EVALUATION OF A DECISION SUPPORT SYSTEM FOR THERAPY PLANNING IN GESTATIONAL DIABETES

M.E. Hernando ¹, E.J. Gómez ¹, R. Corcoy ² and F. del Pozo ¹.


ABSTRACT

DIABNET is a knowledge-based system designed to aid doctors with therapy planning in gestational diabetes. The system core is a qualitative model, implemented by a Causal Probabilistic Network, that is able to detect the insulin effectiveness on a daily basis. DIABNET analyses monitoring data and proposes quantitative changes in insulin therapy and qualitative diet modifications. This paper proposes an evaluation methodology to assess the system performance when working in a real scenario. The methodology manages the absence of a gold standard and includes: a subjective analysis based on questionnaires and an objective analysis based on a quantitative comparison of the system’s and experts’ proposals. The paper also shows the results of two experiments in which expert diabetologists evaluated the therapeutical advice provided by DIABNET during the follow up of 9 patients with gestational diabetes. DIABNET detected the need of a therapy modification in 92% of the cases showing its appropriateness for automatic alarm generation. Around 80% of the proposals were accepted by experts. The evaluation results are encouraging and allow characterisation of the system’s performance when proposing therapy modifications. Evaluation in its turn helps to refine the knowledge managed by DIABNET and enables us to look towards the further clinical use of DIABNET as a decision tool in gestational diabetes integrated in a telemedicine service.

Keywords: Evaluation; Therapy planning; Subjective questionnaires; Gestational diabetes; Causal Probabilistic Networks.

correspondence to:

M. Elena Hernando
Grupo de Bioingeniería y Telemedicina
ETSI Telecomunicación
Universidad Politécnica de Madrid
28040- Madrid- Spain
e-mail: elena@gbt.tfo.upm.es
Ph: +34 91 5495700 (ext. 332)
EVALUATION OF A DECISION SUPPORT SYSTEM FOR THERAPY PLANNING IN GESTATIONAL DIABETES

M.E. Hernando 1), E.J. Gómez 1), R. Corcoy 2) and F. del Pozo 3).


ABSTRACT

DIABNET is a knowledge-based system designed to aid doctors with therapy planning in gestational diabetes. The system core is a qualitative model, implemented by a Causal Probabilistic Network, that is able to detect the insulin effectiveness on a daily basis. DIABNET analyses monitoring data and proposes quantitative changes in insulin therapy and qualitative diet modifications. This paper proposes an evaluation methodology to assess the system performance when working in a real scenario. The methodology manages the absence of a gold standard and includes: a subjective analysis based on questionnaires and an objective analysis based on a quantitative comparison of the system’s and experts’ proposals. The paper also shows the results of two experiments in which expert diabetologists evaluated the therapeutical advice provided by DIABNET during the follow up of 9 patients with gestational diabetes. DIABNET detected the need of a therapy modification in 92% of the cases showing its appropriateness for automatic alarm generation. Around 80% of the proposals were accepted by experts. The evaluation results are encouraging and allow characterisation of the system’s performance when proposing therapy modifications. Evaluation in its turn helps to refine the knowledge managed by DIABNET and enables us to look towards the further clinical use of DIABNET as a decision tool in gestational diabetes integrated in a telemedicine service.

Keywords: Evaluation; Therapy planning; Subjective questionnaires; Gestational diabetes; Causal Probabilistic Networks.

1. INTRODUCTION

1.1 The medical problem

Gestational diabetes (GD) is a condition present in approximately 10% of pregnant women during the last months of pregnancy and usually disappears after delivery. The hallmark of GD is an elevated blood glucose concentration caused by an inadequate adjustment of glucose metabolism to the requirements of pregnancy [1]. Ketonuria may also be present either spontaneously or following carbohydrate restriction. Very tight control of blood glucose levels in pregnancy is necessary to avoid features of diabetic fetopathy, such as macrosomia or neonatal hypoglycaemia. Diet is the main control action at the onset of the illness, but when control deteriorates, insulin therapy must be applied.

The metabolism of gestational diabetic patients changes continuously throughout the pregnancy and increasing levels of glycaemia are usually observed due to the rise in insulin resistance. Home monitoring of blood glucose levels aims to detect these changes in order that the therapy may be adjusted to the new situation. The short duration of the illness prevents patients from being educated in insulin self-management, so patients are strictly controlled by doctors through frequent visits to the hospital (e.g. every one or two weeks).
1.2 The DIABNET system

The characteristics of gestational diabetes highlight the need for information systems to help clinicians in their daily tasks of patient monitoring and therapy planning. As patients’ conditions must be assessed very frequently, the use of a telemedicine service could optimise the number of visits, offering the possibility of making data available when necessary to allow an adequate and continuous patient care. In this context, decision aid tools will be useful to help clinicians to interpret the huge amount of data produced during the management of the illness. These considerations motivated the development of DIABNET [2], a knowledge-based system designed to aid doctors with therapy planning in GD.

