YALE JOURNAL OF BIOLOGY AND MEDICINE 85 (2012), pp.363-377.

FOCUS: EDUCATING YOURSELF IN BIOINFORMATICS

The Future of Medical Diagnostics: Large Digitized Databases

Wesley T. Kerr[a,b*], Edward P. Lau[c], Gwen E. Owens[b,d], and Aaron Trefler[e]

[a]Department of Biomathematics, University of California, Los Angeles, California; [b]UCLA-Caltech Medical Scientist Training Program, Los Angeles, California; [c]Department of Psychiatry, University of California, Los Angeles, California; [d]California Institute of Technology Graduate Program in Biochemistry and Molecular Biophysics, Los Angeles, California; [e]Department of Psychology, University of California, Los Angeles, California

The electronic health record mandate within the American Recovery and Reinvestment Act of 2009 will have a far-reaching affect on medicine. In this article, we provide an in-depth analysis of how this mandate is expected to stimulate the production of large-scale, digitized databases of patient information. There is evidence to suggest that millions of patients and the National Institutes of Health will fully support the mining of such databases to better understand the process of diagnosing patients. This data mining likely will reaffirm and quantify known risk factors for many diagnoses. This quantification may be leveraged to further develop computer-aided diagnostic tools that weigh risk factors and provide decision support for health care providers. We expect that creation of these databases will stimulate the development of computer-aided diagnostic support tools that will become an integral part of modern medicine.

---

*To whom all correspondence should be addressed: Wesley T. Kerr, 760 Westwood Plaza, Suite B8-169, Los Angeles, CA 90095; Email: wesley1610@gmail.com.

†Abbreviations: AD, Alzheimer's disease; ADNI, Alzheimer's Disease Neuroimaging Initiative; ADHD, Attention Deficit and Hyperactivity Disorder; AED, automated electronic defibrillator; ARRA, American Recovery and Reinvestment Act of 2009; CAD, computer-aided diagnostic; CDC, Centers for Disease Control; CT, X-ray computed tomography; EDE, European Database for Epilepsy; EHR, electronic health record; EKG, electrocardiogram; FTLD, fronto-temporal lobar degeneration; HIPAA, Healthcare Insurance Portability and Accountability Act; IRB, institutional review board; MCI, mild cognitive impairment; ML, machine learning; MRI, magnetic resonance imaging; NPCD, National Patient Care Database; NIH, National Institutes of Health; PGP, Personal Genome Project; PPI, Protected Patient Information; RFI, request for information; SVM, Support Vector Machines; TED, Technology, Entertainment, Design; UCLA, University of California, Los Angeles.

Keywords: electronic health record, computer-aided diagnostics, machine learning, databases

Contributions: WTK initiated, organized and wrote the majority of this article. EPL created Figure 3 and contributed significantly to sections regarding computational efficiency, the structure of databases and the content strategy problem. GEO created Figures 1 and 2 and Table 1 and contributed significantly to sections regarding patient and physician attitudes toward databases and computer-aided diagnostics. AT contributed by conducting substantial literature review to support the ideas expressed. All co-authors and one outsider provided substantial editorial support to WTK. WTK is funded by the UCLA-Caltech MSTP, the UCLA SIBTG and the UCLA Department of Biomathematics. EPL is funded by NIH R33 DA026109. GEO is funded by the UCLA-Caltech MSTP, the Caltech Graduate Program in Biochemistry and Molecular Biophysics, and the Hearst Foundation. AT is funded by the UCLA Department of Psychology.

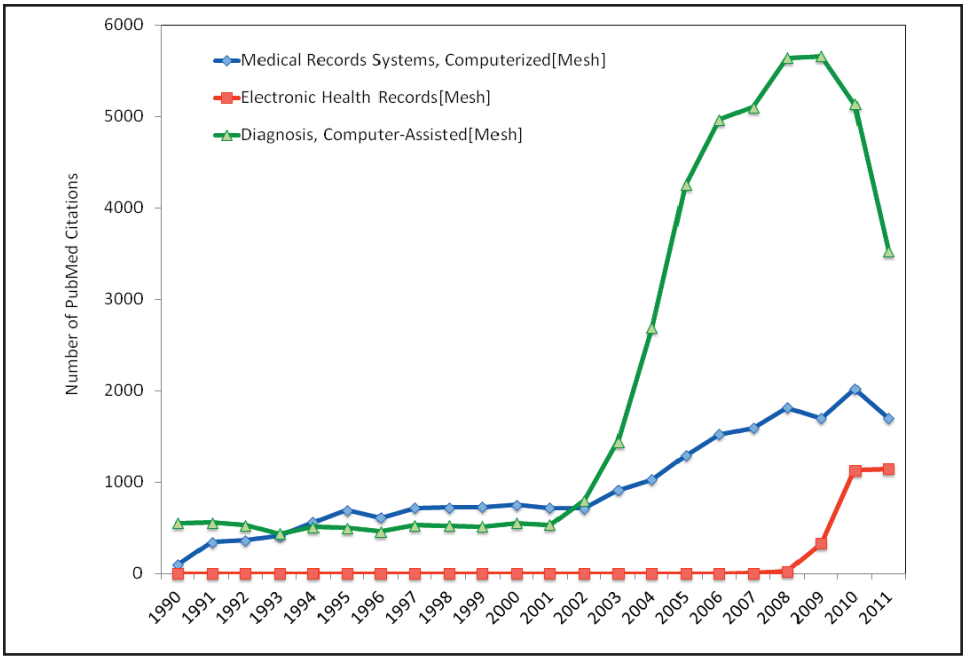

Figure 1. This figure illustrates the number of PubMed citations using each of the Mesh terms listed. Since 2002, the number of publications regarding computer-aided diagnostics has increased substantially. We are already seeing a commensurate increase in the number of publications regarding computerized medical record systems and electronic health records [1].

