An expert system for high-throughput collection and analysis of clinical data

David Zakim
Institut für Digitale Medizin
Am Kriegsburg Turm, 44
70192 Stuttgart
david.zakim@IDM-foundation.org

Abstract: Clinical decision-making in everyday medical care is not evidence-based because the evidence base is too vast to learn and too complex to use. The lack of evidence-based decision-making generates poor outcomes for patients but high costs for care. Separate from issues related to using a very large knowledge base, progress in medicine at the clinical level is inhibited by the high cost of clinical research, poor design of clinical research protocols, and limited tools for data analysis. Clinical research programs for the 21st Century need high-throughput technologies for data collection from large numbers of patients and high-throughput data analysis unconstrained by standard concepts of pathophysiology. We present a scheme to show how an expert medical system can address simultaneously the problem of inefficient use of existing knowledge and inefficient methods for generating new medical knowledge at the clinical level.
1 Introduction

Current resource-use for delivering medical care is not sustainable. The problem is discussed almost exclusively in terms of finance but cannot be solved by financial “reforms” because the high cost of health care is a consequence of poor quality everyday care, where poor quality reflects a discrepancy between outcomes it is possible to achieve and outcomes patients experience. The quality of everyday health care is poor, as measured above [Wo02], because human cognitive function constrains collection of relevant clinical information from patients and interpretation of clinical information in the context of state-of-the-art knowledge. Simply stated, the evidence base for medical practice is too large to learn. It also is too complex to use: short-term memory is inadequate to factor simultaneously the number of clinical variables that typically are significant for an evidence-based clinical decision [Mi56]. But, whereas the existing knowledge base already is too large to learn and use, it is seriously deficient in evidence relating to therapeutic efficacy; and it is inefficient in predicting true risk for diseases with complex etiologies. For example, coronary artery disease can be prevented by lowering levels of low density lipoprotein cholesterol [Ch05]. Clinical application of current risk equations prevents 1 coronary event for treatment of between 100 and 200 patients, depending on the guideline and the population to which it is applied [Ma06]. It is time to acknowledge that the medical profession practices and pursues clinical knowledge with paradigms that evolved at the turn of the 20th Century and that have been left behind by all other applied and theoretical sciences. It appears that the communities of practicing physicians and clinical scientists have to be shown the way into the 21st Century.
2 The Ontology of medical practice

Figure 1 is a schematic representation of how students are taught the practice of medicine. The figure shows 3 sources of clinical information: history acquired by interview of the patient, physical data acquired by examination of the patient and laboratory testing. Data from history and physical are integrated by principles of pathophysiology [the knowledge base for practice] in the box Decision Support to generate Working Diagnoses, which are diagnostic or treatment possibilities compatible with the findings from history and physical examination. Orders for diagnostic testing are generated by considering what additional information might confirm or exclude a specific diagnosis. Orders for treatment are generated by which treatment(s) are appropriate for a given diagnosis. The scheme does not show it explicitly; but collecting the medical history depends itself on immediate integration of each data item into a pathophysiologic concept, i.e., history-taking is dynamic not the filling out of a questionnaire. For example, the questions a physician should ask a patient with a main complaint of “chest pain” would be formulated in the context of knowledge about all possible causes of chest pain and the specific ways in which different causes affect the details of the complaint, e.g., severity, location, radiation, evocative factors, to focus the interview on a limited, most likely set of possible causes of chest pain. The working diagnoses generated from analysis of the clinical data suggest laboratory tests for determining/refining a diagnosis or treating a problem that already is diagnosed. Results from laboratory tests feed-back into the Decision Support box to output Refined Diagnoses, which lead in turn to treatment decisions in the box labeled continuous management. All parts of the scheme are used iteratively to update data sets and possible pathophysiologies of interest as a patient’s clinical status evolves. This scheme, which has been used for more than 100 years, was developed when a single physician could know essentially all the relevant evidence base and could apply the scheme to any clinical setting with which they might be confronted. The evidence-base has changed remarkably since the scheme developed, but it is as applicable today as 100 years ago. The problem is not to invent a new scheme for practicing medicine but to make Figure 1 workable in present circumstances.
2.1 Significance of the medical history

The patient’s medical history is the most important of the sources of clinical information in Figure 1. Diagnoses can be made at least 80% of the time from history alone, assuming of course that an adequate history is available. When history alone is not diagnostic, it nevertheless suggests possible diagnoses and points to tests that could confirm a possible diagnosis. History is critical not only for identifying what a patient perceives as their main problem but for delineating co-morbid conditions, which can be important for treatment decisions. There are no chemical or genetic tests or imaging studies that substitute for the information in the history; and there is nothing to suggest this will change in the foreseeable future. Despite its importance, history-taking usually is done poorly in routine medical practice because it is time-intensive and knowledge-intensive. There is a compelling case, therefore, that the medical history is the starting point for addressing the quality of care issue. It is as well the starting point for modernizing the fundamental approach to the conduct of clinical research.