DIABNET integrates qualitative and quantitative reasoning modules to produce advices including insulin therapy adjustments and diet recommendations. The system is designed for use in health centres, so that through integration into the DIABTel Telemedicine service [3], diabetic patients may be monitored.

DIABNET inputs are the patient’s available self-monitoring data (blood glucose, time and diet modifications and ketonury measurements) that could be incomplete. The DIABNET output is a dietary and insulin therapy adjustment that includes initiation of insulin therapy, quantitative insulin dose changes and qualitative diet and schedule modifications. Over periods of several days monitoring data are analysed to detect the diet or insulin therapy items which may require modification.

The DIABNET system imitates diabetologists’ reasoning processes by integrating qualitative and quantitative modules in a hybrid architecture. The system core is a probabilistic qualitative model which performs the “cause identification” task and is implemented by a Causal Probabilistic Network. This metabolic qualitative model analyses incomplete data from a single day. Later on, data from several days are considered to detect whether anomalies appear regularly or just sporadically, in which case they could be due to deviation from the therapeutic recommendations. To perform this temporal analysis, the temporal reasoning module interprets information from up to seven days, which is the natural assessment period. The temporal analysis obtains the “adjustment needs” during the day and indicates not only whether one or more therapeutic actions are required but also the level of insulin “need” or “excess”. The final step is to translate these required adjustments into quantitative and qualitative therapy modifications, by considering patient characteristics, current therapy and interactions between insulin absorption profiles. Proposals are built in line with the modification strategies usually used by doctors treating patients with gestational diabetes.

1.3 Evaluation of expert systems in diabetes

Although several evaluation theories have been developed for application in different fields of expert systems in medicine [4,5,6,7], there is no consensus on the best way to evaluate advices from expert systems in diabetes. The number of available system prototypes versus the number of formal evaluation studies is clearly unbalanced. This is due to the effort, time and work input that these studies imply as well as to the difficulties in establishing clinical co-operation and access to a significant number of patients [8]. These problems are common to most of the expert systems in medicine, but in diabetes, the lack of a gold standard is an additional difficulty in the evaluation of different therapy adjustment proposals. The therapy of a diabetic patient can be affected by different variables, e.g.: number of injections; types and doses of insulin administrations; waiting time between each injection and the related meal; changes in the injection sites; total daily calory intake; distribution of daily calories between meals; and duration of fasting intervals. Some of these different therapeutical actions can produce similar effects in the control of blood glucose profiles but there is no consensus between clinical groups about the best solution.

The inputs and outputs of an expert system in therapy planning usually consider a subset of the clinical variables that would be considered and modified in practice. As systems generate more limited proposals than humans, it is common for evaluation methodology to require the definition of a translation phase where experts code humans’ and systems’ proposals to express them in the same terms before other experts compare them blindly. The process of information translation can introduce biases or errors and can mask the evaluation of the system performance in a real scenario where human translations will not be possible. Thus, to measure the real benefits of a system in clinical practice, it is important not only to compare systems and
humans in the same conditions but also to compare the differences between the system’s working conditions and the real ones. Diabetes is a typical domain where evaluation under real conditions should be measured carefully because most of the expert systems for therapy planning are able to adjust the insulin protocol but not the other variables involved in the therapy.

Evaluation is a complex task, but, as Bergman and Buchanan highlighted [9], only after evaluation studies doctors and patients will have confidence on system safety. Thus, evaluation is the imperative step to achieve the use of decision support systems in clinical routine. A general classification of the evaluation phases of an expert system can be found in the work of Engelbrecht et al. [10]. Some specific methods must be developed to manage the evaluation of systems for therapy planning in diabetes. The absence of a gold standard means that it is not possible to determine whether two therapy adjustment proposals are equivalent, merely through computational methods so it is necessary to know the degree of experts’ acceptance for each proposal, in terms of “effectivity to correct deviations” or “risk for the patient” [11]. Comparisons cannot be done by using computational methods but they must be performed by experts in the management of diabetic patients.

This paper presents the evaluation methodology applied to assess the therapeutical advices provided by DIABNET. The results of the evaluation of the DIABNET system are also presented.

2. EVALUATION OF THE DIABNET SYSTEM

2.1 Evaluation methodology

The proposed methodology supports the evaluation of decision aid systems for therapy planning by giving feedback on performance when they were working in a real scenario, analysing real cases. The goal of this methodology is to compare the system proposals against the clinical decisions taken by patients’ physicians. Information is also obtained on the cases and on the appropriateness of the inputs and outputs managed by the system. The methodology deals with the lack of a gold standard in diabetes by obtaining subjective information from expert diabetologists which is used to interpret similar patients’ cases to the presented ones (e.g. gestational diabetes, children, etc). No special selection of cases should be done for inclusion in the evaluation experiments.

A subjective questionnaire is designed to record the evaluators’ assessment of the different therapy modification proposals. The same questionnaire is used to evaluate the proposals built by the system and also the proposals prepared by the human expert. Evaluators had to express their agreement to each item in the questionnaire by choosing one of the five possible answers: “completely disagree”, “disagree”, “none”, “agree” or “completely agree”. To avoid experts answering the questionnaires mechanically, items were designed carefully to prevent that the answer “completely agree” was the most positive one for all the items.

