Ethical and Legal Issues for Medical Data Mining

Ashwinkumar.U.M Dr. Anandakumar.K.R

ABSTRACT
This Paper emphasizes on uniqueness and specialty of medical data mining Healthcare related data mining is one of the most rewarding and challenging areas in application of data mining and knowledge discovery. The challenges are due to the data sets which are large, complex, heterogeneous, hierarchical, time series and of varying quality. The available healthcare datasets are fragmented and distributed in nature, thereby making the process of data integration a challenged task. The major issues related to tackle are ethical, legal and social aspects. Due to the lack of domain knowledge on the analyst's behalf it becomes necessary for an active collaboration between domain specialist and data miner with ethical and legal clearance from specialized hospitals. Medical datasets constitute a significant part of medical research. Ethical concerns, especially issues of confidentiality have resulted in the introduction of stringent regulations in doing this form of research. The merits and demerits of these new regulations are debated all over the world. The introduction of regulations for individual informed consent will prove costly to Indian physicians. Attempts are being made to evolve a consensus in which ethical concerns are given due respect without discouraging research.

1. Introduction:
Researchers and doctors have been using medical datasets for research. This research has played a critical role in medical progress. Reviews of medical datasets and publication of these analysis are almost done without revealing patient's identitities. However there is a little debate about the need to obtain informed consent from patients when their identities must be revealed. The use of medical datasets i.e. medical records has conventional taken two forms: systematic record review and record linkage [1, 2].
Systematic record review may be used to review the records of a consecutive series of patients with the same diagnosis to identify common clinical features, response to treatment, or factors influencing prognosis. This form of retrospective analysis constitutes the most common source of medical publication by physicians in India and abroad.
Record linkage means collecting medical information from separate sources on individual patients identified by name and date of birth to identity, among other things, any potential association between drug and a disease. In such research the personal identification in the records is essential for data collection. It entails a greater risk of loss of confidentiality. Such review seldom takes in India. This probably explains the public's relative lack of concern about the use of medical records for research.

2. METHODS
The major points of uniqueness and specialty of medical datasets may be organized under three general headings:

Ethical, legal, and social issues
Heterogeneity of medical data
Special status of medicine

Ethical, Legal and social issues.
Two primary ethical concerns pertaining to research based on medical records are obtaining informed consent maintaining the confidentiality of data [2, 3, 4].ideally patients should understand for what their medical records will be used for, who will have their access to their records, and how their of records will be maintained before they give explicit consent., such consent is so far not required since clinicians and researchers have taken the availability of this information as granted. Thus record linkage has usually been carried out without patient consent and qualifies for exemption from review by most ethics review boards (ERBs) [1, 2].
Numerous surveys outside India have revealed that patients are willing to support and participate in research but first he/she want to be consulted on the use of information from their medical records. They are worried about t their data could be used for marketing and insurance purposes. They are also concerned that sensitive information could be widely used and distributed without their knowledge. [5, 6, 7].These concerns have led to international efforts to enhance the protection afforded to data from medical records. In United States, the health insurance portability and accountability act (HIPPA) [8, 9] directs the secretary of Health and human services to establish safeguards for the privacy of individually identifiable health information. A variety of federal legislative proposals have also been developed to address the issue. The European commission has proposed in its draft directive that explicit patient consent should be obtained before each record can be used-a rule so stringent that record based research would probably stop together. her guidelines, notably those recently proposed by the united kingdom's department of health and British Medical association, are less stringent but nonetheless restrictive[3,4,5]

Current Practice of medical records review and publication of data in research: Indian Scenario.
Until recently most Indian investigators could get retrospective analysis published without an ethics review, as most international journals do not insist on such clearance, Now ERB clearance is mandatory. Indian Researchers specially in using medical data too must get their

retrospective studies reviewed. The guidelines of the Indian council of Medical research (ICMR) provide a waiver of informed consent if the study is of minimal risk or conducted in an emergency. The Medical council of India's code of medical ethics (MCI) [10] also permits such waivers if the patient identity is not revealed. However all such proposals must be clear by ERBs in a formal meeting? As there are very few ERBs in the country. Such research will definitely [11] Slow down.

Ethical and practical arguments against stringent regulations.

In many instances obtaining consent from patients either direct or indirect contact is problematic because such contact may introduce bias in to research process. It may also constitute a breach of privacy. Such contact may cause psychological, social or other harm to the former patient. Undue hardship may be imposed on an organization when additional financial, material, human or other resources are required.

