Achievements, Milestones and Challenges in Biomedical and Health Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI220940

My Journey Through the Field of Medical Informatics

Arie HASMAN[a,1]
[a] *Dept. of Medical Informatics, Amsterdam UMC, location AMC, Amsterdam, the Netherlands*

Abstract. In this contribution some achievements and milestones in the field of medical informatics, especially concerning decision support, as perceived by the author, are presented. The author focuses on those topics with respect to decision support that during his career in medical informatics impressed him and triggered him to convince his PhD students to start research on related topics. Both some of these achievements and the related research of some of his PhD students will be presented. The contribution starts with signal classification. Both ECG classification and sleep EEG classification are discussed. Then the use of Bayes' theorem for diagnostic purposes is discussed and some early applications pass review, among which the AAPHelp system developed by de Dombal and colleagues. Attention is subsequently paid to the advent of expert systems and other knowledge-based systems such as MYCIN and INTERNIST and to guideline-based decision support systems. Finally, the author presents his ideas about challenges for the field.

Keywords. milestones, medical informatics, computer-aided diagnosis, history

1. Introduction

This contribution is about achievements, milestones and challenges in medical informatics. I still use the term medical informatics and not biomedical and health informatics in this contribution because I grew up with this term.

In this contribution I want to discuss a number of what I regard achievements and milestones in the area of computer-aided diagnosis. Of course, this does not mean that these are the only achievements and milestones in this area, but I mention some of those that during my career in medical informatics impressed me and triggered me to convince my PhD students to start research on related topics. Therefore, I will not only refer to articles describing these achievements but also to articles describing the results of related research of some of my PhD students.

After obtaining my PhD I got a job as radiation physicist in the department of Radiotherapy and Nuclear Medicine of the Radboud Hospital of the Catholic University in Nijmegen, the Netherlands. During that period, I became interested in how physicians diagnosed patients. When does a physician, for example order an X-ray or a scan? If he does not have a clue of what the patient suffers from or when he is almost certain that something will be detected? How does the physician arrive at a diagnosis? If the patient is diagnosed, is the treatment then obvious or can different physicians prescribe various

[1] Corresponding Author, Prof. Dr. Arie Hasman, Dept. of Medical Informatics, Amsterdam UMC, location AMC, Amsterdam, the Netherlands; E-mail: hasmanarie@gmail.com.

therapies? At that time, I had the idea that interpreting images was rather straightforward and error-free. I was totally surprised when I read an article of Yerushalmy [1] about the reliability of chest radiography in the diagnosis of pulmonary lesions. He stated for example that in judging a pair of serial roentgenograms for evidence of progression, regression or stability of disease, two competent and experienced physicians are likely to disagree with each other in nearly one-third of the cases, and a single reader is likely to disagree with himself in about one-fifth of the pairs. So, I asked myself whether such a variability also occurs for other types of diagnoses and if so, whether the situation can be improved by better training of the physicians.

After three years I moved to Amsterdam where Jan van Bemmel had just started the department of Medical Informatics. In his former job he was involved in ECG and VCG analysis. I learnt that also the interpretation of ECGs showed interrater variability. So if physicians use different criteria in deciding whether an abnormality is present or not, how can you improve the situation? Appropriate training could be a solution, but how to reduce the variability that trainers will also show? Protocols for managing several situations were developed for nurses and ancillary personnel. The protocols were usually based on consensus. Does consensus lead to the truth, given the inter-rater variability? Could guidelines for physicians reduce variability? The various criteria used by expert physicians should be discussed and unified. I learnt about the Delphi technique with which variability in criteria can be reduced by involving a panel of experts, asking each one individually about their judgements of for example certain problem solutions and feeding back the answers of each member anonymously to all other members. On the basis of this feedback each member can adapt his answers in the next round, etc. It is expected that this procedure will converge and lead to consensus. So, the Delphi technique could be used for reaching consensus. But again, is the consensus indeed the truth? Probably the best approach is to use biomedical literature as a gold standard.

In this contribution I will tell about my journey in medical informatics. What did I learn and what do I expect for the future?

2. Analysis of Electrocardiograms and Electroencephalograms

When in 1974 I entered the Medical Informatics field by joining Jan van Bemmel's department at the Free University in Amsterdam, I soon became acquainted with the research leading to the modular TNO EGG/VCG interpretation system, carried out by his group in Utrecht [2]. Interpretation of an ECG is a complex task that requires knowledge in a number of fields like anatomy, electrophysiology, and pathophysiology.

An advantage for ECG and VCG interpretation is the availability of a (patho) physiological model. No such physiological model is available to support the analysis of EEGs. So compared to ECGs EEGs need a different approach to analyze them. The diagnostic value of EEG abnormalities is limited: different pathologies may produce similar abnormalities. However, the ease with which continuous monitoring can be achieved and the very fact that the EEG is unspecific makes it a valuable tool for the monitoring of many physiological variables, because changes in these variables may lead to changes in the EEG. For example, the EEG can be a valuable tool during open heart surgery, for sleep staging and for assessing the adequacy of dialysis programs.

2.1. ECG Analysis

The use of computers for ECG interpretation was first applied to the orthogonal 3-lead VCG and later the 12 lead ECG. Hubert Pipberger started in 1957 investigating the prospects of computer analysis using three simultaneously recorded orthogonal leads [3]. In 1959 Cesar Caceres and colleagues in the National Institute of Health in Washington started analyzing the 12 lead ECG, initially by processing one lead at a time [4].

Computer analysis of ECGs (VCGs) consists of two parts: a measurement and a classification part [5]. In the measurement part features relevant for diagnosis are measured (time intervals, wave durations and amplitudes of the various deflections, etc.) and in the classification part these features are used for classifying the ECG (VCG) into one or more diagnostic categories.

