# Progress in Biomedical Knowledge Discovery: A 25-year Retrospective

L. Sacchi[1], J. H. Holmes[2]

[1] Biomedical Informatics Laboratory "Mario Stefanelli", Department of Electrical, Computer, and Biomedical Engineering, University of Pavia, Italy

[2] Institute for Biomedical Informatics, Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA

## Summary

**Objectives:** We sought to explore, via a systematic review of the literature, the state of the art of knowledge discovery in biomedical databases as it existed in 1992, and then now, 25 years later, mainly focused on supervised learning.

**Methods:** We performed a rigorous systematic search of PubMed and latent Dirichlet allocation to identify themes in the literature and trends in the science of knowledge discovery in and between time periods and compare these trends. We restricted the result set using a bracket of five years previous, such that the 1992 result set was restricted to articles published between 1987 and 1992, and the 2015 set between 2011 and 2015. This was to reflect the current literature available at the time to researchers and others at the target dates of 1992 and 2015. The search term was framed as: Knowledge Discovery OR Data Mining OR Pattern Discovery OR Pattern Recognition, Automated.

**Results:** A total 538 and 18,172 documents were retrieved for 1992 and 2015, respectively. The number and type of data sources increased dramatically over the observation period, primarily due to the advent of electronic clinical systems. The period 1992-2015 saw the emergence of new areas of research in knowledge discovery, and the refinement and application of machine learning approaches that were nascent or unknown in 1992.

**Conclusions:** Over the 25 years of the observation period, we identified numerous developments that impacted the science of knowledge discovery, including the availability of new forms of data, new machine learning algorithms, and new application domains.

Through a bibliometric analysis we examine the striking changes in the availability of highly heterogeneous data resources, the evolution of new algorithmic approaches to knowledge discovery, and we consider from legal, social, and political perspectives possible explanations of the growth of the field. Finally, we reflect on the achievements of the past 25 years to consider what the next 25 years will bring with regard to the availability of even more complex data and to the methods that could be, and are being now developed for the discovery of new knowledge in biomedical data.

## Keywords

Algorithms; artificial intelligence; databases, factual; data mining, knowledge discovery in databases

**Yearb Med Inform 2016;Suppl1:S117-29**
**http://dx.doi.org/10.15265/IYS-2016-s033**
**Published online August 2, 2016**

## 1 Introduction

It is very evident that the domain of biomedical knowledge discovery has grown on many dimensions over the past 25 years. These include algorithms and systems for discovering knowledge, the burgeoning of many different types of data from many different sources, and the concomitant growth of legal imperatives, which has had a considerable effect on the availability and use of biomedical data. But first, we should consider the term "knowledge discovery", a term which has often been used synonymously with "data mining", and more than occasionally, "data dredging" or "fishing". We categorically reject these latter two terms as pejorative, as they do not accurately reflect the activities associated with what we are calling knowledge discovery.

In this paper, we define knowledge discovery quite literally as the process of discovering knowledge in data, specifically biomedical data, which spans molecules to populations. The process involves the acquisition of data that are appropriate for a given purpose, such as an investigation into a specific disease. The process also includes the preparation of these data such that they can be analyzed using specialized software algorithms that will assist in the identification of patterns in the data, which may (or may not) reflect some biomedical phenomenon. The output of this activity is information; however, it should be clear to anyone familiar with the commonly used data-information-knowledge-wisdom framework [1] that the process of knowledge discovery is not yet complete. Indeed, the mere identification of patterns in data is only one step along this spectrum. To take us to the next step, knowledge discovery, requires substantial human expertise as the results of applying these specialized algorithms must be assessed in light of that expertise. In a word, knowledge discovery is automated, not automatic, and this is a key consideration in this paper.

We present here a brief survey of the past 25 years of development of knowledge discovery in biomedical data and its emergence as a scientific discipline in its own right, as evidenced by the state of the art practice in 1992 and the present. We hope that this survey will provide the foundation for dreaming about the next 25 years, such that when a paper such as this one is written in 2042, its authors will look back and either validate our predictions or refute them.

## 2 Methods

In order to gain an empirical sense of the climate in 1992 and 2015, we conducted a search of PubMed, using the following as text words and subject headings: "Knowledge Discovery" OR Data Mining (MesH term) OR "Pattern Discovery" OR Pattern Recognition, Automated (MesH term).

We restricted the result set using a bracket of five years previous, such that the 1992

result set was restricted to articles published between 1987 and 1992, and the 2015 result set between 2011 and 2015. This was to reflect the current literature available at the time to researchers at the target dates of 1992 and 2015. Publications outside the five-year window were not included. This combination of search terms yielded 538 documents for 1992 and 18,172 documents for 2015. Rather than using all documents in each result set, we selected documents that appeared to represent the state of the art at their respective time period, determined by the prevalent type of knowledge discovery algorithms and availability of data.

# 3 Results

To analyze the results of our search, we decided to focus on two related aspects: the type of data that were available back in 1992 and those that are available in 2015, and the methodologies that were developed to deal with that kind of data. On the basis of the results of this analysis, we have identified a set of emerging research areas, that were not available back in 1992, but have been developed thanks to the continuous research on methodologies and novel data sources availability. In addition, we have identified a set of challenges, still open and to be considered for the next 5 years.

## 3.1 Data Sources Availability

### 3.1.1 Data Sources and Availability in 1992

Nearly all clinical data was handwritten in 1992, thereby making automated knowledge discovery a difficult task. Among the few data already available for automated processing, there were signals and images collected through dedicated devices. For this reason, biomedical informatics at its early stages was much more intertwined with signal processing and image analysis than it is now. This observation is supported by our literature search, where, out of the 538 extracted papers, 72 (13.3%) contained the term ECG, or EEG, or the MesH term "signal processing, computer assisted". Forty papers (7%) included one of the following MesH terms: "image analyses, computer assisted", "image analysis, computer assisted", "computer assisted image analyses", or "computer assisted image analysis".

As regards the analysis of other clinical data to be used by computerized algorithms, one needed to abstract them onto case report forms or similar instruments. These provided a platform for manual coding that transformed data to a representation that was amenable to computer-assisted analysis. Data such as text required interpretation by the abstractor in order to transform complex concepts into discrete representations that could be analyzed by a computer algorithm. Let's consider the example of a radiology report, written in free text and rich in clinical concepts. In 1992, a discrete field indicating the presence of a Colles' fracture was very rarely found on any radiology report. Instead, this information was contained in the text of the report. Interestingly and in a very real sense, the abstractor in 1992 was performing a kind of knowledge discovery, or at least pattern identification, which is a central component of discovery. However, there were very few computerized tools in mainstream, or even research use at that time. As we will see in the following, the challenge of automatically extracting clinical information from texts is a topic that has gained particular interest in the past few years until 2015.

Administrative data, such as insurance claims, are a potentially very rich source of information for research, quality assurance, and health services allocation. These data capture virtually all transactions for which a claim is file, such as a clinical service like a lab test or office visit, and include extensive cost and diagnostic coding data. In 1992, the sources of administrative data, such as insurance claims, were somewhat limited in the United States, primarily available through the Centers for Medicare and Medicaid Services. In other countries, such as the United Kingdom, administrative data that represented a wider swath of the population were not so difficult to obtain, given the existence of the National Health Service. Once obtained, however, these data posed an extraordinarily difficult problem for the user, as they were extremely voluminous. One could say that these data were the first "Big Data" in the biomedical domain. While the datasets were typically "narrow", meaning relatively few columns, they were potentially extremely "long", in that they could contain millions of records, each representing an individual claim for an individual beneficiary. The tools used at this time to discover knowledge in these data were primarily statistical. Software such as SAS [2] and SPSS [3] provided the user with a full palette of statistical techniques and algorithms with which to explore the data; these included basic descriptive statistics and frequency distributions as well as graphical tools such as histograms and scatterplots. It was very unusual to see someone use a machine learning tool in the quest of discovering knowledge from such data during this time, but users of these claims data seemed to fare well in spite of this.

By 1992, surveillance of specific diseases such as cancer or infectious diseases was quite well established. The Surveillance, Epidemiology, and End Results (SEER) database created in 1973 continues to the current day as a national "registry of cancer registries" and is a rich source of data for researchers and others [4]. Infectious disease surveillance by the Centers for Disease Control (CDC) and Prevention has a long history, and involves the collection, management, and analysis of a wide variety of data from many different geographic and clinical sources. Since the 1960s, the US CDC has conducted a number of important national health status and health services surveys that serve a surveillance function. These include the National Health and Nutrition Examination Survey (NHANES) and the Behavioral Risk Factor Surveillance System [5, 6]. While all of these surveillance systems are very rich sources of health-related data, in 1992, analysis and knowledge discovery were nearly always accomplished by traditional statistical methods. This implies that nearly all such activities were hypothesis driven.