The statements comprised in the questionnaire are divided into two sets. The first set acquires information on the degree of acceptance of each proposal:

1. The proposal is effective to correct deviations
2. The proposal implies risk for the patient
3. The proposal is aggressive
4. The proposal is conservative
5. You accept this proposal as valid

The second set of statements acquires information about case characteristics and the suitability of the information contained in a case to prepare a therapy adjustment proposal:

6. The difficulty of data interpretation is high
7. The decision on therapy modification could be made in a shorter period
8. There is enough information to make a decision
Two different experiments are designed to evaluate the performance of the system compared to physicians. The main difference between the experiments is that in the first, the evaluator is the clinician that sees the patients in most of their visits to the hospital and decides on the prescribed therapy adjustments having access to all the patient information. In experiment 2, evaluators are experts who have not participated in the development of the system and do not know the origin of the presented proposals.

The following sections describe in more detail the two evaluation experiments and the process of presenting the case data and questionnaires to the evaluators.

2.1.1 Experiment 1
The main goal of this experiment was for proposals prepared by the system to be judged by the diabetologist treating the patients. The evaluator that participates in this experiment exclusively assesses proposals coming from the system and knows the origin of those proposals.

The experiment is as follows (see Figure 1): each system proposal is presented to the expert together with: the monitoring data associated to the analysis interval; the active therapy; and the therapy modification prescribed to the patient during his/her visit to the hospital. The expert had to fill in a complete evaluation questionnaire associated to each case to express his acceptance or rejection of each proposal and to give information about case characteristics (items from 1 to 7).

A priori, differences between system and expert’s proposals are expected because they are not built under the same conditions: the expert has access to more information than the system: 1) asking patients during their visits to justify abnormal data; 2) knowing patients’ past metabolic evolution; 3) knowing patients’ response to previous modifications and 4) knowing patients’ habits and compliance to the therapy. Usually this type of information is not recorded but is taken into account when analysing patient’ data.

When the evaluator is the same expert that participates in the creation of the system knowledge bases, this experiment also allows evaluation of the degree of similarity between the system and the expert’s reasoning.

2.1.2 Experiment 2
Evaluators were experts who had not participated in the development of the system. The first goal of this experiment was to characterise the cases in terms of difficulty and amount of information to make proper decisions. The second objective of this experiment was for different diabetologists to judge the system proposals and those of the patients’ physician. Experiment 2 was divided into two steps:

Step 1:
A deep analysis of data is necessary in order to characterise the cases. In this step (see Figure 2), evaluators made the effort of analysing each case in detail in order to form their own proposals. Evaluators were provided with the same information as the system. Cases were presented to experts showing information about 1) self-monitoring data between two visits (blood glucose and ketonuria); 2) patient characteristics (weight); and 3) actual insulin therapy. No information about prescribed therapy adjustments was provided in this step.

After analysing each case, evaluators had to fill in a questionnaire to determine the difficulty of the case and whether there was sufficient information available to make a good decision. This questionnaire includes the second set of sentences presented above (items from 6 to 8). An example of case presentation can be found in Annex A.

Step 2:
The goal of this step (see Figure 3) was to compare the system’s performance with the decisions taken in the real scenario. Each case was presented to the evaluators together with the two different therapy
modification proposals: a) the proposal prepared by the system and b) the therapy modification prescribed during the patient’s visit to the hospital. No information about the origin of the proposals was provided to the evaluators.

Evaluators had to fill in a questionnaire to express their acceptance or rejection of each proposal. This questionnaire included the first set of sentences presented above (items from 1 to 5).

(Insert Figure 3 around here)

2.2 Data registration

In the evaluation of any diabetes decision support system, involving comparison with the real clinical decisions, it is important to register the majority of the information considered for therapy modification decisions (weight, self-monitoring data, therapy deviations, illness, etc) as well as all the therapeutical actions decided at patients’ visits.

Diabetic patients register in their logbook a lot of information that will be useful to doctors deciding on therapy adjustments. Usually the logbook information is not complete and doctors have to question patients during their visits to get the necessary information. This complementary information is not registered later in the logbook but it could be taken into account when making the final decision. One of our goals was to compare the system proposals with the decisions taken during the visits, so we could not use old patients’ logbooks because they could contain missing information (e.g. patient’s weight) affecting the associated clinical decision. Only new patients were included in the DIABNET evaluation experiments and the additional information obtained by doctors was also held in the patients’ logbook where patients usually fill in their self-monitoring data.

The process of data registration was as follows: During the visit to the hospital, the diabetologist made a special effort to record in the logbook the weight and other information considered to decide on the therapy modification (e.g. diet and time modifications). At the end of the visit a hard copy of the last logbook pages was kept at the hospital. From the hard copy, the information was entered manually in the DIABNET database to perform the analysis and to obtain the system’s proposals.