Since I was familiar with signal and image analysis, I was especially interested in how ECGs were classified. I learned that predominantly two different approaches were used by existing ECG analysis computer programs: a heuristic and a statistical one. In the heuristic approach the reasoning of the cardiologist is simulated (the earlier mentioned NIH program [4] for example used conventional clinical ECG criteria). For simulating the cardiologist's reasoning decision trees and fuzzy classifiers are among others used. When a database with labeled ECGs is available, decision trees can be automatically constructed [6]. A disadvantage of the use of decision trees is that a small change in a feature value can lead to a different path through the decision tree, possibly leading to a different diagnosis if the feature value is close to a threshold value. Fuzzy set classifiers can be applied to prevent this or to cope with imprecise descriptions, like 'a large Q-wave'.

In the statistical approach multivariate statistical techniques are applied to ECG features. The VCG interpretation program AVA (Automatic Vectorcardiogram Analysis), developed by the group of Pipberger, used the Bayesian approach [3]. The probability density functions of the relevant features needed for disease classification were obtained from a database of VCGs. The Bayesian classification procedure computed the patient's posterior probabilities of various disease categories like normal, various types of hypertrophy and myocardial infarction, etc. The results were promising, suggesting that diagnostic ECG classification can be significantly enhanced through the use of multivariate analysis.

Comparing several computer programs analyzing identical ECGs showed large differences in measurement results. Such large differences limit the possibility of exchanging diagnostic criteria between programs. To overcome some of these problems a concerted action, CSE (Common Standards for Quantitative Electrocardiography), a large international co-operative project, sponsored by the European Commission, was launched in 1980. The project led to standardization of ECG measurement procedures, standardization of diagnostic criteria and to the establishment of an ECG reference library with well annotated wave reference points. A board of cardiologists visually determined the onsets and offsets of the P, QRS, and T waves on highly amplified parts of ECG tracings and by using a modified Delphi approach, individual outlying point estimates were eliminated in four successive rounds [7,8].

The library proved to be a useful instrument. Using a set of ECGs and VCGs from the reference library it was shown that combined cardiologist and program results demonstrated the highest accuracy, higher than the result of any individual reader or program [9]. Another study compared the performance of nine electrocardiographic computer programs with that of eight cardiologists using 1220 ECGs from the library.

The median total accuracy level was 6.6% lower for the computer programs (69.7 percent) than for the cardiologists (76.3 percent). However, the performance of the best computer programs nearly matched that of the most accurate cardiologist [10]. The results of the concerted action have become internationally recognized milestones for the standardization of quantitative electrocardiography.

I was involved in the comparison of serial ECGs of patients, who suffered a myocardial infarction [11]. The two most recent ECGs were compared, and a trend analysis based on all ECG recordings of the patient was performed. It could be concluded that serial ECG comparisons are useful in acute myocardial infarction management.

2.2. Monitoring the EEG

As mentioned earlier the EEG can be a valuable tool for monitoring purposes. Therefore, a method to detect changes in the EEG whenever they occur is valuable. I became involved in research concerning this topic when supervising PhD candidate Ben Jansen [12]. The main goal of his study was to design an objective method that could quantify changes in the EEG and that could be applied in such diverse areas as monitoring the level of anesthesia, the efficacy of perfusion during open heart surgery or automatic sleep staging.

According to Elul [13] short EEG segments (one to five seconds long) can be regarded as stationary. Each short segment represents a specific state of the EEG. Most likely only a limited number of states (and thus differing segments, called elementary patterns) will be encountered in one recording. Stationary intervals can be lumped together into clusters, where each cluster presents a state of the EEG. This results in a description of the EEG (a profile) in terms of the percentage of time the EEG remains in each state.

Several researchers syntactically modeled the EEG as the output of an auto-regressive filter of an appropriate order with random noise as input. Fernando Lopes da Silva used the model for detecting spikes in the EEG [14]. He adapted the model to the first few seconds of a recording and then the remaining part of the tracing was used as input to the inverse model, thus generating random noise as long as no transients occurred. Spikes were detected when the output of this model exceeded some pre-set threshold. Jansen also used an autoregressive filter of order five to simulate a given measured EEG and applied a Kalman filter to compare the output of the auto-regressive filter with the measured EEG. The updated filter coefficients minimized the difference between the measured and the simulated EEG. Because of the earlier mentioned stationarity considerations, the EEG was segmented into 1.28 second intervals. Each interval was represented by a vector, consisting of the five (averaged) filter coefficients and the range of the EEG amplitude in that interval. The vectors representing the EEG intervals of a training set were used in an unsupervised cluster analysis. From this analysis emerged different clusters, representing different states of the EEG. The interval in the center of each cluster was regarded as the elementary pattern representing that cluster. These elementary patterns were used to classify the vectors of the segments in a test set.

After clustering each EEG interval from the test set was assigned to the most similar elementary pattern from the training set. For sleep staging profiles indicating the number of intervals assigned to each elementary pattern in an EEG recording (in sleep staging 30 second epochs were used) obtained from a test set were classified according to their similarity with the average profiles of the various sleep stages determined in a training phase. Using the profiles (per sleep stage) of one subject as a reference, about 80%

correct classifications of the sleep stages were obtained with the profiles of two other subjects. Between 60% and 90% agreement between judges was reported for sleep staging. Automatic sleep staging could reliably be done by means of profile classification. Frequency changes, induced by the starting and stopping of the pump and by the cooling and rewarming cycles during open heart surgery could reliably be detected.

3. Bayes Theorem

At the end of the 1950s Ledley and Lusted wrote several articles about medical diagnosis and decision making (among others [15]). They indicated that a physician in the processes of determining the diagnosis and formulating the treatment plan for a patient is frequently faced with a sequence of complex decisions. For the most part these decisions are made by means of heuristic procedures, on a largely intuitive basis. They suggested that logical analysis for determining a differential diagnosis, probabilistic analysis (Bayes' theorem) for determining the probabilities of the diseases contained in the differential diagnosis and value theory (decision analysis) to assist in the choice of the treatment plan if more options were available, could be successfully applied. Moreover, they advocated the use of computers for supporting physicians.