INTRODUCTION

The impact of nationwide implementation of electronic health record (EHR†) systems will change the daily practice of medicine as we know it. With medical records in their current state, it is extremely difficult to efficiently collate records and mine clinical information to understand trends in and differences between various patient populations. This limits the size of patient groups and thereby reduces the statistical power of many research protocols [2]. The EHR mandate will stimulate institutions to digitize their records in common formats amenable to collating data into large databases. These databases with records from potentially millions of patients can then be processed using sophisticated data mining techniques. There are numerous regulatory, practical, and computational challenges to creating and maintaining these databases that will need to be appropriately addressed. Many groups are already compiling large databases of high quality patient information with great success [3-11]. Based on its previous efforts, we expect the National Institutes of Health (NIH) to fully support researchers who seek to tackle the challenges of creating EHR-based databases that include clinical notes and other data points such as laboratory results and radiological images. Such databases will be invaluable to the development of computer-aided diagnostic (CAD) tools that, we believe, will be responsible for many advances in the efficiency and quality of patient care [2]. CAD tools are automated programs that provide synthesized diagnostic information to providers that are not otherwise readily accessible. The rate of development of CAD tools and the mining of medical record systems has increased markedly since 2002 (Figure 1), and we expect the development of large EHR-based databases will only stimulate this activity further. In this article, we provide an in-depth analysis of the effect of the EHR mandate on the development of databases that could be mined to create high quality CAD tools. Further, we illustrate how computer-aided diagnostics can be integrated efficiently into daily medical practice.

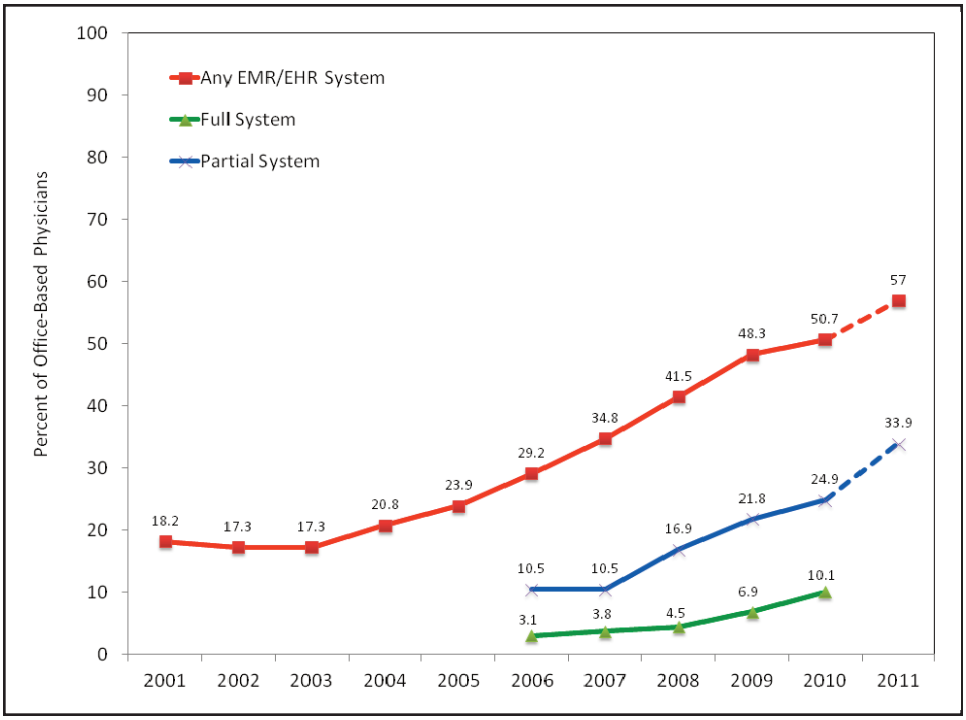

Figure 2. Even before the ARRA in 2009, the number of physicians utilizing EHR systems was increasing. There are already a substantial percent of physicians using electronic records. Consequentially, it is relatively inexpensive to combine and mine these EHR systems for high quality clinical information.

MANDATES AND POLICIES DRIVING THE CHANGE

Although the growth of large digitized databases is stimulated by numerous sources, there are two key policy decisions that have the potential to dramatically speed this growth and change medical diagnostics as we know it: the final NIH statement on sharing research data in 2003 and the EHR mandate in the American Recovery and Reinvestment Act of 2009 (ARRA) [12,13].

The seed for developing large open databases of medical information was planted initially by the NIH statement on sharing research data. In 2003, the NIH mandated that "investigators submitting an NIH application seeking $500,000 or more in direct costs in a single year are expected to include a plan for data sharing" [13]. A large portion of academic medicine research is funded through grants of this type, and therefore, the amount of high quality information about patients in the public domain is growing rapidly. This may be one reason why interest in computerized medical record systems increased in 2003 (Figure 1). Unfortunately, the NIH has identified that this policy has not led to the degree of data sharing it anticipated, as evidenced by NOT-DA-11-021 entitled "Expansion of sharing and standardization of NIH-funded human brain imaging data" [14]. The focus of this request for information (RFI) was to identify the barriers to creating an open-access database for brain imaging data, including medically relevant images. This RFI implies that the NIH likely would support efforts to establish large, open digitized databases that include patient information.

Those who designed the ARRA presumably recognized the potential of digitized medicine and decided to support its development. In the ARRA, $20 billion was provided to establish EHRs for all patients treated in the United States [12]. Health care providers that do not establish an EHR system after 2014 will be subject to fines. This

Table 1. Prominent Medical Databases.

Database	Information Contained	Funding Source(s)	Access	Website
ADHD-200	776 resting-state fMRI and anatomical datasets along and accompanying phenotypic information from 8 imaging sites; 285 of which are from children and adolescents with ADHD aged 7-21	NIH	Research community	fcon_1000.projects.nitrc.org/indi/adhd200/index.html
Alzheimer's Disease Neuroimaging Initiative (ADNI)	Information on 200 control patients, 400 patients with mild cognitive impairment, and 200 with Alzheimer's disease	NIH	Public access	www.adni-info.org/
Australian EEG Database	18,500 EEG records from a regional public hospital	Hunter Medical Research Insitute and the University of Newcastle Research Management Committee	User access required (administrator, analyst, researcher, student)	aed.newcastle.edu.au:9080/AED/login.jsp
Clinical Trials	Registry and results of >100,000 clinical trials	NIH	Public access	clinicaltrials.gov/
Epilepsiae European Database on Epilepsy	Long-term recordings of 275 patients	European Union	Research community	www.epilepsiae.eu/
Healthfinder	Encyclopedia of health topics	Department of Health and Human Services	Public access	healthfinder.gov/
Kaiser Permanente National Research Database	Clinical information on >30 million members of the Kaiser Foundation Health Plan	Kaiser Foundation Research Institute	Kaiser Permanente researchers and collaborating non-KP researchers	www.dor.kaiser.org/external/research/topics/Medical_Informatics/
National Patient Care Database (NPCD)	Veterans Health Administration Medical Dataset	U.S. Department of Veterans Affairs	Research community	www.virec.research.va.gov/DataSourcesName/NPCD/NPCD.htm
Personal Genome Project (PGP)	1,677+ deep sequenced genomes. Goal is 100,000 genomes	NIH and private donors	Open consent	www.personalgenomes.org/
PubMed	Article titles and abstracts	NIH	Public access	www.ncbi.nlm.nih.gov/pubmed/

