35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Merging Medical Informatics and Automated Diagnostic Methods

Donna L. Hudson, *Fellow IEEE*, Maurice E. Cohen, *Member IEEE*

Abstract – In many instances disease diagnosis is more of an art than a science due to the complexity of disease, lack of detailed information on parameters that are indicative of the disease, and lack of sufficient data to apply these parameters to both diagnosis and treatment. Broad-based expansion of electronic health records (EHRs) will produce additional data for improved model development. However many obstacles remain. Patient record content is not broadly available because of privacy concerns and the lack of standardization of EHR formats. If available on a large scale, de-identified medical records can provide a basis for development of disease models by removing privacy concerns. Once comprehensive disease models have been developed that assist in identifying possible diseases and also include parameters that were utilized along with their relative importance, automated analytic methods can be used to indicate the likelihood of the presence of specific diseases. Although the physician will always remain as the final expert, these methods can provide an expanded information set and provide analysis that is too complex for standard methods.

I. Introduction

In the last decade broad-based advances and implementation of electronic health records have opened new possibilities for accessing and utilizing medical data. Although in the United States numerous commercial medical records systems are used, due to HL7 standards these records can be shared among institutions using a broad range of EHR systems [1]. In other locations outside the US the same medical record systems are used for the entire country thus simplifying the process. Currently EHRs are used for both diagnosis and treatment. Potentially these records also can be used to develop comprehensive disease models for a wide range of afflictions improving both diagnosis and treatment of disease. However, a number of obstacles remain. The first consideration is the requirement to protect patient confidentiality. Details on confidentiality concerns are summarized in the US HIPAA code that specifies in detail the handling of patient records and other patient information [2]. Although these regulations complicate the use of medical records for development of disease models, methods are in place to use de-identified data to build disease models using information from patient records [3]. The first component determines the diagnostic findings that define specific diseases and the second is to find treatment options that can moderate the components that contribute to the disease [4].

D. L. Hudson is Professor of Bioengineering (UC Berkeley/UC San Francisco) and Professor of Family and Community Medicine, UCSF, 155 N. Fresno Street, Fresno, CA 93701 (dhudson@fresno.ucsf.edu). M. E. Cohen is Professor of Bioengineering (UCB/UCSF) and Professor of Radiology, UCSF (email: mcohen@fresno.ucsf.edu).

II. Methodology

A. Use of De-Identified Data for Development of Disease Models

A number of components are necessary to successfully develop disease models from patient data. The starting point for disease diagnosis is analysis of currently known indications of the disease. Basic knowledge is found in medical texts. Research articles contribute to the process by adding knowledge gained in studies and observation. In addition, the ability to search very large databases containing patient data for specified diseases can reveal new and important information on common parameters that were not previously known. Effectiveness of treatments can also be confirmed. Building these models relies on automated methods for searching and collating information. In order to have sufficient disease data it is also necessary to seek broad-based contributors who are willing to share their de-identified data.

A first step toward this goal is the use of the Health Ontology Mapper (HOM) which is designed to translate locally coded records from any institution into the RxNorm standard format, forming the basis for analyzing data received from a number of sources to create a large enough dataset to identify commonalities to build both diagnostic models and to identify effective treatments. A cooperative program is currently underway with University of California San Francisco, University of Washington, University of Pennsylvania, University of Rochester, and University of California Davis [5]. Other clinical data repositories have also been established [6].

B. Barriers to Automated Diagnosis

Development of disease models that can be used directly with automated methods of diagnosis is complicated by the presence of non-numeric data. Figure 1 shows a simplified sketch of the diagnostic process [7]. Each of these components brings challenges in model development due to the inclusion of complex data structures. The only components that are in numerical format are laboratory data and the parts of the medical record that contain numeric test results. The family history portion is largely descriptive as are imaging results. Signal analysis has many components, including the pictorial signal itself which may be accompanied by numerical summaries or by descriptive detail. All these formats pose problems for automated analysis methods.

Signal Analysis Data

The most commonly used signal data in diagnosis is the electrocardiogram (ECG). Currently many commercial devices are available that automatically characterize the

ECG during routine physical exams. The automated results are based on the QRS complex that is associated with each heartbeat. Results are presented as normal or abnormal. If specific numerical parameters indicate variations from the normal form of the QRS pattern, for example ST depressions, a numerical value is provided that indicates the degree of the depression. These values can be used directly in a numeric scheme.

