Health Informatics Journal

Original Article

Medical informatics research trend analysis: A text mining approach

Health Informatics Journal
1–21

DOI: 10.1177/1460458216678443
jhi.sagepub.com

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

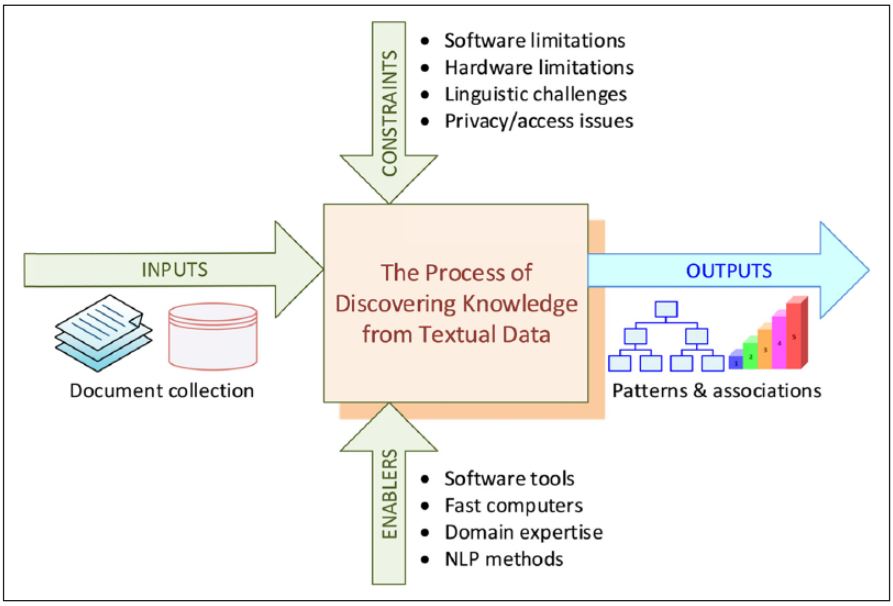

Figure 1. High-level context diagram for text mining of published literature.