A quick summary of notable databases of high quality information that have been developed and are being used for large scale studies.

was intended to further stimulate the trend of increased utilization of EHR systems (Figure 2). As stated in the bill, the reasons for this mandate include reduction of medical errors, health disparities, inefficiency, inappropriate care, and duplicative care. Further, the ARRA EHR mandate has and is meant to improve coordination, the delivery of patient-centered medical care, public health activities, early detection, disease prevention, disease management, and outcomes [12,15]. To facilitate these advances, the knowledge about and methods for bioinformatics must be applied to millions of EHRs to develop automated computer-aided diagnostic (CAD) tools. For example, one efficient way to avoid inappropriate care is for an automated program to produce an alert when a health care provider attempts to provide questionable service. The development of such CAD tools is not trivial; however, large high-quality, open EHR databases will greatly decrease development costs and accelerate testing. Below, we discuss why it is our firm belief that these databases will make the implementation of computer-aided diagnostics virtually inevitable.

LARGE DATABASES

There are a growing number of these large databases populated with clinically relevant information from patients suffering from a diverse range of medical conditions, some already including detailed multimodal information from hundreds to millions of patients. Here we will briefly review the General Practice Research Database (GPRD), the Alzheimer's Disease Neuroimaging Initiative (ADNI), the Personal Genome Project (PGP), the European Database on Epilepsy (EDE), and the Australian EEG Database. These and other databases are summarized in Table 1.

The GPRD includes quality text-based records from more than 11 million patients primarily from the United Kingdom but also includes patients from Germany, France, and the United States [4,16]. The database is used primarily by pharmacoepidemiologists, though other researchers are mining this database actively to create automated tools that extract, at base, the diagnostic conclusions reported in each note [17,18]. Although the recall and precision of these tools was good — 86 percent and 74 percent, respectively, in one study [17] — these tools are constantly improving. We expect the increasing size of this and other databases will further stimulate high quality research in this field and result in highly efficient and effective data extraction tools. This conclusion is supported by the fact that more than 73 scholarly publications utilized the GPRD in the first three quarters of 2011 alone [16]. This database, however, is limited to the text of the clinical notes.

Other databases go further by providing complex data regarding large cohorts of patients. The ADNI database contains data fields that track the rate of change of cognition, brain structure and function from 800 patients, including 200 with Alzheimer's disease (AD) and 400 with mild cognitive impairment (MCI) [7]. Researchers are planning to add 550 more patients to this cohort in ADNI2 [6]. The current ADNI database includes full neuroimaging data from all of these patients in the hope that this data can be used to discover the early warning signs for AD. ADNI has been used already to develop machine learning (ML) tools to discriminate between AD and "normal" aging [19]. Another database compiled by the PGP currently has 1,677 patients, and researchers plan to expand this to include nearly complete genomic sequences from 100,000 volunteers using open-consent [3]. Researchers involved in the PGP anticipate that this sequence information will be used to understand risk profiles for many heritable diseases [8]. Other similarly large databases of complex data already exist; the EDE contains long-term EEG recordings from 275 patients with epilepsy [10,11], and the Australian EEG Database holds basic notes and full EEG results from more than 20,000 patients [5,9]. These databases have been used to develop sophisticated seizure prediction and detection tools. Here at the University of California, Los Angeles (UCLA), we are compiling a database of clinical notes, scalp

EEG, MRI, PET, and CT records from more than 2,000 patients admitted for video-EEG monitoring.

The existence of these databases containing detailed clinically relevant information from large patient cohorts confirms that the international research establishment and the NIH are extremely excited about and supportive of large clinical databases. This suggests that as the EHR mandate simplifies collation of patient data, the limiting factor in generating large databases of thousands to millions of patient records will be for organizations to work through the practical hurdles of consenting patients and making data available for efficient searching and processing.

ANTICIPATED CHALLENGES TO DATABASE CREATION

Our conclusion that large clinical databases will continue to expand is based on key assumptions that important regulatory and computational hurdles will be overcome. These challenges include, but are not limited to: 1) patient consent, 2) IRB approval, and 3) consistent improvements in processing these large datasets. We believe the probability that these potential problems will be solved is high.

Forming open databases requires that patients consent to the sharing of pertinent parts of their medical records. In the development of the Personal Genome Project (PGP), Church et al. established open-consent so that all de-identified records can be shared freely [3]. Patients in EHR databases would likely utilize an identical open-consent process. We have personal experience analyzing datasets that require consenting adult patients admitted for video-EEG monitoring for epilepsy as well as pediatric epilepsy patients undergoing assessment for resective neurosurgery at UCLA. After we explained that consent would have no impact on their care, every patient admitted for these reasons (716/716) consented to their records being used for research. Weisman et al. reported that 91 percent of respondents would be willing to share their records for "health research" and that most would be more comfortable with an opt-in system [20]. Other surveys of patients report a consent rate of approximately 40 percent for providing de-identified information to researchers [21,22]. Even after consenting, patients are relatively uninformed about the safeguards and risks to sharing their health information [23]. A more detailed and careful explanation of these procedures and the potential impact of the research may result in an increased consent rate. Any national patient database is likely to face pushback from a public already concerned about invasions of privacy by corporations and the government; therefore, we suspect consent rates would be lower than what we have experienced. Additionally, the rate of consent is likely to decline, in part, due to media coverage of previous unethical practices in research. A prime example is the book, *The Immortal Life of Henrietta Lacks* by Rebecca Skoot, published in 2010, that recounts how, due to lack of proper regulation in 1951, Ms. Lacks' cells were immortalized without her consent and used widely for important advances in medical research [24]. We expect that patients and regulators sensitive to the concept of information about their care being stored indefinitely for research use may not consent on the basis of this and other salient examples.