In 1992, data users of virtually every persuasion were starting to investigate the importance of data linkage, in which records one data source would be linked, record for record, or patient for patient, in order to investigate a problem that would require such linkage. Such record-level linkage could be

accomplished deterministically, where there existed a unique identifier that was common to two or more datasets, or probabilistically, where linkage was accomplished by matching key variables that taken together could probably match one record with another [7-9]. An example of where such linkage was used, and continues to be, is the so-called SEER-Medicare dataset [10]. In this case, the clinical data of a patient represented in SEER is matched to his or her Medicare claims data. This affords the data user a complete picture of the clinical status and services used by the patient, in a much richer way than either dataset could provide individually. However, the enthusiasm for data linkage was relatively short-lived due to increasing concern about privacy and confidentiality.

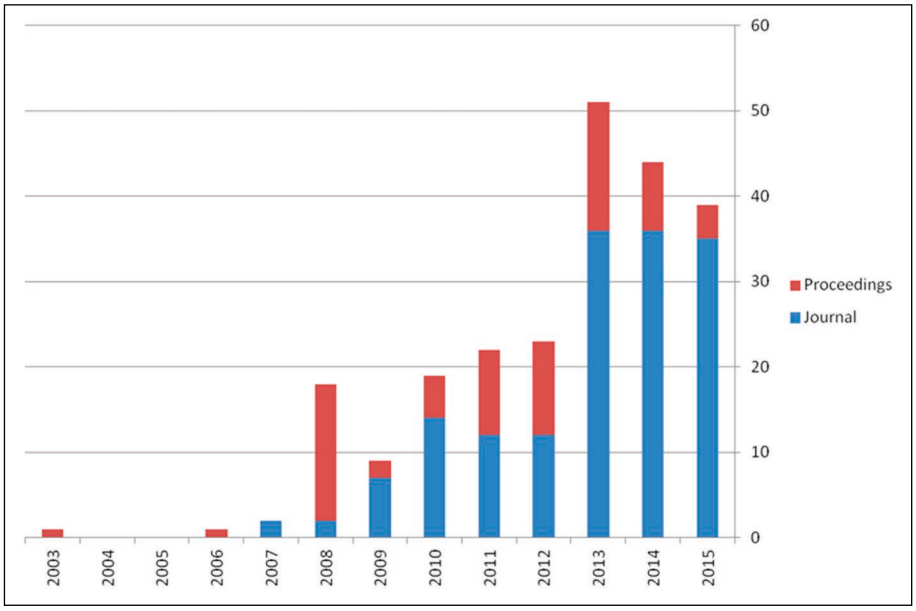


**Fig. 1** Number of papers published per year extracted using the following MEDLINE query: (("Knowledge Discovery" OR Data Mining[MeSH Terms] OR "Pattern Discovery" OR Pattern Recognition, Automated[MeSH Terms]) AND (EMR OR EHR OR "electronic medical record" OR "electronic health record")), divided into conference Proceedings and Journal Papers.

## 3.1.2 Data Sources and Availability in 2015

### 3.1.2.1 Rise of the Electronic Medical Record (EMR) and Other Clinical Systems

One of the most important differences between 1992 and 2015 is related to the widespread use of the EMR [11]. This has generated an increasing interest in the analysis of the data coming from those systems [12-15]. Interestingly, if we constrain our original PubMed search to elements including in text words the following terms: EMR, ("electronic medical record"), EHR ("electronic health record"), without constraints on the publication date, the total number of resulting papers is 229, with the first paper of this list dating back to 2003 (Figure 1). Interestingly, constraining the PubMed search to 2011-2015, the resulting number of papers is 179.

The most important applications that have been covered in the past 5 years were data analytics and predictive modeling, often coupled to visual analytics and visualization solutions, with the goal of delivering decision support to the users [16-19]. In general, we have witnessed a trend towards a translational use of data mining and knowledge discovery. Great interest has been devoted also to the management of longitudinal data, with several methodologies centered on the extraction and visualization of temporal patterns [20-30]. These methods were exploited both for clinical applications and for tackling organizational issues [31, 32]. In this last case, process mining started to be successfully applied [33, 34]. Given the heterogeneity of the data collected in hospital EMRs, and the frequent presence of textual reports, text mining and natural language processing (NLP) have started to gain increasing interest and several works were published to deal with these methodologies. In particular, out of the 179 papers extracted in the search constrained to the years 2011-2015, 38 (21% - 28 journal and 10 proceedings) were dealing with text processing techniques.

### 3.1.2.2 Administrative Data

The integration of data coming from different sources has been one of the topics of main interest in the past five years. As mentioned, a particularly important category of data is administrative data. These data are usually collected for billing purposes and show a different structure with respect to clinical data. Administrative data are in general collected as process data, as they record the occurrence of specific events (e.g. insurance claims or hospital billing claims), but they do not report the clinical information related to the event itself. In our search, 24 papers explicitly consider administrative data in their title or text. All the papers acknowledge the usefulness of this kind of data to improve the data analysis process.

Some studies analyze administrative data alone, highlighting potential analyses that wouldn't be possible by using only clinical data [18-25]. Other studies perform a validation of methodologies that exploit administrative data, using a comparison with clinical data to perform evaluation [43-51]. The results of these studies are not homogeneous: some studies highlight a lower sensitivity of the use of administrative data with respect to using clinical information [43, 44, 46, 48, 50], whereas some others report very good performances. Of course a trade-off needs to be reached between the advantages of administrative data and the drawbacks related to the fact they are collected for different purposes.

Interestingly, not many of the papers address novel methodologies to deal with administrative data. Most deal with state of the art machine learning methods (classification trees, logistic regression, k-nearest neighbor) [35, 37, 38, 52] or traditional statistical analyses, such as chi-squared

tests, sensitivity analyses, and calculation of positive predictive value [36, 43-45, 50].

#### 3.1.2.3 Surveillance Data

After the attacks on the United States on September 11, 2001, many biomedical researchers turned their attention to a new type of disease surveillance that focused on the identification syndromes that could indicate a potential outbreak before it started- an early warning system of sorts. For example, one could monitor the sales of certain over-the-counter medications such as anti-diarrheal preparations for a potential foodborne infection outbreak. Another important source of data is the emergency department chief complaint, which can be a rich source of symptom data that is not captured in quantitative data. It should be evident that such systems handle very large amounts of noisy data in real-time, and novel methods to discover valid patterns in these data are needed. A number of investigators have developed such methods that provide accurate and timely surveillance data to key personnel in the public health and law enforcement professions [53-59].

#### 3.1.2.4 Wearable Technology

In the past five years, the focus shift to personalization increased the interest in considering the environment the patient lives in, also outside the hospital [60, 61]. This has been done mainly by analyzing data coming from wearable sensors and mobile phones. In this scenario, the Center of excellence for Mobile Sensor Data-To-Knowledge (MD2K) was chosen as one of 11 Big Data Centers of Excellence by the National Institutes of Health, as part of its Big-Data-to-Knowledge initiative and it is intended to develop big data solutions for the integration, visualization, and analysis of data generated by mobile and wearable sensors [62].

In the literature, the methodologies that have been proposed span from signal processing (ECG [63-65], gait analysis [66], food ingestion analysis [67], smoking habit monitoring [68]) to methods based on machine learning techniques (for activity detection [69-75], falls detection [76], vital signs [77], cigarette smoking [78], position recognition [79, 80], social influence [82]), with particular relevance of artificial neural networks (ANNs) [82-85] and hidden Markov models (HMMs) [86-88]. Interestingly, few of the proposed methods have already been integrated in the devices [65, 67, 83, 89]; they rather work offline. We could anyway identify a trend in the literature towards the development of lightweight algorithms to be implemented on several sensors for information fusion [67, 90-92].

## 3.2 Evolution of Methodological Approaches

### 3.2.1 The State of the Art of Biomedical Knowledge Discovery: 1992

The slowly increasing availability of data and the ability to use them to formulate hypotheses for further investigation acted as a strong motivation for the development of new tools to automatically extract knowledge from them. In 1992, expectations about the automated discovery of knowledge in databases were quite nascent, and probably more of the "what if" variety. The term "knowledge discovery in databases" (KDD), was first used by Gregory Piatetsky-Shapiro in 1989 (the year of the first KDD conference) so it was still a term, and a concept, in its infancy in 1992 [93].

In 1992, the Internet was just reaching a level of maturity, the World Wide Web was only three years old; Open Database Connectivity (ODBC) was developed; Microsoft Windows 3.1 and Linux were released; Python had been invented the year before, followed shortly by Visual Basic, and the world had to wait another several years before Java and Java Script were available. Sadly, two notable deaths occurred: Grace Hopper (who coined the term computer "bug") and Allan Newell (a giant in the early days of artificial intelligence). Both had contributed mightily to the very foundations of knowledge discovery throughout their career, Hopper in the field of programming languages and software development, and Newell in his contribution to cognitive computing that parallels the way humans think about solving problems.