I wondered whether knowledge about how physicians diagnose a disease could be used to attack diagnostic problems with the help of computers. I learned from Elstein et al. [16] that physicians generate specific hypotheses very early in their encounter with the patient. These provisional hypotheses are generated out of the physician's background knowledge of medicine, including his range of specific experiences, in conjunction with problematic elements which he recognized in the early stages of the encounter with the patient. After hypotheses have been generated and roughly rank-ordered, they are systematically tested in the familiar medical work-up. This strategy, used by physicians, is called the hypothetico-deductive approach and was for example also used in INTERNIST (see further). But eliciting the knowledge and procedures used by expert physicians during diagnosis does not reduce interrater variability given the fact that even experts make diagnostic errors or do not agree between themselves. Randy Miller and colleagues, involved in the design of INTERNIST and QMR (see further) indeed remarked that the standard model for building expert systems (eliciting knowledge through the collaboration of domain expert and knowledge engineer) was not sustainable. They came to the following recommendation: use the biomedical literature as a gold standard for setting up a clinical knowledge base.

The idea of using Bayes' theorem for diagnosis was attractive because it allowed taking into account the uncertainties the diagnostic process has to deal with. Soon after Ledley's and Lusted's publications, articles dealing with the computer-assisted diagnosis of for example congenital heart disease, thyroid function and bone tumors [17-19] appeared. As we saw, also the AVA VCG interpretation program used Bayes' theorem for classification. From the literature [20-22] it is clear that Bayes' theorem was often used in the 1970s for computer-aided diagnosis. The performance of some of these programs was almost as good as those of experts in the respective fields. However, the output of programs using Bayes' theorem were difficult to value by physicians. The physician should know how the program arrives at its results and he should be aware of the quality of the statements made by the computer program. Only then can he take responsibility for his actions that are based upon results of computer programs. Use of

weights of evidence makes the output of programs based on Bayes' theorem easier to interpret [23].

In the beginning of the 1970s Tim de Dombal from Leeds University presented a system for diagnosing acute abdominal pain (AAPHELP) [24]. Their choice of the "acute abdomen" was a deliberate one: it is a common clinical dilemma, the number of possible diagnoses is relatively small, the clinical diagnosis is usually made on the basis of a patient's symptoms and physical signs rather than on biochemical tests and the final diagnosis is usually made at surgery. The program was based on an independence model of Bayes' theorem as were most of the programs using Bayes' theorem. The prior probabilities of the diseases and the conditional probabilities of the symptoms and signs given the diseases were determined from a database of 600 patients. A structured form was developed on which the data needed by the system were documented by the clinician. The performance of the clinician increased during the trial, probably due to the discipline of data collection (the structured form) and feedback about their performance. However, after the trial the performance of the clinicians decreased to the 'normal' level.

The system was validated in a controlled prospective trial in which the diagnostic performance of the unaided clinician was compared with that of the system. It appeared that the system performed better than the clinicians, even the most senior ones. The abdominal pain program could not always be used successfully: sometimes problems were encountered when the system was transferred to another location. The quality of the advice of the system is, among others, dependent on the referral policy in that new location. Different referral strategies may result in different prior probabilities of the diseases. Also, geographical variations in disease probabilities may occur. Although the abdominal pain system performed well, it was not used on a large scale. The problem probably was too specific, given the amount of time it took physicians to obtain a diagnostic prediction. The desktop computer version took the clinician five minutes, which is far too long when 15 to 20 patients are seen daily. So, even though de Dombal et al. could prove that the system performed at expert level, it was not used regularly.

Spiegelhalter and Knill-Jones [23] presented a statistical application for the diagnosis of dyspepsia, GLADYS (Glasgow Dyspepsia System), that made use of weights-of-evidence, being the logarithm of the likelihood ratio of a finding for a given disease. According to independence Bayes these weights, when added, are equal to the logarithm of the likelihood ratio of the posterior probability of the considered disease. The physician can now better interpret the size of the posterior probabilities: the higher the weight the more important the finding influenced the result.

According to Gorry and Barnett [25] calculating the posterior probability distribution for the diseases in question is one aspect of diagnosis but another important aspect is the determination of an appropriate sequence of questions and tests: determining which question should be asked or test be ordered next should be based on the information available after the previous question is answered or test result obtained. A sequential approach leads to a minimum number of questions and/or tests and saves discomfort to the patient, time and money. Bayes' theorem can also be used in this case. Gorry and Barnett describe the procedure for determining the appropriate sequence of questions and tests. For each ordered pair of diseases, the cost of misdiagnosing one disease for the other was specified. The cost of misdiagnosis given the current differential diagnosis can now be determined (when for example the diagnosis with the highest probability is selected as the definitive one) using the provided costs of misdiagnosis. This cost can be compared with the cost that results from asking a new question or requesting a new test. That question is asked or that test ordered for which the cost,

averaged over the differential diagnoses obtained for each possible outcome, is lowest and also lower than the cost of the differential diagnosis before this question is asked or test ordered. Then the posterior probabilities of the diagnoses in the differential diagnosis are updated. The procedure is repeated until no test or question will have a lower average cost than that of the current differential diagnosis. Also, other strategies for sequential diagnosis were introduced. Gleser and Collen [26] calculated the entropy of the distribution of the prior or posterior probabilities of the diseases and selected the test or question that gave rise to the largest reduction in entropy whereas Rector et al. [27] selected the test or question that maximized the weighted difference of the new and current posterior probabilities. For a detailed description the reader is referred to their papers.