The key regulatory challenge to the creation of such large databases, however, is the complex multicenter IRB approval process. The most important concern that current IRBs have expressed is whether the data stream includes adequate de-identification of all records before they are released for research use, as illustrated in Figure 3. This would likely require each contributing institution to develop a reliable and consistent method of de-identifying all records. For written records, this includes removing all protected patient information (PPI) as defined by HIPAA regulations and the Helsinki Declaration [25,26]. In order to do this effectively, numerous safeguards must be put in place. For example, if a nationwide database is updated in real time, malicious individuals could potentially re-identify individual pa-

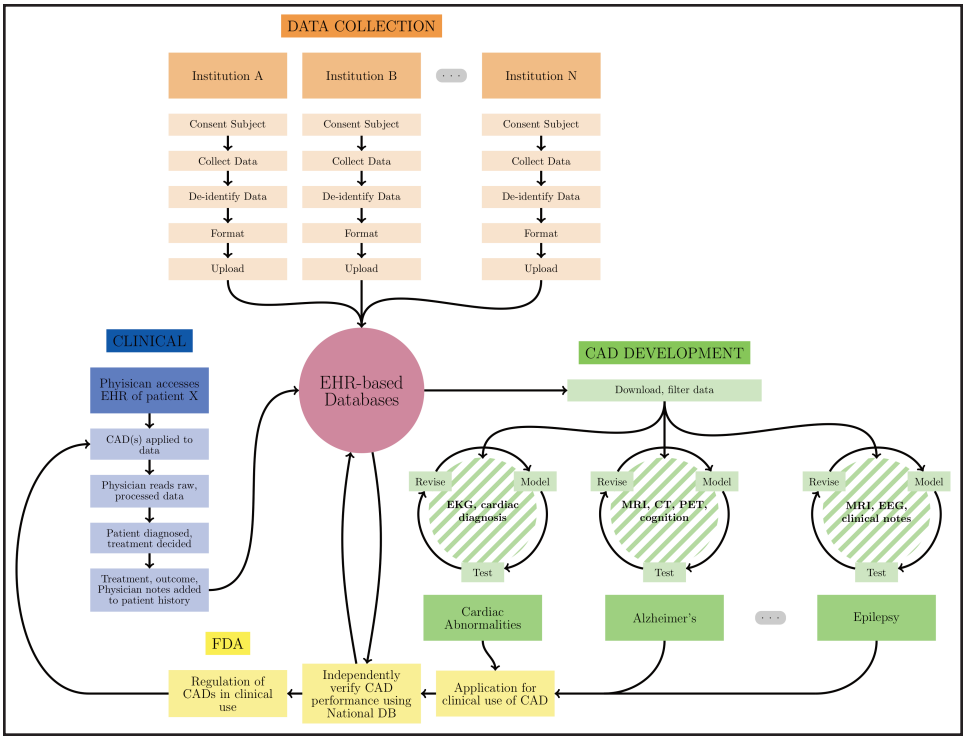

Figure 3. The creation and utilization of EHR databases is complex; however, each of the steps in the data and implementation stream are well defined. We expect that responsible researchers will be capable of tackling each of these steps to create unparalleled databases and develop high quality, clinically applicable CAD tools.

tients by their treatment location, date, and basic information as to what care they received. One solution to minimize these risks, suggested by Malin et al., is to granulize dates and treatment locations to ensure that the potential re-identification rate of patients remains well below 0.20 percent [23]. This granulation may also allow for inclusion of patients older than 89, the maximum reportable age under HIPAA regulations [25]. Although specific dates and locations are important, especially to the Centers for Disease Control (CDC), simply generalizing days to months and towns to counties is required to maintain patient privacy. When dealing with more complex records as in neuroimaging, all centers would be required to be proactive in using the most up-to-date software for de-identification including, but not limited to, the removal of the bone and skin structure of the face that can be used to recreate an image of the patient's face and thereby identify the patient. Automated software to do these complex steps has already been made publicly available by the Laboratory of Neuroimaging (LONI) at UCLA [27]. Due to the unprecedented quality and applicability of these large databases, we are confident that responsible researchers will work to identify and address these regulatory hurdles.

Lastly, the computational burden of utilizing such large databases is not trivial. The question is not if mining this database is possible, it is when. Moore's law has accurately predicted the biennial doubling of computer processing power [28], and, though this rate is showing signs of slowing, growth still is exponential [29]. Current ML methods have been effectively applied to the ADNI database of 800 patients [19,30-32] and as well as the GPRD of almost 12 million patients from the United Kingdom [16]. This suggests that if adequate computational technology does not already exist to effectively mine U.S.-based EHR databases, it will be available soon.

CURRENT APPLICATIONS AND BENEFITS OF CAD

The application of CAD to patient data is not a novel idea. Numerous CAD tools have been demonstrated to be extremely useful to clinical medicine, but few have been approved for routine clinical use [2,33-48]. In general, these tools attempt to predict the outcome of more expensive or practically infeasible gold standard diagnostic assessments. Humans are capable of weighing at most 30 factors at once using only semi-quantitative modeling [49]. The key exception to this is visual processing in which the visual heuristic reliably removes noise from images to readily detect the underlying patterns [50]. This exquisite pattern detection, however, is limited by our inability to detect relationships separated widely in space or time or whose patterns evolve out of changes in randomness. Further, human performance is highly variable due to the effects of expertise, fatigue, and simply human variation [51]. Computational analysis, on the other hand, can integrate complex, objective modeling of thousands to millions of factors to reliably predict the outcome of interest [52]. During validation, the performance of a CAD tool is described in detail to understand its strengths and weaknesses. Unlike manual analysis, given a similar population of test samples, a CAD tool can be expected to perform exactly as it did during validation. In some cases, the constantly updating algorithms inherent in human decision-making may result in deviation from the previously studied ideal. It is not certain that this deviation always results in improved sensitivity and specificity. The cost of expert analysis of clinical information also is increasing continually. Effective implementation of automated screening tools has the potential to not only increase the predictive value of clinical information but also to decrease the amount of time a provider needs to spend analyzing records. This allows them to review more records per day and thereby reduce the cost per patient so that the effective public health impact of each provider is increased [53]. This will complement the numerous potential benefits quoted above. Here we review the success of implemented CAD tools and highly promising new tools that have demonstrated the potential for wider application. In particular, CAD tools have been applied to aid in the diagnosis of three extremely prevalent maladies in the United States: heart disease, lung cancer, and Alzheimer's disease (AD).

