

Clinical Information Systems and Artificial Intelligence: Recent Research Trends

Carlo Combi¹, Giuseppe Pozzi²

¹ Dipartimento di Informatica, Università degli Studi di Verona, Verona, Italy

² Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy

Summary

Objectives: This survey aims at reviewing the literature related to Clinical Information Systems (CIS), Hospital Information Systems (HIS), Electronic Health Record (EHR) systems, and how collected data can be analyzed by Artificial Intelligence (AI) techniques.

Methods: We selected the major journals (11 journals) collecting papers (more than 7,000) over the last five years from the top members of the research community, and read and analyzed the papers (more than 200) covering the topics. Then, we completed the analysis using search engines to also include papers from major conferences over the same five years.

Results: We defined a taxonomy of major features and research areas of CIS, HIS, EHR systems. We also defined a taxonomy for the use of Artificial Intelligence (AI) techniques on healthcare data. In the light of these taxonomies, we report on the most relevant papers from the literature.

Conclusions: We highlighted some major research directions and issues which seem to be promising and to need further investigations over a medium- or long-term period.

Keywords

Survey; clinical information systems; hospital information systems; electronic health records; artificial intelligence; big data analysis

Yearb Med Inform 2019;83-94

<http://dx.doi.org/10.1055/s-0039-1677915>

1 Introduction

Clinical Information Systems (CIS) [1] can be defined as the overall set of resources, techniques, devices, and methodologies that are used by different healthcare and medical organizations in order to support the knowledge needs required to reach the clinical goals of the organization themselves. More particularly, clinical information systems focus explicitly on the informational needs related to clinical activities, as hospitalizations (admission, discharge, transfer), management of patients' follow up, and disease prevention. Generally speaking, a clinical information system is often the part of the wider Hospital Information System (HIS) [1] that specifically focuses on the information needs related to the patient care.

Patient care consists of a set of complex activities usually intertwined and decision-making oriented. CIS have to support the integrated management of both patients' clinical information and clinicians' decision-related information, as diagnoses and prescribed therapies. Moreover, CIS have to store all the information related to the clinical activities performed by healthcare stakeholders. Collecting big data in healthcare differs from collecting big data in other application domains: *e.g.*, Viceconti *et al.* in [2] highlight that the extension to healthcare of the classical 5V definition (Volume, Variety, Velocity, Veracity, Value) of big data moves from problem solving to knowledge discovery.

Thus, with respect to more traditional information systems, knowledge- and data-intensive tasks need to be supported by a CIS, even when considering non-alphanumeric data. On the other hand, clinical

information systems can be considered as a potential source of (new) clinical knowledge both for the best (evidence-based) clinical tasks, such as identifying care flows, validating guidelines and recommendations, or verifying the compliance of a treatment to a research protocol, and for the identification of specific clinical contexts with respect to a given population of patients, such as extracting sub-cohorts of patients from a wider cohort.

Recently, the term “Artificial Intelligence” (AI) attracted attention in many different and popular domains. Among them, medicine has always been considered as an important application domain for AI-based solutions, both for medical knowledge representation and new knowledge extraction from stored clinical data, and for the support to clinical decision making or reasoning. According to Combi [3], artificial intelligence in medicine may be characterized as the scientific discipline pertaining to research studies, projects, and applications that aim at supporting decision-based medical tasks through knowledge- and/or data-intensive computer-based solutions that ultimately support and improve the performance of a human care provider.

According to this overall scenario, the application of AI-based approaches in the design, extension, and management of clinical information systems gained some attention in the recent years.

The overall goal of this paper is to survey the most recent research approaches considering the application of AI-based techniques within CIS. Without any attempt to be complete, we focus on some interesting research directions we found in the recent literature, and discuss some possible future trends.

The following sections introduce and discuss the main specific goals of this work (Section 2), the methods we used for identifying and analyzing the relevant literature (Section 3), the results of our analysis (Section 4), and a final discussion and conclusions (Section 5).

2 Objectives

The main goal of the current survey is that of reviewing the research area of CIS, Health Information Systems (HIS), and Electronic Health Record (EHR) systems, with a particular focus on how data can be not merely stored and retrieved in a patient-oriented manner, typically when the patient is hospitalized, re-hospitalized, or shows up for a periodical follow up visit, but also exploited for knowledge discovery. Stored data can also be successfully exploited and analyzed by AI techniques, aimed at inferring new knowledge about diseases, therapies, clinical pathways, readmission prediction, genomics – just to mention a few of them. Moreover, considered data are not just traditional alpha-numerical ones (*e.g.*, physician's reports, lab tests), but also some complex ones (*e.g.*, bio-signal recordings, bio-medical images, genomic data).

The literature already presents several survey papers - *e.g.*, [4, 5] - on topics close to those we consider here. However, given the rapidly evolving scenario, the recent advances of new techniques (some of them applied for the first time to the health domain), and new application domains, an updated survey is perceived as necessary. The temporal span we consider here is that of the last five years, namely from 2014 to date (2018).

Moving from the collected information, we sketch out a brief taxonomy of major, ongoing research activities described by the literature, so that any scientist can identify the main stream he/she is in, as well as which are the neighbor streams. At the end, we identify the most populated streams, as well as the most promising ones, research directions, also turning the lights on over expectations of results from future medium-term and long-term research activities.

3 Methods

Figure 1 summarizes the activities that set up the process we followed to select the papers to survey using an UML (Unified Modeling Language) Activity Diagram. The same activity diagram can also be depicted by a PRISMA flow diagram [6].

The first step (namely, “Select Major Journals” in Figure 1) we adopted in considering the literature on the topics of CIS, HIS, and EHR is that of selecting the major journals, *i.e.* those journals where the leading scientists of the scientific community publish scientific papers. We search Google by the keywords CIS, HIS, and EHR and sort results by journal count. Some of these top journals have a pure computer-science or computer-engineering approach and are more methodology-oriented; some other top journals have a bio-medical or bio-informatics approach, and are more application-oriented; some more other top journals have a pure, unique medical approach, *i.e.*, those focusing on clinical aspects and not on technological aspects, and they are not considered herein.

As the second step (namely, “Inspect the Issues and Extract papers on CIS, HIS, EHR” in Figure 1), we read and analyzed the list of all the papers (more than 7000) published over the last five years (from 2014 to date) by the top journals we previously selected, and for every paper we considered the title, the keywords, as well as the abstract: we included in the results, *i.e.* the papers we shall then consider in detail, those papers which related to CIS, HIS, and EHR. At this stage, we preferred this manual inspection review rather than directly starting by search engines due to the following reason: the topics of CIS, HIS, and EHR (and, later on, AI) and of their techniques are so big that too many keywords are to be defined. In fact, different types of models, algorithms, techniques, methods, and approaches must be considered and aren't generally grouped within one limited set of descriptions or keywords: moreover, processed data may come from many different types of source (mostly CIS, HIS, and EHR) as well as other collections of data both from the research field and from the everyday practice. Thus, grouping

of techniques and grouping of considered data don't fit suitably in the limited set of keywords a search engine needs.

As the third step (namely, “Select Search Engines and Find papers from Conferences” in Figure 1), we completed the survey by some search engines we fed with the most common, most likely to appear keywords we identified during the second step: as search engines and literature databases, we considered major ones (Google and Medline). This step was also particularly focused on conferences and conference proceedings rather than on journals, as conferences and proceedings are not covered by the second step above. Since we wanted to consider the most prominent research projects, we expected the contributions coming from conferences and proceedings to be minimal.

After completing the literature search, in the fourth step (namely, “Build up a Taxonomy of Papers” in Figure 1) we focused on the papers from the fields of CIS, HIS, EHR and grouped the papers in some categories according to their major features, producing a taxonomy of dimensions according to which we classified the aspects dealt with by the literature. The major dimensions considered by this taxonomy focus on techniques and data, and also on other relevant aspects such as application domain, used technology, and the expectations of the analysis over those data.