3 Computerization of history-taking

History-taking is a rules-based activity that can be formalized in machine-readable form. We have created a web-based history-taking program that interacts directly with people seeking medical care. The program with the acronym CLEOS® [Clinical Expert Operating System] begins by asking why the individual is seeking medical care and proceeds, based on this answer, to acquire a history of their main complaint and a history of their health experience up to the present point in time. Figure 2 is an example of a formalization of the pathophysiology for collecting information from a patient of gynecologic causes. Nodes of the type HITHIA01 ask the patient specific questions. Nodes of the types ROB or PHTHIID are sets of trees representing distinct sets of different pathophysiology. Nodes coded with XC are inferences that interpret the clinical significance of any set of prior answers to direct further questioning. Arrows indicate the direction of questioning for a given answer value. The value z on some arrows does not represent an answer but an instruction not to ask the question if it has not been asked. The z therefore is a function that allows the program to look backward to “see where it has been” without writing an inference. They are not shown in this tree, but nodes also can represent laboratory results so that the direction of questioning can be based in part on laboratory findings. The same applies to physical examination fields. The trees leverage the knowledge of committees of physicians with clinical expertise in different fields of medicine in that any patient with access to the Internet can receive the benefit of the knowledge of the set of experts, who authored the content of the program. The current interview program has more than 12,000 decision nodes and more than 18,000 data fields that cover common, acute illnesses in adults and all chronic illnesses in adults. The detail of coverage varies for different conditions; but the program can be expanded indefinitely.
At the end of an interview, the program uses several thousand inferences to extract the clinical meaning of the information collected and to generate a narrative report of the findings, including possible diagnoses, co-morbid states and recommended actions in regard to diagnosis and treatment. The number of inferences for interpreting the clinical information is infinitely expandable. Indeed, the program can be enriched more quickly at the moment by expanding the number of inferences that interpret the data than by extending the range of data collection. The program out-performs physicians in terms of the accuracy and completeness of historical information. The program is easy for patients to use, and it is highly acceptable to patients [Za08].

4 Clinical data elements beyond the history

Findings from physical examination and laboratory testing must be part of a patient’s digital medical database [Figure 1]. To insure completeness of the clinical information, physical examination and laboratory data should be incorporated directly into the database not through transposing written records. CLEOS®, to this end, has a program that enables direct incorporation of physical findings into a patient’s CLEOS® chart from a hand held device as the physical examination is conducted. The templates for entering physical findings are programmed logically with a scheme like that in Figure 2. The user sees templates for data entry only for positive findings not all possible findings, which speeds up and facilitates data capture. Results from laboratory testing are added directly to a CLEOS® file as long as the program is directed to the testing site and laboratory outputs are in HL7 format. These features do not insure, of course, complete collection of relevant physical and laboratory findings because the program has no direct control over acquiring physical findings. And although CLEOS® can issue orders for indicated laboratory testing, legal constraints prevent action without physician approval. On the other hand, digitization of the clinical information as a CLEOS® file together with the program’s analytic inferences enables a structure that can “proof-read” files to determine whether there are clinically significant data elements missing from a file, based for example on a provisional diagnosis or a prior finding. CLEOS® can generate emails to responsible providers that indicate absence of relevant data elements, the pathophysio-logic basis for relevance of missing elements and that remind the provider to collect the missing elements.
5 Use of computing to solve the quality problem