The tangible accomplishments of computing were primarily felt in the consumer marketplace. The personal computer was becoming ever more affordable and easy to use, and was connected to the world via the Internet and the Web. However, there was growing expectation that the emerging field of artificial intelligence could help to solve complex problems, none more potentially important than finding those relationships and patterns in data that we, as humans, would ordinarily miss. Somehow, machines could do this faster, cheaper, and more accurately.

However, in 1992 the predominant approach to solving such problems in the biomedical field was in fact not automated but accomplished through manual methods of knowledge engineering, such as interviews and talk-aloud protocols [94], in order to elicit from experts the knowledge that could be incorporated into deductive, rule-based ("expert") systems [95]. This was an arduous task that involved many hours of observation and highly detailed documentation of expert knowledge in a limited problem domain. So computer scientists and a few informatics researchers set about investigating the possible role for computer-assisted knowledge discovery. In this effort, they tended to focus on algorithms that were primarily, if not completely, inductive in nature; rather than relying on previously acquired and encoded knowledge in the form of rules, these algorithms learned by experience and formed rules based on that experience. With these inductive algorithms, one could discover patterns in complex data, free of any bias or pre-existing knowledge that could taint the inferences made by such tools. In 1992, the paradigm of semi-automated knowledge discovery found elements in the re-use of the philosophical studies on abduction [96], setting the preliminary ground for the development of cognitive informatics, a field that in 2015 continues to be very popular, especially for understanding the nature of clinical activities and for developing engineering and computer solutions that can improve clinical practice and patient engagement [97]. In the biomedical area, a prevalent example of an inductive learner in 1992 was the artificial neural network [98-107]. Statistical methods were rarely used [108], although there was increasing interest in Bayesian and other probabilistic approaches to knowledge discovery, primarily in the realm of classification and prediction.

The evolution towards automated knowledge discovery was helped by the important methodological advances that had been achieved in those years by the machine learning community. Bayesian approaches had a real boost at that time, and the idea of learning them from data had been around for some years after the publication of Judea Pearl's book in 1988 [109]. One important contribution was the one by Cooper and Herskovits, where Bayesian Networks were transformed from a knowledge representation and inference tool into a machine learning one [110]. Several applications of the methodologies dealt with the development of expert systems, also in the biomedical field, following the idea of automatically derive knowledge from data rather than eliciting it form experts [111-114]. Moreover, inductive logic programming saw some of the major contributions in these years [115-116], and evolutionary computation and genetic programming gained increasing popularity thanks to the work of JR Koza, who proposed hierarchical genetic algorithms to build up computational procedures [117].

### 3.2.2 Methodologies in 2015: Evolution of New Analytic Approaches

As shown in the previous section, 25 years ago, some of the algorithms and methodologies that are now used as standard techniques for data mining and knowledge discovery had just been developed. These "novel" analytic approaches have evolved in the last 25 years, and their application has become more popular with the increase of the volume and variety of the data to be analyzed. In particular, machine learning and artificial intelligence methods have been increasingly considered for application to biomedical problems as an alternative to traditional statistical analysis. The first applications of machine learning methods to biomedical research date back to the early nineties, and after 2011 more than 200 papers have been published every year in this field. Papers can be divided into those that apply one or more known techniques to solve a specific problem, and those that define novel or improved methodologies to be applied to biomedical data. In this section we will focus on a set of selected supervised machine learning techniques, which were the ones that saw an early development and application during the years this review considers as a baseline, and were then applied as standard approaches to biomedical knowledge discovery. Such techniques are ANNs, Support Vector Machines (SVMs), Hidden Markov Models, and Bayesian approaches. To study how the application of these methods to biomedical knowledge discovery has evolved in the past 25 years, we have run our PubMed search several times, each one constraining it to include the following terms:

- "support vector machines" OR SVM
- Bayes
- "Hidden Markov models" OR HMM
- ANN OR NN OR "Neural Networks"

As expected, these queries resulted in a high number of papers, as shown in Figure 2.

Given the high number of extracted papers, to analyze the results of this search we followed a twofold approach: first of all, we reviewed the annual publication trend for each topic to analyze the changes during the past 25 years. As a second step, we chose to apply a text mining algorithm to analyze the most relevant applications of the selected techniques to the biomedical domain. To this end, we have applied latent Dirichlet allocation (LDA) for topic modeling [118] and extracted the most relevant topics from the abstracts of the papers returned when using the PubMed queries. LDA is a robust technique that allows extracting a set of relevant topics from a document corpus. The user sets the desired number of topics and the algorithm extracts them, together with the words that are most relevant to the specific topic. To apply this algorithm, we first exported the abstract of all the papers from PubMed. Since the exporting function produces a single file, we automatically split it into a set of documents, each one containing a single abstract. To do this, we used a rule-based approach based on specific words that delimit the abstracts in the list. We defined the rules as general as possible, but during automatic extraction some of the abstracts might potentially be missed. This is the reason why sometimes the number of abstracts inputted to LDA is slightly lower than the one resulting from the PubMed query. As a final step, we performed several runs of LDA for each group of papers, and the most recurrent topics were identified. Such topics will be reported in the following of this section.

The query that resulted in the highest number of papers is the one related to ANNs (5033 papers). As already pointed out, neural networks were developed in 1940s and saw an early application to biomedical problems. Figure 3 shows the number of papers extract-

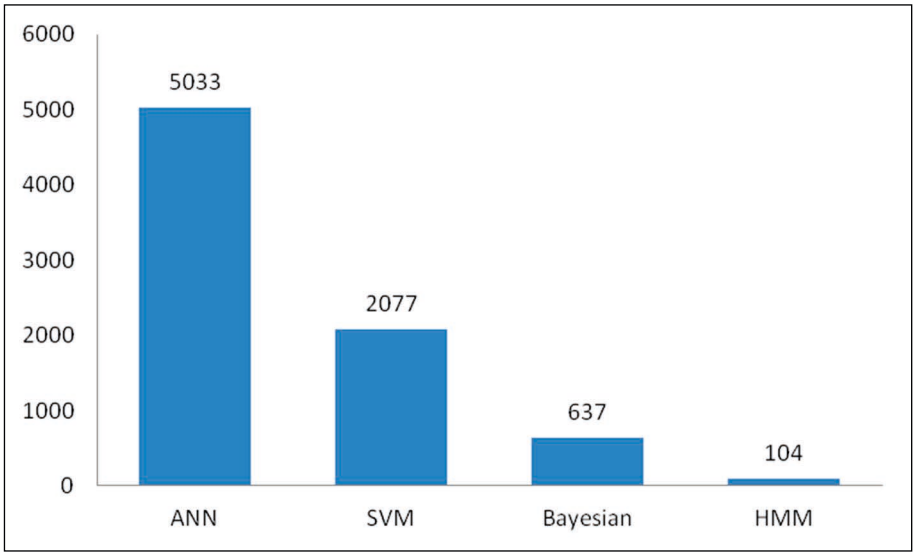


**Fig. 2** New analytic approaches publications in the past five years (2011-2015)

ed from our PubMed search constrained on ANNs and not specifying any limitation on the publication date. The first applications of ANNs to biomedical research date back to the late fifties, but it was during the nineties that we witnessed to a rapid increase of the number of papers exploiting this kind of technique. After this increase, the number of publications reached a plateau during the 2000s and registered a slight decrease after 2010, though keeping a high rate of publications per year.

To interpret the results of the LDA analysis, we have assigned a title to each topic, considering the words that were extracted as representative of the topic itself. For ANNs, we found applications that cover: predictive modeling (502 papers), image and signal analysis (1182 papers), motion control (340 papers), cancer research (479 papers), bioinformatics (338 papers), the analysis of water quality (related to the analysis of environmental data, 502 papers), and applications to control systems (706 papers). The detailed application topics resulting from the LDA analysis are shown in Table 1 of the Appendix.

SVMs were introduced 20 years later than ANNs, and their non-linear extension was proposed in 1992 [118]. Since that year, these algorithms have seen a constant increase in the application in biomedical research (Figure 4). In particular, since 2011, we noticed a boost in publications, resulting in a total of 2077 papers in the past 5 years (as shown in Figure 1).

As regards SVMs, the most important topic that was found was bioinformatics (854 papers), with a particular focus on gene expression analysis for cancer research and structural analysis of proteins. Also for this methodology, image and signal analysis were very well represented, and in particular we found contributions on brain signals analysis (633 papers) and image analysis for cancer research (333 papers). An interesting topic that was not found for ANNs was the analysis of drug compounds (285 papers). The detailed results of this analysis, including all the topics and the representative words are shown in Table 2 of the Appendix.

Searches on Bayesian approaches and HMMs resulted in a lower number of papers, but still all these methods have shown an increased publication rate in the past two decades.