The fifth step (namely, “Refine the Selection with AI Topics” in Figure 1) moved from the results of step four, and restricted the number of papers to those closest to AI and to new techniques to process clinical and health data: we obtained a list of papers which coupled the topics of health data management (CIS, HIS, and EHR) and data analysis by advanced techniques (AI). As a result, we identified the four emerging and most relevant directions of the ongoing research.

The sixth step (namely, “Build up a Refined Taxonomy of Papers” in Figure 1) further restricted the number of papers that we reviewed and described in detail by the present paper: some papers were considered for every direction detected by the fifth step.

Finally (namely, “Select the Most Interesting Papers and Survey Them” in Figure 1), based on the results of the papers of the sixth step, some medium-term and long-term research directions and expectations were described.

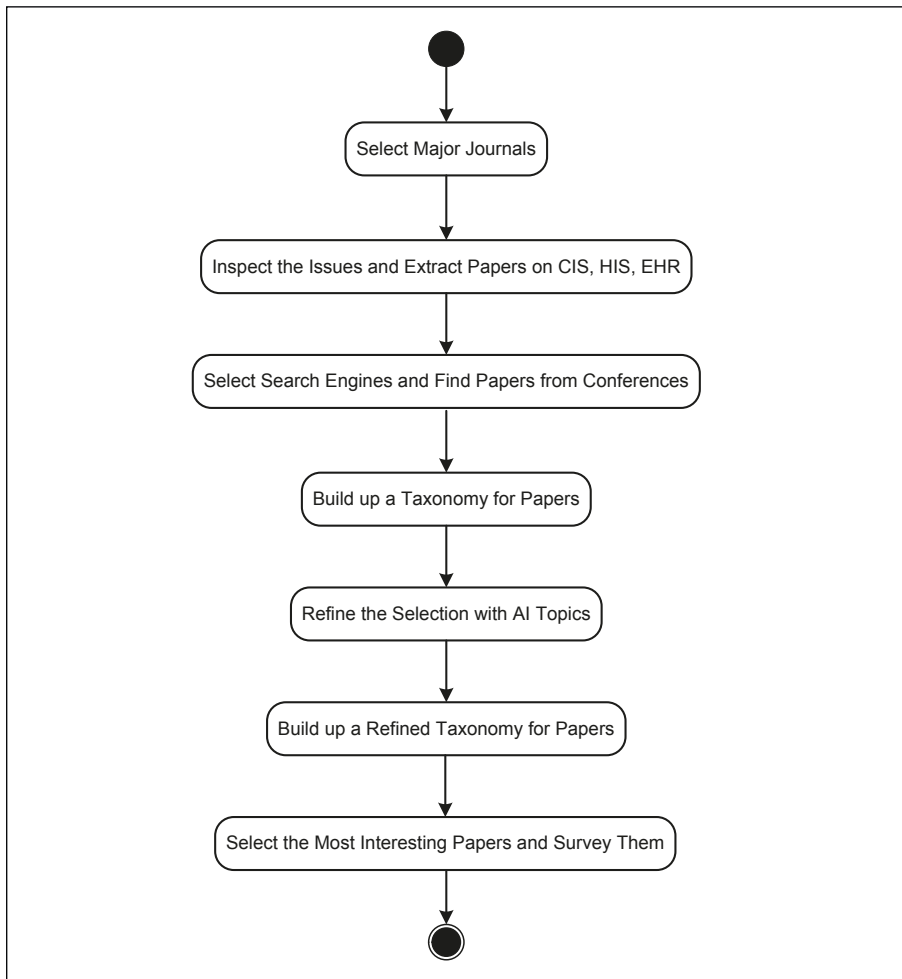


Fig. 1 UML (Unified Modeling Language) activity diagram for the process of selecting the papers to be surveyed in detail.

4 Applying the Proposed Methods

According to Section 3, we selected (first step) the following major journals: Artificial Intelligence in Medicine (AIIM); Applied Clinical Informatics (ACI); Computer Methods and Programs in BioMedicine (CMPB); Computers in Biology and Medicine (CBM); Information Systems (IS); Journal Health Informatics Research (JHIR); IEEE Journal of Biomedical and Health Informatics (JBHI); Journal of Biomedical Informatics (JBI); Journal of Medical Informatics (JMI); Methods of Information in Medicine (MIM); Journal of the American Medical Informatics Association (JAMIA). Reading and analyz-

ing (second step) the issues of the mentioned journals over the last five years requested to go through 7,652 papers: out of them, 201 papers dealt with the topics of CIS, HIS, and EHR or similar.

We then completed (third step) the search by asking the most common search engines: we considered Google Scholar, Medline and its search engine PubMed from the National Library of Medicine, and the computer science bibliography DBLP (Digital Bibliography & Library Project). Particular attention was paid to conferences, annual conferences, and proceedings from the following series: Artificial Intelligence in Medicine (AIME); ACM Bioinformatics, Computational Biology and Health Informatics (BCB); IEEE International Confer-

ence on Healthcare Informatics (ICHI); American Medical Informatics Association (AMIA) Annual Conference.

The papers so far considered and selected were then classified according to a taxonomy (fourth step), discussed in Section 4.1.1. Figure 2 graphically depicts the proposed taxonomy for CIS, HIS, and EHR systems.

The list of papers selected in the fourth step was then further refined (fifth step) to extract those papers that also covered the topics of AI. The resulting some 200 papers were then grouped (sixth step) according to a taxonomy, discussed in Section 4.1.2. Figure 3 graphically depicts the proposed taxonomy.

Table 1 numerically summarizes the papers we considered.

Finally, we considered the most relevant contributions we identified in the literature (seventh step) and surveyed in detail in Sections 4.2 – 4.5. In order to reduce the number of papers from some 200 to some 40, we selected those papers which, according to our evaluation, were introducing some major novelties and innovations.

4.1 Taxonomies

In this paper, we identified two taxonomies: one related to the major dimensions of a CIS, HIS, and EHR, and the other related to the use of Artificial Intelligence techniques over data collected by a CIS, HIS, and EHR.

4.1.1 Taxonomy of Major Dimensions

We describe here a taxonomy including the major dimensions according to which we classified clinical information systems, as depicted in Figure 2. The taxonomy we propose here differs from already published taxonomies, such as [7, 8]. In fact, existing taxonomies either focus on health information technology in general, and do not go in depth with clinical information systems, or focus on the success factors for a clinical information system, and do not provide the reader with a real high level taxonomy for clinical information systems.

According to our approach, the five major dimensions we identified considering the more than 200 papers (step 5 of Figure 1) retrieved from the literature are