The above publications were an incentive to dedicate part of a block course given at the Free University in Amsterdam [28] to the use of computers as an aid to diagnosis [29, 30]. The student was given an overview of what had been done in this field and the benefits and limitations of the approach were explained. As mentioned earlier Bayes' theorem was often used. This was the reason for demonstrating this statistical rule in the block course. Also, attention was paid to objections raised against using this theorem [31].

A database of 277 patients (including 63 normal patients), all referred to the hospital suspected of having congenital heart disease, was established. In total, seventeen questions could be asked about the status of the patient. The questions concerned age, sex, EKG data, radiologic data, presence of heart murmurs, cyanosis, femoral pulsations, and hepatomegaly. For each heart disease and the 'normal' population and for each answer subjective probabilities were available.

One of the tasks the students had to carry out is presented here. A clinician has to determine an optimal sequence of diagnostic tests for a particular patient and computers can be used to obtain such an optimal strategy. In the program this problem was translated into the problem of asking questions efficiently. The student could direct questions at a patient randomly selected from the database. The aim was to arrive at the correct diagnosis by asking a minimal number of questions. After the student had obtained the correct diagnosis (s)he could compare her/his strategy with the outcome of the computer that applied the sequential strategy explained in [27]. In this way (s)he could learn which questions were relevant and which were irrelevant in the present situation. But (s)he would also notice that the sequence of questions asked by the computer looks more erratic than the sequence of questions asked by physicians.

Not only Bayes' theorem was used to support decision making. Several statistical techniques like linear discriminant analysis or logistic regression were also applied. It was shown that these statistical techniques produced similar diagnostic results and had similar performances. The statistical approach to medical decision making was popular until in the early 1970s AI was introduced. Spiegelhalter and Knill-Jones [23] discussed various criticisms the AI community had towards statistical systems like the frequent assumption of conditional independence, the restriction to mutually exclusive and exhaustive diseases, the ignorance of the rich physiological knowledge and judgmental experience of clinicians, the need for large amounts of data, the placement of all shades of inexactness within a single probabilistic framework and an unavoidable loss of comprehensibility to the physician. They conclude that a synthesis between AI and statistical approaches is possible with the logical medical knowledge organized using an AI representation and any inexactness modelled using weights of evidence.

4. Knowledge-based Systems

The advantage of knowledge-based systems is that the knowledge for solving a problem is stored separately from the part that reasons with this knowledge. In order to be useful for physicians a knowledge-based system should be able to explain its knowledge of the domain and the reasoning processes it employs. Because the knowledge base is separate from the rest of the program the knowledge can more easily be updated than in algorithms, where the knowledge is interspersed with the code. There are various ways to represent knowledge, like IF-THEN rules, frames, ontologies, etc. Rules were considered as a model for human cognition and were therefore used for the representation of knowledge in expert systems. However, experts often appeared to have problems in formulating their medical judgements in the form of rules and to keep the rule base consistent. Also, the necessity to specify the context in which the rule is eligible may cause problems.

When I read about MYCIN, an early expert system developed by Ted Shortliffe and colleagues [32], used to identify bacteria causing severe infections and to determine an appropriate therapy, I was impressed. By backward chaining through applicable rules and asking the user medical information about the patient where necessary, the system could determine a diagnosis and suggest a medication. In addition, the system worked with certainty factors that indicated how certain either the antecedents of the rules or the certainty of the conclusions were. So, reasoning under uncertainty was possible in MYCIN.

Expert systems based on IF-THEN rules could explain to a certain extent how they arrived at a conclusion by showing the chain of rules that was followed. For example, when MYCIN asked a question, the user could ask why that question was posed. The system then presented to the user the higher-level goal it was attempting to satisfy. The user also could ask how the system arrived at certain conclusions. MYCIN was an example for many expert systems to come. Several of my PhD candidates developed decision support systems using a rule-based approach. Here I present one example.

To manage test consumption in the region of Maastricht in the Netherlands GPs got bi-yearly written feedback by human experts about their test requesting behavior. The feedback on GPs ordering behavior was highly effective and appreciated by the GPs, but such a form of feedback was laborious. Therefore, the aim of the research of Rianne Bindels [33] was to develop and evaluate an accurate and reliable reminder system that would give GPs immediate feedback about diagnostic test ordering that was not in line with national or regional guidelines and to provide the opportunity, when the GP agreed with the feedback, to change a request immediately after a reminder was shown. The reminder system, GRIF, consisted of a knowledge base, an order entry system and modules to provide passive support when the GP asked for background information and active support in the form of reminders when a test request was not according to the guidelines.

Knowledge was represented in the form of rules. With the help of a GP and an experienced internist the relevant parts of the guidelines were formalized. Not all guidelines could be formalized because of the lack of well-defined conditions like ‘the elderly’, ‘atypical complaints’, etc. These guidelines would be formalized later. The knowledge base used for testing consisted of 149 reminders concerning various medical problems.

The output of the system was validated using 253 randomly selected request forms from GPs with together 1217 test requests, also containing medical information about why the test was requested. Three expert physicians independently indicated whether the

tests were appropriately requested, based on their knowledge of the guidelines. In a first validation round the intra-rater and inter-rater agreement were determined. The intra-rater agreement varied between 72% and 85%, the inter-rater agreement between two raters varied from 67% and 74%. Also, the kappa values were relatively low (around 0.6 for each individual rater validating a number of forms twice and around 0,4 for couples of raters). Therefore, the majority agreement of the three experts was taken as the gold standard and the results again analyzed. In 13% of the requests the system did not react correctly. However, 4.5% of the accepted test requests, where the system accepted these requests using the majority judgement, appeared to be inadequate.