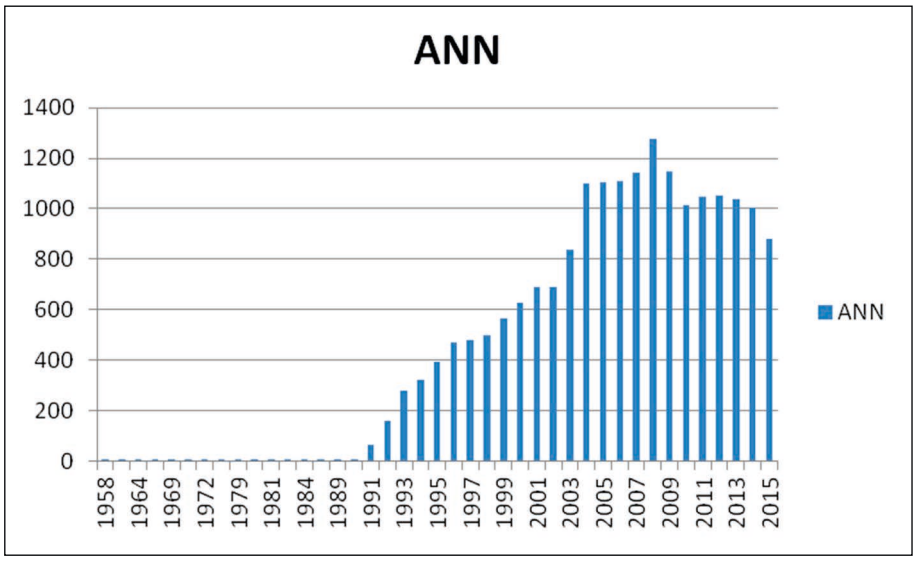


**Fig. 3** Number of papers published per year extracted using the following MEDLINE query: (("Knowledge Discovery" OR Data Mining[MeSH Terms] OR "Pattern Discovery" OR Pattern Recognition, Automated[MeSH Terms]) AND (ANN OR NN OR "Neural Networks"))

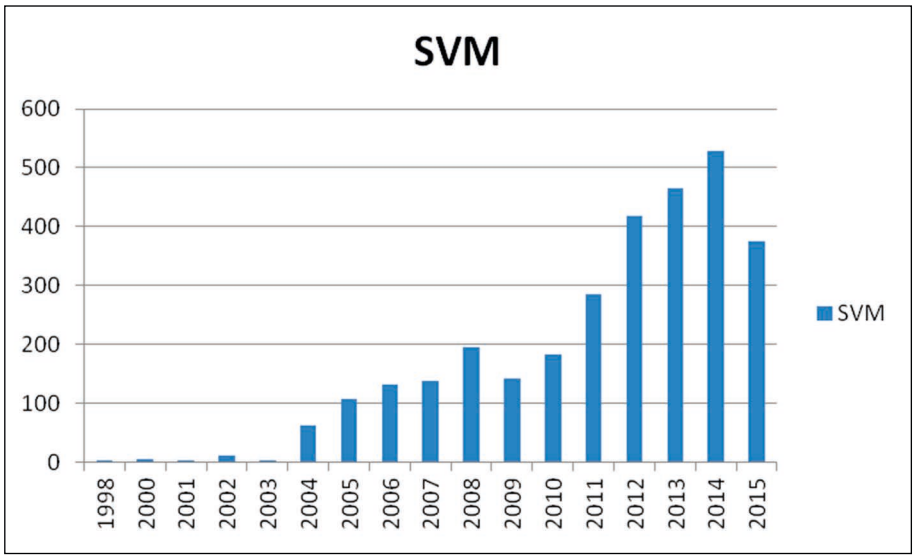


**Fig. 4** Number of papers published per year extracted using the following MEDLINE query: (("Knowledge Discovery" OR Data Mining[MeSH Terms] OR "Pattern Discovery" OR Pattern Recognition, Automated[MeSH Terms]) AND ("support vector machines" OR SVM))

As for SVMs, also for HMMs the most represented topic was bioinformatics (30 papers), with a stress on the recognition of motifs or sequences in protein structures. An important application of HMMs is activity recognition, in particular during training (16 papers). A similar topic involves the analysis of motion and posture (10 papers). Two interesting applications of these techniques are related to the analysis of human and social interactions (12 papers), and to the extraction and prediction of adverse drug reactions (6 papers). Table 3 of the Appendix shows the detailed results for HMMs.

One of the most relevant applications of Bayesian techniques in the past five years is

cancer research with a focus on gene expression data (79 papers). Bayesian approaches have been frequently applied to image analysis, including brain imaging (163 papers). Applications common also to other domains are activity recognition (139 papers) and gene expression analysis (71 papers). The detection of adverse drug events is also a common application of this kind of techniques (128 papers).

## 3.3 Emerging Areas of Research

From the analysis presented in the previous sections, it has been possible to identify some novel areas of research that have been emerging distinctively in the past five years and that were not present back in 1992. These areas are Bioinformatics, NLP, and Visualization. These areas of research could be developed thanks to the introduction of new data sources (bioinformatics), and to the digitalization of data that were previously only paper-based (NLP). Visualization techniques have evolved to offer a support to deal with the increasing amount of available information, both as structured and unstructured data (such as text).

As it clearly appeared from the analysis of the abstracts of the papers on methodologies, one of the most important applications for all of them was bioinformatics. As a matter of fact, performing a rough search on PubMed for papers with Bioinformatics as MesH term and analyzing the trend of this keyword along the years, it was possible to notice that, since 2000, there has been an exponential growth of the number of published research, which reached a maximum of 13,232 papers in 2014. This growth reflects two of the most important breakthroughs that took place after 1992: the successful completion of the Human Genome Project in 2003 [120], and the development of the first high throughput gene expression analysis techniques [121]. From then on, the research in this field continued to improve, following the evolution of new techniques for next generation sequencing and the recent focus on precision medicine [122].

As already observed in the previous sections, with the availability of novel sources of data such as administrative databases and EHRs, it has become very important to be able to mine data stored in narrative reports, and to integrate them to other more structured data sources. This is reflected by the publication trend of papers dealing with NLP, that saw a constant increase in the past five years (1072 papers resulting from our query), reaching its maximum in 2014. To relate NLP to its main applications and study how this relation has evolved over the period 2011-2015, we ran a yearly LDA analysis on the NLP abstracts, to understand what topics have increased their popularity and those that instead lost interest. To perform this analysis, we ran LDA four times each year, and we identified the most recurrent topics in the four runs. In 2011, the scenario was dominated by two topics: bioinformatics applications and discovery of adverse drug events. In bioinformatics, NLP was mainly used to extract interactions between genes and between proteins, to identify relevant biological pathways. Related to this, ontology is also a recurrent topic, since many resources are available and can be exploited for such purposes. These two topics kept being well-represented also in the following years (especially in 2013-2014), when in the bioinformatics community the attention started to be focused also on gene and protein networks, and a particular focus has been given to the discovery of gene-cancer associations from documents. Interestingly, during these years, attention started to be dedicated also to other type of applications, such as discharge summaries and radiology reports (2012), and nursing reports (2014). A very interesting result is related to the novel trends of 2015, when the topic on adverse drug reactions seems to be less addressed, while social media and sentiment analysis appear consistently for the first time. In addition, attention is specifically devoted to the identification of risk factors, both related to cancer and to cardiovascular diseases, from medical records.

An additional consequence of the availability of novel, heterogeneous data sources has been the increasing attention devoted to decision support systems that provide support by proposing to the user an intelligent visualization of the available data, to simplify the summarization process and, indirectly, to allow capturing new information [123]. For this reason, we included in our search also the term "visualization", and we ran LDA on the resulting abstracts. In this case, we identified five topics of interest. The first is the visualization of complex biological networks, and diseases and intervention networks (198 papers). Brain connectivity models and activation maps have also seen in interest in the past five years (114 papers). This was especially true in 2011 and 2012, also thanks to the Human Connectome project, a NIH-funded project that analyzes connectivity data, neuroimaging, behavioral, and genetic data to construct a map of the complete functional and structural neural connections [124]. Visualization of the results of clustering algorithm has been also found as a relevant topic. In this case the applications span from bioinformatics to the analysis of clinical data. A similar observation holds for the topic related to cancer that is often related to –omics analyses, but also to drugs and treatments.

## 3.4 Reflections on Three Perspectives

### 3.4.1 Legal Perspectives

Even before the advent of "Big Data" and sophisticated knowledge discovery tools, substantial concern was raised about the protection of privacy and confidentiality of individuals (and corporate entities such as hospitals) who were represented in data in some way. In the mid-1990s, this concern became starkly manifest and was addressed to some extent in the ratification in the US of the Health Insurance Portability and Accountability Act (HIPAA) [125]. The purpose of HIPAA was to allow the transfer and continuation of health insurance when workers become unemployed. However, the act is probably better known for the constraints it places on sharing identifiable health-related data between parties who have no rights to them. Two rules were promulgated in order to ensure individuals' privacy- the so-called Privacy Rule (2000) - and the security of the data as they are stored or transferred within or between entities - the Security Rule (2003). Both rules have had a constraining effect on the availability and contents of health data, to the extent that data linkage, which engendered so much interest 25 years ago, is very difficult to accomplish. In addition, we now must be circumspect in applying the sophisticated algorithms we now have for identifying patterns in all types of biomedical data in order to avoid compromising privacy and confidentiality through the unintentional re-identification of individuals in data.

