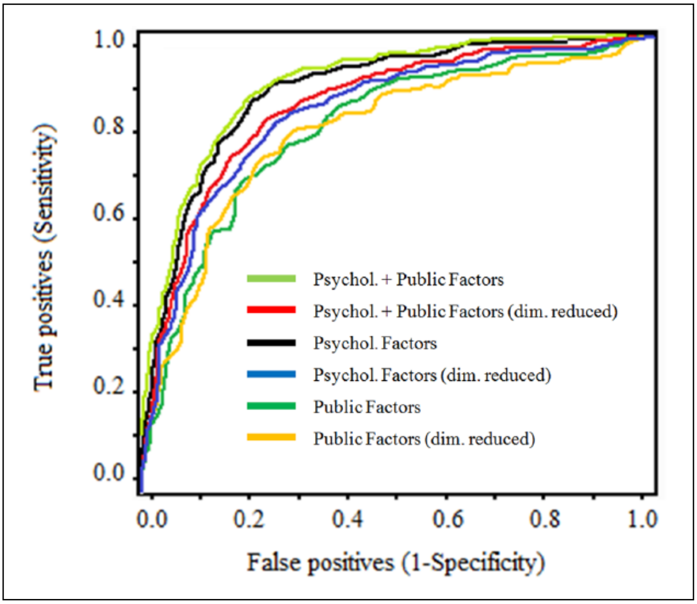


Figure 4. Receiver operating characteristic (ROC) plots produced by the ANN models for different datasets.

Table 2. Comparing the results obtained from the ANN models using different datasets.

	Accuracy (%)	
	Training samples	Testing samples
Psychological + public factors	95.7	91.3
Psychological + public factors (dimension reduced)	93.8	89.7
Psychological factors	95.1	90.2
Psychological factors (dimension reduced)	92.3	88.5
Public factors	90.8	87.5
Public factors (dimension reduced)	88.1	83.4

factors, the most effective factors among psychological features of elderly was *go up or down the slope*, *walking on slippery surfaces*, *walking in a crowded places*, *bathing or showering*, and *walking on an uneven surfaces*, respectively. Based on the PCA on the public factors, *use of drug, fear of falling, metabolic diseases, cardiovascular diseases*, and *employment* are five most important factors with a high impact on the risk of elderly falling.

Conclusion

In this study, we used an ANN model to predict fall risk among elderly adults. Based on the experimental outcomes, the proposed model obtained a fruitful result based on physiological profile of people. The analysis of the physiological profile of the older people revealed that the protection against falling can be maximized by utilizing intelligent prediction tools. The tool can be used as a complement for the medical assessment and management of elders who are at risk for falling. Furthermore, the tool provides useful analysis to identify the critical factors for falling in older people.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This paper is published as part of a research supported by the Research Affair Office of University of Tabriz.

References

1. World Health Organization Regional Office for the Eastern Mediterranean. A strategy for active, healthy ageing and old age care in the Eastern Mediterranean Region 2006–2015, 2006, http://applications.emro.who.int/dsaf/dsa542.pdf?ua=1
2. Lord SR, Menz HB and Tiedemann A. A physiological profile approach to falls risk assessment and prevention. *Phys Ther* 2003; 83(3): 237–252.
3. Fathi-Rezaie Z, Aslankhani MA, Farsi A, et al. A comparison of three functional tests of balance in identifying fallers from non-fallers in elderly. *J Knowl Health* 2010; 2010: 21–26.
4. Rubenstein LZ. Falls in older people: epidemiology, risk factors and strategies for prevention. *Age Ageing* 2006; 35: 37–41.
5. Laessoe U, Hoeck HC, Simonsen O, et al. Fall risk in an active elderly population—can it be assessed? *J Negat Results Biomed* 2007; 6: 2.
6. Pluijm SMF, Smit JH, Tromp EAM, et al. A risk profile for identifying community-dwelling elderly with a high risk of recurrent falling: results of a 3-year prospective study. *Osteoporos Int* 2005; 17: 417–425.
7. Boyd R and Steven J. Falls and fear of falling: burden, beliefs and behaviours. *Age Ageing* 2009; 38(4): 423–428.
8. Haines TP, Hill K, Walshe W, et al. Design-related bias in hospital falls risk screening tool predictive accuracy evaluations: systematic review and meta-analysis. *J Gerontol A Biol Sci Med Sci* 2007; 62: 664–672.
9. Myers H. Hospital falls risk assessment tools: a critique of the literature. *Int J Nurs Pract* 2003; 9: 233–235.
10. Scott V, Votova K, Scanlan A, et al. Multifactorial and functional mobility assessment tools for fall risk among older adults in community, home-support, long-term care and acute settings. *Age Ageing* 2007; 36: 130–140.
11. Gates S, Smith LA, Fisher JD, et al. Systematic review of accuracy of screening instruments for predicting fall risk among independently living older adults. *J Rehabil Res Dev* 2008; 45(8): 1105–1116.
12. Oliver D, McMurdo M, Daly F, et al. Risk factors and risk assessment tools for falls in hospital inpatients: a systematic review. *Age Ageing* 2004; 33: 122–130.
13. Hassankhani H, Darvishpur Kakhki A, Asghari Jafarabadi M, et al. Elders fall risk predictors. *Int Res J Appl Basic Sci* 2012; 3(8): 1662–1672.
14. Marschollek M, Gövercin M, Rust S, et al. Mining geriatric assessment data for in-patient fall prediction models and high-risk subgroups. *BMC Med Inform Decis* 2012; 12(19): 1–6.
15. Rose S. Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol* 2013; 177(5): 443–452.
16. Basheer IA and Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. *J Microbiol Methods* 2000; 43: 3–31.
17. Pearson K. On lines and planes of closest fit to systems of points in space. *Philos Mag* 1901; 2(11): 559–572.
18. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol* 1933; 24: 417–441.

Original Article

Health Informatics Journal

Health Informatics Journal
2018, Vol. 24(4) 419–431
© The Author(s) 2016
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1460458216678442
journals.sagepub.com/home/jhi
SAGE

Virtual environments in cancer care: Pilot-testing a three-dimensional web-based platform as a tool for support in young cancer patients

Mette Terp Høybye
Silkeborg Regional Hospital, Denmark; Stanford University, USA

Pia Riis Olsen
Aarhus University Hospital, Denmark

Helena Eva Hansson
University of Copenhagen, Denmark

David Spiegel
Stanford University, USA

Henrik Bennetsen
Otherlab, USA

Ewen Cheslack-Postava
Stanford University, USA

Abstract

Bringing virtual environments into cancer support may offer a particular potential to engage patients and increase adherence to treatment. Developing and pilot-testing an online real-time multi-user three-dimensional platform, this study tested the use of an early prototype of the platform among adolescent and young adult cancer patients. Data were collected with an online questionnaire and using ethnographic methods of participant observation. The adolescent and young adult patients tested basic features of the virtual environment and some conducted brief in-world interactions with fellow patients during hospitalization. They had no reservations about using the technology and shared their ideas about its use. Our pilot test pointed to a number of areas of development for virtual environment applications as potential

Corresponding author:
Mette Terp Høybye, Interdisciplinary Research Unit, Elective Surgery Centre, Silkeborg Regional Hospital, Falkevej 1-3, 8600 Silkeborg, Denmark.
Email: methoy@rm.dk

platforms for medical or behavioral interventions in cancer care. Overall, the results demonstrate the need for high user involvement in the development of such interventions and early testing of intervention designs.

Keywords
Internet, IT design and development methodologies, pilot test, virtual environment, young cancer patients

Introduction

The virtual space, independent of its technological shape, is characterized by the possibility to navigate the borderlands between the actual and the potential. In the context of cancer therapy, this may carry a prospect of new forms of behavioral intervention. New online social technologies are creating novel opportunities for therapeutic and behavioral intervention with a positive impact on patient empowerment, knowledge, self-efficacy and pain control.[1–7] Virtually mediated communication has the potential to transform our approach to patient education and interaction with the healthcare sector.[8–11] This may offer new ways to approach the psychosocial support of adolescent and young adult (AYA) cancer patients.

Cancer is the leading disease-related cause of death in AYAs.[12] Each year, nearly 70,000 people in the United States between the ages of 15 and 39 are diagnosed with cancer.[12] In Denmark, where this pilot study was conducted, approximately 1500 people (in a population of 5.2 million people) between 15 and 39 years of age are diagnosed with cancer each year.[13] Even though increased survival has been documented among AYA cancer patients,[14,15] a number of studies suggest that the survival of AYA patients still seems to be significantly lower than the survival of children with biologically similar cancer diagnoses.[14–17] Likewise, the survival of AYA patients is lower than that of adults above 40 years of age.[18,19] Several factors are believed to contribute to this disparity, including delay in diagnosis, suboptimal treatment adherence and physiological differences.[16,17,20–22]

AYA cancer patients are a particularly vulnerable group of patients as they are in transition between dependent childhood and independent young adulthood.[23] They fall in between pediatric and adult treatment protocols and have different social, psychological, biological and physiological needs and challenges than other cancer patients.[12,24] Cancer profoundly disrupts the everyday and social life of the AYA patients and the availability of sufficient, appropriate support becomes crucial. However, studies have shown that the social support of AYA cancer patients is challenged by many different factors during treatment.[25,26] Keeping up everyday life with regular activities such as school, sports and friends is difficult and in some periods impossible for patients in active treatment, as the very intense treatment protocols require frequent hospitalizations.

Computer games and other interactive, multi-media tools have been shown in previous studies to positively affect and change behavior in patients with chronic illness, increasing treatment adherence and self-efficacy, as well as relieving stressful aspects of the experience of illness.[27–32] Social identification of self and other via avatars in virtual environments has been found to affect both virtual and real-life behavior, indicating that the vicarious reinforcement and identification with avatars can motivate behavior change that could have an impact on health.[33–35] The avatar is in this study a graphical representation of the patient or user in a three-dimensional (3D) form. The simulation of events and effects found in computer games and virtual technologies make these technologies particularly suitable as educational tools to affect health and treatment behavior, beyond the traditional didactic methods.[10,32,36] A virtual environment has the possibility of mediating the difference between the actual circumstances of a patient (e.g. a state of illness or pain) and a potentially new situation, demonstrating to the patient an opportunity to perform actions and engage in social interactions that are not possible within the actual everyday frame.[2,4]

At the same time, AYAs are very active and skilled users of media, both as consumers and as creators of content,[37] which makes it suitable to develop and assess media-based interventions to this group of patients.

As described, the social challenges facing AYA patients are significant. Finding ways to alleviate social isolation and sustain social interaction with peers in AYA cancer patients during treatments is of critical importance to assist their re-integration into everyday life after treatments. Bringing virtual environments into cancer support may offer a particular potential to engage AYA cancer patients and facilitate such a process.

The goal of this article is to present the results of a pilot study that tested an early prototype of an online virtual environment for psychosocial support of AYA cancer patients, focusing on areas for further development of the environment. The pilot study did not include or evaluate an actual intervention component at this time.

Methods, design and development

Setting

In February and March 2011, 12 young Danish cancer patients were recruited from the Youth Unit (YU) at the Department of Oncology at Aarhus University Hospital to test the feasibility of an early prototype of the online environment. The oncology department is one of five oncology treatment centers in Denmark that undertakes non-surgical cancer treatment and outpatient follow-up. The YU is part of an adult ward and comprises two 2-bedded rooms and a "youth corner" in the sitting room area. A small group of nurses are dedicated to the care of the YU patients. On average, the YU admits 8–10 new patients per year between 15 and 22 years of age.[23]

Participants and procedures

Patients were recruited for the study by two of the researchers (M.T.H and P.R.O) following a weekly review through the 8 weeks of the planned patient flow in the YU. All patients eligible for treatment in the YU were also eligible for the study. All participants were in active treatment for their cancer. However, the status of each patient was assessed by the researcher (M.T.H) and clinical nurse specialist (P.R.O) and the YU nursing staff to protect the patients considered too vulnerable for participation. In the course of the 8 weeks, 16 patients were available for recruitment to the study. Four patients did not wish to participate. All participants were informed of the objectives and procedures of the study and their right to terminate participation at any time. All participants provided written informed consent to the study. The study was approved by the Danish Data Protection Agency and the Danish Regional Committee of Southern Denmark on Biomedical Research Ethics (S-20110037).

An appointment was made between the research team and the young patients as they were recruited for the study to introduce them to the virtual environment during the same hospital admission period in which they were recruited. Two participants were recruited and introduced to the virtual environment during an outpatient follow-up visit to the department. All patients received one session with a researcher. In one instance, two patients preferred to attend this researcher session together. The young patients were provided with a personal log-in to the virtual environment during this session and, if necessary, assisted to upgrade their browser software. The young patients were asked to test the basic features of the virtual environment, such as logging in, moving around their avatar and chatting with other patients' in-world, if others were simultaneously online. Participants accessed the virtual environment using their own laptop or a laptop provided by the research team.

This served as part of a collaborative development process, where the young patients were encouraged to provide feedback on the use of the virtual environment to optimize and target the intervention to their needs and behavior. An interdisciplinary group of programmers, graphic artists and researchers with different social science and health backgrounds collaborated to integrate the responses of the young patients in the further development.

Design and development

This study used a web-based online multi-user 3D environment developed for the project between October 2010 and February 2011. The environment was designed with the low technical requirement of a modern web browser on the patient's computer. The core idea was removing the friction of having to install a separate client or additional graphics drivers to participate, which traditionally had been a barrier to participation in 3D multi-user environments for non-technical users. The patient's web browser acted as a client and contacted a single server administered by one of the researchers (E.C.P) to interact with the environment, similar to the design of many multiplayer games but all it would take to enter this environment was the click on a link on a webpage.