The approach used in CLEOS® solves the problem of acquiring the base of clinical information that is the sine qua non for evidence-based practice of medicine. Since the information CLEOS® collects is subject to machine-based analysis according to rules of pathophysiology, CLEOS® also is a mechanism for insuring that clinical information is interpreted in the context of an up-to-date evidence-base. Although the present version of CLEOS® operates at a single point in time and cannot fulfill all functions of the scheme in Figure 1, the program can be configured to operate iteratively. This is primarily a matter of content, which is not trivial but is doable. The concept and functionality of CLEOS® thus can solve the problem of poor quality, everyday medical care. Formalization of the rules for practice as machine-readable code and application of these rules to complete clinical databases should achieve outcomes for patients that equal what it is possible to achieve by applying best current evidence. We note, however, that many of the projected outcomes from treating disease are based on experience with selected sets of patients, e.g., results generated in double-blind, prospective clinical trials. Results from these trials often do not apply across broad populations of patients because trials almost always recruit as homogeneous a population of affected individuals as it is possible to assemble. It remains to be seen how applicable a great deal of the evidence at hand will be to treating heterogeneous groups of patients with apparently singular diagnoses.
6 CLEOS® as enabling technology for high throughput collection and analysis of clinical information

6.1 Data collection

The double blind, random, controlled, prospective clinical trial is the “gold standard” for discovery of new clinical knowledge. The attraction of this design is control for confounding variables in that all variables affecting outcome(s) of interest are believed to be known for certain. Variables are better controlled in double blind, prospective trials than in patient charts from routine care; but this difference reflects the deficiencies in collecting clinical information data in everyday practice not an inherent flaw in observational research. So-called “observational studies” based on data from everyday medical care delivered to millions of patients can substitute for controlled trials if one can insure that patient records are collected in a standardized manner and that potentially confounding variables can be controlled by the robustness of the data sets. Technology like CLEOS® thus is an important advance for clinical research because the controlled, prospective clinical trial is seriously flawed. Its expense makes it inapplicable to the myriad of unresolved clinical issues; and as mentioned, results from prospective studies often do not apply in practice because they exclude the natural heterogeneity of patients with apparently singular diagnoses [Ch06, On09]. Technology like CLEOS® operates in the stream of everyday practice and can capture for clinical science the 900,000,000 outpatient visits, 35,000,000 hospitalizations, 64,000,000 surgeries and 3.5 billion prescriptions in the U.S/year, all of which is currently wasted information. Technology like CLEOS® can assemble extremely large clinical databases at marginal cost because professional time is not required for generating most of the clinical data.

6.2 Data analysis

Current clinical research uses inadequate methods for correlating clinical variables with outcomes because analyses assume that an outcome of interest is a singular event, i.e., the outcome occurs along a single pathophysiologic pathway, that clinical variables predicting an outcome of interest can be discerned from known pathophysiology, and that outcomes can be modeled from a relatively limited set of variables. By contrast with these assumptions, we know that few if any clinical outcomes are singular events, that pathophysiology is incomplete and that the clinical problems to solve are inherently more complex than can be accommodated by the methods by which clinical data are evaluated. Chronic disease develops over time for example; but the time-dependence in evolution of risk to disease seldom is modeled [Di06, Li06; Ta09]. The assumptions driving data analysis in clinical research remain in place, nevertheless, because they simplify recruitment of patients, decrease cost by limiting heterogeneity of study patients and outcomes and rationalize data analysis in the context of known pathophysiology. Moreover, modeling outcomes with a limited number of variables is predicated on the data that are available as well as the statistical methods in general use. The poor efficiency in predicting outcomes with the existing approach to the problem is well-documented but ignored.
Clearly, the absence of clinical data gathered at an affordable cost impedes effective, efficient clinical research. CLEOS® can free the analytic side of clinical research from the constraint of limited amounts of data as it generates data across heterogeneous populations and across time. One important feature of the clinical information collected by CLEOS® is the data are collected by a standardized protocol determined by the pathophysiology of disease. This means that a question not asked in a CLEOS® history is not asked for a medically indicated reason, which is a surmise not applicable to physician-acquired histories. Unasked questions have medical meaning when history-taking is determined always by a standardized protocol, and the usable data fields from a CLEOS® history number in the thousands.

There is a clear analogy between data gathering by CLEOS® and microarray chips. I raise the idea of this similarity because a mathematical algorithm, not a concept derived from pathophysiology, is the starting point for analysis of microarray experiments. Yet machine learning algorithms when applied to microarray data appear to illuminate new principles of pathophysiology [Hi09]. I propose that operation of technology like CLEOS® as part of everyday medical practice will bring to clinical research the power of mathematical and machine learning algorithms and thereby that technology like CLEOS® will revolutionize the conduct of clinical research.

![Ontology for medical practice](image)

Figure 1: Ontology for medical practice. Significance of each item and relationships are explained in the text.
Figure 2: Formalization of pathophysiology as machine-readable instructions to acquire a history directly from patients.

Literaturverzeichnis


