Evaluating a Multi Modal Reasoning system in diabetes care

Stefania Montani\textsuperscript{1}, Riccardo Bellazzi\textsuperscript{1}, Luigi Portinale\textsuperscript{2}, and Mario Stefanelli\textsuperscript{1}

\textsuperscript{1} Dipartimento di Informatica e Sistemistica
Universit\`a di Pavia, Pavia (Italy)
\textsuperscript{2} Dipartimento di Scienze e Tecnologie Avanzate,
Universit\`a del Piemonte Orientale “A. Avogadro”, Alessandria, (Italy)

Abstract. In the context of Insulin Dependent Diabetes Mellitus care, we developed a decision support system that relies on a tight integration of Case Based Reasoning and Rule Based Reasoning methodologies. In this paper, we aim at presenting the evaluation strategy we have defined to test the system accuracy, safety and reliability, and the first results obtained both on simulated and on real patients data. Reliability was positively judged by a group of expert diabetologists; an increase in the performances of the system is foreseen as new knowledge will be acquired, through its usage in clinical practice.

1 Introduction

An effective management and exploitation of knowledge is a key requisite in the medical domain, where the introduction of Hospital Information Systems (HIS) into clinical practice has led to the necessity of keeping, distributing and reusing a large quantity of data. A proper Knowledge Management (KM) approach is particularly needed in the field of chronic diseases, among which Insulin Dependent Diabetes Mellitus (IDDM) care. IDDM patients suffer from a reduced functionality of the pancreatic beta cells, and need to inject themselves exogenous insulin 3 to 4 times a day to regulate blood glucose metabolism. Such an intensive therapy may lead to hypoglicemic episodes: Blood Glucose Level (BGL) has therefore to be frequently tested and logged. Because IDDM is a life-long condition, the amount of self-monitoring data is huge, and needs to be correctly interpreted by physicians. In particular, patients are normally visited every 2-4 months, to assess their health status, and to eventually revise therapy. The DCCT study \cite{1} has clearly shown that the definition and realization of an appropriate individual therapy, derived from the current metabolic behavior and customized on the single patient’s needs, is the key to an effective diabetes care. Therefore it would be extremely useful to increase the contacts with the diabetologist, and to remind her/him of past situations similar to the current one, belonging to the same patient or to a different person, in order to provide a context into which the present problems can be interpreted and faced. The Case Based Reasoning (CBR) methodology seems to be well suited for performing this task.
As a matter of fact, CBR promises to be a valuable way for managing *implicit* knowledge, i.e. past cases, individual expertise of workers and organizational practices [2,3]. In the IDDM domain, viewing a periodical visit as a case, a large case library can be made available. Keeping track of the problem/solution (i.e. metabolic behavior/therapy) patterns over time, it is possible to maintain expert physicians’ know-how, in order to provide a good quality of care even if they move or retire. Moreover, the use of CBR techniques allows comparisons among different therapies and different physician approaches. Finally, CBR can be useful in providing a significant start point for extremely difficult cases.

Nevertheless, implicit knowledge is not the only useful knowledge type in the IDDM context: well-established and formalized domain knowledge is available as well, and it can be represented through different Artificial Intelligence (AI) formalisms, such as rules. Rule Based Reasoning (RBR), in fact, can be very helpful for supporting *explicit* knowledge exploitation.

To take advantage of these heterogeneous knowledge sources, we have developed a system in which CBR and RBR are integrated in a very tight way [4,5], to provide physicians with a mean for exploiting the available know-how, and with a decision support tool in therapy planning. Our solution aims at overcoming the two paradigms limitations. CBR is used to specialize and dynamically adapt the rules on the basis of the patient’s characteristics and of the accumulated experience, thus avoiding the qualification problem [6]. On the other hand, if a particular patient condition is not sufficiently covered by cases, the use of rules may be exploited to learn a suitable therapy. The new case is then kept, in order to improve the competence of the Case Based component.

This Multi Modal Reasoning (MMR) system has been implemented within the T-IDDM project [7], and can be accessed through its web-based environment. The system is currently used at the Pediatric Department of Policlínico S. Matteo Hospital in Pavia.

Implementation details are described elsewhere [4,5], and are briefly recalled in section 2. The methodological issues addressed to design the system evaluation, together with the study results, are presented in section 3.

## 2 System implementation

In defining the internal structure of cases, we adopted a classical approach, in which a case is described as a triple composed by (i) a set of feature/value pairs, (ii) a solution and (iii) an outcome: in the IDDM domain the case features are the data collected during the visit, the solution is the therapeutic protocol assigned after the features evaluation, and the outcome of such therapy is given by the number of hypoglycemic episodes and by the value of glycated hemoglobin at the following visit.

A case library of past visits has been built, and is automatically upgraded every time a patient undergoes a periodical examination. The case library structure mirrors a taxonomy of prototypical classes, that express typical problems in the age of infancy and puberty [8,9] (see figure 1). Each case belongs to a
leaf in the taxonomy tree. Case Based retrieval is therefore implemented in two steps: classification of the input case, and retrieval of past cases belonging to the identified class.