Through a standard web interface, patients could log in to their accounts, edit their profile information and enter the virtual environment. Patient profiles included options to customize their appearance through a set of pre-configured avatars. All user data, administrative data and logs were stored in the encrypted, password-protected database. The virtual world server, built upon existing open-source software for creating multi-user 3D virtual environments,[38] simulated the environment and allowed the users to interact.

To run within a web browser, the client software leveraged new browser technologies for graphics and networking. Interactive 3D graphics were enabled by WebGL, a system that provides access to the same graphics hardware used by games. To efficiently send and receive a constant stream of updates about the world, the client used WebSockets to access low-level networking functionality normally accessible only to standalone software. At the time of development, both technologies were still under active development but available in stable versions of the Google Chrome and Mozilla Firefox web browsers.

Theoretical and conceptual approach

Our thinking about developing a virtual environment for AYA patients took theoretical outset in an understanding of the virtual space as a space between the actual and the potential.[39] A space carrying a possibility to mediate the difference between the actual circumstances of a patient (e.g. a state of illness or pain) and a potentially new situation in the virtual environment, where the patient has the opportunity to perform actions that may not be possible within the actual everyday frame (reference 2–4), such as engaging peers in conversation or games. Taking an existential approach to the social interaction in the virtual environment, we drew on the thoughts of Arendt emphasizing the crucial role of being heard and seen by others for ensuring and maintaining "humanness."[40] While undergoing cancer treatment, the AYA patients lose their ability to take part in social interaction with peers, to some extent depriving them of relationships with and recognition from others. The stories and interactions shared in a virtual environment are seen and heard by others and thereby constitute a common reality. Storytelling in this sense restores the viability of people's relationship with others.[41] We become visible to others through our stories and

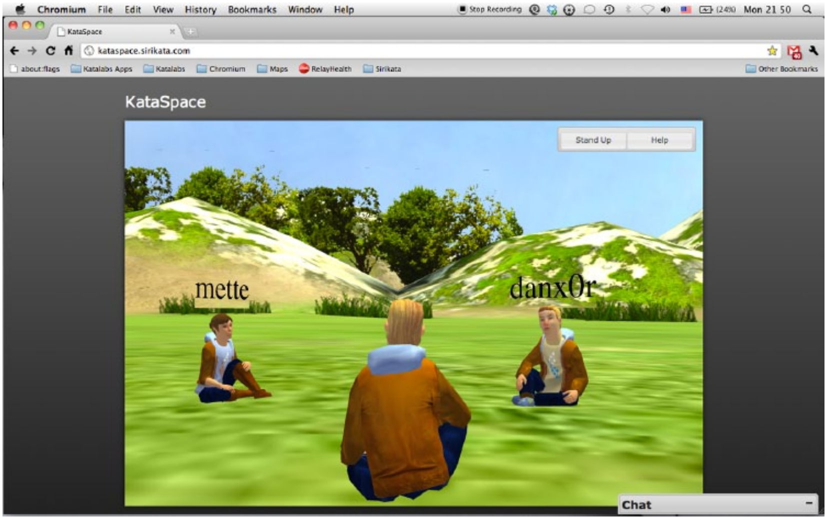


Figure 1. Screen shot of the virtual environment developed for the study. The avatars in this photo represent researchers working on the project in a testing situation.

position ourselves in the world. By focusing on the potential of the virtual environment in this study to create a sense of presence and social interaction between the AYA patients testing the features, we sought to be able to evaluate the potential of the virtual environment as a way to alleviate the social isolation that is often a consequence of cancer.

Data collection and measures

The endpoint for the study reported here was to evaluate additional technological and clinical areas for development of the virtual environment.

This study employed a combination of ethnographic observation and an online survey. The young patients completed an online questionnaire at enrollment in the study comprising socio-demographic questions as well as questions on diagnosis, medical adherence, use of the Internet and social network and support. No post-test measures were used, as the focus was not intervention outcome but the explorative endpoints described above. The data source for this study was the baseline questionnaire, completed by 10 of 12 participants and the ethnographic observations conducted with all participants.

An ethnographic approach[42] was taken to the interaction with the young patients in the researcher sessions. A researcher (M.T.H.) was present with the patient, observing in the actual physical space the use of and interaction with the virtual environment and taking notes of their immediate reactions to moving around in the virtual environment as well as their suggestions for alterations and improvements. In these sessions, a researcher avatar was further employed to interact with the young patients in-world. The same researcher who did the observation controlled this researcher avatar (Figure 1). The sessions were documented through fieldnotes.[43,44]

To test how a small virtual community might be created, we asked the young patients to participate in a professionally facilitated group within the environment to connect with each other and the research team at four specific times, from wherever they might physically be present at the scheduled meeting time. These in-world group sessions were offered in addition to and following the initial researcher session.

Data analysis

Simple, descriptive analyses were conducted on the online questionnaire data, as the population size did not permit more advanced analysis. The ethnographic data were read and organized by categorical indexing[45] to disclose patterns of use and social interaction in the virtual environment.

Results

The young patients tested basic features of the virtual environment and conducted brief in-world interactions with fellow patients during hospitalization. They had no reservations about using the technology, and they shared their ideas about its use during the sessions with the research team.

Participant characteristics

Socio-demographic characteristics. The AYA cancer patients in this study were between 16 and 27 years of age and had been diagnosed with either bone sarcoma or a brain tumor. Two of the 12 participants recruited did not complete the questionnaire. These two however did engage in a session with the research team, testing the virtual environment. Both genders were represented in the group of participants, comprising seven young men and five young women.

The majority (80%) of participants were living with their parents at the time of study. A few indicated that they had temporarily moved back in with their parents during cancer treatments. A third of the participants were enrolled in school. From this group, however, only one was attending classes part-time during treatment, while the rest were on leave from classes. The majority (70%) of participants had full-time or after-school employment, but all were on leave from their jobs at the time of study (Table 1).

Furthermore, the AYA patients reported to have a comprehensive social network of family and friends that they felt able to ask for assistance in matters of practical challenges as well as to talk about their concerns with social and emotional issues of cancer.

Characteristics of Internet usage. All AYA patients reported to use the Internet, and 80 percent used it daily. Their main use of the Internet was to stay in touch with their social network (70%) and communication in online social media networks was the main task performed while online (80%). Furthermore, the main online tasks performed by participants were web browsing for information and writing email. None of the participants, however, participated in Internet discussion groups on cancer (Table 2). During the research sessions, several participants shared that they tried to research the Internet for information on their cancer but often found it very confusing and hard to understand and were generally not sure what to make of it. They had not found Internet sites dedicated to informing about cancer to their age group or with a focus on their special challenges and needs.

Interacting in the virtual environment

Most sessions were carried out with participants and the researcher (M.T.H) all physically located in the youth corner of the hospital ward sitting room. A few sessions were carried out at the bedside of the patient in the YU. No participants had problems understanding how to enter or navigate the virtual environment.

During sessions with the research team, participants engaged with the virtual environment and tested the functions of the avatar and the virtual space. After logging in, they moved around their avatar and interacted with the researcher avatar in-world. On three occasions, AYA participants interacted with each other in-world when more participants were admitted and present at the

Table 1. Socio-demographic characteristics of 10 AYA cancer patients in the virtual environment pilot study.

Characteristics	Men, N = 6 (100%)	Women, N = 4 (100%)
Age		
16–20 years	2 (33)	2 (50)
21–27 years	4 (67)	2 (50)
Diagnosis		
Brain tumor	2 (33)	0 (0)
Bone sarcomas	4 (67)	4 (100)
Current dwelling status		
Living alone	0 (0)	1 (25)
Living with parents	6 (100)	2 (50)
Living with partner	0 (0)	1 (25)
School (high school, college, university)	2 (33)	2 (50)
Currently attending (part-time)	0 (0)[a]	1 (50)[a]
Currently on leave	2 (100)[a]	1 (50)[a]
Employment	4 (67)	3 (75)
Full-time	2 (50)[b]	1 (33)[b]
Part-time	0 (0)[b]	0 (0)[b]
After school job	2 (50)[b]	2 (67)[b]
Work status (if employed)		
Working	0 (0)	0 (0)
Currently on leave	4 (100)	3 (100)

AYA: adolescent and young adult.

[a] Percentage of the number of participants attending school.

[b] Percentage of the number of participants with employment.

Table 2. Reported Internet use of 10 AYA cancer patients in the virtual environment pilot study.

Internet use characteristics	AYA patients, N = 10 (100%)
Overall Internet use	10 (100)
Daily	8 (80)
Several times a week	2 (20)
Less	0 (0)
Main use of time online	
Staying in touch with social network	7 (70)
Research and gathering information	3 (30)
Games, movies and music	0 (0)
Online tasks (several options)	
Email	5 (50)
Online games	1 (10)
Social media networks	8 (80)
Online discussion groups	0 (0)
Blogging	0 (0)
Browsing for information	6 (60)

AYA: adolescent and young adult.

hospital ward at the same time. Those sessions were more playful and of longer duration than the interactions with the researcher avatar, as the AYA participants on those occasions took more control of the interaction. The few sessions engaging more participants provided a suggestion of the feasibility of the platform as a site for social interaction between AYA patients. The immediate feedback and response of participants were positive. They liked the idea of having access to a virtual space of their own, where they could engage with other AYA cancer patients.

To test the creation of a community in the virtual environment, as described above, the AYA cancer patients were asked to connect with each other and the research team and with a cancer counselor as facilitator within the virtual environment at four specific times from their actual location in or out of hospital. Meeting times were scheduled for late afternoons or evenings, as this would avoid interference with treatment appointments. However, these virtual meet-ups were not successful. Only two participants showed up, but not in the same session. Reflecting on their absence in later follow-up responses to the research team, the AYA patients conveyed that they had not had the energy to participate or it had interfered with other appointments they had had with friends or family.

Adverse effects of the use of the virtual environment

We did not register any adverse reactions to the engagement within the virtual environment.

Future directions for development of the virtual environment

The AYA patients made numerous suggestions to the further development of the virtual environment. The key issues of use that emerged from the interactions were concerned with integration with mobile platforms. Most AYA patients had brought a personal laptop to the hospital; however, the device that was always present with them was their mobile phone. Some participants suggested that accessibility would be higher if the virtual environment offered a mobile integration.

Also, participants suggested that a further development of the avatar animations would make it more engaging to move around. In particular, they would be interested in funny animations that would allow for jumping, hugging and dancing.

Exploring the types of interactions favored by the AYA cancer patients for future developments of the virtual environment, participants suggested mundane interactive games over more advanced educational or informational games. Well-known board games (like backgammon or chess) brought in-world were perceived as possible site-s of action to share and meet around, providing an opportunity to talk while playing, but not demanding continuous talking to share meaningful interactions. A few participants also suggested in-world poster boards for sharing photos, notes or poems.

Discussion

The AYA cancer patients in this study were strongly engaged with online media as part of their everyday activity and embraced the potential of an online virtual environment for social interaction when introduced to the idea and testing the virtual environment. The platform, however, did not integrate with the online spaces where the young patients were already present and interacted, such as Facebook. This may have limited the perceived accessibility. Also, it was accessible only through a web browser and did not support a mobile integration. Our observations revealed that the device constantly present with the young patients were their mobile phones, which suggests that a mobile platform may speak more to their use of technology. A future integration with Internet sites that already channel activity from the AYA patients may help drive more activity to a platform like the virtual environment. With the high demand on time and energy made by cancer treatment on the AYA patients, we found that they required additional motivation to check in on the virtual environment. Feedback from the patients suggests that additional information or entertainment might attract them to visit the environment more consistently.

Communication in the virtual environment was strictly synchronous, meaning that it only facilitated communication between users who were present in the environment at the same time. The suggestions made by AYA patients, to include features in a future development that would allow for posting images or poems, are clearly requests for other asynchronous methods of interaction. Such options would make it easier to build up a community within the virtual environment and achieve the critical mass necessary for a virtual environment to be dynamic, as it sustains an imagination of community without requiring simultaneous presence. The integration of asynchronous interactions could increase the success of future group meetings in-world, for example, by developing long-lived connections to the platform and notification services, where a user is still connected though not actually present and can be notified to log in when something happens. Such reminders and services might drive further participation and ensure co-presence in-world.

A crucial point for future design and successful implementation of virtual environment on browser-based platforms and applications is the support of the various devices and operating systems that different patients use. This is an extremely complicated task and a big barrier to applications like these. However, the limitation to a single device or operating system will limit the accessibility and use, as the request for mobile integration in our AYA population suggests.

Most existing studies of the online use and interaction of cancer patients have focused on the Internet use of adult cancer patients,[1,46–50] documenting an extensive use of online cancer support groups in the adult cancer population. Although it is not possible to generalize on the Internet use of the AYA cancer population based on our study, it is interesting to note that no participants in this study reported to use Internet cancer support groups, which may suggest that other types of web-based interventions are more suitable for engaging this group of patients.

The generalizability of this study is limited by the small test population and the specific environment of testing. Also, the AYA patients in this study may have been a selected group of patients with more comprehensive social networks than the general population of AYA patients, if we compare them to reports in the literature.[24,51] This may have affected their lack of perceived need for the social interaction offered in the four in-world group sessions. However, the study provides a first step in the development of the virtual environment technologies to target and serve the particular and specific needs of the AYA cancer population. It is to the best of our knowledge the first study to test the use of online 3D virtual environments with AYA patients. In this sense, this study offers a valuable first step in the use and design of virtual environments for personalized therapeutic treatments.