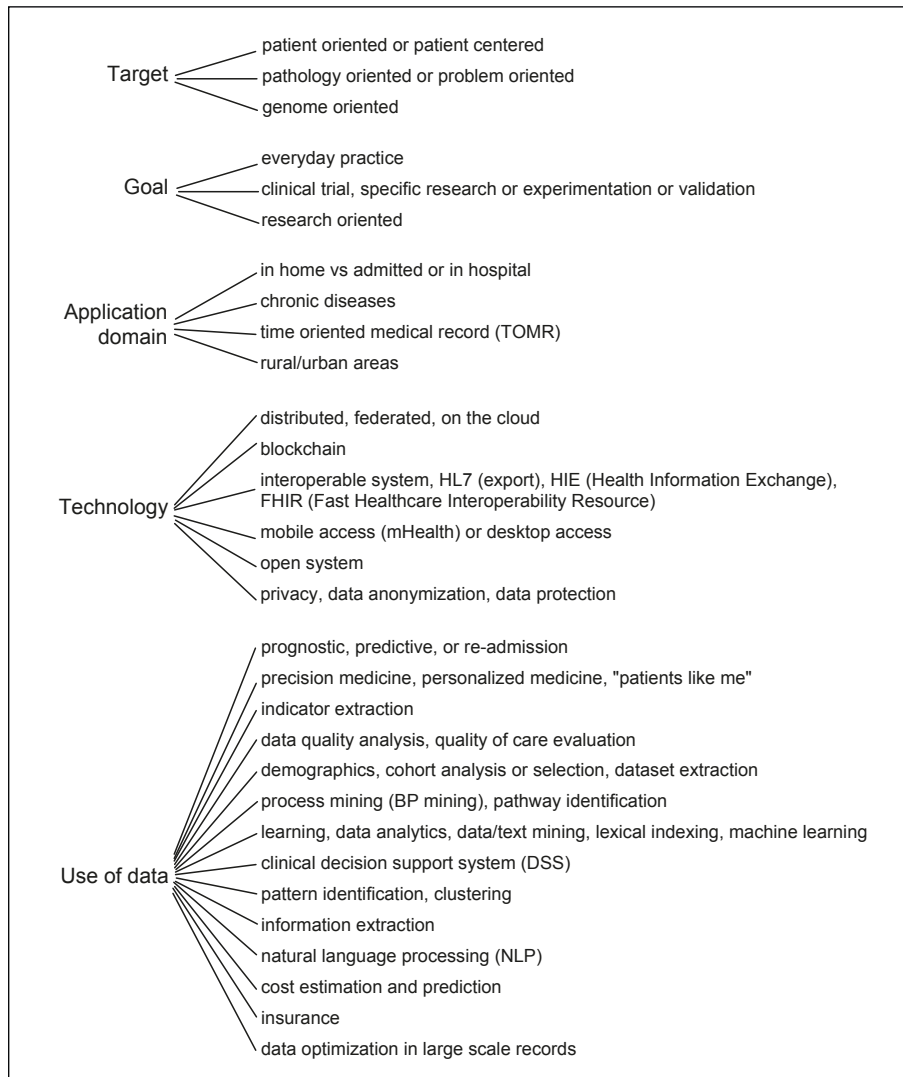


Fig 2 Taxonomy of the major dimensions of Clinical Information Systems (CIS), Hospital Information Systems (HIS), and Electronic Health Records (EHR) as reconstructed from the considered literature.

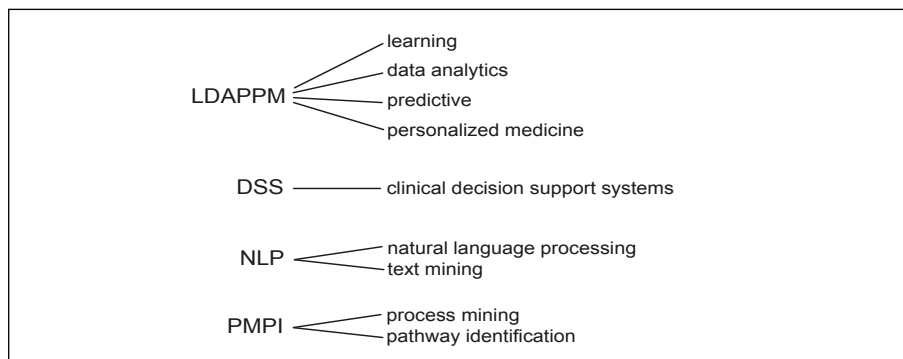


Fig. 3 Taxonomy for the use of Artificial Intelligence (AI) techniques on healthcare data as reconstructed from the considered literature.

the following ones. “*Target*” considers the approach of CIS, HIS, and EHR (patient-oriented; pathology- or problem-oriented; genome-oriented); “*Goal*” considers the main reason for collecting data by a CIS, HIS, EHR (everyday practice; clinical trial, specific research, experimentation, or validation; research-oriented); “*Application domain*” considers the environment and the main characteristics of CIS, HIS, and EHR (in home or admitted patient; chronic disease, time-oriented medical record; rural/urban areas); “*Technology*” considers the architecture of the computer system on top of which the CIS, HIS, or EHR runs (distributed, federated, on the cloud; blockchain; interoperable system and HL7 (Health Level 7), HIE (Health Information Exchange), FHIR (Fast Healthcare Interoperability Resources); mobile or desktop access; open system; privacy, anonymization, and data protection); “*Use of data*” refers to the aim according to which stored data are then processed (prognostic, predictive; personalized and precision medicine; indicator extraction; data quality and care quality evaluation; demographics; process mining and pathway identification; learning, data analytics, data mining, text mining, lexical indexing, machine learning; clinical decision support system (DSS); pattern identification and clustering; information extraction; natural language processing (NLP); cost estimation and prediction; insurance and claims; data optimization in large scale records).

As an example for the taxonomy of Figure 2, one instance of a CIS can be described according to the five dimensions above: some dimensions may assume more than one atomic value, *i.e.* some attributes within one dimension are not mutually exclusive. For example, the CIS can be patient-oriented (according to the “*Target*” dimension, the CIS collects data for every single patient, which may present different pathologies), for the everyday practice (as “*Goal*”, the CIS is for the everyday practice), for follow up visits (according to the “*Application domain*” dimension, the CIS collects data in a time-oriented manner), storing data in the cloud and permitting a mobile access (according to the “*Technology*” dimension, the CIS is cloud-based and with mobile access interfaces – two values for the same

Table 1 Every cell reports data in a X/Y/Z format, where X counts the papers which fit the four high-level topics of Figure 3; Y counts the papers which fit the keywords of “Clinical Information Systems”; and Z counts the grand total of papers published by that journal during that year. Totals are reported per journal (rightmost column) and per year (bottommost row).

Journal name	2014	2015	2016	2017	2018	total per journal
Artificial Intelligence in Medicine - AIIM	-- / 1 / 55	2 / 2 / 59	1 / 4 / 53	2 / 1 / 64	-- / 1 / 53	5 / 9 / 284
Applied Clinical Informatics – ACI	-- / 2 / 74	-- / 2 / 59	-- / 2 / 85	-- / 5 / 96	-- / 7 / 85	0 / 18 / 399
Computer Methods and Programs in BioMedicine – CMPB	-- / 2 / 216	-- / 2 / 122	-- / 4 / 286	-- / 6 / 235	2 / 8 / 299	2 / 22 / 1158
Computers in Biology and Medicine – CBM	-- / 3 / 207	-- / 3 / 320	-- / 3 / 276	1 / 4 / 281	1 / 2 / 306	2 / 15 / 1390
Information Systems – IS	-- / 0 / 76	-- / 1 / 110	-- / 1 / 77	-- / 1 / 117	-- / 0 / 56	0 / 3 / 436
Journal Health Informatics Research – JHIR	DNA	DNA	DNA	1 / 1 / 10	-- / 1 / 21	1 / 2 / 31
IEEE Journal of Biomedical and Health Informatics – JBHI	1 / 3 / 214	2 / 4 / 218	-- / 0 / 170	2 / 5 / 180	1 / 5 / 176	6 / 17 / 958
Journal of Biomedical Informatics – JBI	2 / 10 / 166	1 / 11 / 207	3 / 5 / 202	5 / 6 / 218	3 / 6 / 190	14 / 38 / 983
Journal of Medical Informatics (Elsevier) – JMI	-- / 8 / 97	2 / 5 / 113	-- / 4 / 148	1 / 7 / 182	1 / 4 / 185	4 / 28 / 725
Methods of Information in Medicine – MIM	-- / 0 / 71	-- / 4 / 88	-- / 1 / 74	1 / 5 / 70	-- / 2 / 34	1 / 12 / 337
Journal of the American Medical Informatics Association – JAMIA	2 / 6 / 174	2 / 7 / 188	2 / 10 / 196	1 / 9 / 190	1 / 10 / 203	8 / 42 / 951
total per year	5/35/1350	9/41/1484	6/34/1567	14/50/1643	9/46/1608	43/206/7652

dimension), and using data to support the clinician in decision-making (according to the “Use of data” dimension, the CIS exports data to a decision support system).