### 3.4.2 Social Perspectives

Over the last 25 years, and particularly over the past 15 years, knowledge discovery has been perceived as at once a social good and an agent for social malevolence. The social good of knowledge discovery is manifest in its application in biomedical domains. The increasing availability of data from a wide variety of electronic sources such as the EHR and other clinical systems, from genomic analysis, and environmental and geographical data sources affords us with a great opportunity to explore new corners of the human condition that have been historically impossible to investigate. In this, the public is generally sympathetic; they see the advantages of knowledge discovery in this domain, fueled with the hope that the application of sophisticated knowledge discovery tools will improve health. However, there is another side to knowledge discovery that some find uncomfortable. It has been used in non-medical domains for purposes that some would characterize as malevolent, as "spying" on human behavior. In the court of public opinion, there is a sense that knowledge discovery methods applied to such large and varied data resources could, over time, put people at risk of losing a job, or a reputation. In 1992, we did not think much of these social imperatives of knowledge discovery, but in 2016 they are very real considerations.

### 3.4.3 Political Perspectives

Over the past five years, there has been intense political interest in "Big Data", to the extent that there are special funding programs in the US and elsewhere for training and research in this area. There is even a name for the discipline, Big Data to Knowledge, or BD2K [126]. In 1992, or even 2010, it would have been difficult to imagine such a reaction to the realization that indeed, medical data are mounting daily, and we need new algorithms and computing infrastructures to deal with them. The BD2K movement (as it were) has captured the imagination of lawmakers and leaders, who have passed budgets to support its work and used the term in political speeches and documents. With this zeal comes a great responsibility for those of us who work in biomedical informatics research. In a real sense, we are the practitioners but also the custodians of a new field of inquiry and application that demands our diligence and expertise to ensure its integrity. That field is now called Data Science, and is emerging as a cutting edge discipline not just in biomedical domains, but also in the physical, biological, and social sciences as well.

## 4 The Next 25 Years: 2017-2041

Just as it would have been difficult to imagine in 1992 where we would be in 2016 with regard to biomedical knowledge discovery, it is almost impossible to conceive the future of the field in 2041. Because we are seeing substantial changes in the types of data that are being produced and curated, the quantity and quality of those data, and the potential for new, metaphorically rich algorithms that facilitate analyses like we have never seen. For sure, the scenario is currently dominated by Big Data [127], whose challenge has moved from the perspective of storage to the one of knowledge extraction through data science [128]. As regards methodologies, the Big Data era is pushing towards approaches that would need to be quite different from the current ones, for several reasons. First of all, the need to analyze data coming from different sources potentially located at different centers while preserving privacy and security of the individual patients, will require computation to be moved to the data. Distributed extensions of traditional algorithms have already started to be developed, and this will for sure be a trend in the near future. Another consideration that needs to be made is related to the need to use partial information to extract knowledge and derive a coherent view on the available data, since it won't be possible to process all the data that are potentially generated.

Using the experience of the last 25 years, and considering some of the approaches that were used in 1992, one realizes that there is a certain resurgence of interest in older methods. For example, in 1992, neural networks captured the interest of many researchers and practitioners. By around the turn of the millennium, this interest was waning, in favor of statistical classifiers such as hidden Markov models, support vector machines, and Bayesian methods. For nearly a decade, it seemed that neural computing was relegated to a back seat, no doubt due to its lack of transparency and scalability to large datasets.

Now, over the past several years, there has been something of a neural computing renaissance, in the form of "deep learning". Advances in computing infrastructure, including grid and multiprocessor technology, have facilitated this resurgence, but so too has the development of new thinking about the structure and function of neural networks as well as belief networks. The case for deep learning is still to be made, but there is a degree of promise as evidenced by a nascent but growing literature in the field: a recent PubMed search of the term "deep learning" yielded 170 publications, many of these in top-tier journals [129-134].

Certainly today we see a burgeoning of data sources that include not only clinical data from the electronic health record or administrative data, but also data from a huge variety of physiologic monitoring devices, from geographic systems that provide information about the environment that people encounter and live in, from wearable technology such as activity monitors, and many others of which we are not currently aware. Already, investigators are looking at ways to integrate genomics and metabolomics data with the EHR at the bedside in order to make better-informed clinical decisions. They are also looking at ways to make these heterogeneous and highly complicated data understandable to the clinician. One way is through visualization, an emerging discipline in biomedical informatics, and one which stands to make a unique contribution to data science. One should take it for granted that new data sources will continue to emerge, and that 2041 will see a very different data landscape than the one we have now.

## Conclusion

This survey is unquestionably incomplete, for a number of reasons. First, there have been so many developments in the field of biomedical knowledge discovery over the past 25 years, it is simply impossible to discuss them all in the confines of a survey article. In particular, there are several do-

mains that have been covered only partially by this work. Choosing to focus mainly on supervised machine learning methodologies, we have given less emphasis to unsupervised learning, even if it has been the driver for knowledge discovery in several biomedical areas, such as bioinformatics. Another area that has been only partially covered by this work is the one related to temporal data mining methods, which aim at extracting knowledge from longitudinal data explicitly taking into account the temporal dimension. Another limitation related to the way this survey has been carried out is that, except for a few number of cases, it was not possible to distinguish papers and tools that put into practice the knowledge that was discovered (e.g generating actionable knowledge, guidelines, best practices, etc.) from those published on already available literature data and that remained a purely methodological contribution. To do this, a more restrictive selection of papers, focusing on a less broad plethora of applications, should be performed. Finally, it is equally impossible to consider the future course of the science with the same lens we used to examine the developments of the past 25 years. This is a nonlinear enterprise to be certain, and we already see evidence even in the past decade, that the developments of the field happen at an ever-increasing rate, and as in the case of neural computing, with some degree of iteration. But this is what makes the field of knowledge discovery so exciting; it does not stay still for very long, if at all.