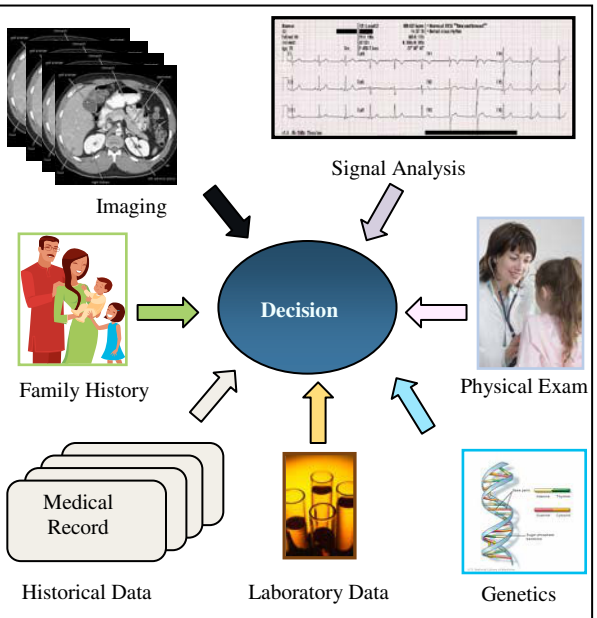

Figure 1: Components of Diagnosis

A more complex analysis of signals is based on a 24-hour Holter ECG recording with the patient going about normal activities [8]. While the automated analysis of the Holter recordings identifies the presence of abnormal occurrences of the QRS complex it provides no numerical components. More complex signals such as the Holter analysis require additional summary measures. One measure is the Central Tendency Measure (CTM) based on the recursive chaotic function in which a_n is based on the previous value a_{n-1}:

$$a_n = A\, a_{n-1}(1 - a_{n-1}) \qquad 2 \le A \le 4 \tag{1}$$

where A is a constant whose value changes the behavior of the function. For increasing values of A, the equation progresses from single value convergence to chaos. Within the chaotic area, regions of stability unexpectedly appear. These regions of stability are a matter of perception when only discrete values of n are considered and disappear in the continuous solution: An approximate solution of the logistic equation at A = 4 shows that chaotic behavior is not apparent when viewed as a continuous rather than a discreet function. The exact solution is:

$$a_n = \tfrac{1}{2}\,[1 - T_{2^n}(1-2a_0)] \tag{2}$$

where $T_n(x)$ is the Chebyshev function and n is assumed to be a real number [9].

Chaotic equations can be used to generate graphs that are known as Poincaré plots. Using the logistic equation a Poincaré plot is obtained by plotting a_{n+1} versus a_n. The resulting plot is a measure of the degree of chaos in the system. Another useful graph for practical applications is the second order difference: $(a_{n+2}-a_{n+1})$ vs. $(a_{n+1}-a_n)$ from which the Central Tendency Measure (CTM) is derived:

$$CTM = \left[\sum_{i=1}^{t-2} \delta(d_i)\right]/(t-2) \tag{3}$$

$$\text{where} \quad \delta(d_i) = \begin{cases} 1 & \text{if } [(a_{i+2}-a_{i+1})^2+(a_{i+1}-a_i)^2]^{.5} < r \\ 0 & \text{otherwise} \end{cases}$$

The CTM measure is useful in the analysis of lengthy time series [10]. An application to diagnosis of congestive heart failure is provided in the next section.

Imaging

For some components in the medical record, development of numerical measures is difficult and/or impractical. An example is medical imaging. Advances over the last forty years have provided many forms of medical imaging including radiographs, CT scans, MRI scans, fMRI scans, and ultrasound, among others. In most cases scans are read by radiologists who write written descriptions of the results or in some cases involving more sophisticated scanners written reports are provided automatically [11]. None of the approaches described above can deal with textual data.

An expert system method based on rules in textual format can analyze findings and determine possible disease presence along with a degree of certainty. A rule-based system developed by the authors for analysis of chest pain can be modified to provide decision support for a variety of conditions based on textual information. The analysis is based on the use of approximate reasoning techniques that allows degrees of importance along with degrees of certainty. The conclusion includes a value for the certainty of the decision. Each rule has a threshold value that must be exceeded to substantiate the rule [12]. Parameters include:

$a_1,\ldots,a_n$	Conditions (textual format)
$d_1,\ldots, s_n$	Degree to which condition n occurred
$w_1,\ldots,w_n$	Relative importance of n^{th} condition

T: Threshold which must be exceeded to fire rule

Information is aggregated to determine if the threshold has been reached. The evidence is aggregated using the equation:

$$E = max[(Q \sum_{i=1}^{n} c_i \wedge s_i) \wedge \min_{i=1,\ldots,n} (w_i^{c_i \wedge s_i})] \quad (4)$$

where ^ indicates min, w_i, s_i are defined above, $c_i \in \{0,1\}$, Q is a linguistic quantifier, n is the number of antecedents. The rule is substantiation if E > T. An example is provided in the next section.