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,

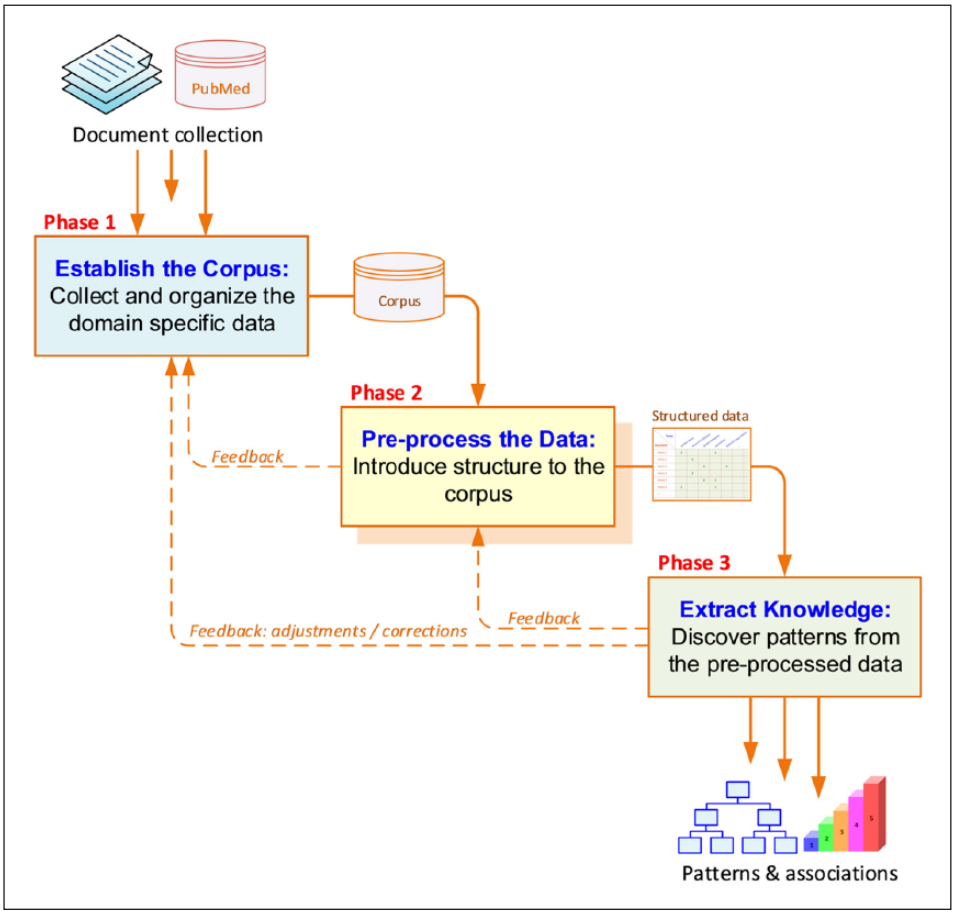

Figure 2. Process for text mining and knowledge discovery.

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

Documents \ Terms	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.

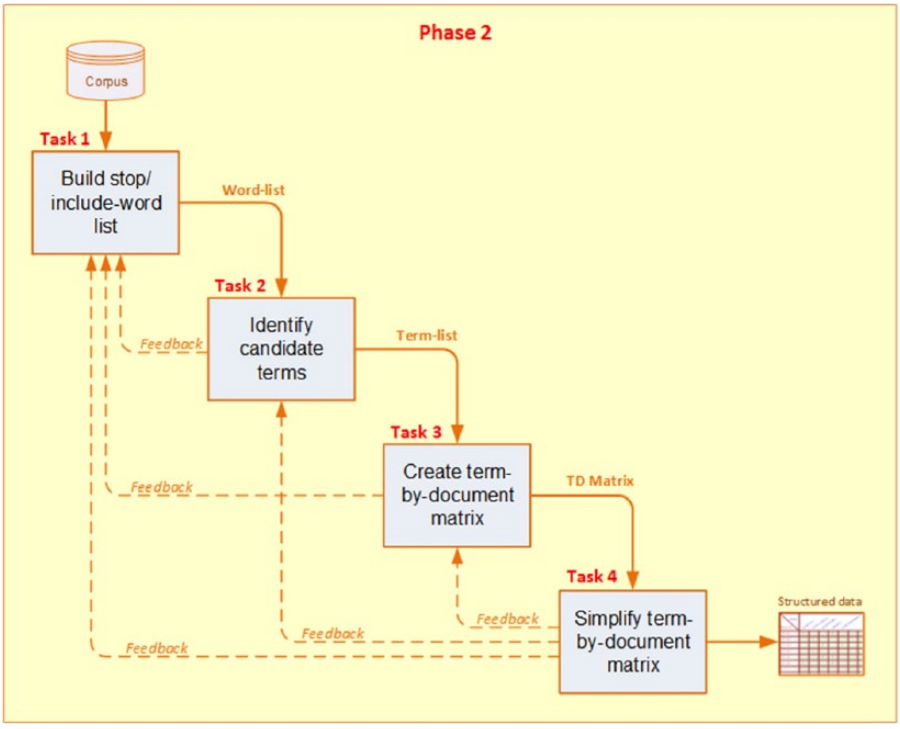

Figure 4. Decomposition of "pre-processing the data" phase.

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.