4.1.2 Taxonomy of the Use of AI Techniques

The dimension “Use of data” from the taxonomy of Figure 2 is then used to further identify the taxonomy of the techniques, taken from Artificial Intelligence.

After reading, analyzing, and grouping the some 200 retrieved papers (step 5 of Figure 1), the taxonomy we propose on the use of AI techniques identifies four major high level topics: the first one is *Learning, Data Analytics, Predictive and Personalized Medicine* (LDAPPM), which includes learning (*i.e.*, extracting new knowledge), data analytics, data mining, lexical indexing, machine learning, pattern identification, clustering, prognostic, predictive, or readmission prediction; then *Decision Support Systems*,

which refers to the use of clinical information to support decision-making activities in clinical contexts; *Natural Language Processing*, which includes processing and mining text-based clinical information; and *Process Mining and Pathway Identification* (PMPI), which includes mining and identification of healthcare/clinical processes and care pathways. Figure 3 graphically depicts the proposed taxonomy.

4.2 Learning, Data Analytics, Predictive and Personalized Medicine (LDAPPM)

This category is very wide, including many related and, sometimes, overlapping topics. We summarized the category as LDAPPM, represented by 44 papers over the grand total of considered papers: the major journals we encountered during our review, and that focus on the topic of LDAPPM, are

JAMIA and JBHI. LDAPPM is the most populated topic among those identified by the taxonomy of Figure 3. Also, LDAPPM is one of the few high level topics for which the literature reports many detailed special issues and survey papers: we consider some of the most relevant ones in the subsection 4.2.1.

4.2.1 Special Issues and Survey Papers

One contribution is a guest editorial from Chiu and Li [9] in the journal *Computer Methods and Programs in Biomedicine*, and it focuses on improving healthcare management with data science. The guest editorial highlights a relevant use of an Electronic Medical Record system to general near real-time estimations from big data analysis: the application domain is that of monitoring flu epidemics, and the interesting aspect is that of predicting admissions to triage and to hospital due to that type of epidemics.

The review article from Mehta and Pandit [10] in the International Journal of Medical Informatics provides an interesting survey on concurrence of big data analytics and healthcare. The paper starts with a collection of definitions about what big data and big data analytics are. The paper, then, provides the reader with a taxonomy of existing systems according to three main criteria: sources of healthcare data (e.g., electronic medical records (EMRs), diagnostics, medical claims, prescription claims, clinical trials, social media, wearable, and sensors); big data analytical techniques (e.g., cluster analysis, data mining, graph analytics, machine learning, neural networks, pattern recognition, and spatial analysis); and big data applications (e.g., genomics, drug discovery, personalized healthcare, precision medicine, elderly care, and many others). The paper also highlights that most of the studies depicted by the literature still have a relatively narrow scope, and limited practical applications. Moreover, most of these studies come from developed countries, and no deployment to data from developing countries is seen in the immediate near future.

Andreu-Perez *et al.* [11] also propose an overview of recent advancements on big data in the IEEE Journal of Biomedical and Health Informatics. The authors extend their analysis and the concept of big data analytics from medical and health informatics to translational bioinformatics, sensor informatics, and imaging informatics. Thus, big data are just not only those stored by traditional EMR and Clinical Information Systems, but include any type of patient's data. The authors also arise some critical issues not to be neglected when dealing with big data, including privacy, security, data ownership, data stewardship, and governance. In fact, such big collections of data are on the one side extremely relevant to progress in clinical research, but on the other side they may interfere with one's private life – and one may want not to share his/her personal details of private life.

Ravi *et al.* [12] in the IEEE Journal of Biomedical and Health Informatics provide the reader with a comprehensive survey on the deployment of deep learning techniques in health informatics. As in [11], the authors highlight potential usage in health informatics, translational bioinformatics, medical

imaging, pervasive sensing, and public health, and they particularly focus on one specific technique for data analytics (deep learning) among the existing ones. The authors also sketch out some limitations and challenges to be faced: convolutional neural networks are deployed in a black box approach, and no modification is applied in case misleading classifications are detected; while several experiments have been performed in the literature, most of them rely on relatively small datasets or focus on rare diseases, and consequently the error on the training set is very small, but results cannot be profitably generalized to new situations which have not been already observed; preprocessing of data still remains a critical step, influencing the overall performances, and the proper dimensioning of the many parameters of a deep neural network still is a blind process which deserves accurate validation; sensitivity to noise (and, as a proof, also to voluntary introduced noise) is still to be improved, as well as in any other data analytics approach.

The guest editorial from Yang and Veltri [13] in the journal Artificial Intelligence in Medicine focuses on intelligent healthcare informatics in big data era. Among the papers presented by the guest editorial, the contribution from Kavuluru, Rios, and Lu [14] describes an empirical evaluation of supervised learning techniques, which read some 71K electronic medical records from in-patients, and assign diagnostic codes. This approach uses a subset of 1,231 ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) codes, out of a full set of 4,723 distinct codes: as expected, better results are achieved for those diseases whose training set includes at least 50 cases, while more rare diseases – or diseases with a smaller training set – are poorly classified.

The methodological review of Parimbelli *et al.* [15] in the Journal of Biomedical Informatics deals with patient similarity for precision medicine. The authors first report a taxonomy of data types used to detect patients' similarity such as molecular, clinical, and laboratory data, as well as imaging/bio signals, data integration, and patient-reported outcomes. The authors, then, report a taxonomy of applications

domains where patients' similarity is investigated: cancer, nervous system/mental health, integumentary/exocrine system, respiratory system, digestive/excretory system, musco/skeletal system, cardiovascular/circulatory system – to mention the most relevant ones. Finally the authors report a taxonomy of approaches used to detect (and, possibly, measure) patients' similarity: clustering, dimensionality reduction, similarity, supervised clustering. Most relevant, the authors envisage concentrating research efforts on the integration of patient similarity measures with decision support system, to boost research on prediction medicine.

Shickel *et al.* [5] in the IEEE Journal of Biomedical and Health Informatics provide the reader with a survey on recent advances in deep learning in analyzing electronic health records. The taxonomy proposed by the authors focuses on machine learning techniques (multilayer prediction, convolutional neural networks, recurrent neural networks, autoencoders, restricted Boltzmann machine) and deep learning applications (electronic health record information extraction, electronic health record representation learning, outcome prediction, conceptual phenotyping, clinical data de-identification). The authors also clearly highlight the major limitations of current research on the topic, which refer to model interpretability, data heterogeneity, and lack of universal benchmarks. This latter one is the most relevant topic the research has to focus on – authors envisage.

Bisaso *et al.* [4] in the journal Computers in Biology and Medicine survey machine learning applications on HIV data from medical records. The authors are from Uganda, and their paper is one of the few papers from developing countries, where the relevance of HIV and HIV-related diseases is extremely high – as well as epidemic. The considered work focuses on papers and data both from medical care and from research communities. The authors clearly demonstrate how the trend of research moved from considering electronic medical records only, to include generic, imaging, and lab data. In fact, up to 2002 data extracted from electronic medical records were the only source of information. Starting from 2008, the key role in providing research

with useful information has been played by genetic data, while information extracted from traditional electronic medical records has become less and less relevant.