Family History and Genetics

While the usage of the EHR is becoming quite widespread, the inclusion of family history in a personal health record (PHR) that includes family history and/or genetic information appears in only about 35% of the records [13]. Traditionally family history has been the sole source of identifying inherited susceptibility to specific disease. As more information is gained regarding the genetic origin of disease the family history can be augmented or replaced by genetic information. The lack of inclusion of either of these sources of information severely limits the use of the EHR as a diagnostic tool for the disease set with inherited susceptibility. With the expansion of details related to genetic disease susceptibility genetic profiles will play a larger role in diagnosis and perhaps also in treatment.

III. Results and Examples

A. Quantifying Signal Analysis Results

A sample application of the method described above uses chaotic analysis to summarize lengthy time series obtained for the analysis of Holter tapes. Figure 1 shows second-order difference plots for a normal individual and a congestive heart failure (CHF) patient. For the normal individual most points are centered near the origin while the CHF patient shows a broad scattering pattern. The numerical values also clearly show the distinction. All normal patients in the study had a CTM greater than 0.70. Some CHF patients overlapped this range. The final diagnostic strategy includes the use of the CTM along with 4 clinical parameters to refine the diagnosis [14].

B Quantifying Imaging Results

Since imaging results are usually textual in nature the rule based-system EMERGE can be used. EMERGE requires either personal input or automated input obtained from written analyses which can be automatically scanned to determine the presence or absence of each diagnostic parameter. EMERGE was originally developed for the analysis of chest pain but the operation of the expert system is independent of the application [15]. However rules must be available for the implementation of the new application. These rules can either be generated from expert input or derived from analysis of standards for analysis for each type of condition for which imaging would be performed. A sample rule for image analysis is given below in Table I. If the threshold is exceeded then the rule is substantiated as determined by equation (4).

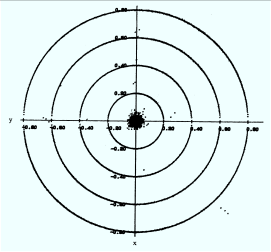

Figure 2a: Normal Individual

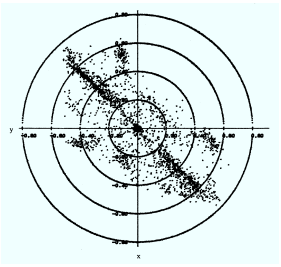

Figure 2b: Congestive Heart Failure Patient

x-axis: a_{n+1}-a_n from equation (2)

y axis a_{n+2}-a_{n+1} from equation (2)

The importance of each parameter is part of the rule. The degree of presence is entered by the user or by an automated system for image analysis. The evidence is then aggregated to determine if the threshold has been reached or exceeded. This method provides a numerical value that can be used in conjunction with other data in the diagnostic process.

Table I: Sample Expert System Rule

Condition	Degree	Importance
Image shows mass	0.8	0.5
Mass appears diffuse	0.8	0.2
Mass > 0.5 CM	0.9	0.2
Patient has symptoms	0.6	0.1

Result

Consider biopsy if E > 0.6

C. Quantifying Family History and Genetic Information

Family history can be automated by identifying diseases by a numerical code and indicating presence or absence by 1 or 0 respectively. With the expansion of specific genetic information it may be possible to establish binary presence or absence of a disease or preferably a measure of the degree to which the patient is at risk for the disease. In most cases genetic information is used in conjunction with other findings to diagnose presence or absence of disease or the likelihood that the patient will develop the disease.

D. Building Disease Models

The availability of large de-identified patient data that contains laboratory tests along with other symptoms and the diagnosis of the patient can assist in building disease models. A second component would rely on the development of automated advice on treatment options. These data can be collected from de-identified medical records that contain symptoms, diagnosis, treatment and outcome that could replace the current practice of basing treatment options on clinical students that may not involve a sufficiently large population.

E. Disease Diagnosis

Establishment of a large base of disease models will assist automated diagnosis by matching findings in the patient record. The process must include temporal analysis identified in the medical record to track changes in parameters that may be indicative of an emerging disease. Components illustrated in Figure 1 will contribute to the basis for diagnosis. The continuously updated EHR will facilitate automated tracking of changes in disease parameters that can send alerts to the physician about negative changes as well as notifications of improvement [16].