2.3 Patients group

DIABNET was evaluated during the follow up of 9 gestational diabetic patients selected at random. Patients were monitored during a period of time ranging from 2 to 9 weeks, with an average of 5 weeks/patient. A total of 313 day’s data were analysed. Patients followed their normal home monitoring protocol of daily blood glucose and ketonuria self-monitoring, with an average of 27 blood glucose samples per patient per week, and two daily ketonuria readings (before breakfast and dinner).

The total number of cases assessed was 50, all of them related to patients visits. Monitoring data for each of the cases varies from 3 to 7 days, with an average of 6 days. Insulin therapy was initiated in 3 patients that followed a basal-bolus insulin program with regular and NPH insulin.

3. RESULTS

The DIABNET system has been evaluated by three diabetologists, specialised in managing patients with gestational diabetes. The evaluator that participated in experiment 1 will be referred to as e1. This experiment allows us to evaluate the degree of similarity between DIABNET and the expert’s reasoning because e1 is the same expert that participates in the creation of the system knowledge bases.

Two evaluators took part in the experiment 2 and they will be referred to as e2 and e3. These two experts did not participate in the development of the system.
Two different methods were used to analyse the results of the experiments: a subjective analysis based on the answers to the questionnaires registering experts’ judgements and an objective analysis based on a quantitative comparison of the system’s and experts’ proposals.

3.1 Subjective analysis

Comparisons performed by human experts are the only way to discover to what degree proposals are equivalent and how effective they are in correcting deviations. This section presents the assessment of DIABNET proposals based on experts’ judgements by means of the subjective questionnaires they filled in during the evaluation experiments. To simplify the results we summarised the five possible responses to the questionnaire items into three sets: “disagreement”, “none” and “agreement”. The most representative item for comparison of a proposal’s appropriateness is number 5 because it shows whether they are finally accepted as valid by the evaluators (even though they were also ranked as not effective, conservative or aggressive). We consider that a proposal is rejected when it is catalogued as “no valid”.

3.1.1 Assessment of DIABNET proposals

Table 1 presents the results of the assessment of DIABNET proposals by e1 in experiment 1. The evaluator catalogued the system proposals as effective in 90% of cases but they are accepted as valid in a lower number of cases (86%). The reasons for rejection were “aggressive and risky” proposals (6%), “conservative” proposals (6%) and “unclear effectivity” (2%). None of the accepted proposals including a therapy adjustment was the same as the modification decided during the patient’s hospital visit but was suitable to correct deviations.

(Insert Table 1 around here)

Evaluators e2 and e3 assessed blindly the DIABNET proposal and the therapy modification decided by the physician during the patient’s visit (experiment 2, step 2). Table 2 shows the distribution of their responses to the questionnaire. The first set of values, in the top lines of the table, refer to the agreements to the sentences and the second set, in the bottom lines of the table, refer to disagreements. Notice that in some cases the addition of value pairs is not 100% because there are responses where evaluators choose the “none” answer. Data show that: 1) evaluators accepted DIABNET proposals in a similar number of cases (e2: 76%, e3: 74%); 2) evaluators accepted clinician’s decisions in a higher number of cases (e2:86%, e3: 92%); 3) evaluators also rejected human proposals (e1: 10% , e2: 2%); 4) evaluators found the system behaviour more “aggressive” and “conservative” than the human one; 5) evaluators found risky proposals coming both from DIABNET and from the patients’ clinician.

(Insert Table 2 around here)

Proposals that can imply risk for the patient have to be considered in a special way. The number of risky proposals should be as low as possible or in fact zero when the system is aimed at providing automatic advises to non expert users (e.g. patients). Table 2 shows that risk is not the main reason for rejections because the total number of the proposals catalogued as risky is much lower than the number of “not valid” proposals. The presence of a reduced number of risky proposals could be accepted in DIABNET because any system decision is supervised later on. The system goal is to help doctors in data analysis by focusing their attention on deviations and giving them one corrective proposal.

It can be observed that evaluators do not interpret the questionnaire items in the same way: e2 found 52% of cases where he/she is not sure whether the proposal will or will not be effective in correcting deviations (no. of answers equal to “none” in the item 1), but he/she accepts as valid a high number of those proposals (92%). This means that even if he/she is not sure of the effectiveness of each proposal, his/her decision would be the same because it is the best according to the information available. This hypothesis is supported by another data presented in next section: evaluator e2 found that the information available was insufficient for data interpretation in a higher number of cases than the evaluator e3.
Another aspect worth mentioning is that the term “conservative” is not always considered negative, because an important number of the “conservative” proposals are accepted as “valid” by the evaluators: e1 accepted as valid 40% of the proposals she catalogued as “conservative”, e2 accepted 60% and e3 accepted 53%. This shows that during the follow up of patients, there are situations where diabetologists also make conservative decisions and wait to see how the patient progresses.