How do these regulations affect Indian doctors and Researchers?

Making it mandatory for researchers to obtain explicit consent from patients before assessing their medical records, as now proposed by European commission, would prevent clinical studies that rely on personal records, with the exception of small case series. In US, HIPPA regulations appear to inhibit medical record and database research (8, 9). Current HIPPA implementation strategies increase the workload for ERBs and researchers and increase the dropout rate for proposed studies when investigators are unable to meet the requirements [8, 9].Researchers also feel that public money from government agencies and charitable organizations is wasted by ERBs when innocuous retrospective studies are required to do through multiple ethical reviews [12].The majority of publications from Indian institutes are related to medical records review. Only a few major institutes have ERBs and most of this form of research is not subjected to ethics review.

Heterogeneity of medical data

Raw medical data are voluminous and heterogeneous. Medical data may be collected from various images, interviews

With the patient, laboratory data, and the physician's observations and interpretations. All these components may bear upon the diagnosis, prognosis, and treatment of the patient, and cannot be ignored. The major areas of heterogeneity of medical data may be organized under these headings:

- Volume and complexity of medical data
- Physician's interpretation
- Sensitivity and specificity analysis
- Poor mathematical characterization

Volume and complexity of medical data

Raw medical data are voluminous and heterogeneous. Medical data may be collected from various images, interviews with the patient, and physician's notes and interpretations. All these data-elements may bear upon the diagnosis, prognosis, and treatment of the patient, and must be taken into account in data mining research and more medical procedures employ imaging as a preferred diagnostic tool. Thus, there is a need to develop methods for efficient mining in databases of images, which are more difficult than mining in purely numerical databases. As an example, imaging techniques like MRI, PET, and collection of ECG or EEG signals, can generate gigabytes of data per day. A single cardiac SPECT procedure on one patient may contain dozens of two dimensional images. In addition, an image of the patient's organ will almost always be accompanied by other clinical information, as well as the physician's interpretation (clinical impression, diagnosis). This heterogeneity requires high capacity data storage devices and new tools to analyze such data. It is obviously very difficult for an unaided human to process gigabytes of records,

Although dealing with images is relatively easier for humans because we are able to recognize patterns, grasp basic trends in data, and formulate rational decisions. The stored information becomes less useful if it is not available in an easily comprehensible format. Visualization techniques will play an increasing role in this setting, since images are the easiest for humans to comprehend, and they can provide a great deal of information in a single snapshot of the results.

Importance of physician's interpretation

The physician's interpretation of images, signals, or any other clinical data, is written in unstructured free-text English, that is very difficult to standardize and thus difficult to mine. Even specialists from the same discipline cannot agree on unambiguous terms to be used in describing a patient's condition. Not only do they use different names (synonyms) to describe the same disease, but they render the task even more daunting by using different grammatical constructions to describe relationships among medical entities. It has been suggested that computer translation may hold part of the solution for processing the physician's interpretation (Manning and Schuetze, 2000; Ceusters, 2000). Principles of computer translation may be summarized as follows (Nagao, 1992):

- Machine translation is typically composed of the following three steps: analysis of a source language sentence; transfer ... from one language to another; and generation of a target language sentence.
- Natural language can be regarded as a huge set of exceptional expressions ... as many expressions as possible must be collected in the dictionary ... It is an endless job.
- Current translation systems can analyze and translate sentences composed of less than ten words.... A reason for such failure is the ambiguity.... Even a human cannot understand the meaning of a long sentence at the first reading.
- Grammatical rules in machine translation can be regarded as (artificial intelligence) production rules."
- These principles, suitably customized for medical text, may be required for future medical data mining applications that depend upon the physician's free-text interpretation as part of the data mining analysis.