Classification and retrieval perform what is called the “situation assessment” step by using CBR terminology, and clarify the relevant context to work with, by making explicit the implicit knowledge embedded in the cases stored in the memory.

Classification relies on a Naive Bayes strategy, a method that assumes conditional independence among the features given a certain class, but that is known to be robust in a variety of situations [10, 11], even in the presence of conditional dependencies. Prior probabilities have been derived from medical knowledge, while posterior probabilities have been learnt from the available case library (147 real cases), using a standard Bayesian updating approach [12]. The most similar cases belonging to the most probable class(es) found in the classification step are retrieved resorting to Nearest Neighbor techniques [8, 9]. Distances are calculated exploiting classical metrics, able to treat numeric and symbolic variables, and to cope with the problem of missing data [13]. When dealing with a large case base, the system performs a non exhaustive search procedure by applying an anytime algorithm called Pivoting-Based Retrieval (PBR) [14]. The idea is to restrict the case search space to the relevant part of the library, by means of the Bayesian classification step, and then to further reduce search through bound computation on the distance of stored cases with respect to the target one.

As anticipated in the previous section, the established domain knowledge is maintained in a set of production rule classes; RBR provides the possibility of
identifying the eventual metabolic alterations experienced by the patient, and of generating a therapeutic suggestion able to cope with them.

In more detail, RBR performs a sequence of reasoning tasks, each one obtained by firing the rules in the following rule classes:

- **problem identification rules.** In order to temporally contextualize the large amount of time-stamped data collected by patients during home monitoring, we have subdivided the day into seven non-overlapping time-slices, centered on the injection and/or meal times. The raw data can then be abstracted through a Temporal Abstractions (TA) technique [15]; in particular, STATE abstractions (e.g. low, normal, high values) are extracted and aggregated into intervals called episodes. From the most relevant episodes, it is possible to derive the BGL modal day [15], an indicator able to summarize the average response of the patient to a certain therapy. When the frequency (called minimum probability) of a certain BGL abstraction is higher than the \( \alpha \) threshold, and when the number of missing data (called ignorance) is sufficiently small to rely on such information (i.e. it is smaller than the \( \beta \) threshold), a problem is detected. For example, the following rule detects a hypoglycemia problem in a generic time-slice \( Y \) using the information contained in the relative modal day component \( X \):

\[
\text{IF } X \text{ IS A BGL-MODAL-DAY-COMPONENT} \\
\text{AND THE TIME-SLICE OF } X \text{ IS } Y \\
\text{AND THE BGL-LEVEL OF } X \text{ IS LOW} \\
\text{AND THE MINIMUM-PROBABILITY OF } X \geq \alpha \\
\text{AND THE IGNORANCE OF } X \leq \beta \\
\text{THEN GENERATE-PROBLEM HYPOGLYCEMIA AT } Y
\]

where the \( \alpha \) and \( \beta \) default values, derived from medical knowledge, are equal to 0.3 and 0.8 respectively;

- **suggestion generation rules.** For each detected problem, a set of alternative suggestions on insulin administration, to be applied to the therapeutic protocol, is produced;

- **suggestion selection rules.** The most suitable and effective suggestions are chosen among all the available ones;

- **protocol revision rules.** Finally, the current protocol is adjusted on the basis of the selected suggestions, and is listed with other library protocols that would fit the current patient’s condition.

In the MMR methodology, the problem identification and suggestion generation rule classes are affected by the integration procedure. Integration takes place within the problem solving cycle, in the following way:

- the input case is classified;

- the most probable class information is used to specialize the problem identification rules parameters. For example, in the hypoglycemia detection rule described above, when dealing with patients suffering from anorexia, \( \alpha \) has to be set to 0.2, and \( \beta \) to 1 [4];

- the retrieval step is performed;
some simple statistics are calculated on the retrieved cases, to set the suggestion generation rules parameters (i.e. the number of insulin injections in a day, the variation in the daily insulin requirement, and the variation of a single insulin dose);

the RBR cycle is hence normally completed, by firing the suggestion selection and the protocol revision rules.

As the case library grows, the MMR tool is able to perform decision support at various levels of complexity: first, when the stored information is of poor quality (i.e. it presents significant gaps in the knowledge space domain), RBR can be applied to propose a therapy, without resorting to CBR. The new case is then kept in the memory in order to learn from past experience a possible way for filling a competence gap. As far as new information becomes available in the case library, making its competence grow, retrieval results become more reliable. Rule parameters are then tuned on the basis of Case Based retrieval results; the therapeutic advice is dynamically adapted to the proper interpretation context and to the single patient’s needs. In other words, general rules are tailored on specific patients, by refining the suggested therapy in an appropriate way.

3 System evaluation

Before implementing the integrated methodology into a real clinical setting, it has been necessary to evaluate some key factors of the system, i.e. the CBR classifier accuracy, the MMR suggestions safety on real and simulated cases, and the MMR suggestions reliability, through a comparison with experts opinions.

As previously noted, the usefulness of retrieval results is strongly influenced by the case library content, because the presence of competence gaps may lead to wrong or unfit rule specialization. RBR can hence be viewed as an application of MMR, when the implicit information is of poor quality. Therefore, to take into account the key role of the case library content, in our evaluation methodology we compared MMR suggestions with RBR ones.