Figure 5 summarises the evaluators’ assessment of the DIABNET proposals (number of experts’ “agreements” to the questionnaire items) by extracting the numerical values presented in Table 2 for evaluators e2 and e3 and adding the assessments of evaluator e1 presented in Table 1. The histogram shows that DIABNET proposals were accepted as valid in a high number of cases (e1:86%, e2:76% and e3:74%) and that it is evaluator e1 who rates the system proposals more positively. One reason for this difference is that evaluator e1 is the same expert that co-operated in the definition of the DIABNET knowledge base, so the system follows e1 preferences for therapy modifications.

Figure 6 shows the coincidences between physicians when rejecting the DIABNET proposals. There are only two cases (4%) where the three evaluators agree in the rejection and both were catalogued as “risky”. In both cases, DIABNET detected the need of insulin and proposed initiation of the insulin therapy but the clinical decision was to wait for more data confirming the evolution in the following visit. In both cases the insulin need was a true fact but it was not present again in the following visits, thus doctors did not start the insulin protocol. Discussing this situation with doctors, they stated that if two gestational diabetic patient present a similar insulin need, diabetologists are more conservative in initiating insulin therapy than in adjusting an existing one. The actual implementation of DIABNET decided to act whenever an insulin need is detected and the existence or absence of an insulin therapy is considered later on to quantify the final insulin doses. DIABNET performance has to be improved by adding some sort of restriction to the system to delay decisions of insulin initiation when insulin needs are detected.

3.1.2 Information characterisation and case difficulty

From all the cases analysed, evaluators found some situations where it was difficult to make any decision from a medical point of view (see item 6 in Table 4). There is a high coincidence rate between cases catalogued as “difficult” and the DIABNET proposals that were rejected (the cases catalogued as “difficult” from the set of rejected proposals are: e1:71%, e2:20%). However, DIABNET also built valid proposals for difficult cases (3 cases for e1 and 6 cases for e2).

All evaluators found cases where the decision of a therapy modification could be taken in a lower number of days (see item 7 in Table 4). These results support the hypothesis of the usefulness of an intelligent system such as DIABNET integrated in Telemedicine to reduce the time taken before initiating a therapeutical action. The idea is to complement the functionality of a telemedicine service that makes data available for diabetologists more frequently than patients’ visits, minimising the doctors’ workload for data interpretation.

The two evaluators of experiment 2 found that in some cases the information given was not sufficient to make a decision (see item 8 in Table 4). In such situations, diabetologists asked patients during their visits to confirm or reject their suspicions of treatment transgressions. In the present studies, DIABNET suffered the
same deficient information as evaluators, but DIABNET can detect such situations and it produces focused
questions to get more information. This feature was not evaluated yet because the system had no access to
patients during the evaluation process. In the future, a deeper analysis should be done to evaluate DIABNET’s
performance in cases with insufficient information.

3.2 Objective analysis

We performed a differential analysis based on a quantitative cross-comparison of DIABNET and experts’
proposals. The objective evaluation complements the subjective analysis as it allows:
- Characterisation of the DIABNET proposals
- Quantification of the performance of DIABNET when detecting the need of a therapy adjustment

It must be noted that experts e2 and e3 were in the same conditions as DIABNET when they built their own
proposals. This means that the most accurate comparisons between DIABNET and experts’ proposals are
those that include evaluators e2 and e3.

3.2.1 Characterisation of proposals

From an objective analysis it is possible to compare different proposals in terms of type and distribution of
therapeutical actions. Figure 4 extracts the information to characterise therapy modification proposals by
measuring the following variables:
- Number of cases where a therapy modification is proposed
- Number of cases where an insulin modification is proposed
- Number of cases where a diet modification is proposed
- Number of cases where proposed modifications affect diet exclusively

The sources of the proposals are: 1) patient’s clinician; 2) DIABNET; 3) evaluator e2 and 4) evaluator e3

(Insert Figure 4 around here)

Figure 4 shows differences between the four sets of proposals, including the proposals from different experts.
DIABNET proposed more therapeutical actions than the experts. The differences between experts decisions
ratify the absence of a unique clinical criterion in diabetes therapy planning. The actual evaluation studies do
not cover the comparison between experts’ proposals but complementary experiments could be performed to
discover to what degree the proposals are equivalent.

3.2.2 DIABNET performance when detecting the need of a therapy modification.

The goal of this analysis is to quantify the performance of DIABNET when detecting the need of a therapy
adjustment. Table 5 shows the number of advices recommending modifications in the four sets of proposals
(patient’s physician, evaluator e2, evaluator e3 and DIABNET) and also shows the number of cases where the
clinician decided on a modification at a patient’s visit and DIABNET, e2 and e3 also decided on a
modification.