The most widely recognized CAD tool in clinical medicine is built into electrocardiogram (EKG) currently available software and reads EKG records and reports any detected abnormalities. These algorithms are responsible for the lifesaving decisions made daily by automated electronic defibrillators (AEDs). The diagnosis of more complex cardiac abnormalities is an extremely active area of research [33-44,54-56]. In one recent example, a CAD tool differentiated between normal beats, left and right bundle block (LBBB and RBBB), and atrial and ventricular premature contraction (AVP, PVC) with more than 89 percent accuracy, sensitivity, specificity and positive predictive value [35]. This and other automated algorithms detect subtle changes in the shape of each beat and variations in the spectral decomposition of each beat over an entire EKG recording that often includes thousands of beats. As a result of this accuracy, conventional EKG readouts in both hospitals and clinics frequently include the results of this entirely automated analysis. When taught to read EKGs, providers are instructed that the automated algorithm is largely correct, but to better understand the complex features of the waveforms, providers must double check the algorithm using their knowledge of the clinical context. This CAD tool was the first to be widely applied because, in part, EKG analysis is simplified by the presence of the characteristically large amplitude QRS wave that can be used to align each beat. Other modalities do not necessarily have features that are as amenable to modeling.

One example of overcoming this lack of clear features is the semi-automated analysis of thoracic X-ray computed tomography (CT) images to detect malignant lung cancer nodules. This tool segments the CT into

bone, soft tissue, and lung tissue, then detects nodules that are unexpectedly radiolucent and assesses the volume of the solid component of non-calcified nodules [48]. This method effectively detected 96 percent of all cancerous nodules with a sensitivity of 95.9 percent and a specificity of 80.2 percent [48]. Even though this tool is not part of routine care, Wang et al. demonstrated that when radiologists interpret the CTs after the CAD tool, they do not significantly increase the amount of cancer nodules detected [48]. In fact, they only increase the number of false positive nodules, indicating that the CAD tool is operating on meaningful features of the nodules that are not reliably observable even by trained radiologists. This suggests that in some cases, computer-aided diagnostics can reduce the number of images that radiologists have to read individually while maintaining the same high quality of patient care.

The success of CAD tools in Alzheimer's disease (AD) shows exactly how automated tools can utilize features not observable by trained radiologists by reliably discriminating AD from normal aging and other dementias. Because of its unique neuropathology, AD requires focused treatment that has not been proven to be effective for other dementias [57]. The gold standard diagnostic tool for AD is cerebral biopsy or autopsy sample staining of amyloid plaques and neurofibrillary tangles [57]. The clear drawback of autopsy samples is that they cannot be used to guide treatment and cerebral biopsy is extremely invasive. An alternative diagnostic is critical for reliably distinguishing between the two classes of patients at a stage that treatment is effective. In 2008, Kloppel et al. demonstrated how a support vector machine (SVM)-based CAD tool performed similarly to six trained radiologists when comparing AD to normal aging and fronto-temporal lobar dementia (FTLD) using structural magnetic resonance imaging (MRI) alone [58]. Numerous other applications of ML on other datasets all have achieved similar accuracies ranging from 85 to 95 percent [19,31,32,59,60]. All of these tools do not require expertise to read; therefore, they can be applied both at large research institutions and in smaller settings as long as the requisite technology is available. These tools, with appropriate validation using large databases, could indicate which patients would benefit most from targeted treatment and therefore substantially reduce morbidity.

These cases are exemplary; however, many other attempts to develop CAD tools have had more limited success. In particular, the automated analysis of physician's notes has proven particularly difficult. In a 2011 publication using a total of 826 notes, the best precision and recall in the test set were 89 percent and 82 percent, respectively [61]. These values are extremely encouraging when considering a similar study in 2008 that attempted to measure the health-related quality of life in 669 notes and achieved only 76 percent and 78 percent positive and negative agreement between the automated algorithm and the gold standard [62]. When viewing these accuracies in terms of the potential of applying these tools to patients, these accuracies are far from adequate. Physicians can quickly scan these notes and immediately understand the findings within them, and therefore, these CAD tools would not improve upon the standard of care if used to summarize the note. Nevertheless, note summaries are useful in an academic setting. It is possible that these tools can be used to interpret thousands of notes quickly and without using any physician time. Even though more than 10 percent of the interpretations are inaccurate, the findings of the CAD tool could be used in a research setting to estimate the risk of other outcomes in these patients, including their risk for cardiovascular disease and even death.

BENEFITS AND CHALLENGES OF DATABASES IN THE DEVELOPMENT OF CAD TOOLS

The establishment of databases made possible by the EHR mandate has enormous potential for the development of CAD tools. A telling quotation from Rob Kass, an expert in Bayesian statistics, reads: "the accumulation of data makes open minded

observers converge on the truth and come to agreement" [63]. In this setting, the accumulation of a gigantic body of clinical data in the form of EHR databases will be invaluable for the description of numerous clinical syndromes and disease. If these databases are unbiased, high quality samples of patients from the general population, there will be no better dataset with which to apply bioinformatics methods to understand the epidemiology, co-morbidities, clinical presentation, and numerous other features of most syndromes and diseases. In addition to quantifying what is known already, these large databases can facilitate the development of new hypotheses regarding neurobiological and genetic underpinnings of these conditions through machine learning approaches [64]. One of the constant factors that limit many clinical and research studies is the steep cost of obtaining high quality data that can be used to develop and test continually updated hypotheses. EHR databases would drastically reduce this cost and thereby allow more funds to be dedicated to the development of models that better elucidate the biology underlying each condition.

In addition to facilitating more applicable and statistically powerful modeling, increased sample size also results in increased machine learning performance. In theory, as sample size increases, the amount of detected signal grows, resulting in an accuracy that is a sigmoid function of sample size. Each feature would therefore have a maximum discriminatory yield that can only be achieved with a sufficiently large training sample size. Using the ADNI database, Cho et al. confirmed this theoretical result by demonstrating that the accuracy of all tested discriminations increased monotonically with the number of training subjects [19]. Therefore, in order to develop the most accurate and therefore applicable CAD tool, one must train it on as large a representative sample size as can be obtained. As noted by van Ginneken et al. [2], if one CAD tool is already FDA approved, securing adequate funding to prove a new tool performs better is a major hurdle. Large EHR databases would lower this barrier and foster innovation that will benefit patient care. If even 10 percent of U.S. patients consented to the addition of their records to databases, millions of cases would be available. It is important to note, however, that the accuracy of a tool developed on an infinite sample is not 100 percent. Instead, it is limited by the ability of the model to understand trends in the data and the discriminatory power of the features used in the model. This discriminatory power, and thereby CAD tool performance, is based on a few key assumptions about the databases.

The most important assumption is that the gold standard reported in the database is definitive. At best, supervised machine learning can only recreate the performance of the gold standard. If, for example, clinicians consistently misdiagnose bipolar disorder as depression, then any database would confuse the two disorders and replicate this misdiagnosis. Thereby, any CAD tool can only be as good as the experts used to train it. This suggests than when training CAD tools, the developers should limit the training and validation sets to clear examples of each condition to minimize but not eliminate this bias. This limitation also leaves space for research groups to develop improved gold standards or clinical procedures that could outperform the CAD tool. Thereby, we expect that CAD tools cannot replace the great tradition of continual improvement of clinical medicine through research or the advice of the national and international experts that study and treat specific conditions.