4.2.2 Promising Results in Some Application Domains

The literature includes some papers on specific application domains where the use of LDAPPM seems to lead to promising results. The paper by Lara *et al.* [16] in the Journal of Biomedical Informatics aims at deploying data mining techniques on data related to events in time series. The paper focuses on EEG (electroencephalography) data, and it uses data from a publically available data source (EEG recordings), while extension of the approach to data from clinical information systems is straightforward. Deployed data mining techniques mainly include adaptive fuzzy inference neural networks - AFINN neural networks. The general framework is based on an event definition language, which improves the overall performance of the approach, as well as its applicability to other medical domains such as electrocardiography (ECG).

Kipnis *et al.* [17] in the Journal of Biomedical Informatics target the alerting of inpatient deteriorations. The goal of the paper is that of detecting the evidence of physiologic derangements for a given patient with a reasonable advance (6 to 24 hours according to the authors) prior to actually observing the deterioration for that patient. This prediction is based on data collected from the electronic health record of the patient: it is not a predictive medicine system based on genomic or ancestral data analysis. The alerting system, namely Advanced Alert Monitor, has been developed from the analysis of some 650 K hospitalization episodes and some 48 M hourly observations. While the system currently performs well over cases with abundant information, the authors are planning to face the challenge of achieving good results also in the cases where data availability is much poorer.

Monsalve-Torra *et al.* [18] in the Journal of Biomedical Informatics focus on the application domain of patients who underwent a surgery for abdominal aortic aneurysm. This disease features a high

rate of mortality and complications with consequent reduced quality of life and higher costs of treatment. Consequently, the estimation of mortality risk is extremely important. The authors deploy machine learning methods based on neural networks and Bayesian networks to build a predictive system which could predict hospital mortality. The considered dataset is made of 57 attributes from 310 cases coming from clinical information systems. The attributes were pre-processed and then fed into the WEKA (Waikato Environment for Knowledge Analysis) system.

Bourne *et al.* [19] and Margolis *et al.* [20] in their papers published in the Journal of American Medical Informatics Association focus on an extremely relevant project by the National Institute of Health (NIH) on the Big Data to Knowledge (BD2K) initiative. The BD2K project aims at identifying solutions of biological problems in the shape of methods, tools, software, and training to be shared within the biomedical research community at large. Thus, BD2K wants to maximize the use of biomedical data to extract value from that data, also developing and disseminating data analysis methods. As highlighted by the authors, scientists have to face many challenges, such as expanding the availability and the use of EHRs spread in different formats in different research and care centers. The use of federated data catalogs, as also described by Brisimi *et al.* [21], is one step in this challenge.

Lo and Li [22] in their editorial in the journal Computer Methods and Programs in Biomedicine extend the concept of machine learning from alphanumeric data to various image modalities, where images are taken from CIS devoted to patients affected by liver or breast cancer, which are among the most common cancer types for men and women, respectively.

The topic of precision medicine is well described by the paper of Frey, Bernstam, and Denny [23] in the Journal of American Medical Informatics Association. Precision medicine aims at matching genomics to therapeutics for an individual. Such a challenge requires considering big data and learning systems in order to properly identify the optimal treatment of that individual. The typical disease in which

precision medicine is applied is cancer, and the authors report about a database from clinical information systems storing some 160 K patients, 64 K of them being tracked across 134 research cohorts.

Machine learning over data from CIS can also help to predict the evolution of the disease of a patient. As an example, Swain and Kharrazi [24] in their paper in the International Journal of Medical Informatics describe a prediction model to estimate the readmission probability over a 30-day period. The model belongs to the category of Readmission Risk Prediction Models (RRPM), and it considers 297 prediction variables. These variables are extracted from HL7 messages transmitted by Health Information exchange Organizations. The model helps in preventing unplanned hospital readmissions.

Turgeman and May in their paper published in the journal Artificial Intelligence in Medicine [25] describe a mixed-ensemble model for hospital readmission prediction. Their predictive model is based on a C5.0 tree classifier, coupled to a Support Vector Machine (SVM) to increase the performance of the classifier. The model has been applied to data of some 20 K inpatient admissions for some 4,8 K patients suffering from congestive heart failure. The model reaches a total accuracy of about 85% of the cases, thus proving its efficacy.

Zhao *et al.* [26] in their paper published in the Journal of Biomedical Informatics describe how machine learning can benefit from considering heterogeneous temporal data coming from EHRs. Traditional machine learning algorithms work on data collected with tables. The approach of the authors moves from the considerations that clinical events are unevenly distributed over time. Temporal machine learning, *i.e.*, machine learning which leverages on the temporality of considered data, can exploit this feature and substantial improvements can be achieved by better focusing on collected data and the temporal distance among events,

Miotto and Weng [27] in their paper published in the Journal of American Medical Informatics Association describe how reasoning on data from EHRs about diagnosis, medications, lab results, and clinical notes,

can be used to identify patients eligible for clinical trials. The approach described, known as cohort selection, moves from data of patients already enrolled in clinical trials, and reasons to profile the ‘target patient’. Patients who comply with this ‘target patient’ can then be enrolled for the trial. The approach was tested on 262 patients already enrolled, and used to select new patients for the trial from a population of some 30 K patients.

Moskovitch *et al.* [28] in their paper published in the Journal of Biomedical Informatics propose a framework (namely, Maitreya) for predicting medical events in order to prevent disease, to understand disease mechanism, and to increase patient quality of care. The approach moves from data stored in clinical information systems, considering some 4.5 M patients, and focuses on duration and gaps of events, which are sparse in time, to discover frequent time interval related patterns (TIRP): patterns are then used as prognostic markers. The approach has been successfully applied to 28 frequent, clinically relevant procedures.

Shknevsky, Shahar and Moskovitch [29] in their paper in the Journal of Biomedical Informatics deal with frequent interval-based temporal patterns to be discovered in clinical data of patients suffering from chronic diseases (cancer, hepatitis, and diabetes). Detected patterns (TIRP - frequent time interval related patterns) are then used to cluster patient clinical trajectories, thus predicting the evolution of the disease. The authors performed a deep consistency check, to ensure that similar TIRPs are constantly and repeatedly discovered in similar groups of patients.

Zhang *et al.* [30] in the journal Methods of Information in Medicine¹ use logistic regression, natural language processing, and neural networks techniques over clinical data from emergency departments (ED) to predict hospital admissions or transferring. The goal is that of predicting the care pathway of a patient, after some basic data have been collected as the patient presented to the ED and underwent a triage process. Moving from an archive of some 47 K ED visits in 642 hospitals, 48 principal components were extracted

and used for prediction. The authors claim a relevant improvement in prediction by the mixed deployment of the three techniques.

4.3 Clinical Decision Support Systems (DSSs)

The category DSS is represented by nine papers over the grand total of considered papers: major journals we encountered during our review and that focus on the topic of DSS are JAMIA and JBHI.

Bennett and Hardiker [31] in the Journal of American Medical Informatics Association review the literature on the use of computerized clinical decision support systems (CCDSSs) in EDs. The use of CCDSS is extremely relevant in EDs, where time to decision must be as shorter as possible and physicians can really benefit from a CCDSS: the efficacy of the treatment also relies on the time required to start the treatment itself. According to the survey, patients over 70 years are five times more likely to be admitted than patients younger than 30 years, meaning that a CCDSS for an ED must be particularly sensitive to chronic diseases. The survey detects 23 studies from the literature which evaluate the impact of CCDSSs in EDs. Surprisingly, only 13 out of the 23 studies identify a significantly positive impact on the clinical care, the authors write.

Ohno-Machado [32] in the Journal of American Medical Informatics Association in 2014 recalls that the informatics community has addressed the structure of EHRs to be the basis of clinical decision support systems (CDSSs) for many decades. The author highlights that, at time of writing, there still is a separation between research and clinical health systems (most of them are problem-oriented or pathology-oriented) and translational research involving genomic and clinical data.