Also, the potential effect of the system on the test ordering behavior of 24 GPs was assessed. The GPs reviewed a random sample of 30 request forms they filled in earlier that year. If deemed necessary, they could make changes in the tests requested. Next, the system displayed critical comments about their non-adherence to the guidelines as apparent from the (updated) request forms. Both the number of requested diagnostic tests and the fraction of tests ordered that were not in accordance with the practice guidelines decreased due to the comments of the automated feedback system. The GPs accepted 50% of the reminders.

I was also very much interested in the program INTERNIST [34] that could diagnose complex cases and covered some 600 diagnoses in internal medicine. The technique used was similar to the hypothetico-deductive approach used by physicians.

The program was developed by Jack Myers, Harry Pople and Randy Miller and coworkers at the University of Pittsburgh in the beginning of the 1970s. In the knowledge base of INTERNIST the disease profiles of more than 600 diseases from internal medicine were stored. Contrary to expert systems philosophy prevalent at that time, where a “knowledge engineer” debriefed a “domain expert” to subjectively construct a knowledge base it was decided to use the peer-reviewed published literature as the independent gold standard source of knowledge on which to build the knowledge base.

A disease profile consists of all manifestations (patient's history, physical exam, and laboratory data) that are associated with that disease. From the disease profiles a differential diagnosis list for each manifestation can be determined. Each manifestation in the disease profile was characterized by the values of two parameters: evoking strength, a kind of positive predictive value (how probable is the diagnosis when the manifestation is present), and frequency, a kind of sensitivity (the probability that patients with this diagnosis have the manifestation). The clinical importance of the manifestation, independent of the disease, is described by the value of a third parameter, import. The parameter values were – as a result of medical judgement - divided into classes and each class had a certain weight.

For each of the patient’s manifestations for each item present in the manifestation’s differential diagnosis list a disease hypothesis is created. The disease hypotheses of all manifestations are stored in a master list. For each disease hypothesis from the master list the weights of the evoking strength parameter of all manifestations present in the patient that also appear in the profile of that disease hypothesis were added. From this total score the sum of the weights of the frequency parameter of those manifestations that appeared in this disease profile but are not present in the patient was subtracted (because it reduces the probability of the disease to be present). Also the sum of the weights of the import parameter of manifestations that do appear in the patient but not in the disease profile was subtracted from the total score. For each diagnosis hypothesis the resulting score is an indication of the probability that it is present. The hypotheses are now ranked according to their total score and the disease hypotheses whose scores are more than a

specified amount less than the highest score are temporarily discarded. The diagnosis with the highest score qualifies as a definitive candidate for the diagnosis of the patient. But there may be competitors of which the set of manifestations is a subset of the set of manifestations of the diagnosis with the highest score. Of this group of competitors only one can be the definitive diagnosis and therefore these competitors form the differential diagnosis. Different strategies were used to determine the definitive diagnosis, depending on the difference between the score of the diagnosis with the highest score and that of the competitors. When the system cannot determine further questions or tests and the differential diagnosis contains more than one disease, the program displays these diagnoses together with their scores as tentative diagnoses. When the remaining manifestations have an import of 2 or less, the program ends. As soon as a definitive diagnosis is established all manifestations explained by the definitive diagnosis are removed and a new differential diagnosis is determined in the same way as explained above for the remaining disease hypotheses, including the ones that were temporarily discarded.

The program was evaluated on 19 clinicopathological complex cases published in the New England Journal of Medicine. A comparison was made with human experts. There were 43 possible correct diagnoses for the 19 cases. INTERNIST made 17 correct definitive diagnoses and 8 correct tentative diagnoses (when INTERNIST presented a differential diagnosis instead of a definitive diagnosis and the tentative diagnosis with the highest score was the correct diagnosis), whereas the clinicians made 23 correct definitive diagnoses and five correct tentative diagnoses (if the physicians presented a differential diagnosis with the correct diagnosis on top). In addition to demonstrating the impressive capabilities of the system, the evaluation identified several shortcomings of its approach. First, the knowledge base and diagnostic algorithms did not adequately represent disease and finding severity. Second, the temporal course of a patient's illness could not be fully described. Also, the program could not "reason" anatomically regarding aspects of the patient's presentation [35].

The role of the physician, when using INTERNIST, was limited to the entry of data, the system did the diagnostics part. Later it was realized that this approach was wrong: the approach was called the Greek oracle approach, clearly indicating the role of the system. In the 1980s QMR (Quick Medical Reference) was developed. The system functioned as an information tool, providing users with multiple ways of reviewing and manipulating the diagnostic information in the program's knowledge base [36].

5. Computer-interpretable Guidelines

Not only reminder systems were developed, but also interest also existed in the formalization of clinical practice guidelines, that describe how to diagnose or treat a patient. Protocols were already in use for a long time to support nurses and ancillary personnel. Protocols can be seen as directives of how the user should approach a problem and were usually displayed as flowcharts or decision tables. Guidelines are less restrictive than protocols and serve as recommendations that may be rejected by the physician as long as the physician documents the reasons why he did not follow the guideline. From the early 1970s several studies demonstrated wide variations in medical practice among physicians, hospitals and different geographical areas. Gradually physicians started to use paper-based guidelines to ensure consistent high-quality care. However, these guidelines were written down in large documents in a textual format,

were often cumbersome to read and difficult to integrate in the patient care process. Updating paper-based guidelines in addition required the production of new documents. Grimshaw and Russell reviewed published evaluations of the application of clinical guidelines [37]. All but 4 out of 59 guidelines resulted in improvements in the process of medical care and all but 2 out of 11 evaluations that also measured the outcome of care reported improvements in outcome. Their conclusion was that guidelines can change clinical practice if they are appropriately developed, disseminated, and implemented.