Sensitivity and specificity analysis Validation

Nearly all diagnoses and treatments in medicine are imprecise, and are subject to rates of error. The usual paradigm in medicine for measuring this error is *sensitivity and specificity analysis*. One should distinguish between a *test* and a *diagnosis* in medicine. A test is one of many values used to characterize the medical condition of a patient; a diagnosis is the synthesis of many tests and observations that describes a pathophysiologic process in that patient. Both tests and diagnoses are subject to sensitivity/specificity analysis. Typically, the test-results are a proposed, inexpensive new test, whereas the hypothesis is either a more expensive test, regarded as definitive, or else a complete medical workup of the patient. The *accuracy of a test*, on the other hand, compares how close a new test value is to a value predicted by if…then rules. To classify a test example, the rule that matches it best determines the example's class membership. An accuracy test is defined as:

Accuracy = **(TP+TN)/(TP+TN+FP+FN)**

Sensitivity = **TP/(TP+FN)**

Specificity =**TN/TN+FP**

Poor mathematical characterization of medical data

Another unique feature of medical data mining is that the underlying data structures of medicine are poorly characterized mathematically, as compared to many areas of the physical sciences. Physical scientists collect data which they can put into formulas, equations, and models that reasonably reflect the relationships among their data. On the other hand, the conceptual structure of medicine consists of word-descriptions and images, with very few formal constraints on the vocabulary, the composition of images, or the allowable relationships among basic concepts. The fundamental entities of medicine, such as inflammation, ischemia, or neoplasia, are just as real to a physician as entities such as mass, length, or force are to a physical scientist; but medicine has no comparable formal structure into which a data miner can organize information, such as might be modeled by clustering, regression models, or sequence analysis. In its defense, medicine must contend with hundreds of distinct anatomic locations and thousands of diseases.

Privacy and security of human data

Another unique feature is privacy and security concerns. For instance, U. S. federal rules set guidelines for concealment of individual patient identifiers. At stake is not only a potential breach of patient confidentiality, with the possibility of
Ensuing legal action; but also erosion of the physician-patient-relationship, in which the patient is extraordinarily candid with the physician in the expectation that such private information will never be made public. By some
Guidelines, concealment of identifiers must be irreversible. A related privacy issue may apply if, for example, crucial diagnostic information were to be discovered on patient data, and a patient could be treated if one could only go back and inform the patient about the diagnosis and possible cure. In some cases, this action may not be taken. Another issue is data security in data handling, and particularly in data transfer. Before the identifiers are concealed, only authorized persons should have access to the data. Since transferring the data electronically via the Internet is insecure, the identifiers must be carefully concealed even for transfers within a single medical institution from one unit to another. On the other hand, it has been noted in recent U. S. federal documents (U.S. 1999), that there are at least two legitimate research needs for re -identification of de-identified medical data: first, there is a need to prevent accidental duplicate records on the same patient from skewing research conclusions; second, there may be a compelling need to refer to original (re-identified) medical records to verify the correctness or to obtain additional information on specific patients. These special requirements could be managed by appropriate regulatory agencies, but they could not be met at all if the data are completely anonymous. There are four forms of patient data identification:

- Anonymous data are data that were collected so that the patient-identification was removed at the time the information was collected. For example, a block of tissue may be taken from an autopsy on a patient with a certain disease, to serve as control tissue-block in the histology laboratory. The patient's identifiers are not recorded at the time of specimen collection, and thus can never be recovered.

Anonymized data are data that are collected initially with the patient-identifiers, which are subsequently, irrevocably removed. That is, there can never be a possibility of returning to the patient's record and obtaining additional information. This research practice has been common in the past. However, anonymized data, as described above, could be accidentally duplicated, and could not be verified for corrections or additional data.

De-identified data are data that are collected initially with the patient-identifiers, which are subsequently encoded or encrypted. The patient can be re-identified under conditions stipulated by an appropriate agency, typically an Institutional Review Board (IRB).

Identified data can only be collected under significant review by the institution, federal guidelines, etc., with the patient giving written informed consent. Even for public Internet distribution, identifier-encrypted data which enter the database only once are fairly safe from
Attackers. For example, in the Johns Hopkins Autopsy Resource (Moore et al, 1996), a publicly-posted internet resource that lists over 50,000 deceased patients, each deceased patient enters the database only once, and is contributed by a single institution with an IRB-approved encryption procedure. On the other hand, data from multiple institutions are only as secure as the procedures from the least-secure contributing institution. Also, data from a single institution, in
Which there are multiple updates of the public database over time, are also less secure from a determined attacker? There are a variety of encryption protocols suitable for such purposes (Berman et al., 1996; Schneier, 1996):
double-brokered encryption
one-time -pad encryption (lookup table)

public-private encryption
The emerging U. S. federal paradigm for using de-identified medical data for research purposes is minimal risk. That is, if one employs only data that are collected in the ordinary diagnosis and treatment of patients, and there is no change in patient management as a result of the research, including no pressure on the patient to accept or refuse certain management, and no call-back for additional data that might upset the patient or next -of-kin, then the only risk of using such data is the loss of confidentiality to the patient. This is called minimal risk data, and may be possible to use in research projects with a simple exemption from the IRB. There was a well-publicized case of a prominent researcher at a major institution a few years ago who called a family in order to verify certain data regarding a deceased patient under study; this is not allowed under the minimal risk paradigm.