Wright *et al.* [33] in the Journal of American Medical Informatics Association survey and classify the reasons which lead to CDSS alert malfunctions. The survey detects 68 cases of alert malfunctions (the rules of the CDSS do not fire and the CDSS doesn't send out alerts to physicians to highlight abnormal situations) in 14 sites through the US. Detected malfunctions are then classified according to a taxonomy the authors propose.

The taxonomy includes four major dimensions for malfunctions: cause of the malfunction (build errors; release of new codes which make the rules of the CDSS obsolete; defect of the EHR; computer environment migration); mode of discovery (mainly user reporting); start of the malfunction (which starts as the CDSS is deployed); and effect on rule firing (wrong rule action or system slows down due to some rules). The major result of the paper is in the effort of identifying malfunctions, so that further releases and/or systems could avoid the pitfalls.

Rahulamathavan *et al.* [34] in their paper published in the IEEE Journal of Biomedical and Health Informatics consider the problem of preserving the privacy of clinical data. The problem occurs when a physician sends patient's data to an outsourced DSS via the Internet, to check the answer from the DSS: the outsourced system may be unreliable or not compliant to the policies of the CIS where patient's data are originally stored. The authors propose a new encryption algorithm which fits such privacy needs, and which can be enriched and extended to cover the needs of transferring patient's data to cloud computing systems.

Yoon, Davtyan, and van der Schaar [35] in their paper published in the IEEE Journal of Biomedical and Health Informatics consider the problem of predicting the evolution of the disease of a patient to suggest the optimal therapy. In fact, the paper proposes a discovery engine (DE) which moves from the patient's characteristics and data stored in the clinical information systems to perform a personalized prediction. The engine detects which are the most relevant characteristics to exploit. The main feature of the DE is that performance remains good also in case of large number of contexts, the authors claim. As application domain, the DE has been applied to breast cancer and related therapies, providing the physicians with an average improvement of 2.18 – 4.20% with respect to traditional DSSs.

4.4 Natural Language Processing

Natural language processing (NLP) is needed to manage clinical information, as different kinds of clinical information are acquired in an unstructured or semi-struct-

¹ No PDF publically available on the web.

tured form. Accordingly, NLP in medicine is a wide and vital area of research, where one of the underlying goals consists of extracting knowledge from natural language texts coming from different sources, as medical records, reports, or social media. Within this category, we have both survey papers that confirm the high liveliness of this research community, and research papers ranging on different clinical and methodological issues.

In Kreimer *et al.*, [36], the authors follow a sound and systematic approach to consider and discuss existing NLP systems specifically developed for clinical domains. The analyzed systems allow the extraction of structured information from unstructured free texts. Different bibliographic databases were considered for the survey and, at the end of an articulated screening and selection phase, 86 papers were considered in detail. They describe 71 different clinical NLP systems. Such systems range over different clinical and research tasks by adopting different techniques. While some tasks are suitably acknowledged and sound solutions have been proposed, such as, for example, the identification of medication information and the extraction of cancer features from pathology reports, some other challenges remain unsolved, as, for example, the extraction of temporal information or the mapping of concepts expressed through natural language expressions to standard terminologies.

In Ford *et al.* [37], the focus is on the role of EMRs with respect to health-related research. Accordingly, the authors propose a survey narrower in the scope with respect to the previous one, and analyze papers that deal with incorporating information from texts into algorithms that support the identification of clinical cases. The authors, after a systematic search through literature, identified 67 papers focusing on the extraction of information from free text of EMRs, with the explicit goal of detecting cases having a specific clinical condition. This survey highlights that the considered papers mainly deal with US EMRs. Both rule-based and machine learning methods have been adopted without any clear difference with respect to the accuracy of the proposed approaches. Moreover, it is quite evident that including information from text significantly improves the performance

of such algorithms, with respect to only considering coded information. Quite interestingly, the authors underline the need to standardize the result reporting of algorithm performances, as for accuracy metrics. The topic of using NLP to process EHRs has also been dealt with by Wang *et al.* [38] and applied to congestive heart failure patients.

In Goldstein *et al.* [39], the authors evaluate the effectiveness of a new methodology for the automated creation of meaningful free-text summaries from longitudinal clinical records. Moreover, they consider the potential benefits to the clinical decision-making process, when applying the proposed method to build draft letters that can be manually improved by clinicians. The general knowledge-based system, named CliniText, has been applied for the automated summarization in free text of longitudinal Intensive Care Unit (ICU) medical records, using an ICU clinical knowledge base, created by involving two ICU clinical experts. CliniText generated free-text summary letters for 31 different patients, and such letters were compared with respect to the original discharge letters, written by physicians. The comparison was performed according to different measures as, for example, relative completeness, readability, and “semantic accessibility” (*i.e.*, how fast and correctly other physicians understood the clinical content of the letter). After some interesting quantitative results that confirmed the soundness of the proposed methodology, the authors underline that the use of summarization systems would allow the enhancement of such letters, as for standard structure, completeness, and time required for their composition. Such enhancements would positively influence the quality of decisions made by other clinicians, by considering such summaries.

In [40], Agarwal *et al.* consider an emerging problem in many healthcare institutions worldwide. Indeed, both for the quality of patients’ life and for reducing the healthcare-related costs, hospital readmission rates have to be monitored by healthcare institutions. Particularly, chronic diseases account for many hospital readmissions which have to be taken under control. In this paper, chronic obstructive pulmonary disease (COPD) has been explicitly considered, as it is highly relevant in many countries and it

requires a continuous long-lasting monitoring of affected patients. This work focuses on the use of unstructured clinical notes to statistically predict the patients most in danger of readmission. A framework is proposed, which uses natural language processing for the analysis of clinical notes. The prediction of readmissions is based on the selection of most suitable algorithms within the field of data mining and machine learning.

The paper [41] by Safari and Patrick considers natural language issues with respect to the capability of clinical users to simply perform complex research-oriented queries on EMRs. More specifically, authors focus on the support of research questions involving internal time-event dependencies, through cascaded queries. The proposed approach is based on an extension of the recently proposed Clinical Data Analytics Language (CliniDAL). Different aspects of research-oriented queries have been considered, from the elicitation of subjects to be considered in a study to the time span of the experiment, to the control group definition. Three different scenarios have been considered for evaluation. Such evaluation confirms that the proposed system can support the expression of complex queries involving also temporal aspects by users not aware of Structured Query Language (SQL) details.

4.5 Process Mining and Pathway Identification

Recently, there has been an increasing attention to represent and reason about complex and coordinated execution of clinical and healthcare activities. Such intertwined activities compose complex processes that may stem from the application of clinical practice guidelines, from care pathways, which may be viewed as the application of guidelines and internal good practice policies to specific domains and clinical environments, or, more generally, from organizational and clinical plans. According to one of the most common definitions of care pathways (CPs), we may consider them as complex interventions for the mutual decision-making and organization of care processes for a specified group of patients during a given period. Such CPs, often characterized by complex

decision-making tasks and data-intensive activities, need to be represented, designed, managed, analyzed, and discovered [42, 43]. According to the focus of this survey paper, we consider here clinical process and care pathway mining, where AI-based approaches have been widely applied.

In Rojas *et al.* [44], the authors provide a survey of healthcare process mining, which consists of deriving knowledge from data generated and stored in healthcare/hospital information systems, to analyze/discover the executed processes. Seventy-four papers with associated case studies have been considered and discussed. In particular, eleven main features characterize such analysis: process and data types; frequently asked questions; process mining techniques, perspectives, and tools; methodologies; implementation and analysis strategies; geographical analysis; and medical fields. Such survey underlines and suggests different techniques and tools to adopt for healthcare process mining. Moreover, it underlines the importance of process mining for supporting process-aware information systems. Adopting process-aware information systems could provide benefit both for the quality of the performed healthcare processes and for the optimal use of the related resources.