Another key assumption is that the training sample is an unbiased representation of the population in which the CAD tool will be applied. Correction of this bias is critically important because a supervised CAD tool is only as applicable as its training and validation set is unbiased. We expect that these databases will endow modern statistical methods the power needed to identify, quantify, and control for possible sources of bias that have not been appreciated in smaller databases [65]. In many clinical research protocols, it is common practice to ignore this assumption because

the practical cost of obtaining a truly unbiased sample is prohibitive. For example, it is often the case that patients recruited at large academic medical centers have more severe disease than at other centers. This assumption of an unbiased sample is justified because, in most cases, there is little evidence that the pathophysiology underlying disease in research subjects or patients with severe disease differs from the full population. Because of their size, EHR-based databases would be expected to include patients who would not ordinarily be recruited into research studies. Research based on these databases would then be more representative of the affected population than current research methods.

Current experimental design methods produce high quality clinical information that minimizes noise in the sampled data. As the number of patients increases, so does the number of independent health care providers and institutions that collect data associated with each patient. This in turn substantially increases the number of possible sources of uninformative noise that must be adequately controlled. Controlling for some of these sources of noise is simply a statistical exercise, but others require more complex biostatistical modeling. One particularly egregious source of noise is if providers at particular institutions do not write clinical notes that fully represent the patient's symptoms and the provider's findings. No matter how effective CAD tools become, providers will always need to speak to patients, ask the right questions, and provide consistent, high quality care. Patients are not trained, unbiased observers. Patients frequently omit pertinent details regarding their complaints unless they trust the provider and the provider asks the right question in the right way. On the scale of the entire database, detecting low quality or biased information is difficult because it requires testing if the data from each institution varies significantly from the trends seen in the rest of the dataset. These differences, however, could reflect unique characteristics of the patient population being treated at that institution. The development of reliable techniques to identify and control for these sources of noise will be critical to the effective mining of the EHR databases.

THE FUTURE OF MEDICAL DIAGNOSTICS

The key hurdle to deploying CAD tools in academic and clinical medicine is the efficient implementation of these tools into software already utilized by clinicians. As stated by van Ginneken et al., the requirements of a CAD are that it has sufficient performance, no increase in physician time, seamless workflow integration, regulatory approval, and cost efficiency [2]. We have already discussed how the sheer size of the EHR database will substantially improve the performance and applicability of CAD tools. The improvements that were the basis for the ARRA EHR mandate — which we believe will be implemented using computer-aided diagnostics — provide clear evidence for the issue of cost effectiveness. Each of the improvements from the reduction of duplicative or inappropriate care to the increase in early detection, will decrease the cost of health care nationwide [12]. Given these benefits and improved performance, it would only be a matter of time before these tools would be given regulatory approval. The only facet of CAD implementation left would be efficient implementation that does not increase physician time. This is a content strategy problem.

Before seeing a patient, many providers scan the patient note for information such as the primary complaint given to the intake nurse, if available, and the patient's history. A CAD tool could provide a formatted summary of such notes, making it more accessible. Reviewing other test data is also routine. A CAD tool that pre-reads radiological images could simply display the predicted result as part of the image header. Radiologists could then see and interpret the results of the CAD tool as well as confirm these results and provide additional details in their subsequent clinical note. Outputs similar to these could be provided at the top or bottom of reports for EEGs, metabolic panels, and

other medical procedures. Regardless, physicians should have access to the raw data so that they can delve deeper if they desire more detailed information [2].

During a patient visit, the CAD tool could help remind the physician of key issues to cover that are related to previous clinical notes to address patterns that the computer notices but the physician may have overlooked. The Agile Diagnosis software is already exploring how best to design this type of tool [66].

After the visit, the tool could then operate on the aggregate information from this patient and provide recommendations and warnings about medications and treatments. The inclusion of citations that verify the evidence-based efficacy of the recommended medications and warnings is simple and requires very little space and processing power though frequent updating may be necessary.

Although the CAD reminders would likely be ignored by experienced providers, their constant presence could serve as a quality assurance measure. As discussed by Dr. Brian Goldman, MD, at his TED talk, all providers make mistakes [67]. These CAD-based reminders have the potential to improve upon the rate at which these mistakes are made and important details are missed. The most impactful benefits of CAD, however, are not in improving the care given by experienced providers who rarely make mistakes or miss details. Instead, these CAD tools will help inexperienced providers, those with limited medical training or special expertise, or experienced practitioners who lack current expertise to provide basic health care information to underserved populations. In this way, the development of CAD tools could reduce the magnitude of health disparities both inside the United States and worldwide.

CONCLUSIONS AND OUTLOOK

The EHR mandate will likely have widespread beneficial impacts on health care. In particular, we expect that the creation of large-scale digitized databases of multimodal patient information is imminent. Based on previous actions of the NIH, we expect it to substantially support the development of these databases that will be unprecedented in both their size and quality. Such databases will be mined using principled bioinformatics methods that have already been actively developed on a smaller scale. In addition to other potential impacts, these databases will substantially speed up the development of quality, applicable CAD tools by providing an unprecedented amount of high quality data at low cost upon which models can be built. We believe that these tools will be responsible for many of the improvements quoted in the motivation for passing ARRA, including the reduction of medical errors, inefficiency, inappropriate care, and duplicative care while improving coordination, early detection, disease prevention, disease management, and, most importantly, outcomes [12].

The development of widespread CAD tools validated on large representative databases has the potential to change the face of diagnostic medicine. There are already numerous examples of CAD tools that have the potential to be readily applied to extremely prevalent, high profile maladies. The major limiting factor is the validation of these methods on large databases that showcase their full potential. The development, validation, and implementation of these tools, however, will not occur overnight. Important regulatory, computational, and scientific advances must be achieved to ensure patient privacy and the efficacy of these automated methods. The problem of mining large databases also introduces numerous statistical problems that must be carefully understood and controlled.

The goal of these methods is not to replace providers but to assist them in delivering consistent, high quality care. We must continue to respect the science and art of clinical medicine. Providers will always be needed to interact with patients, collect trained observations, and interpret the underlying context of symptoms and findings. In addition, providers will have the unique ability to understand the applicability of computer-aided diagnostics to each patient.