DIABNET and evaluators e2 and e3 used the same information to form their proposals, so it is also interesting
to compare the number of cases where they decided on therapeutical actions versus the real therapy
adjustment prescribed during the patient’s visit. Table 5 shows that DIABNET detected the need for a therapy
modification in a higher number of cases than evaluators e2 and e3 when comparing with the patients’
clinician decisions (92%). It can be noticed that DIABNET has a higher coincidence with the clinician than
the two evaluators of the experiment 2 that built their proposals under the same conditions as the system
(DIABNET: 92%, e2: 83%, e3: 75%). The main reason for these differences is that e1 is the expert that
collaborated in the creation of DIABNET so the results show a high degree of similarity between the system
and the expert’ reasoning. We can conclude that DIABNET is capable of a reliable detection of therapy
modification needs and that the system is therefore suitable to perform automatic data analysis for alarm
generation. Working in a telemedicine scenario, DIABNET could perform an automatic processing of patients’ data whenever they are received and it will activate an alarm if a therapy modification is needed. Alarms will be useful to focus doctors’ attention on abnormal data when a huge amount of information has to be processed and also to give automatic feedback to patients.

(Insert Table 5 around here)

4. DISCUSSION

An evaluation methodology for decision aid tools in diabetes therapy planning has been proposed and applied to evaluate DIABNET showing information about its performance when it was working in a real scenario analysing real cases. The evaluation of expert systems in medicine is complex [8] and the absence of a gold standard means that comparisons cannot be done simply by using computational methods but must be performed by human experts in the management of diabetic patients. For this reason, the proposed methodology requires the co-operation of expert diabetologists who express their degree of acceptance by means of a subjective questionnaire.

The best evaluation of a system should perform comparisons with the real modifications associated to patients’ cases, but in this situation it is almost impossible to capture all the information a clinician considers when evaluating the patient’s state during a visit to the hospital, and so, the system would not be in the same conditions as the experts. As a solution to this difficulty, the proposed methodology starts at the process of data acquisition and combines two experiments to allow two types of evaluation to be carried out: 1) against real therapy modifications and 2) against proposals elaborated by experts using the same information as the system. External experts are required to compare the clinician decision against the system proposal for each case. This evaluation methodology assures that evaluators have a deep knowledge about each case when facing the case characterisation process because they are asked to form their own proposals in a previous step. Comparisons of system’ proposals versus clinicians’ proposals are done blindly assuring that evaluators are not provided with information on the origin of each decision.

The results of the DIABNET evaluation are encouraging and help to refine the system and look towards its further clinical use as a decision tool in gestational diabetes integrated in a telemedicine service. The main result of the subjective analysis is that evaluators accepted as valid the DIABNET proposals in a high number of cases (e1:86%, e2:76% and e3:74%). As expected, the external evaluators (e2 and e3) ranked the clinician’s decisions more positively than the DIABNET proposals. Evaluators found that some decisions could be made in a shorter period (e1:25%, e2: 12%, e3: 66.6%). In actual clinical practice it is extremely difficult to reduce the number of days between visits below the values used in these experiments (an average of 6 days), but this situation justifies the use of a Telemedicine service, where patients can be asked to send data as frequently as needed, allowing the system to perform a preliminary patient assessment without increasing the number of visits and improving patient-clinician communications.

The main result of the objective analysis is that DIABNET detects the need for a therapy modification in 92% of the cases. This data shows its appropriateness for automatic alarm generation in a Telemedicine service were patients collect data in an ambulatory scenario and the system informs doctors about the presence of control deviations.

The available information in patients’ cases could be insufficient for a proper decision to be made, not only due to missing data but also because more evidence is needed to ratify the observed deviations. Doctors manage this situation by delaying their decisions until the following visit (in the case of gestational diabetes it will be some days later). In order to characterise the cases used in the evaluation experiments, the questionnaire proposed in this paper also acquires information on the difficulty, completeness and quality of the available cases. The evaluation results show that in some cases there is an absence of relevant information for the interpretation of monitoring data: Evaluators indicated that in some situations the collected information was not enough to make a proper decision (e2:34%, e3:12%). This limitation could be solved by asking patients to register “all” data and daily actions that could affect their blood glucose levels (detailed
intake composition and timing of actions, etc). However, from previous clinical experiences [12], we learned that patients do not comply with a system that requires the registration of a huge amount of variables. This constraint can be faced by a patient-oriented system, capable of optimising the amount of extra information demanded from patients by interacting with them during the process of data acquisition and asking only for new information when needed to justify abnormal measurements.

On the basis of the evaluation results presented in this paper we detected that the ways to increase the number of accepted proposals are: 1) to delay the decision of initiating the insulin therapy; 2) to consider the knowledge about patients’ evolution and patients’ responses to previous therapy modifications and 3) to improve the process of data acquisition. A further step in the evaluation process would be to test the system in clinical routine facing the practical difficulties of installing the information system at the hospital. New evaluation methods should be added to analyse not only the system impact on diabetic patient’s health care but also its impact on the health center organisation and the users’ acceptance of the information system looking at usability.

5. ACKNOWLEDGEMENTS

This work has been partially funded by the Spanish Research Grant CICYT SAF98-0170 “DIABTel”. We would like to give our special thanks to Prof. Dr. A. de Leiva, “Hospital Sant Pau” of Barcelona, Spain, for his support and contribution to this research and to Apolonia García-Patterson and Mercè Albareda from the same center for performing the blind evaluations. We also thank Lifescan (Johnson& Johnson) for their support of this work.