2.3 Expected benefits

Any use of patient data, even de-identified, must be justified to the IRB as having some expected benefits. Legally and ethically one cannot perform data analysis for frivolous or nefarious purposes. However, the Internet is the cheapest and most convenient way to distribute data, and the most accessible to the public which may have legitimate reasons for access. For example, there may be rare -disease interest groups, medical watchdog groups, or even investigators with
Unconventional scientific perspectives, who have reasonable claims to mine the data, but who could not mount the financial and administrative resources to mine privately-held databases. How is this conflict between public access and frivolous use of public human data to be resolved? There is as yet no answer to this question. If we are to make progress.

3 Special status of medicine

Finally, medicine has a special status in science, philosophy, and daily life. The outcomes of medical care are life or death, and they apply to everybody. Medicine is a necessity, not merely an optional luxury, pleasure, or convenience. Among all the professions, medicine has the longest apprenticeship. Most medical specialists in the USA require at least eleven years of training after high school graduation, and some surgical subspecialties require up to sixteen. In the USA, medical care costs consume one-seventh of the GDP. Licensed physicians represent about 0.2% of the U.S. population; the incomes for fulltime physicians are in the top several percent; and the average physician causes seven times his/her income to be spent on services ordered. The average citizen has high expectations of medicine and its practitioners. A sick person is expected to recover. Physicians are expected to be ethical, caring, and not too greedy. Medicine is a popular subject for the popular media. Medical care is sometimes risky, but when it fails, the desire for legal revenge is intense and punitive. Medical information about the individual patient is considered highly private, and the general public is extremely fearful about disclosure (U.S., 1999). We all enjoy the benefits of medical research conducted on other patients, but we are very often reluctant to contribute or release our own information for such purposes. When medical data are published it is expected that the researchers will maintain the dignity of the individual patient, and that the results will be used for socially beneficial purposes (Saul, 2000). It has been suggested that scientific truths are fundamentally amoral; they can be used for good or evil (Changeux and Connes, 1995). Yet although medicine is based upon science, there are certain tests that may not be performed, certain questions that may not be asked, and certain conclusions that may not be drawn, because of medicine's special status. There has been a vigorous public debate, for example, on whether data obtained from human experimentation, such as those obtained in Nazi Germany, should be published and used. Data from similar experiments, performed on laboratory animals, would be regarded as valid biological data without any further consideration. As we have seen in this article, this special status of medicine pervades our attitudes about medical data mining, as well as our attitudes about medical diagnosis and treatment.

4. CONCLUSION

In summary, data mining in medicine is distinct from that in other fields, because the data are heterogeneous; special ethical, legal, and social constraints apply to private medical information; statistical methods must address these heterogeneity and social issues; and because medicine itself has a special status in life. Data from medical sources are voluminous, but they come from many different sources, not all of commensurate structure or quality. The physician's interpretations are an essential component of these data. The accompanying mathematical models are poorly characterized compared to the physical sciences. Medicine is far, far from the intellectual gold-standard of a canonical form for its basic concepts. The ethical, legal, and social limitations on medical data mining relate to privacy and security considerations, fear of lawsuits, and the need to balance the expected benefits of research against any inconvenience or possible injury to the patient Methods of medical data mining must address the heterogeneity of data sources, data structures, and the pervasiveness of missing values for both technical and social reasons.

Suggestions and Recommendations

No survey has been done in India to study the views of patients about the use of personal data for research. However issues of confidentiality are likely to gain importance with wider insurance coverage. The investigator should anticipate this plan for the future. The ICMR guidelines allow ERBs to waive informed consent in appropriate cases where the study carries only minimal risk or in cases of emergency. However the guidelines should also provide allowances for expedited reviewer exemption from the review process. Study proposals involving medical records would be included under this category of
Review. The ICMR should resist the move to universalize the new set of stringent guidelines proposed by the European commission. It would be ideal for India to adopt guidelines of the working group .where the ERBs are responsible for assessing the potential importance of research proposal and deciding whether or not waive the requirement for informed consent. Circumstances under which ERBs may opt to do this include the following situations.
Access to the clinical record is essential for completion of research and consent is not applicable

The research is likely to yield information of sufficient merit

The research pertains to some future planning, preventive or therapeutic initiatives which may benefit the patients whose records are studied.