Thereby, we believe that bioinformatics and machine learning will likely support high quality providers in their pursuit of continual improvements in the efficiency, consistency and efficacy of patient care.

Acknowledgments: We thank our reviewers for their helpful comments. The authors thank the UCLA-Caltech Medical Scientist Training Program (NIH T32 GM08042), NIH R33 DA026109, the Systems and Integrative Biology Training Program at UCLA (NIH T32-GM008185), the UCLA Departments of Biomathematics and Psychology, the Caltech Graduate Program in Biochemistry and Molecular Biophysics, and the Hearst Foundation for providing funding and course credit that made this work possible.

REFERENCES

1. United States Department of Health and Human Services. Centers for Disease Control and Prevention. National Ambulatory Medical Care Survey [Internet]. 2009. Available from: http://www.cdc.gov/nchs/ahcd.htm.
2. van Ginneken B, Shaefer-Prokop CM, Prokop M. Computer-aided Diagnosis: How to Move from the Laboratory to the Clinic. Radiology. 2011;261(3):719-32.
3. Lunshof JE, Chadwick R, Vorhaus DB, Church GM. From genetic privacy to open consent. Nat Rev Genet. 2008;9(5):406-11.
4. Chen YC, Wu JC, Haschler I, Majeed A, Chen TJ, Wetter T. Academic impact of a public electronic health database: bibliometric analysis of studies using the general practice research database. PLoS One. 2011;6(6):e21404.
5. Hunter M, Smith RL, Hyslop W, Rosso OA, Gerlach R, Rostas JA, et al. The Australian EEG database. Clin EEG Neurosci. 2005;36(2):76-81.
6. Aisen PS. ADNI 2 Study [Internet]. 2008 [cited 2012 April 12]. Available from: http://adcs.org/Studies/ImagineADNI2.aspx.
7. Weiner MW. Letter of welcome from the ADNI Principal Investigator [Internet]. 2009 [cited 2012 April 12]. Available from: http://www.adni-info.org.
8. Church GM. Personal Genome Project Mission [Internet]. 2012 [cited 2012 April 12]. Available from: http://www.personalgenomes.org.
9. Provost A (Department of Science and Information Technology, University of Newcastle, Australia). Electronic mail to: Wesley Kerr (Department of Biomathematics, University of California, Los Angeles, CA). 2011 Oct 17.
10. European database on epilepsy 2007 [Internet]. [cited 2012 April 12]. Available from: http://www.epilepsiae.eu.
11. Schrader D, Shukla R, Gatrill R, Farrell K, Connolly M. Epilepsy with occipital features in children: factors predicting seizure outcome and neuroimaging abnormalities. Eur J Paediatr Neurol. 2011;15(1):15-20.
12. American Recovery and Reinvestment Act [Internet]. 2009. Available from: http://www.recovery.gov/about/pages/the_act.aspx.
13. NIH. Final NIH statement on sharing research data grants [Internet]. [cited 2012 Mar 12]. Available from: http://grants.nih.gov/grants/guide/notice-files/not-od-03-032.html.
14. NIH. Expansion of sharing and standardization of NIH-funded human brain imaging data grants [Internet]. [cited 2011 Nov 20]. Available from: http://grants.nih.gov/grants/guide/notice-files/not-da-11-021.html.
15. Brockstein B, Hensing T, Carro GW, Obel J, Khandekar J, Kaminer L, et al. Effect of an electronic health record on the culture of an outpatient medical oncology practice in a four-hospital integrated health care system: 5-year experience. J Oncol Pract. 2011;7(4):e20-4.
16. GPRD. General Practice Research Database London: National Institute for Health Research [Internet]. 2012. [cited 2012 March 20]. Available from: http://www.gprd.com.
17. Wang Z, Shah AD, Tate AR, Denaxas S, Shawe-Taylor J, Hemingway H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. PLoS One. 2012;7(1):e30412.
18. Rodrigues LAG, Gutthann SP. Use of the UK General Practice Research Database for pharmacoepidemiology. Br J Clin Pharmacol. 1998;45:419-25.
19. Cho Y, Seong JK, Jeong Y, Shin SY. Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. Neuroimage. 2012;59(3):2217-30.
20. Weitzman ER, Kaci L, Mandl KD. Sharing medical data for health research: the early personal health record experience. J Med Internet Res. 2010;12(2):e14.
21. Whiddett R, Hunter I, Engelbrecht J, Handy J. Patients' attitudes towards sharing their health information. Int J Med Inform. 2005;75:530-41.
22. Teixeira PA, Gordon P, Camhi E, Bakken S. HIV patients' willingness to share personal health information electronically. Patient Educ Couns. 2011;84(2):e9-12.
23. Malin B, Benitez K, Masys D. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. J Am Med Inform Assoc. 2011;18(1):3-10.
24. Skloot R. The immortal life of Henrietta Lacks. New York: Crown Publishers; 2010.
25. Health Insurance Portability and Accountability Act 1996 [Internet]. Available from: https://www.cms.gov/Regulations-and-Guidance/HIPAA-Administrative-Simplification/HIPAAGenInfo/downloads/HIPAALaw.pdf.

26. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects [Internet]. 2008 [updated 2008; cited 2012 April 12]. Available from: http://www.wma.net/en/30publications/10policies/b3/17c.pdf.
27. Neu SC, Crawford K. LONI Deidentification Debablet 2005 [Internet]. [cited 2012 April 12]. Available from: http://www.loni.ucla.edu/Software/DiD.
28. Schaller RR. Moore's Law: Past, Present and Future. IEEE Spectrum. 1997;34(6):52-9.
29. Rupp K. The Economic Limit to Moore's Law. IEEE Trans Semiconductor Manufacturing. 2011;24(1):1-4.
30. Chu C, Hsu A-L, Chou K-H, Bandettini P, Lin C-P. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. Neuroimage. 2012;60(1):59-70.
31. Coupe P, Eskildsen SF, Manjon JV, Fonov V, Collins DL. Simultaneous segmentation and grading of anatomical structures for patient's classification: Application to Alzheimer's disease. Neuroimage. 2012. 59(4):3736-47
32. Liu M, Zhang D, Shen D. Ensemble sparse classification of Alzheimer's disease. Neuroimage. 2012;60(2):1106-16.
33. Ince T, Kiranyaz S, Gabbouj M. A generic and robust system for automated patient-specific classification of electrocardiogram signals. IEEE Trans Biomed Eng. 2009;56:1415-526.
34. Ebrahimzadeh A, Khazaee A, Ranaee V. Classification of electrocardiogram signals using supervised classifiers and efficient features. Comput Methods Programs Biomed. 2010;99:179-94.
35. Zadeh AE, Khazaee A. High efficient system for automatic classification of the electrocardiogram beats. Ann Biomed Eng. 2011;39(3):996-1011.
36. Langerholm M, Peterson C, Braccini G, Edenbrandt L, Sornmo L. Clustering ECG complexes using Hermite functions and self-organizing maps. IEEE Trans Biomed Eng. 2000;47:839-47.
37. Chazal R, O'Dwyer M, Reilly RB. Automated classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Trans Biomed Eng. 2004;51:1196-206.
38. Shyu LY, Wu YH, Hu WC. Using wavelet transform and fuzzy neural network for VPC detection from the Holter ECG. IEEE Trans Biomed Eng. 2004;51:1269-73.
39. Andreao RV, Dorizzi B, Boudy J. ECG signal analysis through hidden Markov models. IEEE Trans Biomed Eng. 2006;53:1541-9.
40. de Chazal F, Reilly RB. A patient adapting heart beat classifier using ECG morphology and heartbeat interval features. IEEE Trans Biomed Eng. 2006;53:2535-43.
41. Mitra S, Mitra M, Chaudhuri BB. A rough set-based inference enginer for ECG classification. IEEE Trans Instrum Meas. 2006;55:2198-206.
42. Lin CH. Frequency-domain features for ECG beat discrimination using grey relational analysis-based classifier. Comput Math Appl. 2008;55:680-90.
43. Joy Martis R, Chakraborty C, Ray AK. A two-stage mechanism for registration and classification of ECG using Gaussian mixture model. Pattern Recognit. 2009;42:2979-88.
44. Yu SN, Chou KT. Selection of significant for ECG beat classification. Expert Syst Appl. 2009;36:2088-96.
45. Cuthill FM, Espie CA. Sensitivity and specificity of procedures for the differential diagnosis of epileptic and non-epileptic seizures: a systematic review. Seizure. 2005;14(5):293-303.
46. Hoeft F, McCandliss BD, Black JM, Gantman A, Zakerani N, Hulme C, et al. Neural systems predicting long-term outcome in dyslexia. Proc Natl Acad Sci USA. 2011;108(1):361-6.
47. San-juan D, Claudia AT, Maricarmen GA, Adriana MM, Richard JS, Mario AV. The prognostic role of electrocorticography in tailored temporal lobe surgery. Seizure. 2011;20(7):564-9.
48. Wang Y, van Klaveren RJ, de Bock GH, Zhao Y, Vernhout R, Leusveld A, et al. No benefit for consensus double reading at baseline screening for lunch cancer with the use of semiautomated volumetry software. Radiology. 2012;262(1):320-6.
49. Cowan N. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behav Brain Sci. 2001;24(1):87-114; discussion 114-85.
50. Veneri G, Pretegiani E, Federighi P, Rosini F, Federico A, Rufa A, editors. Evaluating Human Visual Search Performance by Monte Carlo methods and Heuristic model. 2010 10th IEEE International Conference; 3-5 Nov 2010. Information Technology and Applications.
51. Harrison Y, Horne JA. The impact of sleep deprivation on decision making: a review. J Exp Psychol Appl. 2000;6(3):236-49.
52. Kohl P, Noble D, Winslow LR, Hunter PJ. Computational Modeling of Biological Systems: Tools and Visions. Philosophical Transactions: Mathematical, Physical and Engineering Sciences. 2000;358(1766):579-610.
53. Wang SJ, Middleton B, Prosser LA, Bardon CG, Spurr CD, Carchidi PJ, et al. A cost-benefit analysis of electronic medical records in primary care. Am J Med. 2003;114(5):397-403.
54. Balli T, Palaniappan R. Classification of biological signals using linear and nonlinear features. Physiol Meas. 2010;31(7):903-20.
55. Gelinas JN, Battison AW, Smith S, Connolly MB, Steinbok P. Electrocorticography and seizure outcomes in children with lesional epilepsy. Child's nervous system. Childs Nerv Syst. 2011;27(3):381-90.

56. Rodrigues Tda R, Sternick EB, Moreira Mda C. Epilepsy or syncope? An analysis of 55 consecutive patients with loss of consciousness, convulsions, falls, and no EEG abnormalities. Pacing Clin Electrophysiol. 2010;33(7):804-13.
57. Fita IG, Enciu A, Stanoiu BP. New insights on Alzheimer's disease diagnostic. Rom J Morphol Embryol. 2011;52(3 Suppl):975-9.
58. Kloppel S, Stonnington CM, Barnes J, Chen F, Chu C, Good CD, et al. Accuracy of dementia diagnosis: a direct comparison between radiologists and a computerized method. Brain. 2008;131(Pt 11):2969-74.
59. Lee J-D, Su S-C., Huang C-H, Wang JJ, Xu W-C, Wei Y-Y, et al. Combination of multiple features in support vector machine with principle component analysis in application for Alzheimer's disease diagnosis. Lecture Notes in Computer Science. 2009;5864:512-9.
60. Dai Z, Yan C, Wang Z, Wang J, Xia M, Li K, et al. Discriminative analysis of early Alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier (M3). Neuroimage. 2012;59(3):2187-95.
61. Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. A Am Med Inform Assoc. 2011;18(5):601-6.
62. Pakhomov SV, Shah N, Hanson P, Balasubramaniam S, Smith SA, editor. Automatic Quality of Life Prediction Using Electronic Medical Records. American Medical Informatics Association Symposium; 2008.
63. McGrayne SB. The theory that would not die: How Bayes' rule cracked the Enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy. New Haven, CT: Yale University Press; 2011.
64. Oquendo MA, Baca-Garcia E, Artes-Rodriguez A, Perez-Cruz F, Galfalvy HC, Blasco-Fontecilla H, et al. Machine learning and data mining: strategies for hypothesis generation. Mol Psychiatry. 2012. Epub ahead of print.
65. Huelsenbeck JP, Suchard MA. A nonparametric method for accommodating and testing across-site rate variation. Syst Biol. 2007;56:975-87.
66. Agile Diagnosis [Internet]. 2012 [cited 2012 April 12]. Available from: http://www.agile-diagnosis.com.
67. Goldman B. Doctors make mistakes: Can we talk about that? Ted.com [Internet]. Toronto, Canada: TED; 2011. Available from: http://www.ted.com/talks/brian_goldman_doctors_make_mistakes_can_we_talk_about_that.html.