To improve access to the paper-based guidelines they were entered into the computer. But most of the guidelines were still presented as large documents in a textual format. Marieke Vissers started a research project with the goal to develop and implement a prototype information system that would present guidelines in a well-organized and user-friendly way. Users should be able to familiarize themselves with the system in a short time and limited data input by keyboard and an easy to control user interface should enhance the acceptability of the system. The system, ProtoVIEW, had the characteristics of a reference system, provided solicited advice and guided the user through the protocol [38]. Among others the value of the system for assisting inexperienced residents in the management of common medical problems in the A&E department was evaluated [39]. The residents stated that they found ProtoVIEW easy to use. However, although consultation of ProtoVIEW under routine circumstances took only one and a half minute, residents doubted whether the use of ProtoVIEW would be faster than consultation of other information sources, like colleagues. Later a Web-based version of ProtoVIEW was developed that contained all its functionalities plus several new ones. The web version contained an X-ray viewer and provided a great deal of interactivity such as validation of electronic patient data forms. The most important additional function was the context sensitive protocol support that may lead to improved protocol adherence. Finally, the web-based version could be accessed from any working place since patient data and protocols were stored centrally [40].

Implementing executable guidelines in a computer-based decision support system could improve the application of guidelines still more because the actions and observations of care providers can be monitored, and advice related to the individual patient is generated when needed. Many parties developed decision support systems that incorporated guidelines, covering a wide range of clinical settings and tasks [41]. However, only a few systems progressed beyond the prototype stage. Building systems that were both effective in supporting clinicians and accepted by them proved to be a difficult task. Yet, of the few systems that were evaluated by a controlled trial, the majority showed impact [42]. To make the advice patient specific the system must be able to access clinical data. The guideline system therefore should be interfaced with the EPR system, otherwise the physician has to enter data twice.

Various difficulties were encountered with respect to the guideline development process ranging from the development of a guideline representation model to the implementation of actual decision support systems that operate in daily practice. Existing paper-based guidelines had to be formalized and expressed in a common representation language, using a common terminology for expressing clinical data. The interpretation of the content of guidelines and therefore their formalization could be difficult: the exact meaning of terms was not always defined; recommendations were not always clearly articulated, and sometimes vague wording was used. In the Netherlands new cardiac rehabilitation guidelines were being drafted during the time Rick Goud started developing the decision support system CARDSS (cardiac rehabilitation decision support system) to support the entire process of rehabilitation, with a focus on needs

assessment [43]. Goud could participate in all the meetings of the cardiac rehabilitation guidelines development committee and co-authored the flowchart summarizing the needs assessment procedure. The concurrent development and formalization of the guideline helped to identify in the narrative guidelines both vague, inconsistent as well as difficult to apply recommendations [44]. CARDSS was adopted in practice and was used in over 30 Dutch outpatient clinics.

A number of research projects started developing generic methodologies that could solve many of the problems related to the guideline development process. We mention some important approaches: GLIF [45], PROforma [46], Asbru [47] and EON [48]. Also PhD candidate Paul de Clercq developed and evaluated a generic approach that addressed questions such as how to represent, acquire and implement computer-based guidelines. The approach led to the development of the Gaston framework [49]. The project started in 1996. A number of systems were developed using the Gaston approach: the earlier introduced GRIF reminder system [33], that provided feedback on test ordering in general practice; CritiCIS, a real-time critiquing system used in critical care environments such as intensive care units [50], M-PADS, a psychopharmacological advisory system that supports the process of selecting the most suited psycho-active drug [51], a consumer health record system for managing chronic diseases [52], GASTINE [53] for intention-based decision support (see below) and CARDSS [43] for cardiac rehabilitation guidelines, presented above.

Decision support systems can issue reminders or alarms when the EPR shows that physicians are not working according to the guideline. But the physician may have executed an action that was in the spirit of the guideline and still get a warning. Such warnings of the system will annoy the user. If next to the suggested actions the guideline also contains information about why these actions are carried out (the intention behind the action) and which actions are in line with the intentions, the physician will not receive such warnings. Moreover, the intentions can be used to explain to the interested physician why a certain recommendation was given. Agnes Latoszek-Berendsen started research on intention-based decision support. The representation formalism for intentions and their implementation in guidelines was called GASTINE (Gaston intentional expressions) [53]. The formalism was used to formalize and implement the Dutch heart failure guideline. She demonstrated that the use of intentions offers the flexibility needed to avoid unnecessary error messages and warnings. When the system was used in the pro-active mode it provided the user with actions mentioned in the guideline. Only when the system was reacting to information entered by the physician in the EPR it would check whether the action that was different from the action in the guideline was in the spirit of the guideline and if so it would not present a warning.

6. Conclusions

During my journey I learnt a lot. For example, I came gradually to the conclusion that we have to live with uncertainties in medicine. Inter-rater and intra-rater variability for example will not disappear and therefore basing the gold standard on raters will not be error-free. Use of consensus between experts as a gold standard may help, but we saw above that 4.5% of the test requests were incorrectly accepted because the system agreed with the majority judgement of experts. Evidence-based knowledge used in decision support systems should come from validated studies reported in scientific literature.

Many studies have shown the human judgement to be unreliable. Moreover, the capacity of man as an effective problem-solver is very limited. Man tends to gather information indiscriminately, although he is only capable of combining a limited number of facts simultaneously. Furthermore, men are conservative information processors, who do not extract all the material inherent in the information. Therefore, systems that provide pro-active reminders (before actions are taken) or reactive warnings (when actions have been taken) are necessary tools. In my opinion we should not try to design systems that can solve everything as long as the user provides the relevant data. The INTERNIST project showed that such a Greek oracle approach is unwanted. The physician should get support in a way that interferes with his work as little as possible. Decision support systems with executable guidelines that can follow the physicians' actions via the EPR and that provide warnings when the physician does not work according to the guideline or give reminders so that a physician does not forget to take certain actions are in my view most wanted. They will reduce errors.