As was documented by Tsangaris et al.[51] in a study of the supportive care needs of AYA cancer patients, the social needs of patients are often not well met. Participants in this study in general reported to have comprehensive social networks and the majority was living with their parents. It may however be difficult for the AYA patients to sustain such a network and its support throughout treatment, as shown by Zebrack et al.[24] Systematic work with network-focused nursing to facilitate the involvement of the social network of AYA patients has been shown in previous studies to strengthen the AYA patient's sense of self and help them navigate their interactions with family and friends.[23,25]

Although our pilot study was not successful to drive activity to in-world group sessions, the one-on-one researcher test sessions with AYA patients suggested a potential for further developing and using virtual environments in behavioral interventions with AYA patients. Such interventions could benefit from how the virtual environment makes it possible to simulate certain interactions in cancer treatment. This could increase knowledge and understanding in the individual patient of his or her situation, possibly affecting adherence, as has been documented in game-based interventions.[32] Furthermore, the AYA patients were highly motivated by thinking of social interaction features that would make it possible to just hang out and play together in-world. A supplementary approach to strengthening the social network and support of AYA patients may be a further development of the potential of the virtual environment interactions that may serve as a platform for alleviating social isolation by interacting with peers. The importance of such social integration has previously been documented in text-based Internet interactions.[1,52–56] The continuous input from AYA patients to further and future development of such technologies is crucial to ensure they are perceived to be useful in terms of providing spaces for social interaction.

The user-test of the online virtual environment in this study encourages further work with similar types of virtual platforms and interventions as personalized treatment options. This may help to meet the needs of the AYA patients and strengthen their involvement in their care and follow-up. A particular focus point for such development may be to use in-world simulations as personalized intervention strategies to engage patients and increase their adherence to treatment.

It is an important finding of this study that AYA patients are motivated and eager to participate in the testing and development of interventions and technologies, which will serve to increase and ensure that they meet the needs of this particular patient population, as has also been argued in previous e-health development projects.[57] Online virtual environments may offer a particular potential in the future development of comprehensive healthcare programs that address such needs for personalized medicine and care, holding a particular potential for engagement and social interaction.

Conclusion

Our pilot test identified a number of areas of development for virtual environment applications as platforms for medical or behavioral interventions in personalized cancer therapies. Overall, the results demonstrate that engaging the user in the development and early testing helps identify the needs of the intended population, pointing to additional technological areas for development, such as integration with a mobile platform and asynchronous methods of interaction.

Moving toward-s employing such virtual environments in cancer care, issues concerning design, presence and usability are important to consider, as they situate and affect utilization and potential future health outcomes.

Acknowledgements

The virtual environment and the study design were developed in collaboration between Department of Computer Science, Stanford University; Katalabs Inc.; Center on Stress and Health, Department of Psychiatry and Behavioral Sciences, Stanford School of Medicine, Stanford University; and the Pediatric Oncology Research Laboratory, Bonkolab, at the Pediatric Oncology Unit at University Hospital Rigshospitalet, Denmark. The authors wish to acknowledge the valuable input in this process from Professor Kjeld Schmiegelow, University Hospital Rigshospitalet. The authors further wish to acknowledge the unique support and collaboration with Department of Oncology, Aarhus University Hospital, in implementing the pilot test and recruiting patients. The authors wish to thank the young cancer patients who participated in the study.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

This study was generously funded by the Danish Child Cancer Foundation (Grant Number: 2010-12) and the Danish Agency for Science, Technology and Innovation (Grant Number: 10-098405).

References

1. Høybye MT, Johansen C and Tjørnhøj-Thomsen T. Online interaction. Effects of storytelling in an internet breast cancer support group. *Psychooncology* 2005; 14: 211–220.
2. Stewart S, Hansen TS and Carey TA. Opportunities for people with disabilities in the virtual world of second life. *Rehabil Nurs* 2010; 35: 254–259.
3. Beard L, Wilson K, Morra D, et al. A survey of health-related activities on second life. *J Med Internet Res* 2009; 11: e17.
4. Morie JF, Antonisse J, Bouchard S, et al. Virtual worlds as a healing modality for returning soldiers and veterans. *Stud Health Technol Inform* 2009; 144: 273–276.
5. Bender JL, Radhakrishnan A, Diorio C, et al. Can pain be managed through the Internet? A systematic review of randomized controlled trials. *Pain* 2011; 152: 1740–1750.
6. Keim-Malpass J, Baernholdt M, Erickson JM, et al. Blogging through cancer: young women's persistent problems shared online. *Cancer Nurs* 2013; 36: 163–172.
7. Griffiths F, Cave J, Boardman F, et al. Social networks—the future for health care delivery. *Soc Sci Med* 2012; 75: 2233–2241.
8. Van den Brink JL, Moorman PW, de Boer MF, et al. An information system to support the care for head and neck cancer patients. *Support Care Cancer* 2003; 11: 452–459.
9. Tufano JT and Karras BT. Mobile eHealth interventions for obesity: a timely opportunity to leverage convergence trends. *J Med Internet Res* 2005; 7: e58.
10. Kakinuma A, Nagatani H, Otake H, et al. The effects of short interactive animation video information on preanesthetic anxiety, knowledge, and interview time: a randomized controlled trial. *Anesth Analg* 2011; 112: 1314–1318.
11. Høybye M, Vesterby M and Jørgensen L. Producing patient-avatar identification in animation video information on spinal anesthesia by different narrative strategies. *Health Informatics J* 2016; 22: 370–382.
12. Nass SJ and Patlak M; National Cancer Policy Forum. *Identifying and addressing the needs of adolescents and young adults with cancer: workshop summary*. Washington, DC: The National Academies Press, 2013, www.nap.edu
13. Sundhedsdatastyrelsen. Cancerregisteret. Find tal og analyser, 2016. http://sundhedsdatastyrelsen.dk/da/registre-og-services/om-de-nationale-sundhedsregistre/sygedomme-laegemidler-og-behandlinger/cancerregisteret (accessed November 2016).
14. Tai E, Pollack LA, Townsend J, et al. Differences in non-Hodgkin lymphoma survival between young adults and children. *Arch Pediatr Adolesc Med* 2010; 164: 218–224.
15. Gatta G, Capocaccia R, De AR, et al. Cancer survival in European adolescents and young adults. *Eur J Cancer* 2003; 39: 2600–2610.
16. Bleyer A. Older adolescents with cancer in North America deficits in outcome and research. *Pediatr Clin North Am* 2002; 49: 1027–1042.
17. Bleyer WA. Cancer in older adolescents and young adults: epidemiology, diagnosis, treatment, survival, and importance of clinical trials. *Med Pediatr Oncol* 2002; 38: 1–10.
18. Albritton K, Barr R and Bleyer A. The adolescence of young adult oncology. *Semin Oncol* 2009; 36: 478–488.
19. Bleyer A, Barr R, Hayes-Lattin B, et al. The distinctive biology of cancer in adolescents and young adults. *Nat Rev Cancer* 2008; 8: 288–298.
20. Spinetta JJ, Masera G, Eden T, et al. Refusal, non-compliance, and abandonment of treatment in children and adolescents with cancer: a report of the SIOP Working Committee on psychosocial issues in pediatric oncology. *Med Pediatr Oncol* 2002; 38: 114–117.
21. Tebbi CK. Treatment compliance in childhood and adolescence. *Cancer* 1993; 71: 3441–3449.
22. Schmiegelow K, Heyman M, Gustafsson G, et al. The degree of myelosuppression during maintenance therapy of adolescents with B-lineage intermediate risk acute lymphoblastic leukemia predicts risk of relapse. *Leukemia* 2010; 24: 715–720.

23. Olsen PR and Harder I. Keeping their world together—meanings and actions created through network-focused nursing in teenager and young adult cancer care. *Cancer Nurs* 2009; 32: 493–502.
24. Zebrack B, Kent EE, Keegan TH, et al. "Cancer sucks," and other ponderings by adolescent and young adult cancer survivors. *J Psychosoc Oncol* 2014; 32: 1–15.
25. Olsen PR and Harder I. Caring for teenagers and young adults with cancer: a grounded theory study of network-focused nursing. *Eur J Oncol Nurs* 2011; 15: 152–159.
26. Enskar K, Carlsson M, Golsater M, et al. Symptom distress and life situation in adolescents with cancer. *Cancer Nurs* 1997; 20: 23–33.
27. Krishna S, Francisco BD, Balas EA, et al. Internet-enabled interactive multimedia asthma education program: a randomized trial. *Pediatrics* 2003; 111: 503–510.
28. Griffiths M. Video games and health. *BMJ* 2005; 331: 122–123.
29. Lieberman DA. Management of chronic pediatric diseases with interactive health games: theory and research findings. *J Ambul Care Manage* 2001; 24: 26–38.
30. Brown SJ, Lieberman DA, Germeny BA, et al. Educational video game for juvenile diabetes: results of a controlled trial. *Med Inform (Lond)* 1997; 22: 77–89.
31. Suzuki LK and Kato PM. Psychosocial support for patients in pediatric oncology: the influences of parents, schools, peers, and technology. *J Pediatr Oncol Nurs* 2003; 20: 159–174.
32. Kato PM, Cole SW, Bradlyn AS, et al. A video game improves behavioral outcomes in adolescents and young adults with cancer: a randomized trial. *Pediatrics* 2008; 122: e305–e317.
33. Fox J, Bailenson JN and Binney J. Virtual experiences, physical behaviors: the effect of presence on imitation of an eating avatar. *Presence: Teleop Virt* 2009; 18: 294–303.
34. Fox J and Bailenson JN. Virtual self-modeling: the effects of vicarious reinforcement and identification on exercise behaviors. *Media Psychol* 2009; 12: 1–25.
35. Bailenson JN, Blascovich J and Guadagno RE. Self representations in immersive virtual environments. *J Appl Soc Psychol* 2008; 38: 2673–2690.
36. Leiner M, Handal G and Williams D. Patient communication: a multidisciplinary approach using animated cartoons. *Health Educ Res* 2004; 19: 591–595.
37. Ito M, Horst H, Bittanti M, et al. *Living and learning with new media: summary of findings from the digital youth project*. Berkeley, CA: The MacArthur Foundation, 2008.
38. Cheslack-Postava E, Azim T, Mistree BF, et al. A scalable server for 3D metaverses. In: *Proceedings of the USENIX annual technical conference*, Boston, MA, 13–15 June 2012, https://www.usenix.org/system/files/conference/atc12/atc12-final18.pdf
39. Massumi B. *Parables for the virtual*. Durham, NC: Duke University Press, 2002.
40. Arendt H. *The human condition*. Chicago, IL: The University of Chicago Press, 1958.
41. Jackson MD. *The politics of storytelling: violence, transgression and intersubjectivity*. Copenhagen: Museum Tusculanum Press, University of Copenhagen, 2002.
42. Hammersley M and Atkinson P. *Ethnography: principles in practice*. London: Routledge, 1996.
43. Emerson R, Fretz RI and Shaw L. *Writing ethnographic fieldnotes*. Chicago, IL: The University of Chicago Press, 1995.
44. Sanjek R. On ethnographic validity. In: Sanjek R (ed.) *Fieldnotes: the makings of anthropology*. Ithaca, NY: Cornell University Press, 1990, pp. 385–418.
45. Mason J. *Qualitative researching*. London: SAGE, 2002.
46. Eysenbach G, Powell J, Englesakis M, et al. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *BMJ* 2004; 328: 1166.
47. Dickerson SS, Boehmke M, Ogle C, et al. Seeking and managing hope: patients' experiences using the Internet for cancer care. *Oncol Nurs Forum* 2006; 33: E8–E17.
48. Høybye MT, Dalton SO, Deltour I, et al. Effect of Internet peer-support groups on psychosocial adjustment to cancer: a randomised study. *Br J Cancer* 2010; 102: 1348–1354.
49. Leykin Y, Thekdi SM, Shumay DM, et al. Internet interventions for improving psychological well-being in psycho-oncology: review and recommendations. *Psychooncology* 2012; 21: 1016–1025.
50. Eysenbach G. The impact of the Internet on cancer outcomes. *CA Cancer J Clin* 2003; 53: 356–371.

51. Tsangaris E, Johnson J, Taylor R, et al. Identifying the supportive care needs of adolescent and young adult survivors of cancer: a qualitative analysis and systematic literature review. *Support Care Cancer* 2014; 22: 947–959.
52. Hoey LM, Ieropoli SC, White VM, et al. Systematic review of peer-support programs for people with cancer. *Patient Educ Couns* 2008; 70: 315–337.
53. Meier A, Lyons EJ, Frydman G, et al. How cancer survivors provide support on cancer-related Internet mailing lists. *J Med Internet Res* 2007; 9: e12.
54. Tichon J and Yellowlees P. Internet social support for children and adolescents. *J Telemed Telecare* 2003; 9: 238–240.
55. Van Uden-Kraan CF, Drossaert CH, Taal E, et al. Participation in online patient support groups endorses patients' empowerment. *Patient Educ Couns* 2009; 74: 61–69.
56. Winefield HR. Support provision and emotional work in an Internet support group for cancer patients. *Patient Educ Couns* 2006; 62: 193–197.
57. Petersen LS, Bertelsen P and Bjornes C. Cooperation and communication challenges in small-scale eHealth development projects. *Int J Med Inform* 2013; 82: e375–e385.