## References

1. Ackoff RL. From Data to Wisdom. J Appl Syst Anal 1989;16:3–9.
2. Analytics, Business Intelligence and Data Management | SAS [Internet]. [cited 2016 Jan 25]. Available from: http://www.sas.com/en_us/home.html
3. IBM SPSS software [Internet]. [cited 2016 Jan 25]. Available from: http://www-01.ibm.com/software/analytics/spss/
4. Surveillance, Epidemiology, and End Results Program [Internet]. [cited 2016 Jan 25]. Available from: http://seer.cancer.gov/
5. NHANES - National Health and Nutrition Examination Survey Homepage [Internet]. [cited 2016 Jan 25]. Available from: http://www.cdc.gov/nchs/nhanes.htm
6. CDC - BRFSS [Internet]. [cited 2016 Jan 25]. Available from: http://www.cdc.gov/brfss/
7. Fair ME, Lalonde P, Newcombe HB. Application of exact ODDS for partial agreements of names in record linkage. Comput Biomed Res 1991 Feb;24(1):58–71.
8. Shannon HS, Jamieson E, Walsh C, Julian JA, Fair ME, Buffet A. Comparison of individual follow-up and computerized record linkage using the Canadian Mortality Data Base. J Public Health 1989 Feb;80(1):54–7.
9. Shapiro S. Automated record linkage: a response to the commentary and letters to the editor. Clin Pharmacol Ther 1989 Oct;46(4):395–8.
10. SEER-Medicare Linked Database [Internet]. [cited 2016 Jul 19]. Available from: http://healthcaredelivery.cancer.gov/seermedicare/
11. Blumenthal D, Tavenner M. The "Meaningful Use" Regulation for Electronic Health Records. N Engl J Med 2010 Aug 5;363(6):501–4.
12. Ohno-Machado L. Mining electronic health record data: finding the gold nuggets. J Am Med Inform Assoc 2015 Sep;22(5):937.
13. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, et al. Using electronic health records for clinical research: the case of the EHR4CR project. J Biomed Inform 2015 Feb;53:162–73.
14. Ross MK, Wei W, Ohno-Machado L. "Big Data" and the Electronic Health Record. Yearb Med Inform 2014 Aug 15;9(1):97–104.
15. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013;20(1):117–21.
16. Soulakis ND, Carson MB, Lee YJ, Schneider DH, Skeehan CT, Scholtens DM. Visualizing collaborative electronic health record usage for hospitalized patients with heart failure. J Am Med Inform Assoc 2015 Mar 1;22(2):299–311.
17. Warner JL, Denny JC, Kreda DA, Alterovitz G. Seeing the forest through the trees: uncovering phenomic complexity through interactive network visualization. J Am Med Inform Assoc 2015 Mar 1;22(2):324–9.
18. Perer A, Wang F, Hu J. Mining and exploring care pathways from electronic medical records with visual analytics. J Biomed Inform 2015 Aug;56:369–78.
19. Viangteeravat T, Nagisetty NSVR. Giving Raw Data a Chance to Talk: A Demonstration of Exploratory Visual Analytics with a Pediatric Research Database Using Microsoft Live Labs Pivot to Promote Cohort Discovery, Research, and Quality Assessment. Perspect Health Inf Manag [Internet] 2014 Jan 1 [cited 2015 Nov 15];11(Winter). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3995483/
20. Hripcsak G, Albers DJ, Perotte A. Parameterizing time in electronic health record studies. J Am Med Inform Assoc 2015 Jul 1;22(4):794–804.
21. Hajihashemi Z, Popescu M. A Multidimensional Time Series Similarity Measure with Applications to Eldercare Monitoring. IEEE J Biomed Health Inform 2015;PP(99):1–1.
22. Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. Identifying and Mitigating Biases in EHR Laboratory Tests. J Biomed Inform 2014 Oct;0:24–34.
23. 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 1;48:148–59.
24. Hripcsak G, Albers DJ. Correlating electronic health record concepts with healthcare process events. J Am Med Inform Assoc 2013 Dec;20(e2):e311–8.
25. Warner JL, Zollanvari A, Ding Q, Zhang P, Snyder GM, Alterovitz G. Temporal phenome analysis of a large electronic health record cohort enables identification of hospital-acquired complications. J Am Med Inform Assoc 2013 Dec;20(e2):e281–7.
26. Sengupta D, Naik PK. SN algorithm: analysis of temporal clinical data for mining periodic patterns and impending augury. J Clin Bioinform 2013 Nov 28;3:24.
27. Batal I, Valizadegan H, Cooper GF, Hauskrecht M. A Temporal Pattern Mining Approach for Classifying Electronic Health Record Data. ACM Trans Intell Syst Technol [Internet] 2013 Sep [cited 2015 Nov 15];4(4). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4192602/
28. Hanauer DA, Ramakrishnan N. Modeling temporal relationships in large scale clinical associations. J Am Med Inform Assoc 2013;20(2):332–41.
29. Wang F, Lee N, Hu J, Sun J, Ebadollahi S, Laine AF. A framework for mining signatures from event sequences and its applications in healthcare data. IEEE Trans Pattern Anal Mach Intell 2013 Feb;35(2):272–85.
30. Hripcsak G, Albers DJ, Perotte A. Exploiting time in electronic health record correlations. J Am Med Inform Assoc 2011 Dec;18(Suppl 1):i109–15.
31. Chen Y, Lorenzi N, Nyemba S, Schildcrout JS, Malin B. We Work with Them? Healthcare Workers Interpretation of Organizational Relations Mined from Electronic Health Records. Int J Med Inform 2014 Jul;83(7):495–506.
32. Murphy DR, Laxmisan A, Reis BA, Thomas EJ, Esquivel A, Forjuoh SN, et al. Electronic health record-based triggers to detect potential delays in cancer diagnosis. BMJ Qual Saf 2014 Jan;23(1):8–16.
33. Perimal-Lewis L, Teubner D, Hakendorf P, Horwood C. Application of process mining to assess the data quality of routinely collected time-based performance data sourced from electronic health records by validating process conformance. Health Informatics J 2015 Oct 11;
34. Ben-Assuli O, Shabtai I, Leshno M. Using electronic health record systems to optimize admission decisions: the Creatinine case study. Health Informatics J 2015 Mar;21(1):73–88.
35. Sung S-F, Hsieh C-Y, Kao Yang Y-H, Lin H-J, Chen C-H, Chen Y-W, et al. Developing a stroke severity index based on administrative data was feasible using data mining techniques. J Clin Epidemiol 2015 Nov;68(11):1292–300.
36. DeFraia GS. Psychological Trauma in the Workplace: Variation of Incident Severity among Industry Settings and between Recurring vs Isolated Incidents. Int J Occup Environ Med 2015 Jul 1;6(3):155–68.
37. Zhang-Salomons J, Salomons G. Determine the therapeutic role of radiotherapy in administrative data: a data mining approach. BMC Med Res Methodol 2015 Feb 3;15(1):11.
38. Hassani S, Lindman AS, Kristoffersen DT, Tomic O, Helgeland J. 30-Day Survival Probabilities as a Quality Indicator for Norwegian Hospitals: Data