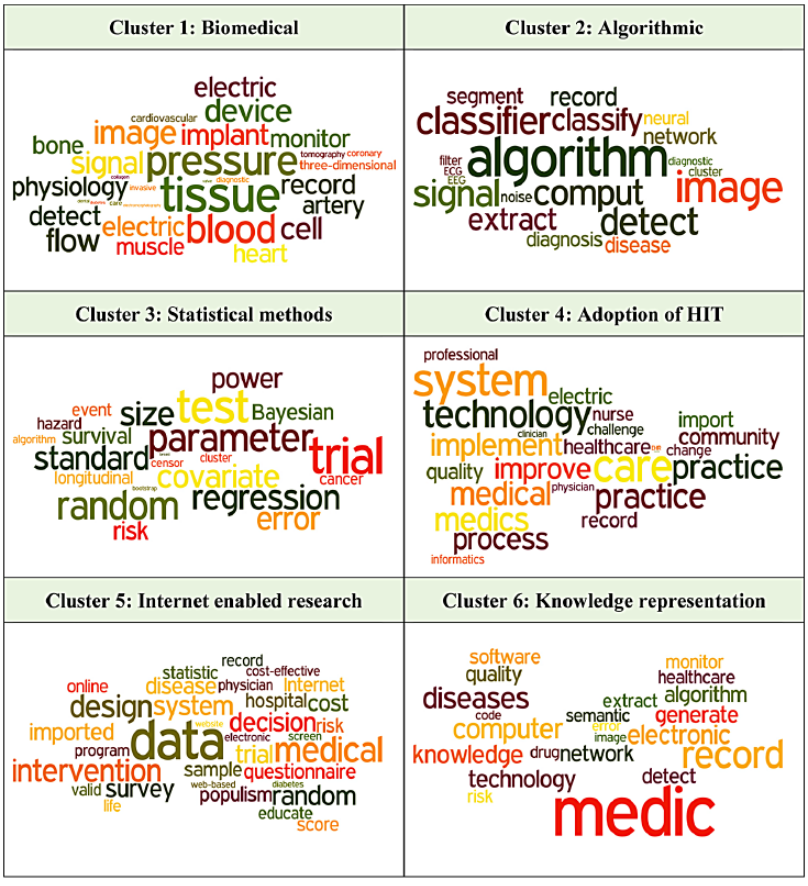

Figure 5. Frequently surfacing words in cluster.

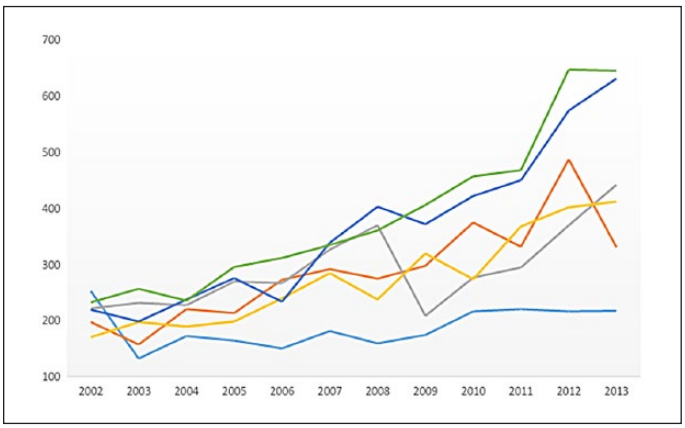

Figure 6. Research trends by count.

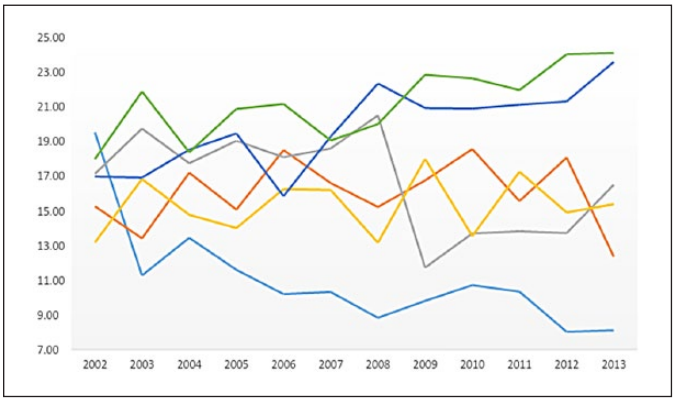

Figure 7. Research trends by percentage.

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.