Health Informatics Journal

Original Article

Health Informatics Journal
2018, Vol. 24(4) 432–452
© The Author(s) 2016
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1460458216678443
journals.sagepub.com/home/jhi
SAGE

Medical informatics research trend analysis: A text mining approach

Yong-Mi Kim
The University of Oklahoma—Tulsa, USA

Dursun Delen
Oklahoma State University—Tulsa, USA

Abstract

The objective of this research is to identify major subject areas of medical informatics and explore the time-variant changes therein. As such it can inform the field about where medical informatics research has been and where it is heading. Furthermore, by identifying subject areas, this study identifies the development trends and the boundaries of medical informatics as an academic field. To conduct the study, first we identified 26,307 articles in PubMed archives which were published in the top medical informatics journals within the timeframe of 2002 to 2013. And then, employing a text mining -based semi-automated analytic approach, we clustered major research topics by analyzing the most frequently appearing subject terms extracted from the abstracts of these articles. The results indicated that some subject areas, such as biomedical, are declining, while other research areas such as health information technology (HIT), Internet-enabled research, and electronic medical/health records (EMR/EHR), are growing. The changes within the research subject areas can largely be attributed to the increasing capabilities and use of HIT. The Internet, for example, has changed the way medical research is conducted in the health care field. While discovering new medical knowledge through clinical and biological experiments is important, the utilization of EMR/EHR enabled the researchers to discover novel medical insight buried deep inside massive data sets, and hence, data analytics research has become a common complement in the medical field, rapidly growing in popularity.

Keywords

cluster analysis, electronic health records, medical informatics, PubMed, text mining

Corresponding author:
Yong-Mi Kim, School of Library and Information Studies, Schusterman Center, The University of Oklahoma—Tulsa, 4502 East 41st Street, Tulsa, OK 74135, USA.
Email: yongmi@ou.edu

Introduction

Medical informatics is an interdisciplinary research field that applies information technology (IT) to the medical field for creating and analyzing data, information, and knowledge to improve healthcare and medicine.[1–3] Despite the recently heightened interest, medical informatics is not a new field. Rather, it has been a long-running endeavor of applying IT to medical research for creating and processing data, information, and knowledge. According to Masic,[4] the bulk of the related research activity within this field began in the 1950s due to the rise of information communication technology (ICT). Masic noted that the first medical informatics article was authored by Ledley and Lusted in 1959, and it was the medical application of the electronic computer in diagnostics and therapy which established medical decision-making. Masic further claimed that the falling price of computers in the 1980s opened the door for an extensive development of health information technology (HIT) at all levels of the healthcare systems. He states that in the 1990s, medical/health research on the improvement of methods and techniques of artificial intelligence was actively conducted. The development and application of expert systems in medical diagnostics and therapy was also widely researched during this period. As such, the discussion of medical informatics cannot be separated from the development of technology.

Recently, the development of technology which enabled the healthcare industry to collect massive data and analyze for quality care and cost-effectiveness has prompted a big stream of research, which is data mining using electronic medical/health records (EMR/EHR).[5,6] Because of the hope of what data mining can offer to the healthcare industry, President Bush called for nationwide use of EMR by 2014, and the Department of Health and Human Services (HHS) is involved in various aspects of achieving this goal.[7] Accordingly, the Administration and Congress have both been requiring the adoption, connectivity, and interoperability of HIT.[7] This effort will increase the adoption of EMR/EHR and is expected to increase research using the collected data. As such, it is not surprising to observe that this field grew very fast and became a mature independent field of study.[1] In fact, medical informatics is now the fastest growing field based on the number of publications in PubMed.[8]

Couple with the policy and the advancement of technology, the academia observed the rapidly growing research interests, and scholars have attempted to understand a boundary for the field, but it is limited to a specific topic area[9] or a specific journal.[10] Although existing review studies assist in the understanding of the field of medical informatics, they were written before the wide spread of the EMR/EHR deployment or narrowly focused.[1,11] Because EMR-related medical informatics articles have recently seen a rapid increase, and because those articles are the most cited in the field,[10,12] it is important to investigate how these recent governmental efforts to adopt EMR/EHR and emerging technological capabilities to store and analyze big data have changed the field, and how they impact medical informatics as an academic discipline. As such, the purpose of this article is to comprehensively investigate the major subject areas over the last decade, as well as how those subject areas have changed and where they are heading.

In order to achieve the research objective, we identified the top 23 journal publications in the field of Medical Informatics within the past 12 years. Journals are identified based on the Institute for Scientific Information's (ISI) Web of Knowledge Journal Citation Report (JCR) 2012. These journals are considered to be leading and shaping the field of study because of their impacts. The total number of articles amounts to 26,307 articles. With such a large number of articles, if one attempted to discover the trends of the field by manual means, it would be an overwhelming, if not impossible, task to process and categorize this vast quantity of articles in order to identify the trends of major subject areas in general and in a time-series by year. Even if it is possible, the outcome could be inaccurate or incomplete. The recent advances in data mining technologies and related software development efforts enable researchers to discover underlying patterns and trends from massive quantities of documents, as is the case in this study. Text mining is the automated or partially automated processing of text. It involves imposing meaningful structure upon text so that relevant information can be extracted from it.[13–15] In this study, following the general guidelines presented in Delen and Crossland's article from 2008, we used a hybrid methodology (text mining followed by data mining) to apply semi-automated text categorization to organize available research abstracts into logical categories (or topic clusters) by published years in order to observe the trends of each subject area over time.

The rest of the article is organized as follows: in section "Materials and methods," we will discuss research methods that include how we identified the articles, the analytical techniques used, and the processes of those techniques. Section "Results" reports the findings followed by an interpretation of the findings in section "Discussion." We conclude this article with some recommendations and future research directions.

Materials and methods

This section offers explanations on how sample articles are identified, searched, and finalized. The second part of this section deals with text mining techniques that were used for this article. We offered detailed explanations on the text mining process that we used for this manuscript in order to introduce the emerging technique to the field.

Data acquisition and preprocessing

The sample journals are identified using the ISI's Web of Knowledge JCR 2012 in the field of medical informatics. This method is chosen because it is commonly used in academia as an indicator to assess journal rankings across disciplines. Scholars and publishers perceive highly ranked JCRs as prestigious or of such an esteemed quality that they commonly use them as their own research outlet.[16] As such, those journals shape and guide the directions of the field. Following the practice, we used all journals listed in JCR 2012 in the medical informatics area as our journal sample, and the criteria produced 23 journals shown in Table 1.

The dates of the included journals range between 2002 and 2013. The beginning year, 2002, was chosen because the application of HIT to the medical field became widely utilized in the timeframe. This was due to the demands for decreasing paperwork through the adoption of HIT and preventing medical errors via evidence-based treatments.[17–21] Despite the popular application of HIT to medical informatics, sparse research has been conducted after this timeframe. The ending timeframe, 2013, was set because many of the journals' summaries for 2014 were unavailable when we searched the articles in March 2014 as PubMed does not make journal articles available until 1 year after publication.

Based on the 23 journals, we went through each journal in the PubMed website and retrieved the title, abstract, publication date, and journal name.[22] Two of the journals began included in PubMed after 2002. More specifically, *Health Informatics Journal* recorded in PubMed from 2006 and *Informatics for Health and Social Care* started from 2008. Those two journals are searched from their inclusion years in PubMed.

After going through each journal's search process, 26,307 articles were identified. Among them, editor's notes, book reviews, and commentaries were excluded from the dataset as the purpose of

Table 1. List of journals included in the study.

Journal name	2012 total cites	Impact factor	5-year impact factor
J Med Internet Res	2421	3.768	4.728
J Am Med Inform Assn	5012	3.571	3.959
Med Decis Making	3335	2.890	3.190
IEEE Eng Med Biol	1508	2.727	1.526
Stat Methods Med Res	2044	2.364	3.142
J Biomed Inform	1899	2.131	2.434
Int J Med Inform	2411	2.061	2.700
Stat Med	15,994	2.044	2.789
IEEE T Inf Technol B	2232	1.978	2.327
Med Biol Eng Comput	3889	1.790	1.986
J Med Syst	1412	1.783	1.863
BMC Med Inform Decis	966	1.603	2.185
Method Inform Med	1341	1.600	1.402
Comput Meth Prog Bio	2461	1.555	1.589
Int J Technol Assess	1637	1.551	1.806
J Eval Clin Pract	2093	1.508	1.642
Artif Intell Med	1281	1.355	1.767
Inform Health Soc Ca	112	1.273	1.493
Biomed Tech	476	1.157	0.871
Health Inform J	212	0.830	N/A
Cin-Comput Inform Nu	388	0.816	0.945
Health Inf Manag J	86	0.704	0.826
Biomed Eng-Biomed Te	13	N/A	N/A

this project was to cluster the words in the abstract of research papers. Also, the above items are not research papers, and thus they are not peer reviewed. Furthermore, they do not have abstracts. After removing those items, the data sample was reduced to 21,464. All searched articles were directly transferred from the PubMed website to EndNote and then to an Excel file in order to analyze clusters in SAS Enterprise.

Text mining methodology

Text mining is the semi-automated process of extracting patterns to discover knowledge from large amounts of unstructured data sources.[23] Text mining is closely related to data mining in that it has the same purpose and uses the same processes, but with text mining, the input to the process is a collection of unstructured text files such as Word documents, PDF files, text excerpts, and XML files. The benefits of text mining are in areas where large amounts of textual data are being generated, such as academic research literature (the one that is used in this study), finance (quarterly reports, media commentaries), medicine (discharge summaries, doctor notes), law (court orders), biology (molecular interactions), technology (patent files), and marketing (customer comments).

Text mining process

In order to succeed, text mining studies should follow a sound methodology based on lessons learned and best practices. A standardized process, such as CRoss Industry Standard Process for Data Mining (CRISP-DM), is also needed for text mining. At a very high level, the text mining

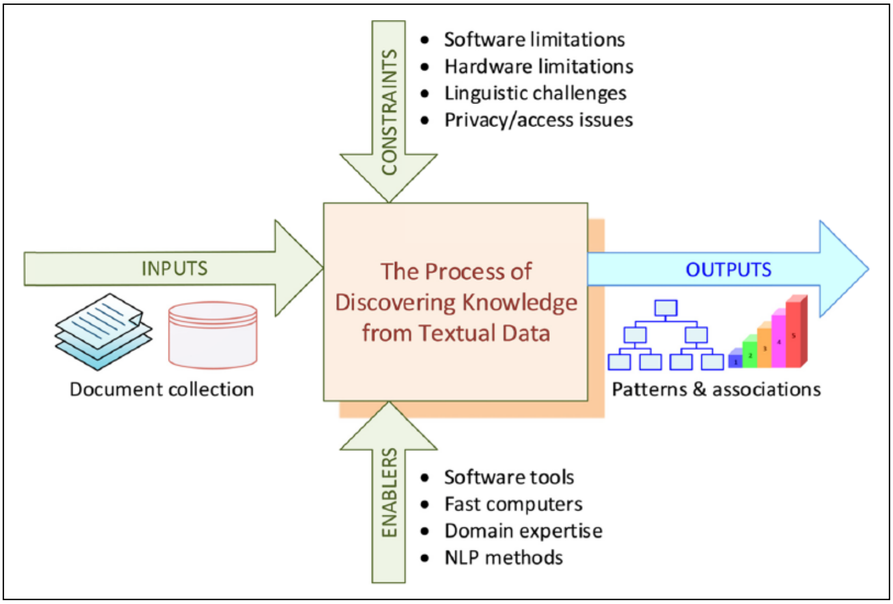


Figure 1. High-level context diagram for text mining of published literature.

process can be represented with a context diagram where the inputs, output, controls (i.e. constraints), and mechanisms (i.e. enablers) are captured as directed arrows as in Figure 1.

The context diagram can be decomposed into three consecutive activities/phases, each of which having specific inputs to generate certain outputs (see Figure 2). If, for some reason, the output of a task is not that which is expected, a backward redirection to the previous task execution is necessary. Figure 1 provides a graphical presentation.

Phase 1: establish the corpus. The main purpose of the first task activity is to collect all of the documents related to the context (domain of interest) being studied. In this specific study, we retrieved article summaries of the selected sample from the PubMed website. The collected documents were then transformed and organized in a manner in which they are all in the same representational form (e.g. ASCII text files) for computer processing.

Phase 2: pre-process the data (create the term-by-document matrix). Upon the establishment of the corpus, the term-by-document matrix (TDM) is created using the corpus. In the TDM, rows represent the documents and columns represent the terms (Figure 3). The relationships between the terms and documents are characterized by indices (i.e. a relational measure that can be as simple as the number of occurrences of the term in respective documents). As can be seen in Figure 4, there are several consecutive tasks that need to be carried out to create the clusters.

Task 1. The first task generates stop-terms whose terms do not discriminate across documents. In this task, the terms appearing in almost every article such as research method, propose, author, or findings are removed from the analysis.