Management and Analysis. PLoS One 2015 Sep 9;10(9):e0136547.
39. He D, Mathews SC, Kalloo AN, Hutfless S. Mining high-dimensional administrative claims data to predict early hospital readmissions. J Am Med Inform Assoc 2014 Mar 1;21(2):272–9.
40. Guiney H, Felicia P, Whelton H, Woods N. Analysis of a Payments Database Reveals Trends in Dental Treatment Provision. J Dent Res 2013 Jul 1;92(7 suppl):S63–9.
41. Unnikrishnan KP, Patnaik D, Iwashyna TJ. Spatio-temporal Structure of US Critical Care Transfer Network. AMIA Summits Transl Sci Proc 2011 Mar 7;2011:74–8.
42. Concaro S, Sacchi L, Cerra C, Fratino P, Bellazzi R. Mining Health Care Administrative Data with Temporal Association Rules on Hybrid Events. Methods Inf Med 2011;50(2):166–79.
43. Hussey K, Siddiqui T, Burton P, Welch GH, Stuart WP. Understanding Administrative Abdominal Aortic Aneurysm Mortality Data. Eur J Vasc Endovasc Surg 2015 Mar;49(3):277–82.
44. Thigpen JL, Dillon C, Forster KB, Henault L, Quinn EK, Tripodis Y, et al. Validity of International Classification of Disease Codes to Identify Ischemic Stroke and Intracranial Hemorrhage Among Individuals With Associated Diagnosis of Atrial Fibrillation. Circ Cardiovasc Qual Outcomes 2015 Jan;8(1):8-14.
45. Benchimol EI, Guttmann A, Mack DR, Nguyen GC, Marshall JK, Gregor JC, et al. Validation of international algorithms to identify adults with inflammatory bowel disease in health administrative data from Ontario, Canada. J Clin Epidemiol 2014 Aug;67(8):887–96.
46. Grams ME, Waikar SS, MacMahon B, Whelton S, Ballew SH, Coresh J. Performance and Limitations of Administrative Data in the Identification of AKI. Clin J Am Soc Nephrol 2014 Apr 7;9(4):682–9.
47. Widdifield J, Bernatsky S, Paterson JM, Tu K, Ng R, Thorne JC, et al. Accuracy of Canadian Health Administrative Databases in Identifying Patients With Rheumatoid Arthritis: A Validation Study Using the Medical Records of Rheumatologists. Arthritis Care Res 2013 Oct 1;65(10):1582–91.
48. Mähönen M, Jula A, Harald K, Antikainen R, Tuomilehto J, Zeller T, et al. The validity of heart failure diagnoses obtained from administrative registers. Eur J Prev Cardiol 2013 Apr 1;20(2):254–9.
49. Marrie RA, Yu BN, Leung S, Elliott L, Caetano P, Warren S, et al. Rising prevalence of vascular comorbidities in multiple sclerosis: validation of administrative definitions for diabetes, hypertension, and hyperlipidemia. Mult Scler J 2012 Sep 1;18(9):1310–9.
50. Molodecky NA, Myers RP, Barkema HW, Quan H, Kaplan GG. Validity of administrative data for the diagnosis of primary sclerosing cholangitis: a population-based study. Liver Int 2011 May 1;31(5):712–20.
51. Schultz SE, Rothwell DM, Chen Z, Tu K. Identifying cases of congestive heart failure from administrative data: a validation study using primary care patient records. Chronic Dis Inj Can 2013 Jun;33(3):160–6.
52. Gregori D, Petrinco M, Bo S, Rosato R, Pagano E, Berchialla P, et al. Using Data Mining Techniques in Monitoring Diabetes Care. The Simpler the Better? J Med Syst 2009 Sep 10;35(2):277–81.
53. Wu T-SJ, Shih F-YF, Yen M-Y, Wu J-SJ, Lu S-W, Chang KC-M, et al. Establishing a nationwide emergency department-based syndromic surveillance system for better public health responses in Taiwan. BMC Public Health 2008;
54. Mikosz CA, Silva J, Black S, Gibbs G, Cardenas I. Comparison of two major emergency department-based free-text chief-complaint coding systems. MMWR Suppl 2004 Sep 24;53:101–5.
55. Gesteland PH, Gardner RM, Tsui F-C, Espino JU, Rolfs RT, James BC, et al. Automated syndromic surveillance for the 2002 Winter Olympics. J Am Med Inform Assoc 2003 Dec;10(6):547–54.
56. Silva JC, Shah SC, Rumoro DP, Bayram JD, Hallock MM, Gibbs GS, et al. Comparing the accuracy of syndrome surveillance systems in detecting influenza-like illness: GUARDIAN vs. RODS vs. electronic medical record reports. Artif Intell Med 2013 Nov;59(3):169–74.
57. Eyre DW, Walker AS. Clostridium difficile surveillance: harnessing new technologies to control transmission. [Review]. Expert Rev Antiinfective Ther 2013 Nov;11(11):1193–205.
58. Brownstein JS, Mandl KD. Reengineering real time outbreak detection systems for influenza epidemic monitoring. AMIA Annu Symp Proc 2006:866.
59. Bourgeois FT, Olson KL, Brownstein JS, McAdam AJ, Mandl KD. Validation of syndromic surveillance for respiratory infections. Ann Emerg Med 2006 Mar;47(3).
60. Andreu Perez J, Leff D, Ip H, Yang G-Z. From Wearable Sensors to Smart Implants - Towards Pervasive and Personalised Healthcare. IEEE Trans Biomed Eng 2015 Apr 13;
61. Rehman MH ur, Liew CS, Wah TY, Shuja J, Daghighi B. Mining Personal Data Using Smartphones and Wearable Devices: A Survey. Sensors 2015 Feb 13;15(2):4430–69.
62. Kumar S, Abowd GD, Abraham WT, al'Absi M, Gayle Beck J, Chau DH, et al. Center of excellence for mobile sensor data-to-knowledge (MD2K). J Am Med Inform Assoc 2015 Nov;22(6):1137–42.
63. Zhang Z, Pi Z, Liu B. TROIKA: a general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. IEEE Trans Biomed Eng 2015 Feb;62(2):522–31.
64. Deepu CJ, Lian Y. A joint QRS detection and data compression scheme for wearable sensors. IEEE Trans Biomed Eng 2015 Jan;62(1):165–75.
65. Noh YH, Jeong DU. Implementation of a Data Packet Generator Using Pattern Matching for Wearable ECG Monitoring Systems. Sensors 2014 Jul 15;14(7):12623–39.
66. Lopez-Meyer P, Fulk GD, Sazonov ES. Automatic Detection of Temporal Gait Parameters in Post-stroke Individuals. Ieee Trans Inf Technol Biomed 2011 Jul;15(4):594–601.
67. Fontana JM, Farooq M, Sazonov E. Automatic Ingestion Monitor: A Novel Wearable Device for Monitoring of Ingestive Behavior. IEEE Trans Biomed Eng 2014 Jun;61(6):1772–9.
68. Sazonov E, Lopez-Meyer P, Tiffany S. A Wearable Sensor System for Monitoring Cigarette Smoking. J Stud Alcohol Drugs 2013 Nov;74(6):956–64.
69. Seiter J, Derungs A, Schuster-Amft C, Amft O, Tröster G. Daily life activity routine discovery in hemiparetic rehabilitation patients using topic models. Methods Inf Med 2015;54(3):248–55.
70. Garcia-Ceja E, Brena RF, Carrasco-Jimenez JC, Garrido L. Long-Term Activity Recognition from Wristwatch Accelerometer Data. Sensors 2014 Nov 27;14(12):22500–24.
71. Cole BT, Roy SH, De Luca CJ, Nawab SH. Dynamical Learning and Tracking of Tremor and Dyskinesia From Wearable Sensors. IEEE Trans Neural Syst Rehabil Eng 2014 Sep;22(5):982–91.
72. Gao L, Bourke AK, Nelson J. Evaluation of accelerometer based multi-sensor versus single-sensor activity recognition systems. Med Eng Phys 2014 Jun;36(6):779–85.
73. Tang W, Sazonov ES. Highly Accurate Recognition of Human Postures and Activities Through Classification With Rejection. IEEE J Biomed Health Inform 2014 Jan;18(1):309–15.
74. Teichmann D, Kuhn A, Leonhardt S, Walter M. Human motion classification based on a textile integrated and wearable sensor array. Physiol Meas 2013 Sep;34(9):963–75.
75. Ganea R, Paraschiv-Ionescu A, Aminian K. Detection and Classification of Postural Transitions in Real-World Conditions. IEEE Trans Neural Syst Rehabil Eng 2012 Sep;20(5):688–96.
76. Özdemir AT, Barshan B. Detecting Falls with Wearable Sensors Using Machine Learning Techniques. Sensors 2014 Jun 18;14(6):10691–708.
77. Clifton L, Clifton DA, Pimentel MAF, Watkinson PJ, Tarassenko L. Predictive Monitoring of Mobile Patients by Combining Clinical Observations With Data From Wearable Sensors. IEEE J Biomed Health Inform 2014 May;18(3):722–30.
78. Lopez-Meyer P, Tiffany S, Patil Y, Sazonov E. Monitoring of Cigarette Smoking Using Wearable Sensors and Support Vector Machines. IEEE Trans Biomed Eng 2013 Jul;60(7):1867–72.
79. Barsocchi P. Position Recognition to Support Bedsores Prevention. IEEE J Biomed Health Inform 2013 Jan;17(1):53–9.
80. Sun H, Yuao T. Curve aligning approach for gait authentication based on a wearable accelerometer. Physiol Meas 2012;33(6):1111.
81. Alshamsi A, Pianesi F, Lepri B, Pentland A, Rahwan I. Beyond Contagion: Reality Mining Reveals Complex Patterns of Social Influence. PLoS One [Internet] 2015 Aug 27 [cited 2015 Nov 15];10(8). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4551670/
82. Kher R, Pawar T, Thakar V, Shah H. Physical activities recognition from ambulatory ECG signals using neuro-fuzzy classifiers and support vector machines. J Med Eng Technol 2015 Feb 17;39(2):138–52.
83. Lin H-C, Chiang S-Y, Lee K, Kan Y-C. An Activity Recognition Model Using Inertial Sensor Nodes in a Wireless Sensor Network for Frozen Shoulder Rehabilitation Exercises. Sensors 2015 Jan 19;15(1):2181–204.
84. Lin C-W, Yang Y-TC, Wang J-S, Yang Y-C. A Wearable Sensor Module With a Neural-Network-Based Activity Classification Algorithm for Daily Energy Expenditure Estimation. IEEE Trans Inf Technol Biomed 2012 Sep;16(5):991–8.
85. Wang Z, Jiang M, Hu Y, Li H. An Incremental Learning Method Based on Probabilistic Neural Networks and Adjustable Fuzzy Clustering for