IV. CONCLUSION

The wide-spread availability of electronic health records forms a foundation for changing the practice of medicine on a number of fronts. If EHRs are compatible and/or adhere to HL7 standards, patients will benefit through the physician's ability to access medical records from any location. It also opens the way for the creation of a personal health record for the lifetime of a patient. This record could include not only the details for the patient but also important information on family history.

Access to large de-identified medical databases will assist in the development of disease models that can aid in accurate diagnosis. These models can be used in conjunction with an automated system or on an individually basis in a physician's office. The extension of technology into more areas of medicine can help not only in diagnosis and treatment, but also in prevention and early diagnosis of disease which is many cases will allow more effective treatment. The tasks described above cannot be accomplished by individuals or isolated entities. It is important to establish joint projects and large data repositories to move analytic techniques forward into practical and important diagnostics.

REFERENCES

[1] T Viangteeravat, MN Anyanwu, VR Nagisetty, E Kuscu, ME Sakauye, D Wu, Clinical data integration of distributed data sources using Health Level Seven (HL7) v3-RIM mapping, Journal of Clinical Bioinformatics, 1:32, doi:10.1186/2043-9113-1-32, 2011.

[2] SA Collins, DK Vawdrey, R Kukafka, G J Kuperman, Policies for patient access to clinical data via PHRs: current state and rcommendations, J Am Med Inform Assoc., 18: i2-i7, 2011.

[3] R Wynden, DL Hudson, Lab Norm: Automated Clinical Lab Data Normalization, ISCA Software Engineering and Data Engineering, 19:349-354, 2010.

[4] R Chen, GO Klein, E Sundvall, D Karlsson, H Åhlfeldt, Archetype-based conversion of EHR content models: pilot experience with a regional EHR system, BMC Medical Informatics and Decision Making, 9:33 doi:10.1186/1472-6947-9-33, 2009.

[5] R Wynden MG Weiner MG, I Sim, D Gabriel, M Casale, S Carini, S,Hastings, D Ervin, S Tu, JH, Gennari, N Anderson, K Mobed, P Lakshminarayanan, M Massary, RJ Cucina, Ontology mapping and data discovery for the translational investigator, AMIA Summits Transl Sci Proc. 2010 Mar 1;2010:66-70.

[6] GM Web. SN McMurry, D. Macfadden, D Nigrin, S Churchill, OS Kohane, The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories, J Am Med Inform Assoc, 6(5):624-30, 2009.

[7] DL Hudson, ME Cohen, Computational Methods for Personal Healthcare, ISCA Computer Applications in Industry and Engineering, 22:237-242, 2009.

[8] DW Rowley, S Glagov, Heart Attacks, Heart Rate, and Gel Electrodes: The Invention of Ambulatory Cardiology, Perspective in Biology and Medicine, 49, 346-356, 2006.

[9] ME Cohen, DL Hudson, Nonlinear Analysis using Continuous Chaotic Modeling, Cell and Molecular Biology, 50(3):291-295, 2004.

[10] DL Hudson, ME Cohen, Technologies for Personalized Heathcare, ISCA Computer Applications in Industry and Engineering, 24:121-126, 2012

[11] CH Lien, T-L Yang, C-H Hsiao, T Kao, Realizing Digital Signatures for Medical Imaging and Reporting in a PACS Environment, J Med Syst, 37:9924, 2013.

[12] DL Hudson, ME Cohen, Diagnostic Models Based on Personalized Analysis of Trends (PAT), IEEE Transactions on Information Technology in Biomedicine, 14(4), 941-948, 2010.

[13] C Widmer, J.P. Deshazo, J.Bodurtha, J. Quilloin, H. Creswick, Genetic Counselors' Current Use of Personal Health Records-Based Family Histories in Genetic Clinics and Consideration of Their Future Adoption, J. Genet. Couns, 2012.

[14] DL Hudson, ME Cohen, Neural Networks and Artificial Intelligence for Biomedical Engineering, Wiley, 1999.

[15] DL Hudson, ME Cohen, Overcoming Barriers to Development of Cooperative Medical Decision Support Models, IEEE EMBS, 34:2194-2197, 2012

[16] DL Hudson, ME Cohen, Temporal Trend Analysis in Personal Health Records, IEEE EMBS, 30:3811-3814, 2008.