In Huang *et al.* [45], the discovery of CP patterns is faced through clinical event logs, which record various treatment activities. The authors propose a novel approach to CP pattern discovery by representing CPs through an extension of the Latent Dirichlet Allocation family that jointly models various treatment activities and their occurrence. In order to evaluate both applicability and soundness of the proposed approach, two real-world scenarios have been considered, namely that of unstable angina, and of oncology. The obtained results show the feasibility of the proposed CP pattern mining approach.

In Gotz *et al.* [46], the authors present a methodology for interactive pattern mining and analysis from past clinical patient data. The method supports an ad-hoc visual exploration of patterns mined. The proposed approach combines the support of visual queries for the interactive specification of clinical episodes to look for with pattern mining techniques that allow the discovery of significant intermediate events inside a clinical

episode. Moreover, interactive visualization techniques are integrated that allow the user to identify event patterns that are associated to specific outcomes together with their temporal behavior. A prototype implementation is presented as a proof-of-concept of the proposed methodology and its successful application to some real world clinical domains, namely that of heart failure, hypothyroidism, and hypertensive patients, is described.

Mining and visualizing CPs from EHR data is also the main topic of [47] where Zang *et al.* propose a practice-based CP development process and a data-driven methodology for deriving common clinical pathways from EHR data. Such patient-centered approach aims at facilitating evidence-based care. Indeed CPs helps translate best available evidence into clinical practice, suggesting the most suitable treatment sequences for specific therapy-based goals. The authors focus on visit data of chronic kidney disease patients who developed acute kidney injury in some given years. Such data were extracted from the EHR and mapped into one-dimensional sequences using novel constructs designed to capture information related to different visit facets, as purpose, procedures, medications, and diagnoses. Clustering visit sequences allows the identification of distinct patient subgroups. Markov chains have been used to characterize visit sequences. Significant transitions are extracted and visualized into CPs across subgroups. According to clustering results, CPs provide insights about the evolution of patients' conditions and medication prescriptions over time. Pathways associated to typical disease progression have been identified, as well as practices consistent with guidelines. Visualization of pathways depicts the likelihood and direction of disease progression within different contexts.

5 Discussion and Conclusions

In this paper we propose a short survey of the literature and we describe contributions dealing with clinical information systems and, at the same time, introducing some AI-based techniques and approaches. After the first analysis of bibliographic databases, we pro-

pose some taxonomy criteria for classifying and interpreting the different contributions. Then, we analyze in more details some contributions we consider representative and relevant for future research directions.

As any other survey, we identify the contributions from the literature and select the most relevant ones, thus introducing some obvious limitations to our survey due to authors' personal evaluations. We conduct the survey according to how we observe and perceive the current scenario of the domain. Visibility, prestige of the journal, as well as the relevance of the topic described by the papers have been the motivating guide to the work.

Among the different topics and approaches we found, some major research directions and issues seem to be promising and need further investigations. Interestingly, some of the highlighted research trends may involve different research topics, according to the proposed taxonomy:

1. *EHR as a source/destination of data and knowledge.* EHR systems benefit, from one side, of the reasoning capabilities of DSSs. Such systems require a tight link between decision-supporting knowledge and the specific patient data, which need to be analyzed and suitably used. On the other side, EHR data are becoming a relevant source of information, to be used to derive new knowledge, to support sound prediction of clinical outcomes, and to build more abstract representations of patients' states. Considering personalized medicine and genomic patient data, EHR systems need to switch from a problem-oriented (or pathology-oriented) approach to a patient-oriented approach, to boost the cooperation between clinical information systems and translational/genomic research;
2. *Healthcare related social web data.* Data from social web platforms are becoming of interest for many possible research directions. Such kind of data requires specific research actions, as there are many challenges related to the dimensions of such data, to the NLP processing techniques required for them, to the different levels of trust with respect to that of data stored in clinical information systems;
3. *Intertwined clinical data and processes.* A third, often underestimated, research

trend concerns the study of process-related aspects of clinical information systems. Both clinical data and processes, expressed as CPs or as the instantiation of clinical practice guidelines, need to be considered in an intertwined way. On one hand, clinical data are often related to care plans and need to be accessed and interpreted for the correct execution of care actions. Thus, a sound co-design of both CPs and related clinical database (even conceptual) design tools. On the other hand, clinical data often represent either implicitly or explicitly the result of the execution of some clinical actions. Thus, reasoning on (and extracting from different data sources) process-related data could be helpful to better understand the most effective actions for specific categories of patients, as well as for improving the quality of the provided care. Some features of both clinical processes and data need to be deeply addressed. Indeed, clinical tasks are often both decision- and data-intensive, and such features are often neglected in the current design and analysis approaches. Moreover, many temporal related aspects have to be considered for both data and processes, such as temporal constraints in process execution, temporal validity of clinical data, temporal freshness of monitoring data, to mention a few of them.

To make more concrete the three research directions we mention here, we would like to conclude by discussing an application-oriented and interdisciplinary contribution. In Dupont *et al.* [48], the focus is on using EHR data to support clinical research distributed among possibly many healthcare institutions. The authors propose a technological platform to allow the sound and secure reuse of hospital EHR data for clinical research (EHR4CR). The EHR4CR platform can support and foster clinical research scenarios, from protocol feasibility assessment, to patient recruitment for clinical trials, to clinical data exchange. The final goal is to have a multi-stakeholder ecosystem that enables the scalable use of the platform in Europe, while evaluating its economic sustainability. A market analysis was conducted by a multidisciplinary task force to define the

ecosystem and the multi-stakeholder value chain. Different requirements were highlighted from heterogeneous stakeholders. Using simulation-modeling techniques, the potential financial outcomes of the business model were forecasted from the perspective of a service provider over an horizon of five years. As for data mining and care-pathway applications and for EHR data re-use, the increasing exploitation of EHRs will facilitate population-based clinical research, and, among other benefits, will enhance healthcare pathway identification, modeling, and optimization.

Acknowledgements

This work has been partially funded by the University of Verona, within 2017-2019 RIBA project “Extending OLAP data analysis with temporal and statistical operators and its application to pharmacovigilance data”.