Task 2. The term list is created by *stemming* or *lemmatization*, which refers to the reduction of words/terms to their simplest forms (i.e. roots). An example of stemming is to identify and index different grammatical forms or declinations of a verb as the same term. For example,

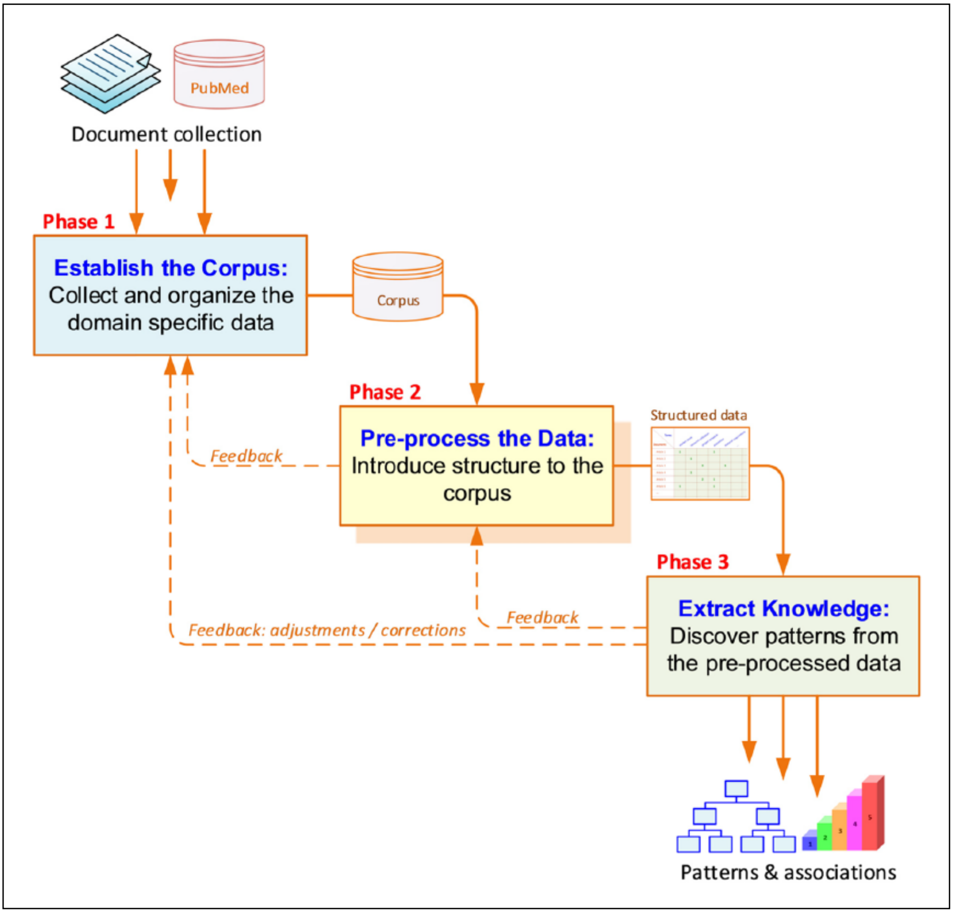

PubMed
Document collection
Phase 1
Establish the Corpus: Collect and organize the domain specific data
Corpus
Phase 2
Pre-process the Data: Introduce structure to the corpus
Structured data
Feedback
Phase 3
Extract Knowledge: Discover patterns from the pre-processed data
Feedback
Feedback: adjustments / corrections
Patterns & associations

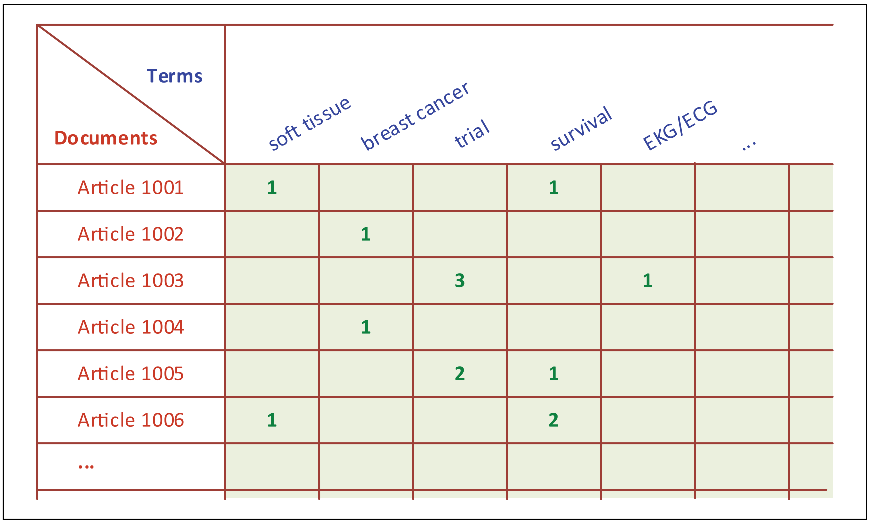
Figure 2. Process for text mining and knowledge discovery.

stemming will ensure that *model, modeling* and *modeled* will be recognized as the term *model*. In this way, stemming will reduce the number of distinct words/terms and increase the frequency of some terms.

Task 3. Create the TDM. In this task, a numeric two-dimensional matrix representation of the corpus is created. Generation of the first form of the TDM includes three steps:

1. Specifying all the documents as rows in the matrix;
2. Identifying all of the unique terms in the corpus (as its columns), excluding the ones in the stop term list;
3. Calculating the occurrence count of each term for each document (as its cell values).

Commonly, the corpus includes a rather large number of documents, which is time-consuming and, more importantly, it might lead to the extraction of inaccurate patterns. These dangers of large matrices and time-consuming operations pose two questions.[24] The first question asks how does a researcher identify the best representation of the indices for optimal processing by text mining



Terms Documents	soft tissue	breast cancer	trial	survival	EKG/ECG	...	
Article 1001	1			1			
Article 1002		1					
Article 1003			3		1		
Article 1004		1					
Article 1005			2	1			
Article 1006	1			2			
...							

Figure 3. Sample term-by-document matrix.

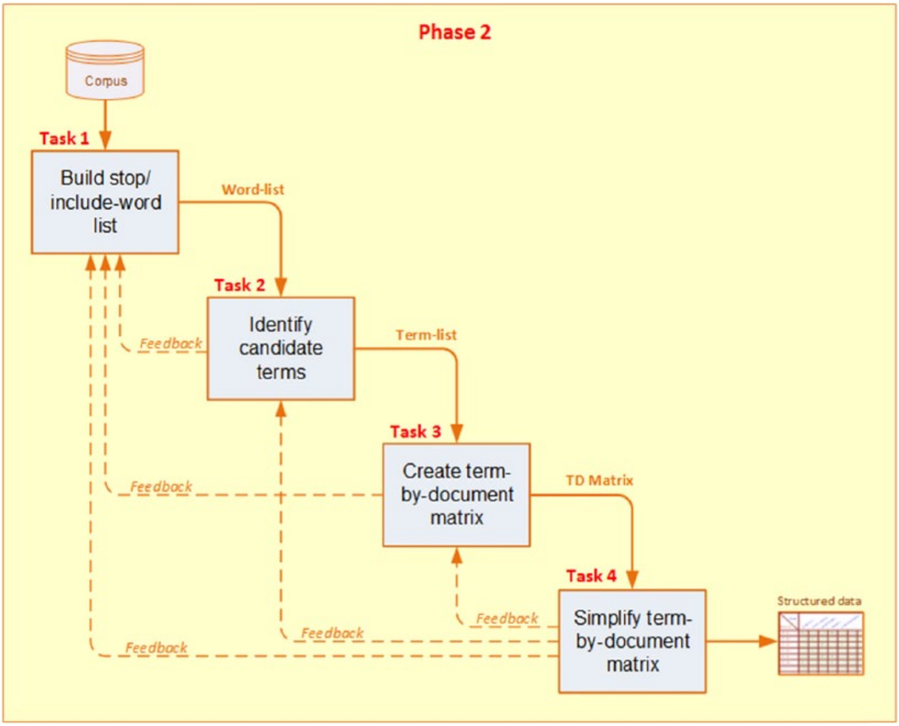


Figure 4. Decomposition of "pre-processing the data" phase.

algorithms. The most commonly used method for this question is to transform the term frequencies. For this method, the input documents are indexed and the initial word/term frequencies (by

Table 2. Simple example of TF/IDF.

Word	Term frequency	Documents with the term
Tissue	1124	484
Bone	989	243

document) are computed in order to summarize and aggregate the extracted information. Specifically, terms that occur with greater frequency in a document may be the best descriptors of the contents of that document. It is not, however, reasonable to assume that the term counts themselves are proportional to their importance as descriptors of the documents. For example, even though a term occurs three times more often in document A than in document B, it is not necessarily reasonable to conclude that this term is three times more important as a descriptor for document A as it would be for document B.

In order to have a more consistent TDM for further analysis, these raw indices should be normalized. In a statistical analysis, normalization consists of dividing multiple sets of data by a common value in order to eliminate differing effects of different scales among the data elements to be compared. Raw frequency values can be normalized using a number of alternative methods that include log frequency, binary frequency, and term frequency (TF) into inverse document frequency (IDF).

The most commonly used representation is TF/IDF, which is the one used in this study (Table 2). It works as follows: a term such as *guess* may occur frequently in all documents, whereas another term, such as *software*, may appear only a few times. The reason is that one might make *guesses* in various contexts, regardless of the specific topic, whereas *software* is a more semantically focused term that is likely to occur only in documents that deal with computer software. A common and very useful transformation that reflects both the specificity of terms (relative document frequencies) and the overall frequencies of their occurrences (transformed term frequencies) is the so-called IDF.[25] This transformation for the *i*th term and *j*th document can be written as

$$idf(i,j) = \begin{cases} 0 & \text{if } tf_{ij} = 0 \\ (1+\log(tf_{ij}))\log\dfrac{N}{df_i} & \text{if } tf_{ij} \geq 1 \end{cases}$$

where tf_{ij} is the normalized frequency of the *i*th term in the *j*th document, df_i is the document frequency for the *i*th term (the number of documents that include this term), and *N* is the total number of documents. You can see that this formula includes both the dampening of the simple-term frequencies via the log function (described above) and a weighting factor that evaluates to 0 if the term occurs in all documents (i.e. $\log(N/N=1)=0$) and to the maximum value when a term only occurs in a single document (i.e. $\log(N/1)=\log(N)$). It also shows how this transformation will create indices that reflect both the relative frequencies of occurrences of terms, as well as their document frequencies representing semantic specificities for a given document. In order to remove the high-frequency words, terms are cut out if they appear more than 80 percent across all papers. An example using the calculation method stated above is as follows: In this document, Cluster 1 is composed of 2253 documents.

Following the formula above, TF for tissue is 1+ (log(1124/484)) = 1.37 and the TF for bone is 1+ (log(989/243)) = 1.61. IDF for tissue is log(2253/484) = 0.67 and for bone is log(2253/243) = 0.97. The TF/IDF for tissue is 1.37*0.67 = 0.92, and for bone, it is 1.61*0.97 = 1.56. As shown in this calculation, although the word tissue appears more frequent than that of bone, because tissue appears more frequently in other documents in Cluster 1, its weight is lower (0.67) than that of bone (0.97), and thus the relative importance of the term bone is higher in this example.

The second question asks how would a researcher reduce the dimensionality of TDM. Because, the TDM is often very large and rather sparse (most of the cells are filled with zeros), this answer is more tractable to handle. While several options are available for reducing such matrices to a manageable size, singular value decomposition (SVD) is a method of representing a matrix as a series of linear approximations that expose the underlying meaning–structure of the matrix. The goal of SVD is to find the optimal set of factors that best predict the outcome. This article adopted the SVD method.

> *Task 4*. The last task is about simplification of the generated TDM to a computer manageable data format. Usually, the TDM is created as a flat file where columns represent the key terms and rows represent the document (in this case, the journal articles). Most data mining algorithms prefer this type of flat-file format; however, some may require the data to be transposed (columns and rows exchanged) before it can be properly processed.

Phase 3: extract the knowledge. Using the well-structured TDM, we then extracted patterns, which are represented in clusters. Clustering is an unsupervised process, whereby objects or events are placed into "natural" groupings. An unsupervised process is one that uses no pattern or prior knowledge to guide the clustering process. The unsupervised clustering process groups an unlabeled collection of objects (e.g. documents, customer comments, and web pages) into meaningful clusters without any prior knowledge. The basic underlying assumption is that relevant documents tend to be more similar to each other than to irrelevant ones. If this assumption holds, the clustering of documents based on the similarity of their content improves search effectiveness.[26]

Finding the "optimal" number of clusters is not an easy task, since there is not a mathematical formula (a closed-form algorithm) developed for it. It is still an experimental heuristic process where the number of clusters are gradually increased from a small number to a large number (or decreased the other way around) to reach a point where the number of clusters are the "optimal" representation of the underlying multi-dimensional dataset. The optimality is determined by measuring the balance between in-cluster similarities and intro-cluster dissimilarities using the Euclidean distance. The Euclidean method is popularly used because of its capability to extract distinct and yet meaningful clusters.[11] This is the heuristic experimental process that we followed in determining a six-cluster representation of the dataset.

Results

Figure 5 shows the most frequently surfaced words/terms in each of the six clusters using a word cloud representation. The counts and percentages of each cluster over time are provided in Figures 6 and 7.

The words in Figure 5 are captured using the entire collection of abstracts of each cluster. As with other cluster analyses, each word in each cluster is not totally independent. For example, the two words of "blood" and "pressure" in Cluster 1 can be two independent words or a combined term. Therefore, the multi-word selections are usually identified as a single unit and are called *terms* in text mining. Figure 6 is a graphical presentation of each cluster by the count of each year. In order to present the relative growth of each cluster over time, the percentage of each cluster is provided in Figure 7.