6. REFERENCES


ANNEX A: Example of case presentation

**EXPERIMENT 2: ELABORATION OF THERAPY MODIFICATIONS**

**DOCTOR:**

<table>
<thead>
<tr>
<th>PATIENT:</th>
<th>hc 0</th>
<th>PAC-XXX</th>
</tr>
</thead>
</table>

Starting date: XXXX/XXXXX
weight: 71.9
Active therapy: NIGHT BREAK LUNCH DINNER

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
</table>

**SELF-MONITORING DATA**

<table>
<thead>
<tr>
<th>BLOOD GLUCOSE</th>
<th>before break</th>
<th>after break</th>
<th>before lunch</th>
<th>after lunch</th>
<th>before dinner</th>
<th>after dinner</th>
<th>night</th>
</tr>
</thead>
<tbody>
<tr>
<td>Th.</td>
<td>78</td>
<td>-</td>
<td>80</td>
<td>-</td>
<td>63</td>
<td>113</td>
<td>-</td>
</tr>
<tr>
<td>F.</td>
<td>128</td>
<td>61</td>
<td>-</td>
<td>72</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sat.</td>
<td>78</td>
<td>123</td>
<td>61</td>
<td>115</td>
<td>69</td>
<td>108</td>
<td>-</td>
</tr>
<tr>
<td>Sun.</td>
<td>79</td>
<td>-</td>
<td>64</td>
<td>107</td>
<td>58</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>M.</td>
<td>79</td>
<td>-</td>
<td>65</td>
<td>-</td>
<td>88</td>
<td>84</td>
<td>-</td>
</tr>
<tr>
<td>Tu.</td>
<td>76</td>
<td>-</td>
<td>66</td>
<td>103</td>
<td>77</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>W.</td>
<td>83</td>
<td>134</td>
<td>64</td>
<td>-</td>
<td>81</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>KETONURY (++, +, -)</th>
<th>break.</th>
<th>lunch</th>
<th>snack</th>
<th>dinner</th>
<th>night</th>
</tr>
</thead>
<tbody>
<tr>
<td>Th.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>F.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sat.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sun.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>M.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tu.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>W.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**PLEASE, FILL IN YOUR MODIFICATION PROPOSAL**

**INSULIN:**
NIGHT BREAK LUNCH DINNER

(in increments in units)

**WAITING TIME:**
(in minutes)

**INTAKES:**
(in increments in grs of carboh.)

**PLEASE, CLASSIFY THE FOLLOWING STATEMENTS:**

<table>
<thead>
<tr>
<th>completely disagree</th>
<th>disagree</th>
<th>none</th>
<th>agree</th>
<th>completely agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The difficulty of data interpretation is high</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. There is enough information to make a decision</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. The decision on therapy modification could be made in a shorter period. How many days?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Keys to Figures:

Figure 1: Description of the evaluation experiment number 1 where the evaluator was the clinician following the patients.

Figure 2: Description of the evaluation experiment number 2, step 1

Figure 3: Description of the evaluation experiment number 2, step 2

Figure 4: Type and frequency of therapy modifications proposed by the patient’s clinician, DIABNET system and evaluators e2 and e3

Figure 5: Histogram showing the evaluators’ assessment of DIABNET proposals summarising answers to the questionnaire equal to “completely agree” and “agree”.

Figure 6: DIABNET proposals rejected by the evaluators. The reasons of disagreement between experts and DIABNET (A, B, C and D) are presented in Table 3.
Keys to Tables:

Table 1: Assessment of DIABNET proposals done by evaluator e1

Table 2: Blind assessment of DIABNET proposals and clinician’s decisions done by external evaluators e2 and e3

Table 3: Reasons of disagreement between experts and DIABNET (CHs: carbohydrates). The frequency and coincidences in proposal types (A, B, C and D) are presented in Figure 6.

Table 4: Subjective case characterisation

Table 5: Performance of DIABNET system versus patients’ clinician, e2 and e3 when detecting the need of a therapy modification
Figure 1: Description of the evaluation experiment number 1 where the evaluator was the clinician following the patients.
Figure 2: Description of the evaluation experiment number 2, step 1
Figure 3: Description of the evaluation experiment number 2, step 2
Figure 4: Type and frequency of therapy modifications proposed by the patient's clinician, DIABNET system and evaluators e2 and e3.
Assessment of DIABNET proposals

Figure 5: Histogram showing the evaluators’ assessment of DIABNET proposals summarising answers to the questionnaire equal to “completely agree” and “agree”.