Researchers who are non-clinicians are formally instructed about their duty of confidentiality and they enlist a clinical supervisor who formally accepts professional responsibility for any breach of confidentiality, should it occur

Excessive restrictions on access to medical data for research could harm large number of people and hamper in medical care. A consensus policy respecting the rights of individuals and responsibilities of investigators are needed in India.

Data mining has become an integral part of health care delivery, planning and management. There have been many studies reporting various data mining models and their effectiveness in managing huge medical database. These studies in general use existing patient information. However there seems to be a breach in that, the studies which had access to crucial patient information, have not undergone any ethical review process. There is a need for the ethical consideration with prompt review process before any study is undertaken within the realm of data mining research.

Now how do we give an evidence of this?
One simple approach could be that we do a internet search, (not GOOGLE) generally on literature databases… either Pubmed or could be Anything from IEEE as well…
We would search the internet using certain "key words" something like "Medical Data mining"; whatever be the search words please note it down (a max of 2-3 words can be searched). Also note down the URL, date of search.
And we can limit our search to few more things like
Publications from India
Time frame from say between Jan 1st 2009 to June 30th 2009 (something like 6 months.. can make it an year also)

Report the number of hits you got/ how many were screened to be included, what made you exclude others.
Finally using the hits one obtained, information available regarding the ethical review can be looked at.
Report the statistics of what you see..
"An internet search undertaken using the terms "Medical data mining" from India between Jan 1st 2009 to June 30th 2009; revealed 400 hits out of which 300 fitted our criteria based on the information needed. We found that there were __ % who reported having undergone ethical review. This suggests that the ethical considerations while practicing data mining is very much undermined. Steps must be taken to ensure a mandatory requirement of Ethical consideration to be part of the data mining research.

5. REFERENCES

1. Doyal Len. Informed consent in medical Research: Journals should not publish Research to which patients have not given Fully informed consent-with three exemptions.BMJ 1997; 314:1107
2. Wald Nicholas, Law Macolm,Meade Tom,Miller George,Alberman Eva,Dickinson John.Use of personal Medical Records for Research purposes.BMJ 1994;309:1422-1424.
3. Parkes S E. legal aspects of record based medical research.Arch Dischild 2001;89:899-901.
4. Taube Daniel O, Burkhardt S. Ethical and legal risks associated with archival research.Ethics Behav 1997;7:59-67.
5. Baker R, Shiels C, Stevenson K,Fraser R,Stone M.What proportion of patients refuse consent to data collection from their records for research purposes. Br J Gen Pract 2000;50(457):655-6.
6. Willison Donald J, Keshavjee Karim,Nair Kalpana, Goldsmith Charlie,Holbrook Anne M. Patients consent preferences for research uses information in electronic medical records:interview and survey data. BMJ 2003;15;326:373.
7. Robling M R, Hood K, Houston H,Pill R,Fay J and Evans H M.Public attitudes towards the use of primary care patient record data in medical research without consent: a qualitative study. J Med Ethics 2004;30:104-109.
8. Kulynych Jennifer,Korn David.The new HIPAA(Health Insurance Portability and Accountability Act of 1996) Medical privacy rule: help or hindrance for clinical research. Circulation 2003;108:912-4.
9. O'Herrin Jacquelyn K,Fost Norman, Kudsk Kenneth A. Health Insurance Portability Accountability Act(HIPPAA) Regulations;effect on medical record research.Ann surg 2004;239(6);discussion 776-8
10. Medical Council of India. Indian Medical Council(Professional Conduct,Etiquuette and Ethics) Reulations,2002.Gazette of india dated 06-04-02,part III,Section 4.{cited 2006 March 25} Available from:http://mohfw.nic.in/code.htm
11. Nundy Samiran,Gulhati Chandra M .A new colonialism: conducting clinical trials in india.NEJM 2005;352:1633-1636.
12. Wagner Richard M.Ethical review of research involving human subjects:When and why is IRB necessary. Muscle nerve 2003;28:27-39.