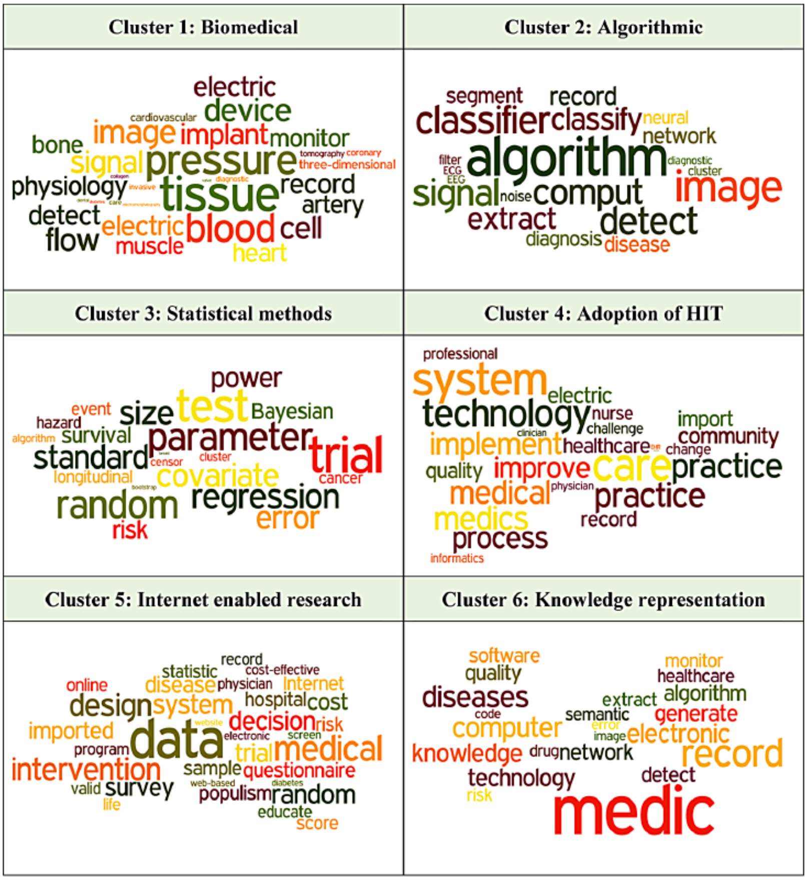


Figure 5. Frequently surfacing words in cluster.

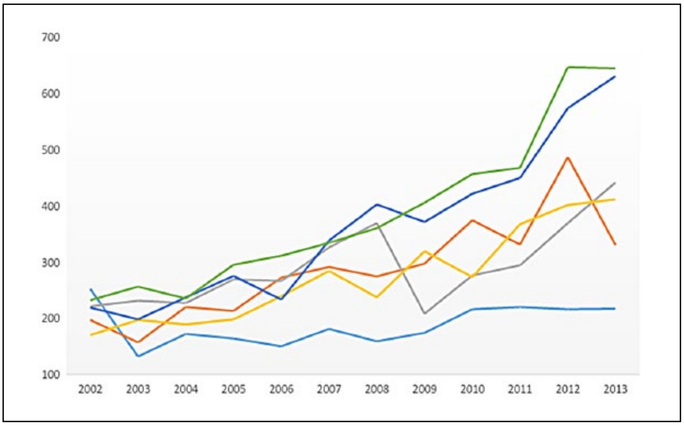


Figure 6. Research trends by count.

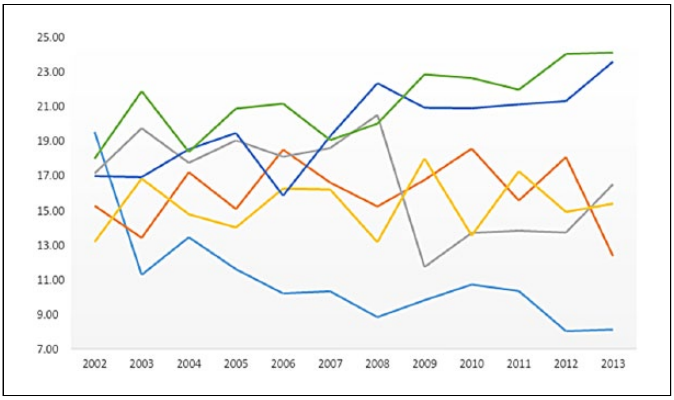


Figure 7. Research trends by percentage.

Cluster 1: biomedical

The most frequently surfaced words in Cluster 1 include tissue, pressure, blood, image, flow, device, and signal. Those words propose that medical informatics research in this cluster directly deals with utilizing the capability of technology to experiment in order to improve patient treatments. More specifically, the term "tissue" is used in the context of tissue differentiation in fractured vertebra with and without fixation devices, the numerical optimization of gene electrotransfer into muscle tissue for minimizing muscle tissue damage, the simulation of tissue differentiation and bone regeneration, and a patient-specific tumor and normal tissue for prediction of the response to radiotherapy.[27–30]

The major context of "pressure" is used in conjunction with blood and artery. Included examples are blood pressure long-term regulation, oscillometric measurement of systolic and diastolic blood pressures, a procedure for the evaluation of non-invasive blood pressure simulators, neural set points for the control of arterial pressure, the estimation of mean arterial pressure from the oscillometric cuff pressure, and the feasibility of measuring coronary blood flow.[31–36]

The term "image" is mainly used in the context of guiding treatments. Some examples include image-guided oral implantology, design of the image-guided biopsy marking system for gastroscopy, extraction of three-dimensional (3D) information on bone remodeling in the proximity of titanium implants in SRmuCT image volumes, 3D patient-specific geometry of the muscles involved in knee motion from selected magnetic resonance imaging (MRI) images, and jaw tissues segmentation in dental 3D computed tomography (CT) images using fuzzy-connectedness and morphological processing.[37–41]

Other frequently appearing words are "electric," "device," and "implant." This category of research includes electromagnetic interference with active implantable devices, electromagnetic interference of cardiac rhythmic monitoring devices to radio frequency identification, implantable devices for long-term electrical stimulation of denervated muscles, and computer methods for automating preoperative dental implant planning.[42–45]

These results reveal that medical informatics research in this category mainly deals with discovering knowledge and improving the ability to treat patients through experimentation. The research by count in this category receives a stable academic interest over time in Figure 5. However, Figure 7 shows that this type of research had the highest percentage of interest (19.52) in 2002, yet its academic interest has decreased since that time. This was especially the case when it abruptly dropped in 2003. In fact, since 2003, the number of publications within this discipline has shown to be the lowest among the six categories. This indicates that the recent application of technology in the medical field is used for much more than just understanding a patients' physical body.

Cluster 2: algorithmic

Cluster 2 captures many computer-aided technical languages such as algorithm, image, signal, comput [compute, computer], classifier, classify, extract, detect, disease, clinic, record, segment, and network. Those clustered words propose that scholars in this cluster use algorithms and neural networks to extract images and detect diseases, and the collected images and recorded data are further classified in order to better diagnose diseases. For example, an algorithm is used to classify images, genes, micro-array data of ovarian cancer, epilepsy diseases, and an effective cardiac arrhythmia.[46–51] Algorithms are also frequently used to analyze signals.[52]

Computerized images are also used to diagnose tumors, cancer, bone disease, and dental treatments. Those collected images are further analyzed and classified. Examples are bone disease classification using collected 3D image analysis and the artificial intelligence diagnosis of dental deformities in cephalometry images using a support vector machine.[53,54] Neural networks or computers are used to record and categorize patterns using the collected and recorded information, which in turn is the basis for data mining analysis and anomaly detections. Examples are electroencephalogram (EEG) recordings for a better description of sleep, automatic classification of long-term ambulatory electrocardiogram (ECG) records according to type of ischemic heart disease, automated detection of neonate EEG sleep stages, epileptic seizure detection in EEGs using time-frequency analysis, central sleep apnea detection from ECG-derived respiratory signals, a detection of Alzheimer's disease using independent component analysis (ICA)-enhanced EEG measurements, and epileptic EEG signal detection using time-frequency distributions.[55–61]

The main subject matter in this cluster is the utilization of algorithms, neural networks, and computational technology to group or categorize diagnoses or symptoms so that one can discover patterns of symptoms and identify anomalies. Figure 5 shows that the number of publications has increased; however, in terms of relative research interests compared to the other clusters in Figure 6, it shows a stable research interest over time.

Cluster 3: statistical methods

Frequently surfaced words in Cluster 3 are trial, test, random, parameter, regression, standard, error, covariate, size, Bayesian, power, risk, and longitudinal. Those clustered words suggest that the medical informatics research in this group mainly deals with various statistical methods applied to medical trials. This may be the reason that "trial" and "test" appear most frequently in this cluster. Also, terms such as standard, error, covariate, size, and power are all closely related to statistical analyses. A few example articles derived from Cluster 3 are further discussed below.

Statistical methods applied to medical informatics research are Bayesian strategies for monitoring clinical trial data, Bayesian analysis of multicenter trial outcomes, Bayesian approach to phase I cancer trials, techniques for incorporating longitudinal measurements into analyses of survival data from clinical trials, estimation and testing based on data subject to measurement errors, regression analysis for multiple-disease group testing data, power and sample size calculation for log-rank testing with a time lag in treatment effect, joint modeling of multivariate longitudinal measurements and survival data with applications to Parkinson's disease, and regression analysis for the examination of the benefits of group testing.[62–69]

The trend of this research group shows an interesting pattern. It steadily increased from around 12 percent in 2002 to 21 percent in 2008 and then rapidly dropped to about 12 percent in 2009; however, the research interest of this group steadily increased since then.

Cluster 4: adoption of HIT

The frequently surfaced words in medical informatics research for Cluster 4 are system, care, technology, implement, improve, practice, process, medics, and medical. As the frequently surfaced words suggest, the research in this cluster mainly deals with the adoption and effectiveness of HIT including electronic patient record systems, electronic prescriptions, data sharing, and electronic reminders in a healthcare setting. A few examples are the determinants of primary care nurses' intention to adopt an electronic health record in their clinical practice, introduction of eHealth to nursing homes, implementation of health information exchange for public health reporting, HIT implementation in critical access hospitals, hospital implementation of HIT and quality of care, improvement of HIT adoption and implementation through the identification of appropriate benefits, integration of a nationally procured electronic health record system into user work practices, and prediction of the influence of the electronic health record on clinical coding practice in hospitals.[70–77]

Another group of frequently surfaced terms are "process" and "improve." A major context of those words is commonly captured with the adoption of HIT: the process of developing technical reports for healthcare decision makers, the role of methodologies in improving the efficiency and effectiveness of care delivery processes, the measurement of healthcare process quality, and a process for the consolidation of redundant documentation forms.[78–81] Also examined is the effectiveness of HIT for patient care delivery, as well as the communications between nurses, physicians, and patients.[82,83]

In sum, this group of research deals with the effective adoption and use of HIT and EMR, and the question of whether or not HIT improves quality. Although some fluctuations are observed within the sample period, they are within the range of 13–18 percent. As such, this topic shows a relatively stable research interest during the research period.

Cluster 5: Internet-enabled research

Cluster 5 is composed of data, medical, intervention, design, decision, system, random, survey, cost, trial, Internet, and online. Based on the frequently surfacing words, this group of medical informatics research utilizes the Internet or online services to improve quality care, treat patients, and use online resources for medical research. More specifically, design and intervention are often used in conjunction with research design; however, unlike Cluster 3, this group of research utilizes the Internet. A few examples are the design of a website on nutrition and physical activity for adolescents, an Internet-based intervention to promote mental fitness for mildly depressed adults using a randomized controlled trial, evaluation of a community-based intervention to enhance breast cancer screening practices, and technology-based interventions for mental health in tertiary students.[84–87]

The Internet is also used for interventions to promote seeking treatment for mental health, online alcohol intervention, and online intervention for a health behavior change campaign.[88,89] Further examined using the Internet in this cluster is sample randomization. Included examples are the accuracy of geographically targeted Internet advertisements on Google AdWords for recruitment in a randomized trial, recruitment to online therapies for depression using a pilot cluster randomized controlled trial, comparison of questionnaires via the internet to pen-and-paper in patients with heart failure, and reach, engagement, and retention in an Internet-based weight loss program in a multi-site randomized controlled trial.[90–93]

Based on the research in this cluster, the Internet and the web revolutionized the ways in which medical informatics' research is conducted. This group of research concerns how the healthcare industry can better leverage the Internet and the web in order to provide better treatments for patients. This group of research consistently increased from 2002 until 2013. Since 2009, it has become one of the two clusters which contain the most prominent topics. This may be attributed to the general public's accessibility of the Internet, their skill sets, the demands from the public to deliver information online, and cost-saving pressures.