Human Activity Recognition by Using Wearable Sensors. IEEE Trans Inf Technol Biomed 2012 Jul;16(4):691–9.
86. Yurtman A, Barshan B. Automated evaluation of physical therapy exercises using multi-template dynamic time warping on wearable sensor signals. Comput Methods Programs Biomed 2014 Nov;117(2):189–207.
87. Taborri J, Rossi S, Palermo E, Patanè F, Cappa P. A Novel HMM Distributed Classifier for the Detection of Gait Phases by Means of a Wearable Inertial Sensor Network. Sensors 2014 Sep 2;14(9):16212–34.
88. Mannini A, Sabatini AM. Accelerometry-Based Classification of Human Activities Using Markov Modeling. Comput Intell Neurosci [Internet] 2011 [cited 2015 Nov 21];2011. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3166724/
89. Valenza G, Nardelli M, Lanata A, Gentili C, Bertschy G, Paradiso R, et al. Wearable Monitoring for Mood Recognition in Bipolar Disorder Based on History-Dependent Long-Term Heart Rate Variability Analysis. IEEE J Biomed Health Inform 2014 Sep;18(5):1625–35.
90. Koshmak G, Linden M, Loutfi A. Dynamic Bayesian Networks for Context-Aware Fall Risk Assessment. Sensors 2014 May 23;14(5):9330–48.
91. Nam Y, Park JW. Child Activity Recognition Based on Cooperative Fusion Model of a Triaxial Accelerometer and a Barometric Pressure Sensor. IEEE J Biomed Health Inform 2013 Mar;17(2):420–6.
92. Liu S, Gao RX, John D, Staudenmayer JW, Freedson PS. Multisensor Data Fusion for Physical Activity Assessment. IEEE Trans Biomed Eng 2012 Mar;59(3):687–96.
93. Piatetsky-Shapiro, G. Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. AI Mag 1989;11(5):68–70.
94. Ericsson KA, Simon HA. Protocol analysis: Verbal reports as data. MIT Press; 1992.
95. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain. An evolving diagnostic decision-support system. JAMA 1987 Jul 3;258(1):67–74.
96. Evans DA, Patel VL, editors. Advanced Models of Cognition for Medical Training and Practice [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 1992 [cited 2016 Jul 9]. Available from: http://link.springer.com/10.1007/978-3-662-02833-9
97. Patel VL, Kannampallil TG. Cognitive informatics in biomedicine and healthcare. J Biomed Inform 2015 Feb;53:3–14.
98. Barnes FS. Some engineering models for interactions of electric and magnetic fields with biological systems. [Review] [63 refs]. Bioelectromagnetics 1992;67–85.
99. Cavallo V, Giovagnorio F, Messineo D. Neural networks in diagnostic imaging. Rays 1992 Dec;17(4):556–61.
100. Clark BD, Leong SW. Crossmatch prediction of highly sensitized patients. Clin Transpl 1992;435–55.
101. Howells SL, Maxwell RJ, Peet AC, Griffiths JR. An investigation of tumor 1H nuclear magnetic resonance spectra by the application of chemometric techniques. Magn Reson Med 1992 Dec;28(2):214–36.
102. Maclin PS, Dempsey J. Using an artificial neural network to diagnose hepatic masses. J Med Syst 1992 Oct;16(5):215–25.
103. Miller AS, Blott BH, Hames TK. Review of neural network applications in medical imaging and signal processing. [Review] [81 refs]. Med Biol Eng Comput 1992 Sep;30(5):449–64.
104. Schaberg ES, Jordan WH, Kuyatt BL. Artificial intelligence in automated classification of rat vaginal smear cells. Anal Quant Cytol Histol 1992 Dec;14(6):446–50.
105. Vieth M, Kolinski A, Skolnick J, Sikorski A. Prediction of protein secondary structure by neural networks: encoding short and long range patterns of amino acid packing. Acta Biochim Pol 1992;39(4):369–92.
106. Wittenberg G, Kristan WBJ. Analysis and modeling of the multisegmental coordination of shortening behavior in the medicinal leech. II. Role of identified interneurons. J Neurophysiol 1992 Nov;68(5):1693–707.
107. Wu C, Whitson G, McLarty J, Ermongkonchai A, Chang TC. Protein classification artificial neural system. Protein Sci 1992 May;1(5):667–77.
108. Kernan WJ, Meeker WQ. A statistical test to assess changes in spontaneous behavior of rats observed with a computer pattern recognition system. J Biopharm Stat 1992;2(1):115–35.
109. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1988.
110. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Mach Learn 9(4):309–47.
111. Lauritzen SL, Spiegelhalter DJ. Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. J R Stat Soc Ser B Methodol 1988;50(2):157–224.
112. Shwe MA, Middleton B, Heckerman DE, Henrion M, Horvitz EJ, Lehmann HP, et al. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. I. The probabilistic model and inference algorithms. Methods Inf Med 1991 Oct;30(4):241–55.
113. Cowell RG, Dawid AP, Hutchinson T, Spiegelhalter DJ. A Bayesian expert system for the analysis of an adverse drug reaction. Artif Intell Med 1991 Oct 1;3(5):257–70.
114. Spiegelhalter DJ. Probabilistic Reasoning in Expert Systems. Am J Math Manage Sci 1991 Jan;9(3–4):191–210.
115. Muggleton S. Inductive logic programming. New Gener Comput 1991 Feb;8(4):295–318.
116. Lavrac N, Dzeroski S. Inductive Logic Programming: Techniques and Applications. New York, NY, 10001: Routledge; 1993.
117. Koza JR. Genetic programming: on the programming of computers by means of natural selection [Internet]. MIT Press; 1992 [cited 2016 Jul 9]. Available from: http://dl.acm.org/citation.cfm?id=138936
118. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. J Mach Learn Res 2003 Mar;3:993–1022.
119. Boser BE, Guyon IM, Vapnik VN. A Training Algorithm for Optimal Margin Classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory [Internet]. New York, NY, USA: ACM; 1992 [cited 2015 Nov 22]. p. 144–152. (COLT '92). Available from: http://doi.acm.org/10.1145/130385.130401
120. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004 Oct 21;431(7011):931–45.
121. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996 Dec;14(13):1675–80.
122. Precision Medicine Initiative [Internet]. National Institutes of Health (NIH). [cited 2016 Jul 15]. Available from: https://www.nih.gov/precision-medicine-initiative-cohort-program
123 Sacchi L, Lanzola G, Viani N, Quaglini S. Personalization and Patient Involvement in Decision Support Systems: Current Trends. Yearb Med Inform 2015 Aug 13;10(1):106–18.
124. Toga AW, Clark KA, Thompson PM, Shattuck DW, Van Horn JD. Mapping the Human Connectome. Neurosurgery 2012 Jul;71(1):1–5.
125. Health Information Privacy | HHS.gov [Internet]. [cited 2016 Jan 24]. Available from: http://www.hhs.gov/hipaa/
126. BD2K Home Page | Data Science at NIH [Internet]. [cited 2016 Jul 15]. Available from: https://datascience.nih.gov/bd2k
127. Murdoch TB, Detsky AS. THe inevitable application of big data to health care. JAMA 2013 Apr 3;309(13):1351–2.
128. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. Nat Rev Cardiol 2016 Jun;13(6):350–9.
129. Schmidhuber J. Deep learning in neural networks: an overview. [Review]. Neural Netw 2015 Jan;85–117.
130. LeCun Y, Bengio Y, Hinton G. Deep learning. [Review]. Nature 2015 May;521(7553):436–44.
131. Guclu U, van Gerven MAJ. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. J Neurosci 2015 Jul;35(27):10005–14.
132. Yan Z, Zhan Y, Peng Z, Liao S, Shinagawa Y, Metaxas DN, et al. Bodypart Recognition Using Multi-stage Deep Learning. Inf Process Med Imaging 2015;449–61.
133. Ibrahim R, Yousri NA, Ismail MA, El-Makky NM. Multi-level gene/MiRNA feature selection using deep belief nets and active learning. Conf Proc 2014;3957–60.
134. Lake BM, Salakhutdinov R, Tenenbaum JB. Human-level concept learning through probabilistic program induction. Science 2015 Dec;350(6266):1332–8.

Correspondence to:
John H Holmes
Institute for Biomedical Informatics
University of Pennsylvania School of Medicine
717 Blockley Hall
423 Guardian Drive
Philadelphia, PA 19104, USA
Tel: 215-898-4833
Fax: 215-573-5325
E-Mail: jhholmes@mail.med.upenn.edu

# Appendix – Detailed results of LDA analysis

**Table 1** Results of LDA analysis applied to ANN abstracts and considering eight topics. For each topic, we show the number of extracted papers and the five most representative words.

| ANNs | | | | | | | |
|---|---|---|---|---|---|---|---|
| Water quality (502 papers) | Cancer research (479 papers) | Control systems (706 papers) | Motion control (340 papers) | Predictive Modeling (502 papers) | Analysis of brain signals (731 papers) | Image analysis (451 papers) | Bioinformatics (338 papers) |
| water | compounds | control | patients | patients | neurons | image | protein |
| quality | cancer | stability | subjects | diagnosis | brain | images | sequence |
| concentration | descriptors | state | control | risk | memory | detection | cell |
| temperature | gene | numerical | brain | cancer | synaptic | signals | sequences |
| concentrations | genes | controller | motion | logistic | spiking | recognition | cells |

**Table 2** Results of LDA analysis applied to SVM abstracts and considering eight topics. For each topic, we show the number of extracted papers and the five most representative words

| SVMs | | | | | | | |
|---|---|---|---|---|---|---|---|
| Bioinformatics (854 papers) | | | | Brain signal analysis (633 papers) | | Image analysis (333 papers) | Analysis of drug compounds (285 papers) |
| gene | genes | protein | spectroscopy | eeg | brain | images | compounds |
| protein | gene | amino | variables | signals | fmri | cancer | drug |
| genes | expression | sites | calibration | brain | matter | image | descriptors |
| cancer | cancer | structural | quality | signal | mri | segmentation | chemical |
| expression | microarray | composition | mci | visual | connectivity | breast | drugs |

**Table 3** Results of LDA analysis applied to abstracts of papers dealing with Bayesian techniques and considering eight topics. For each topic, we show the number of extracted papers and the five most representative words

| HMMs | | | | | | | |
|---|---|---|---|---|---|---|---|
| Social Interaction (12 papers) | Bioinformatics (30 papers) | | | Activity Recognition (16 papers) | Speech Recognition (14 papers) | Motion and Posture Analysis (10 papers) | Prediction of adverse drug events (6 papers) |
| Social | Enzyme | Protein | Protein | Recognition | Sequences | Temporal | Drug |
| Human | Genes | Motifs | Annotation | Activity | Speech | Motion | Substrate |
| Interaction | Evolutionary | Sequences | Motifs | System | Emotional | Pose | Predicion |
| Classification | Hisa | Time | Recognition | Training | Training | Subcellular | Adverse |
| Behavioral | Identify | Zinc | Coral | Distributed | Health | Templates | Specificity |

**Table 4** Results of LDA analysis applied to abstracts of papers dealing with Bayesian techniques and considering eight topics. For each topic, we show the number of extracted papers and the five most representative words.

| Bayesian Techniques | | | | | | | |
|---|---|---|---|---|---|---|---|
| Image Analysis (163 papers) | | Adverse drug events (128 papers) | | Activity recognition (139 papers) | | Analysis of gene expression data (79 papers) | Cancer Research (71 papers) |
| image | functional | adverse | diagnosis | tracking | activity | gene | cancer |
| segmentation | fmri | reporting | reports | recognition | risk | kernel | genes |
| images | brain | events | expert | subjects | optimization | protein | risk |
| inference | variables | signals | vaers | injury | fusion | expression | lung |
| distribution | quality | reports | tcm | gene | interactions | genes | liver |