20
Figure 6: DIABNET proposals rejected by the evaluators. The reasons of disagreement between experts and DIABNET (A, B, C and D) are presented in Table 3.
Table 1: Assessment of DIABNET proposals done by evaluator e1

<table>
<thead>
<tr>
<th>Questionnaire items</th>
<th>agree</th>
<th>none</th>
<th>disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>The proposal is <strong>effective</strong> to correct deviations</td>
<td>90%</td>
<td>2%</td>
<td>8%</td>
</tr>
<tr>
<td>The proposal implies <strong>risk</strong> for the patient</td>
<td>6%</td>
<td>0%</td>
<td>94%</td>
</tr>
<tr>
<td>The proposal is <strong>aggressive</strong></td>
<td>6%</td>
<td>4%</td>
<td>90%</td>
</tr>
<tr>
<td>The proposal is <strong>conservative</strong></td>
<td>10%</td>
<td>2%</td>
<td>88%</td>
</tr>
<tr>
<td>You accept this proposal as <strong>valid</strong></td>
<td>86%</td>
<td>0%</td>
<td>14%</td>
</tr>
</tbody>
</table>
Table 2: Blind assessment of DIABNET proposals and clinician’s decisions done by external evaluators e2 and e3

<table>
<thead>
<tr>
<th>Questionnaire items</th>
<th>Evaluator e2</th>
<th>Evaluator e3</th>
</tr>
</thead>
<tbody>
<tr>
<td>The proposal is <strong>effective</strong> to correct deviations</td>
<td>40%</td>
<td>38%</td>
</tr>
<tr>
<td>The proposal implies <strong>risk</strong> for the patient</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>The proposal is <strong>aggressive</strong></td>
<td>8%</td>
<td>0%</td>
</tr>
<tr>
<td>The proposal is <strong>conservative</strong></td>
<td>30%</td>
<td>18%</td>
</tr>
<tr>
<td>You accept this proposal as <strong>valid</strong></td>
<td>76%</td>
<td>86%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>DIABNET proposals</th>
<th>Clinician’s decisions</th>
<th>DIABNET proposals</th>
<th>Clinician’s decisions</th>
</tr>
</thead>
<tbody>
<tr>
<td>The proposal is NOT effective to correct deviations</td>
<td>8%</td>
<td>2%</td>
<td>24%</td>
<td>2%</td>
</tr>
<tr>
<td>The proposal implies NO risk for the patient</td>
<td>96%</td>
<td>96%</td>
<td>90%</td>
<td>98%</td>
</tr>
<tr>
<td>The proposal is NOT aggressive</td>
<td>92%</td>
<td>98%</td>
<td>92%</td>
<td>98%</td>
</tr>
<tr>
<td>The proposal is NOT conservative</td>
<td>68%</td>
<td>60%</td>
<td>70%</td>
<td>80%</td>
</tr>
<tr>
<td>You do NOT accept this proposal as <strong>valid</strong></td>
<td>20%</td>
<td>10%</td>
<td>26%</td>
<td>2%</td>
</tr>
</tbody>
</table>
Table 3: Reasons of disagreement between experts and DIABNET (CHs: carbohydrates). The frequency and coincidences in proposal types (A, B, C and D) are presented in Figure 6.

<table>
<thead>
<tr>
<th>Proposal type</th>
<th>System proposal</th>
<th>Clinician’s decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Starting of insulin administration with NPH</td>
<td>No changes</td>
</tr>
<tr>
<td>B</td>
<td>Decrease of 1 unit in one insulin dose</td>
<td>No changes</td>
</tr>
<tr>
<td>C</td>
<td>Increase CHs in lunch and afternoon snack</td>
<td>Increase CHs only in afternoon snack</td>
</tr>
<tr>
<td>D</td>
<td>Increase in insulin</td>
<td>Higher increase in insulin</td>
</tr>
</tbody>
</table>
Table 4: Subjective case characterisation

<table>
<thead>
<tr>
<th>Questionnaire</th>
<th>e1</th>
<th>e2</th>
<th>e3</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>The <strong>difficulty</strong> of data interpretation is high</td>
<td>16%</td>
<td>16%</td>
</tr>
<tr>
<td>7</td>
<td>The decision could be taken in a <strong>shorter</strong> period</td>
<td>25%</td>
<td>12%</td>
</tr>
<tr>
<td>8</td>
<td>There is <strong>enough information</strong> to make a decision</td>
<td>-</td>
<td>66%</td>
</tr>
</tbody>
</table>
Table 5: Performance of DIABNET system versus patients’ clinician, e2 and e3 when detecting the need of a therapy modification

<table>
<thead>
<tr>
<th>Origin of the advise</th>
<th>No. of advises recommending modifications (n=50)</th>
<th>Comparison of advises proposing modifications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>DIABNET e2 e3</td>
</tr>
<tr>
<td>Clinician</td>
<td>12</td>
<td>92% (11/12) 83% (10/12) 75% (9/12)</td>
</tr>
<tr>
<td>Evaluator e2</td>
<td>17</td>
<td>65% (11/17)</td>
</tr>
<tr>
<td>Evaluator e3</td>
<td>15</td>
<td>100% (15/15)</td>
</tr>
<tr>
<td>DIABNET</td>
<td>18</td>
<td>-</td>
</tr>
</tbody>
</table>