Cluster 6: knowledge representation

Cluster 6 includes medic, record, diseases, computer, electronic, knowledge, technology, and network. As the frequently surfacing words propose, this cluster of research mainly deals with strategic use of collected medical data such as EMR/EHR to improve quality and reduce errors. A few commonly appearing subjects are the use of EMR to enhance detection and reporting of vaccine adverse events, the use of EMR data for quality improvement in schizophrenia treatment, and the discovery of notifiable diseases using EMR.[94–97]

Text mining is actively utilized to categorize various diseases. Examples are medical text classification for the Vaccine Adverse Event Reporting System, semantic classification of diseases in discharge summaries using a context-aware rule-based classifier, and automated evaluation of electronic discharge notes to assess quality of care for cardiovascular diseases using Medical Language Extraction and Encoding System (MedLEE).[98–100] Also, using EHR clinical notes on heart-related symptoms (the Framingham heart failure diagnostic criteria),[101] used a text mining technique to detect early heart failure and improved the understanding of the early detection of heart failure. Data mining is also used for clinical oral health documents to analyze the longevity of different restorative materials.[102] Some of the articles explicitly use the term knowledge discovery. Examples are knowledge-based biomedical word sense disambiguation using document classification, the role of domain knowledge in automating medical text report classification, Mayo clinical text analysis and knowledge extraction system, a knowledge discovery and reuse pipeline for information extraction in clinical notes, and providing concept-oriented views for clinical data using a knowledge-based system.[103–107]

In sum, this cluster of research deals with utilizing EMR/EHR for categorizing or discovering treatment-related knowledge. Also, data and text mining is used in order to discover knowledge buried within text and data. Like Cluster 1, this group of research strives to find better treatments, but this research group utilizes EMR/EHR. This group of research also adopts text and data mining techniques to categorize diseases. The academic interest of this cluster is consistently increasing as each year goes by and has become one of the two most researched topics since 2008.

Discussion

Interestingly enough, the counts and the percentages of each cluster in Figures 5 and 6 between 2002 and 2004 did not show notable differences across clusters; however, distinctively different patterns among clusters began to emerge in 2005. The most notable publication increases are in the clusters, *Internet-enabled research* and *knowledge representation*. The clusters, *algorithmic, statistical methods*, and *adoption of HIT*, are somewhat stable in this research time period, and it may be because there is no momentum taking place like there is in *Internet-enabled research* and *knowledge representation*, which are spurred on as a result of the rapid adoption of EMR/EHR and a widespread use of the Internet. Furthermore, the research in *Internet-enabled research* incorporates the traditional research methods in *statistical methods* by leveraging the power of the Internet and web capabilities. The research in *knowledge representation* somewhat integrates research in *biomedical* and *algorithmic*. More specifically, the research in *knowledge representation* focuses on better diagnoses and treatments using collected data, while research in *biomedical* advances medical knowledge via experimentation using individual cases. *Adoption of HIT* mainly concerns the adoption and implementation of HIT whose research interests are now stabilized as HIT in the healthcare industry matures. As *Internet-enabled research* and *knowledge representation*, which are the utilization of Internet and EMR/EHR, have each recently gained a heightened interest among medical informatics scholars, the focus of discussion will be given to those two clusters.

Knowledge representation shows the most notable increase in publications in medical informatics. The increasing adoption of EMR/EHR associates with many factors for this increase. First, a requirement of adopting EMR/EHR by the American Recovery and Reinvestment Act of 2009 may have contributed to the growth of the cluster of *knowledge representation*. According to the Act, healthcare providers should adopt a form of EMR/EHR by 2014 and imposes penalties for noncompliance.[108] As a result, the EMR adoption rate increased by 31 percent over the period from 2001 to 2005 and by 50 percent over the period from 2005 to 2008.[109,110] Most recently, 71.8 percent of office-based physicians reported the adoption of an EHR system in 2012, which is up from 34.8 percent in 2007.[111] The American Recovery and Reinvestment Act laid the foundation for a transition to more outcome-based reimbursement (i.e. a "pay-for-performance" model), as opposed to the traditional "fee-for-service" model.[112] Outcome-based reimbursement is closely related to evidence-based reimbursement.[113] Massive health data collected from EMR/EHR offers a promising approach to personalized healthcare and evidence-based treatment.[114–116] Because of the recent demands (e.g. evidence-based treatment and cost-effectiveness) and the availability of technology and medical data, these groups of research are expected to grow.

Internet-enabled research, the adoption and utilization of the Internet and the web, gained great academic interest. It may be because the Internet has become a ubiquitous tool for gathering information. The Pew Research Internet Project in 2014 reported that Internet usage was 14 percent in 1995, 61 percent in 2003, and 87 percent in 2014.[117] As a result of widespread Internet use, the ways that the healthcare industry conducts research and interacts with stakeholders (e.g. patients, pharmaceuticals, nurses, and staff) have fundamentally changed. Computer literate individuals prefer online communications and gain health benefits from online information delivery systems.[118] As such, the healthcare industry increasingly relies on the Internet as a mode of health-related communication with their patients, which in turn increases research interests.

Conclusion and future research

This study explored the trends of medical informatics research using articles published between 2002 and 2013. The main objective of this study was to understand where the field of medical informatics has been, where it is heading, and identify a boundary of the field as its subject area has matured as an academic field. According to the findings, as with any other field that goes through different academic interests in different times, an increase or decrease in medical informatics publications corresponds to the needs of the field and the opportunities enabled by the capabilities of HIT. Despite the advantages, the utilization of EMR/EHR data is still in a nascent stage that is necessary for the support of evidence-based medicine,[7] and thus we anticipate continual and rapid growth of Internet-based and data-driven, evidence-based research.

The purpose of this study was to identify the scope and the trend within this field. In order to achieve these research objectives, we focused on big streams of research. As such, for future research, it is recommended to investigate smaller clusters, which can provide insights on emerging fields of study.

The sample of this study is drawn from the major medical informatics journals provided by the ISI's Web of Knowledge JCR 2012. Because smaller journals often publish innovative or non-traditional ideas about the field, in future studies, it is recommended to include smaller journals in the field and investigate how those journals integrate new emerging ideas, and by doing so, define/redefine the field.

This study did not include the statistical significance of the changes in publications over the study period in order to focus on the identification of the scope and boundary of the field. It will be, however, an interesting idea to calculate statistical significances of the publication changes over the time period and discover what drives such changes.

This article has some limitations. Because we chose 23 journals as a representative sample, the research findings do not collectively represent all medical informatics journals. Therefore, readers need to be cautious when they apply these research findings to a bigger sample or a different sample drawing.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Haux R. Medical informatics: past, present, future. *Int J Med Inform* 2010; 79: 599–610.
2. Dalrymple P and Roderer NK. Education for health information professionals: perspectives from health informatics in the U.S. *Edu Inform* 2010–2011; 28: 45–55.
3. Weigel FK, Rainer RK, Hazen BT, et al. Uncovering research opportunities in the medical informatics field: a quantitative content analysis. *Comm Assoc Inform Syst* 2013; 33(2): 15–32.
4. Masic I. Five periods in development of medical informatics. *Acta Inform Med* 2014; 22(1): 44–48.
5. Tolar M and Balka E. Caring for individual patients and beyond: enhancing care through secondary use of data in a general practice setting. *Int J Med Inform* 2012; 81: 461–474.
6. Spasic I, Livsey J, Keane JA, et al. Text mining of cancer-related information: review of current status and future directions. *Int J Med Inform* 2014; 83: 605–623.
7. National Research Council. *Computational technology for effective healthcare: immediate steps and strategic directions*. Washington, DC: National Academies Press, 2009.
8. DeShazo J, LaVallie D and Wolf F. Publication trends in the medical informatics literature: 20 years of "medical informatics" in MeSH. *BMC Med Inform Decis Mak* 2009; 9(7): 1–13.
9. Jiang LC, Wang ZZ, Peng TQ, et al. The divided communities of shared concerns: mapping the intellectual structure of e-Health research in social science journals. *Int J Med Inform* 2015; 84: 24–35.
10. Jiang X, Tse K, Wang S, et al. Recent trends in biomedical informatics: a study based on JAMIA articles. *J Am Med Inform Assoc* 2013; 20(e2): e198–e205.
11. Schuemie MJ, Talmon JL, Moorman PW, et al. Mapping the domain of medical informatics. *Methods Inf Med* 2009; 48: 76–83.
12. Kim HE, Jiang X, Kim J, et al. Trends in biomedical informatics: most cited topics from recent years. *J Am Med Inform Assoc* 2011; 18(Suppl. 1): i166–i170.
13. Delen D and Crossland M. Seeding the survey and analysis of research literature with text mining. *Expert Syst Appl* 2008; 34: 1707–1720.

14. Miller TW. *Data and text mining: a business applications approach*. Upper Saddle River, NJ: Pearson/Prentice Hall, 2005.
15. Romero C and Ventura S. Educational data mining: a survey from 1995 to 2005. *Expert Syst Appl* 2007; 33(1): 135–146.
16. Journal of Medical Internet Research. JMIR now ranked the #1 health informatics journal, #2 in health care services & sciences category, by Impact Factor, http://www.jmir.org/announcement/view/24 (accessed 20 October 2014).
17. Jamal A, McKenzie K and Clark M. The impact of health information technology on the quality of medical and health care: a systematic review. *HIM J* 2009; 38(3): 26–37.
18. Schoen C, Osborn R, Huynh PT, et al. On the front lines of care: primary care doctors' office systems, experiences, and views in seven countries. *Health Aff* 2006; 25(6): 555–571.
19. Hillestad R, Bigelow J, Bower A, et al. Can electronic medical record systems transform healthcare? Potential health benefits, savings and cost. *Health Aff* 2005; 24(5): 1103–1117.
20. Georgiou A. Data, information and knowledge: the health informatics model and its role in evidence-based medicine. *J Eval Clin Pract* 2002; 8(2): 127–130.
21. Smith M, Halvorson G and Kaplan G. What's needed is a health care system that earns: recommendations from an IOM report. *JAMA* 2012; 308(16): 1637–1638.
22. National Center for Biotechnology Information (NCBI). NLM catalog: journals referenced in the NCBI databases, http://www.ncbi.nlm.nih.gov/nlmcatalog/journals (accessed 10 May 2014).
23. Delen D. *Real-world data mining: applied business analytics and decision making*. Upper Saddle River, NJ: Pearson Education, 2015.
24. Miner G, Delen D, Fast A, et al. *Practical text mining and statistical analysis for non-structured text data applications*. Oxford: Academic Press, 2013.
25. Manning CD and Schütze H. *Foundations of statistical natural language processing*. Newport Beach, CA: MIT Press, 1999.
26. Feldman R and Sanger J. *The text mining handbook: advanced approaches in analyzing unstructured data*. New York: Cambridge University Press, 2007.
27. Boccaccio A, Kelly DJ and Pappalettere C. A model of tissue differentiation and bone remodelling in fractured vertebrae treated with minimally invasive percutaneous fixation. *Med Biol Eng Comput* 2012; 50(9): 947–959.
28. Zupanic A, Corovic S, Miklavcic D, et al. Numerical optimization of gene electrotransfer into muscle tissue. *Biomed Eng Online* 2010; 9: 66.
29. Lacroix D, Prendergast PJ, Li G, et al. Biomechanical model to simulate tissue differentiation and bone regeneration: application to fracture healing. *Med Biol Eng Comput* 2002; 40(1): 14–21.
30. Stamatakos G, Antipas VP and Ozunoglu NK. A patient-specific in vivo tumor and normal tissue model for prediction of the response to radiotherapy. *Methods Inf Med* 2007; 46(3): 367–375.
31. Zanutto BS, Frias BC and Valentinuzzi ME. Blood pressure long term regulation: a neural network model of the set point development. *Biomed Eng Online* 2011; 10: 54.
32. Babbs CF. Oscillometric measurement of systolic and diastolic blood pressures validated in a physiologic mathematical model. *Biomed Eng Online* 2012; 11: 56.
33. Gersak G, Zemva A and Drnovsek J. A procedure for evaluation of non-invasive blood pressure simulators. *Med Biol Eng Comput* 2009; 47(12): 1221–1228.
34. Sandoz B, Badina A, Laporte S, et al. Quantitative geometric analysis of rib, costal cartilage and sternum from childhood to teenagehood. *Med Biol Eng Comput* 2013; 51(9): 971–979.
35. Zheng D, Amoore JN, Mieke S, et al. Estimation of mean arterial pressure from the oscillometric cuff pressure: comparison of different techniques. *Med Biol Eng Comput* 2011; 49(1): 33–39.
36. Baltes C, Kozerke S and Boesiger P. Coronary flow quantification by Fourier velocity encoded MRI. *Biomed Tech* 2002; 47(Suppl. 1): 412–415.
37. Xiaojun C, Ming Y, Yanping L, et al. Image guided oral implantology and its application in the placement of zygoma implants. *Comput Methods Programs Biomed* 2009; 93(2): 162–173.
38. Sun D, Hu W, Wu W, et al. Design of the image-guided biopsy marking system for gastroscopy. *J Med Syst* 2012; 36(5): 2909–2920.