Computer systems combine observed facts and interpret them, based upon the existing scientific knowledge available in the programs. We have to remind ourselves that this knowledge is generalized knowledge in the sense that it is pertinent to 'the patient' and that it only describes quantifiable aspects of real patients. The physician is responsible for the management of an individual patient: he has to combine the information delivered by the computer program with other available information about this patient, e.g., non-quantifiable data. Only the doctor as a human being can make decisions about the management of the disease of another human being. The role of the computer is to remind the doctor of possibilities overlooked by him and to furnish him with scientific knowledge pertinent to the patient under consideration. A computer can never be responsible, the physician is. Therefore, the computer can be a useful tool, but can never replace the doctor.

Our goal should be to make tools that can support physicians and do not replace them, as we saw with INTERNIST. We have to admit that the physician is the pilot of the system. Therefore, we should build tools that can be used by the physicians almost in the same way as they use the results from laboratory tests or ECGs. They can use the information of the tools to make up their minds. Also, patients can use the information and be involved in deciding what to do. As a last remark: I did not go into detail about the prospects of NLP and machine learning. I think that they will get an important role in the future. But again, 5we should be modest and not try to build artificial physicians.

References

[1] Yerushalmy J. Reliability of chest radiography in the diagnosis of pulmonary lesions. Am J Surg. 1955;89(1):231-40.

[2] Talmon JL, van Bemmel JH. Modular software for computer-assisted ECG-VCG interpretation, In: Anderson J, Forsighte JM, editors. Medinfo 1974; Amsterdam: North-Holland; 1974. p. 653-58.

[3] Pipberger HV, Freis ED, Taback L, Mason H. Preparation of electrocardiographic data for analysis by digital electronic computer. Circulation 1960;21:413–18.

[4] Caceres CA, Steinberg CA, Abraham S, Carbery WJ, McBride JM, Tolles WF, Rikli AE. Computer extraction of electrocardiographic parameters. Circulation 1962;25:356–362.

[5] Kors JA, van Bemmel JH. Classification methods for computerized interpretation of the electrocardiogram. Methods Inf Med. 1990 Sep;29(4):330-6.

[6] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Ebook: New York: Routledge: 2017.

[7] Willems JL, Arnaud P, van Bemmel JH, Degani R, Macfarlane PW, Zywietz C. Common standards for quantitative electrocardiography: goals and main results. Methods Inf Med 1990;20:263–71.
[8] Willems JL, Arnaud P, van Bemmel JH, Bourdillon PJ, Degani R, Denis B, Graham I, Harms FM, Macfarlane PW, Mazzocca G, et al. A reference data base for multilead electrocardiographic computer measurement programs. J Am Coll Cardiol. 1987 Dec;10(6):1313-21.
[9] Willems JL, Abreu-Lima C, Arnaud P, van Bemmel JH, Brohet C, Degani R, Denis B, Graham I, van Herpen G, Macfarlane PW, et al. Effect of combining electrocardiographic interpretation results on diagnostic accuracy. Eur Heart J. 1988;9(12):1348-55.
[10] Willems JL, Abreu-Lima C, Arnaud P, van Bemmel JH, Brohet C, Degani R, Denis B, Gehring J, Graham I, van Herpen G, et al. The diagnostic performance of computer programs for the interpretation of electrocardiograms. N Engl J Med. 1991;325(25):1767-73.
[11] van Haelst AC, Donker DK, Visser FC, de Cock C, Hasman A, Talmon JL. A computer program for the analysis of serial electrocardiograms from patients who suffered a myocardial infarction. Int J Biomed Comput. 1985 Nov;17(3-4):273-84.
[12] Jansen BH, Hasman A, Lenten R. Piece-wise EEG analysis: an objective evaluation. Int J Biomed Comput. 1981;1:17-27.
[13] Elul R. The genesis of the EEG, In: Pfeiffer CC, Smithies JR, editors. Int Rev Neurobiol. 1971;15:227-72.
[14] Lopes da Silva FH, Dijk A, Smits H. Detection of non-stationarities in EEGs using the autoregressive model – An application to EEGs of epileptics. In: Dolce G and Kuenkel H, editors. CEAN: Computerized EEG Analysis; Stuttgart: Fischer. 1975;180-199.
[15] Elstein AS, Shulman LS, Sprafka SA. Medical problem solving. In: An analysis of clinical reasoning. Cambridge, MA: Harvard University Press; 1978.
[16] Ledley RS, Lusted LB. Reasoning foundation of medical diagnosis: Symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science 1959;130:9.
[17] Warner HR, Toronto AF, Veasy LG. Experience with Bayes' theorem for computer diagnosis of congenital heart disease. Ann N Y Acad Sci. 1964 Jul 31;115:558-67.
[18] Overall JE, Williams CM. Conditional probability program for diagnosis of thyroid function. JAMA. 1963 Feb 2;183:307-13.
[19] Lodwick GS, Haun CL, Smith WE, Keller RF, Robertson ED. Computer diagnosis of primary bone tumors. Radiology. 1963;80:273–5.
[20] Wardle A, Wardle L. Computer-aided diagnosis: A review of research. Methods Inf Med. 1978 Jan;17(1):15-28.
[21] Shortliffe EH, Buchanan BG, Feigenbaum EH. Knowledge engineering for medical decision making: A review of computer-based clinical decision making. Proc. IEEE; 1979 September. p. 1207-1224.
[22] Rogers W, Ryack B, Moeller G. Computer-aided medical diagnosis: literature review. Int J Biomed Comput. 1979 Aug;10(4):267-89.
[23] Spiegelhalter DJ, Knill-Jones RP. Statistical and knowledge-based approaches to clinical decision-support systems with an application in gastroenterology, J.R. Statist. Soc. A. 1984;147(1):35-77.
[24] de Dombal FT, Leaper J, Staniland JR, McCann AP, Horrocks JC. Computer-aided Diagnosis of Acute Abdominal Pain. Br Med J. 1972;2(5804):9–13.
[25] Gorry GA, Barnett GO. Experience with a model of sequential diagnosis. Comput Biomed Res. 1968;1(5):490-507.
[26] Gleser MA, Collen MF. Towards automated medical decisions. Comput Biomed Res. 1972;5(2):180-9.
[27] Rector AL, Ackerman E. Rules for sequential diagnosis. Comput Biomed Res. 1975 Apr;8(2):143-55.
[28] van Bemmel JH. The system behind medical computer applications - guiding principles for courses and training. In: Lindberg DAB, Kaihara S, editors. Proceedings of Medinfo 80, Tokyo;1980. p. 353-357.
[29] Hasman A. Description of a blockcourse in medical informatics. Methods Inf. Med. 1989 Nov;28(4):239-42.
[30] Hasman A. Training in medical informatics. The use of computers for diagnostic purposes. Int J Biomed Comput. 1982 Mar;13(2):109-18.
[31] Feinstein AR. XXXIX. The haze of Bayes, the aerial palaces of decision analysis and the computerized Ouija board. Clin Pharmacol Ther. 1977 Apr;21(4):482-96.
[32] Shortliffe EH. Computer-assisted medical consultations: MYCIN. New York: Elsevier; 1976.
[33] Bindels R, de Clercq PA, Winkens RAG, Hasman A. A test-ordering system with automated reminders for primary care based on practice guidelines. Int J Med Inform. 2000 Sep;58-59:219-33.
[34] Miller RA, Pople HE Jr, Myers JD. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med. 1982 Aug 19;307(8):468-76.
[35] Miller RA. A history of the INTERNIST-1 and Quick Medical Reference (QMR) computer-assisted diagnosis projects, with lessons learned. Yearb Med Inform. 2010;121-36.