References

1. Degoulet P, Luna D, Quiros FG. Chapter 6 - Clinical Information Systems. In: Fatima Marin H, Massad E, Gutierrez MA, Rodrigues RJ, Sigulem D, editors. *Global Health Informatics*. Academic Press; 2017. p. 129-51.
2. Viceconti M, Hunter PJ, Hose RD. Big Data, Big Knowledge: Big Data for Personalized Healthcare. *IEEE J Biomed Health Inform* 2015 Jul;19(4):1209-15.
3. Combi C. Editorial from the new Editor-in-Chief: Artificial Intelligence in Medicine and the forthcoming challenges. *Artif Intell Med* 2017;76:37-9.
4. Bisaso KR, Anguzu GT, Karungi SA, Kiragga A, Castelnuovo B. A survey of machine learning applications in HIV clinical research and care. *Comput Biol Med* 2017 Dec 1;91:366-71.
5. Shickel B, Tighe P, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J Biomed Health Inform* 2018 Sep;22(5):1589-604.
6. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;:339.
7. Dixon BE, Zafar A, McGowan JJ. Development of a Taxonomy for Health Information Technology. *MEDINFO 2007 - Proceedings of the 12th World Congress on Health (Medical) Informatics - Building Sustainable Health Systems*, 20-24 August, 2007, Brisbane, Australia; 2007. p. 616-20.
8. Meulendijks A, Batenburg R, Wetering R. A classification framework for clinical information system implementation in hospitals. *International Journal of Organisational Design and Engineering* 2012;2:402-19.
9. Chiu H-W, Li Y-C. Improving healthcare management with data science. *Comput Methods Programs Biomed* 2018 Feb;154:A1.
10. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: A systematic review. *Int J Med Inform* 2018 Jun;114:57-65.
11. Pérez JA, Poon CC, Merrifield RD, Wong ST, Yang G-Z. *IEEE J Biomed Health Inform* 2015 Jul;19(4):1193-208.
12. Ravi D, Wong C, Deligianni F, Berthelot M, Pérez JA, Lo B, et al. Deep Learning for Health Informatics *IEEE J Biomed Health Inform* 2017 Jan;21(1):4-21.
13. Yang CC, Veltri P. Intelligent healthcare informatics in big data era. *Artif Intell Med* 2015 Oct;65(2):75-7.
14. Kavuluru R, Rios A, Lu Y. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artif Intell Med* 2015 Oct;65(2):155-66.
15. Parimbelli E, Marini S, Sacchi L, Bellazzi R. Patient similarity for precision medicine: A systematic review. *J Biomed Inform* 2018 Jul;83:87-96.
16. Lara JA, Lizcano D, Pérez-Pérez A, Valente JP. A general framework for time series data mining based on event analysis: Application to the medical domains of electroencephalography and stabilometry. *J Biomed Inform* 2014 Oct;51:219-41.
17. Kipnis P, Turk BJ, Wulf DA, LaGuardia JC, Liu V, Churpek MM, Romeroet al. (2016). Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the ICU. *J Biomed Inform* 2016 Dec;64:10-9.
18. Monsalve-Torra A, Fernández DR, Mari n-Alonso O, Payá AS, Camacho-Mackenzie J, Carreño-Jaimes M. Using machine learning methods for predicting inhospital mortality in patients undergoing open repair of abdominal aortic aneurysm. *J Biomed Inform* 2016 Aug;62:195-201.
19. Bourne PE, Bonazzi V, Dunn M, Green ED, Guyer M, Komatsoulis GA, et al. The NIH Big Data to Knowledge (BD2K) initiative. *J Am Med Inform Assoc* 2015 Nov;22(6):1114.
20. Margolis R, Derr L, Dunn M, Huerta MF, Larkin J, Sheehan J, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J Am Med Inform Assoc* 2014 Nov-Dec;21(6):957-8.
21. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated Electronic Health Records. *Int J Med Inform* 2018 Apr;112:59-67.
22. Lo C-M, Li Y-C. Machine learning based cancer detection using various image modalities. *Comput Methods Programs Biomed* 2018 Mar;156:A1.
23. Frey LJ, Bernstam EV, Denny JC. Precision medicine informatics. *J Am Med Inform Assoc* 2016 Jul;23(4):668-70.
24. Swain MJ, Kharrazi H. Feasibility of 30-day hospital readmission prediction modeling based on health information exchange data. *Int J Med Inform* 2015 Dec;84(12):1048-56.
25. Turgeman L, May JH. A mixed-ensemble model for hospital readmission. *Artif Intell Med* 2016 Sep;72:72-82.

26. Zhao J, Papapetrou P, Asker L, Boström H. Learning from heterogeneous temporal data in electronic health records. *J Biomed Inform* 2017 Jan;65:105-19.
27. Miotto R, Weng C. Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials. *J Am Med Inform Assoc* 2015 Apr;22(e1):e141-50.
28. Moskovitch R, Polubriaginof F, Weiss A, Ryan PB, Tatonetti NP. Procedure prediction from symbolic Electronic Health Records via time intervals analytics. *J Biomed Inform* 2017 Nov;75:70-82.
29. Shknevsky A, Shahar Y, Moskovitch R. Consistent discovery of frequent interval-based temporal patterns in chronic patients' data. *J Biomed Inform* 2017 Nov;75:83-95.
30. Zhang X, Kim J, Patzer RE, Pitts SR, Patzer A, Schrage JD. Prediction of Emergency Department Hospital Admission Based on Natural Language Processing and Neural Networks. *Methods Inf Med* 2017 Oct 26;56(5):377-89.
31. Bennett P, Hardiker NR. The use of computerized clinical decision support systems in emergency care: a substantive review of the literature. *J Am Med Inform Assoc* 2017 May 1;24(3):655-68.
32. Ohno-Machado L. Focusing on the patient: mHealth, social media, electronic health records, and decision support systems. *J Am Med Inform Assoc* 2014 Nov-Dec;21(6):953.
33. Wright A, Ai A, Ash JS, Wiesen J, Hickman TT, Aaron S, et al. Clinical decision support alert malfunctions: analysis and empirically derived taxonomy. *J Am Med Inform Assoc* 2018 May 1;25(5):496-506.
34. Rahulamathavan Y, Veluru S, Phan RC-W, Chambers JA, Rajarajan M. Privacy-Preserving Clinical Decision Support System Using Gaussian Kernel-Based Classification. *IEEE J Biomed Health Inform* 2014 Jan;18(1):56-66.
35. Yoon J, Davtyan C, Schaar M. Discovery and Clinical Decision Support for Personalized Healthcare. *IEEE J Biomed Health Inform* 2017 Jul;21(4):1133-45.
36. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J Biomed Inform* 2017 Sep;73:14-29.
37. Ford E, Carroll JA, Smith HE, Scott D, Cassell J. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc* 2016 Sep;23(5):1007-15.
38. Wang Y, Luo J, Hao S, Xu H, Shin AY, Jin B, et al. NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records. *Int J Med Inform* 2015 Dec;84(12):1039-47.
39. Goldstein A, Shahar Y, Orenbuch E, Cohen MJ. Evaluation of an automated knowledge-based textual summarization system for longitudinal clinical data, in the intensive care domain. *Artif Intell Med* 2017 Oct;82:20-33.
40. Agarwal A, Baechle C, Behara RS, Zhu X. A Natural Language Processing Framework for Assessing Hospital Readmissions for Patients With COPD. *IEEE J Biomed Health Inform* 2018 Mar;22(2):588-96.
41. Safari L, Patrick JD. Complex analyses on clinical information systems using restricted natural language querying to resolve time-event dependencies. *J Biomed Inform* 2018 Jun;82:13-30.
42. Combi C, Gambini M, Migliorini S, Posenato R. Representing Business Processes Through a Temporal Data-Centric Workflow Modeling Language: An Application to the Management of Clinical Pathways. *IEEE Trans Syst Man Cybern Syst* 2014;1182-203.
43. Combi C, Oliboni B, Zardini A, Zerbato F. A Methodological Framework for the Integrated Design of Decision-Intensive Care Pathways: an Application to the Management of COPD Patients. *J Healthc Inform Res* 2017;1:157-217.
44. Rojas E, Munoz-Gama J, Sepúlveda M, Capurro D. Process mining in healthcare: A literature review. *J Biomed Inform* 2016 Jun;61:224-36.
45. Huang Z, Dong W, Ji L, Gan C, Lu X, Duan H. Discovery of clinical pathway patterns from event logs using probabilistic topic models. *J Biomed Inform* 2014 Feb;47:39-57.
46. Gotz D, Wang F, Perer A. A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *J Biomed Inform* 2014 Apr;48:148-59.
47. Zhang Y, Padman R, Patel N. Paving the COWpath: Learning and visualizing clinical pathways from electronic health record data. *J Biomed Inform* 2015 Dec;58:186-97.
48. Dupont D, Beresniak A, Sundgren M, Schmidt A, Ainsworth JD, Coorevits P, et al. Business analysis for a sustainable, multi-stakeholder ecosystem for leveraging the Electronic Health Records for Clinical Research (EHR4CR) platform in Europe. *Int J Med Inform* 2017 Jan;97:341-52.

Correspondence to:

Giuseppe Pozzi
 Dipartimento di Elettronica, Informazione e Bioingegneria,
 Politecnico di Milano
 p.za L. da Vinci 32
 I-20133 Milano, Italy
 E-mail: giuseppe.pozzi@polimi.it