39. Sarve H, Lindblad J, Borgefors G, et al. Extracting 3D information on bone remodeling in the proximity of titanium implants in SRmuCT image volumes. *Comput Methods Programs Biomed* 2011; 102(1): 25–34.
40. Sudhoff I, De Guise JA, Nordez A, et al. 3D-patient-specific geometry of the muscles involved in knee motion from selected MRI images. *Med Biol Eng Comput* 2009; 47(6): 579–587.
41. Llorens R, Naranjo V, Lopez F, et al. Jaw tissues segmentation in dental 3D CT images using fuzzy-connectedness and morphological processing. *Comput Methods Programs Biomed* 2012; 108(2): 832–843.
42. Angeloni A, Barbaro V, Bartolini P, et al. A novel heart/trunk simulator for the study of electromagnetic interference with active implantable devices. *Med Biol Eng Comput* 2003; 41(5): 550–555.
43. Ogirala A, Stachel JR and Mickle MH. Electromagnetic interference of cardiac rhythmic monitoring devices to radio frequency identification: analytical analysis and mitigation methodology. *IEEE Trans Inf Technol Biomed* 2011; 15(6): 848–853.
44. Lanmuller H, Ashley Z, Unger E, et al. Implantable device for long-term electrical stimulation of denervated muscles in rabbits. *Med Biol Eng Comput* 2005; 43(4): 535–540.
45. Galanis CC, Sfantsikopoulos MM, Koidis PT, et al. Computer methods for automating preoperative dental implant planning: implant positioning and size assignment. *Comput Methods Programs Biomed* 2007; 86(1): 30–38.
46. Albrecht A, Hein E, Steinhofel K, et al. Bounded-depth threshold circuits for computer-assisted CT image classification. *Artif Intell Med* 2002; 24(2): 179–192.
47. Chandra B and Gupta M. An efficient statistical feature selection approach for classification of gene expression data. *J Biomed Inform* 2011; 44(4): 529–535.
48. Gamberger D, Lavrac N, Zelezny F, et al. Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. *J Biomed Inform* 2004; 37(4): 269–284.
49. Lee ZJ. An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer. *Artif Intell Med* 2008; 42(1): 81–93.
50. Kocer S and Canal MR. Classifying epilepsy diseases using artificial neural networks and genetic algorithm. *J Med Syst* 2011; 35(4): 489–498.
51. Asl BM, Setarehdan SK and Mohebbi M. Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal. *Artif Intell Med* 2008; 44(1): 51–64.
52. Keshri AK, Das BN, Mallick DK, et al. Parallel algorithm to analyze the brain signals: application on epileptic spikes. *J Med Syst* 2011; 35(1): 93–104.
53. Akgundogdu A, Jennane R, Aufort G, et al. 3D image analysis and artificial intelligence for bone disease classification. *J Med Syst* 2010; 34(5): 815–828.
54. Banumathi A, Raju S and Abhaikumar V. Diagnosis of dental deformities in cephalometry images using support vector machine. *J Med Syst* 2011; 35(1): 113–119.
55. Lewandowski A, Rosipal R and Dorffner G. Extracting more information from EEG recordings for a better description of sleep. *Comput Methods Programs Biomed* 2012; 108(3): 961–972.
56. Smrdel A and Jager F. Automatic classification of long-term ambulatory ECG records according to type of ischemic heart disease. *Biomed Eng Online* 2011; 10: 107.
57. Piryatinska A, Terdik G, Woyczynski WA, et al. Automated detection of neonate EEG sleep stages. *Comput Methods Programs Biomed* 2009; 95(1): 31–46.
58. Tzallas AT, Tsipouras MG and Fotiadis DI. Epileptic seizure detection in EEGs using time-frequency analysis. *IEEE Trans Inf Technol Biomed* 2009; 13(5): 703–710.
59. Maier C and Dickhaus H. Central sleep apnea detection from ECG-derived respiratory signals. Application of multivariate recurrence plot analysis. *Methods Inf Med* 2010; 49(5): 462–466.
60. Melissant C, Ypma A, Frietman EE, et al. A method for detection of Alzheimer's disease using ICA-enhanced EEG measurements. *Artif Intell Med* 2005; 33(3): 209–222.
61. Guerrero-Mosquera C, Trigueros AM, Franco JI, et al. New feature extraction approach for epileptic EEG signal detection using time-frequency distributions. *Med Biol Eng Comput* 2010; 48(4): 321–330.
62. Dmitrienko A and Wang MD. Bayesian predictive approach to interim monitoring in clinical trials. *Stat Med* 2006; 25(13): 2178–2195.

63. Gould AL. Bayesian analysis of multicentre trial outcomes. *Stat Methods Med Res* 2005; 14(3): 249–280.
64. Neuenschwander B, Branson M and Gsponer T. Critical aspects of the Bayesian approach to phase I cancer trials. *Stat Med* 2008; 27(13): 2420–2439.
65. Troxel AB. Techniques for incorporating longitudinal measurements into analyses of survival data from clinical trials. *Stat Methods Med Res* 2002; 11(3): 237–245.
66. Vexler A, Tsai WM and Malinovsky Y. Estimation and testing based on data subject to measurement errors: from parametric to non-parametric likelihood methods. *Stat Med* 2012; 31(22): 2498–2512.
67. Zhang B, Bilder CR and Tebbs JM. Regression analysis for multiple-disease group testing data. *Stat Med* 2013; 32(28): 4954–4966.
68. Zhang D and Quan H. Power and sample size calculation for log-rank test with a time lag in treatment effect. *Stat Med* 2009; 28(5): 864–879.
69. He B and Luo S. Joint modeling of multivariate longitudinal measurements and survival data with applications to Parkinson's disease. *Stat Methods Med Res* 2016; 25: 1346–1358.
70. Leblanc G, Gagnon MP and Sanderson D. Determinants of primary care nurses' intention to adopt an electronic health record in their clinical practice. *Comput Inform Nurs* 2012; 30(9): 496–502.
71. Breen GM and Zhang NJ. Introducing ehealth to nursing homes: theoretical analysis of improving resident care. *J Med Syst* 2008; 32(2): 187–192.
72. Merrill JA, Deegan M, Wilson RV, et al. A system dynamics evaluation model: implementation of health information exchange for public health reporting. *J Am Med Inform Assoc* 2013; 20(e1): e131–e128.
73. Bahensky JA, Ward MM, Nyarko K, et al. HIT implementation in critical access hospitals: extent of implementation and business strategies supporting IT use. *J Med Syst* 2011; 35(4): 599–607.
74. Sa Couto J. Project management can help to reduce costs and improve quality in health care services. *J Eval Clin Pract* 2008; 14(1): 48–52.
75. Leonard KJ and Sittig DF. Improving information technology adoption and implementation through the identification of appropriate benefits: creating IMPROVE-IT. *J Med Internet Res* 2007; 9(2): e9.
76. Cresswell KM, Worth A and Sheikh A. Integration of a nationally procured electronic health record system into user work practices. *BMC Med Inform Decis Mak* 2012; 12: 15.
77. Robinson K and Shepheard J. Predicting the influence of the electronic health record on clinical coding practice in hospitals. *HIM J* 2004; 32(3): 102–108.
78. Patwardhan MB, Sarria-Santamera A and Matchar DB. Improving the process of developing technical reports for health care decision makers: using the theory of constraints in the evidence-based practice centers. *Int J Technol Assess Health Care* 2006; 22(1): 26–32.
79. Stefanelli M. The role of methodologies to improve efficiency and effectiveness of care delivery processes for the year 2013. *Int J Med Inform* 2002; 66: 39–44.
80. Yildiz O and Demirors O. Measuring healthcare process quality: applications in public hospitals in Turkey. *Inform Health Soc Care* 2013; 38(2): 132–149.
81. Cowden S and Johnson LC. A process for consolidation of redundant documentation forms. *Comput Inform Nurs* 2004; 22(2): 90–93.
82. Georgiou A, Prgomet M, Markewycz A, et al. The impact of computerized provider order entry systems on medical-imaging services: a systematic review. *J Am Med Inform Assoc* 2011; 18(3): 335–340.
83. Edwards A, Fitzpatrick LA, Augustine S, et al. Synchronous communication facilitates interruptive workflow for attending physicians and nurses in clinical settings. *Int J Med Inform* 2009; 78(9): 629–637.
84. Thompson D, Cullen KW, Boushey C, et al. Design of a website on nutrition and physical activity for adolescents: results from formative research. *J Med Internet Res* 2012; 14(2): e59.
85. Bolier L, Haverman M, Kramer J, et al. An internet-based intervention to promote mental fitness for mildly depressed adults: randomized controlled trial. *J Med Internet Res* 2013; 15(9): e200.
86. Thuler LC and Freitas HG. Evaluation of a community-based intervention to enhance breast cancer screening practices in Brazil. *J Eval Clin Pract* 2008; 14(6): 1012–1017.

87. Farrer L, Gulliver A, Chan JK, et al. Technology-based interventions for mental health in tertiary students: systematic review. *J Med Internet Res* 2013; 15(5): e101.
88. Gulliver A, Griffiths KM, Christensen H, et al. Internet-based interventions to promote mental health help-seeking in elite athletes: an exploratory randomized controlled trial. *J Med Internet Res* 2012; 14(3): e69.
89. Cugelman B, Thelwall M and Dawes P. Online interventions for social marketing health behavior change campaigns: a meta-analysis of psychological architectures and adherence factors. *J Med Internet Res* 2011; 13(1): e17.
90. Jones RB, Goldsmith L, Williams CJ, et al. Accuracy of geographically targeted internet advertisements on Google AdWords for recruitment in a randomized trial. *J Med Internet Res* 2012; 14(3): e84.
91. Jones RB, Goldsmith L, Hewson P, et al. Recruitment to online therapies for depression: pilot cluster randomized controlled trial. *J Med Internet Res* 2013; 15(3): e45.
92. Wu RC, Thorpe K, Ross H, et al. Comparing administration of questionnaires via the internet to pen-and-paper in patients with heart failure: randomized controlled trial. *J Med Internet Res* 2009; 11(1): e3.
93. Glasgow RE, Nelson CC, Kearney KA, et al. Reach, engagement, and retention in an Internet-based weight loss program in a multi-site randomized controlled trial. *J Med Internet Res* 2007; 9(2): e11.
94. Hinrichsen VL, Kruskal B, O'Brien MA, et al. Using electronic medical records to enhance detection and reporting of vaccine adverse events. *J Am Med Inform Assoc* 2007; 14(6): 731–735.
95. Owen RR, Thrush CR, Cannon D, et al. Use of electronic medical record data for quality improvement in schizophrenia treatment. *J Am Med Inform Assoc* 2004; 11(5): 351–357.
96. Decullier E, Dupuis-Girod S, Plauchu H, et al. How to improve specific databases for clinical data in rare diseases? The example of hereditary haemorrhagic telangiectasia. *J Eval Clin Pract* 2012; 18(3): 523–527.
97. Lazarus R, Klompas M, Campion FX, et al. Electronic Support for Public Health: validated case finding and reporting for notifiable diseases using electronic medical data. *J Am Med Inform Assoc* 2009; 16(1): 18–24.
98. Botsis T, Nguyen MD, Woo EJ, et al. Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. *J Am Med Inform Assoc* 2011; 18(5): 631–638.
99. Solt I, Tikk D and Gal V. Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier. *J Am Med Inform Assoc* 2009; 16(4): 580–584.
100. Chiang JH, Lin JW and Yang CW. Automated evaluation of electronic discharge notes to assess quality of care for cardiovascular diseases using Medical Language Extraction and Encoding System (MedLEE). *J Am Med Inform Assoc* 2010; 17(3): 245–252.
101. Byrda RJ, Steinhublb SR, Suna J, et al. Automatic identification of heart failure diagnostic criteria,using text analysis of clinical notes from electronic healthrecords. *Int J Med Inform* 2014; 83: 983–992.
102. Kakilehto T, Saloa S and Larmasa M. Data mining of clinical oral health documents for analysis of the longevity of different restorative materials in Finland. *Int J Med Inform* 2009; 78: e68–e74.
103. Garla VN and Brandt C. Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. *J Am Med Inform Assoc* 2013; 20(5): 882–886.
104. Wilcox AB and Hripcsak G. The role of domain knowledge in automating medical text report classification. *J Am Med Inform Assoc* 2003; 10(4): 330–338.
105. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17(5): 507–513.
106. Patrick JD, Nguyen DH, Wang Y, et al. A knowledge discovery and reuse pipeline for information extraction in clinical notes. *J Am Med Inform Assoc* 2011; 18(5): 574–579.
107. Zeng Q, Cimino JJ and Zou KH. Providing concept-oriented views for clinical data using a knowledge-based system: an evaluation. *J Am Med Inform Assoc* 2002; 9(3): 294–305.
108. University Alliance. Federal mandates for healthcare: digital record-keeping will be required of public and private healthcare providers, 8 February 2013, http://www.usfhealthonline.com/news/healthcare/electronic-medical-records-mandate-january-2014/#.VG9ULE10w5s (accessed 21 November 2014).

109. Burt CW, Hing E and Woodwell D; National Center for Health Statistics. Electronic medical record use by office-based physicians: United States, 2005, http://www.cdc.gov/nchs/data/hestat/electronic/electronic.htm (accessed 10 February 2015).
110. DesRoches CM, Campbell EG and Rao SR. Electronic health records in ambulatory care—a national survey of physicians. *N Engl J Med* 2008; 359(1): 50–60.
111. Hsiao C, Hing E and Ashman J. *Trends in electronic health record system use among office-based physicians: United States, 2007–2012*. National Health Statistics Reports No 75, 20 May 2014, pp. 1–18, http://www.cdc.gov/nchs/data/nhsr/nhsr075.pdf
112. Glaser J. HITECH lays the foundation for more ambitious outcomes-based reimbursement. *Am J Manag Care* 2010; 16: 19–23.
113. Diamond GA and Kaul S. Evidence-based financial incentives for healthcare reform. *Circ Cardiovasc Qual Outcomes* 2009; 2: 134–140.
114. Kerr WT, Lau EP, Owens GE, et al. The future of medical diagnostics: large digitized databases. *Yale J Biol Med* 2012; 85: 363–377.
115. Chawla NV and Davis DA. Bringing big data to personalized healthcare: a patient-centered framework. *J Gen Intern Med* 2013; 28(Suppl. 3): S660–S665.
116. Simpao AF, Ahumada LM, Gálvez JA, et al. A review of analytics and clinical informatics in health care. *J Med Syst* 2014; 38(4): 1–7.
117. Pew Research Center. Internet use over time, http://www.pewinternet.org/ (accessed 10 January 2015).
118. Kim Y. Is seeking health information online different from seeking general information online? *J Inf Sci* 2015; 41(2): 228–241.