[36] Miller RA, McNeil MA, Challinor SM, Masarie FE Jr, Myers JD. The INTERNIST-1/QUICK MEDICAL REFERENCE project--status report. West J Med. 1986 Dec;145(6):816-22.
[37] Grimshaw JM, Russell IT. Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. Lancet. 1993 Nov 27;342(8883):1317-22.
[38] Vissers MC, Hasman A, Donkers HH, vd Linden CJ. Development, implementation and a first evaluation of a protocol processing system (ProtoVIEW). Comput Methods Programs Biomed. 1995 Jun;47(1):81-9.
[39] Vissers MC, Hasman A, vd Linden CJ. Impact of a protocol processing system (ProtoVIEW) on clinical behaviour of residents and treatment. Int. J. Biomed. Comput. 1996;42(1-2):143-50.
[40] Vissers MC, Hasman A. Building a flexible protocol information system with ready for use' web-technology. Int J Med Inform. 1999 Feb-Mar;53(2-3):163-74.
[41] van der Lei J, Talmon JL. Clinical decision support systems. In: Van Bemmel JH, Musen MA, editors. Handbook of Medical Informatics. Houten: Bohn Stafleu van Loghum; 1997.
[42] Johnston ME, Langton KB, Haynes RB, Mathieu A. Effects of computer-based clinical decision support systems on clinician performance and patient outcome. A critical appraisal of research. Ann Intern Med. 1994 Jan 15;120(2):135-42.
[43] Goud R, Hasman A, Peek N. Development of a guideline-based decision support system with explanation facilities for outpatient therapy. Comput Methods Programs Biomed. 2008 Aug;91(2):145-53.
[44] Goud R, Hasman A, Strijbis AM, Peek N. A parallel guideline development and formalization strategy to improve the quality of clinical practice guidelines. Int J Med Inform. 2009 Aug;78(8):513-20.
[45] Ohno-Machado L, Gennari JH, Murphy SN, Jain NL, Tu SW, Oliver DE, Pattison-Gordon E, Greenes RA, Shortliffe EH, Barnett GO. The guideline interchange format: a model for representing guidelines. J Am Med Inform Assoc. 1998 Jul-Aug;5(4):357-72.
[46] Fox J, Johns N, Rahmanzadeh A. Disseminating medical knowledge: the PROforma approach. Artif Intell Med. 1998 Sep-Oct;14(1-2):157-81.
[47] Shahar Y, Miksch S, Johnson P. The Asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines. Artif Intell Med. 1998 Sep-Oct;14(1-2):29-51.
[48] Musen MA, Tu SW, Das AK, Shahar Y. EON: a component-based approach to automation of protocol-directed therapy. J Am Med Inform Assoc. 1996 Nov-Dec;3(6):367-88.
[49] de Clercq PA, Hasman A, Blom JA, Korsten HH. Design and implementation of a framework to support the development of clinical guidelines. Int J Med Inform. 2001 Dec;64(2-3):285-318.
[50] de Clercq PA, Blom JA, Hasman A, Korsten HH. A strategy for developing practice guidelines for the ICU using automated knowledge acquisition techniques. J Clin Monit Comput. 1999 Feb;15(2):109-17.
[51] van Hyfte DM, de Vries Robbé PF, Tjandra-Maga TB, van der Maas AA, Zitman FG. Towards a more rational use of psychoactive substances in clinical practice. Pharmacopsychiatry. 2001 Jan;34(1):13-8.
[52] de Clercq PA, Hasman A, Wolffenbuttel BH. Design of a consumer health record for supporting the patient-centered management of chronic diseases. Stud Health Technol Inform. 2001;84(Pt 2):1445-9.
[53] Latoszek-Berendsen A, de Clercq P, van den Herik J, Hasman A. Intention-based expressions in GASTINE. Methods Inf Med. 2009;48(4):391-6.