3.1 Classification accuracy

The Bayesian classifier devoted to the context identification task has been tested through a leave-one-out cross validation technique [9], applied to a case library composed by 147 real cases, extracted from the histories of 29 pediatric patients. Moreover, we asked a diabetologist other than those who provided the knowledge to judge how the system was able to classify 14 new real patients’ cases.

The classifier performed a correct classification on the 83% of the whole case base. In the 98% of the examples the correct class was one of the two most probable classes [6]. The expert physician who provided a judgment on the classifier soundness applied to the 14 real patients cases confirmed the overall positive outcome.
3.2 MMR safety

RBR and MMR suggestions have been compared in two fashions:
- *retrospectively* by running the two systems independently on 37 cases, belonging to 6 real pediatric patients, and by providing a qualitative judgment on the soundness of the answers;
- *prospectively* by running the two systems on simulated cases. Through a realistic diabetic patient simulator [16], we were able to reproduce the behavior of pediatric patients experiencing the problems described by the taxonomy. The systems performances were judged in terms of number of therapy adjustments required to reach normoglycemia, simulating new BGL data after each revision. Safety results were satisfactory in both the evaluation procedures:
- *retrospective evaluation*: in 55% of the 37 real cases MMR suggested more substantial changes to the insulin administration in comparison to RBR. In this way, it was more prompt to react against hypoglycemia or hyperglycemia, depending on what was the most relevant problem at hand. In an additional 40% MMR and RBR outcomes were comparable, and suitable to the input situation. In the remaining 5%, the CBR tool succeeded in retrieving only one past case, unfit for the examples under examination. This is the typical situation due to the presence of a competence gap: a larger case library could increase reliability (see section 3.4). Nevertheless, as previously noted, every time the retrieval information is poor our system applies RBR without integration;
- *prospective evaluation*: all the tests on simulated data confirmed the ability of MMR to suggest sharper changes in insulin therapy. A lower number of adjustment was hence required to stabilize the sample patient when adopting MMR in comparison to RBR [4, 5, 17].

3.3 MMR reliability

In several clinical domains, especially when dealing with therapy definition, experts may disagree on how to proceed in a certain situation. Moreover, it may be unrealistic to determine a gold standard to be followed through a sufficiently general clinical trial [18]. Diabetes therapy belongs to this category. In absence of a therapeutic behavior that is right in an absolute sense, a decision support system correctness can not be assessed. Only reliability may be addressed, making a comparison between the system's advice and the opinion of a group of experts.

Our methodology for reliability evaluation consisted of the following two steps:
1 - *therapy advice generation*: two independent specialists, not familiar with our tool, were asked to formulate a therapy for 30 of the 37 real patient cases already used for safety assessment. Their answers were indicated as A and B. MMR and RBR therapeutic suggestions were called C and D respectively;
2 - *peer review*: two other experts (E1 and E2), again unfamiliar with the tool, were asked to make a fully-crossed blind review of the A, B, C and D answers.
Each therapeutic advice, produced at step 1, was assigned one of the following scores: (1) ideal; (2) acceptable; (3) weakly acceptable; (4) unacceptable.

A total of 119 judgments were formulated by each peer reviewer (A provided only 29 answers, while B, C and D provided 30 each).

The results have been expressed as 4x4 tables, one for every expert, whose columns report the first level therapy advisors (A, B, C, D) and whose rows report the four qualitative judgments (1, 2, 3, 4) completed at step 2.

Two tests were selected to assess whether there were dependencies between therapy advisors and the judgment they obtained from each of the peer reviewers. First, $X^2$ calculated as in (1) [19], was compared with $X^2$ distribution with 9 degrees of freedom.

$$X^2 = \sum_i \sum_j \frac{f_{ij} - e_{ij}}{e_{ij}}$$  (1)

where

$$e_{ij} = \frac{f_{i0} f_{0j}}{f_{00}}$$

In formula (1), $f_{i0}$ is the sum of all the elements in row i, in a 1xJ table; $f_{0j}$ is the sum of the elements in column j, and $f_{00}$ is the sum of all the cells.

Additionally, we applied Goodman and Kruskal’s $\tau$ measures [19]. $\tau_b$ in (2) involves a comparison of the following two situations: we are asked to guess what is the score of a peer review judgment (i) without any further information, or (ii) knowing whether the author of the therapeutic advice at step 1 was A, B, C or D. If the two answers are unrelated, knowing the identity of the first level expert (physician or system) does not provide any additional indication on the score. Otherwise, there is an association between the categories on columns and on rows in the table. $\tau_a$ quantifies the same concept, switching the roles of scores and first level experts. For both the statistics, 0 means total independence, while 1 stays for a strong association.

$$\tau_b = \frac{f_{00} \sum_i \sum_j (\frac{f_{ij} - e_{ij}}{e_{ij}})^2}{f_{00}^2 - \sum_i \sum_j \frac{e_{ij}^2}{f_{0j}}}$$  (2)

The results of the experts blind review on the 30 real patients cases are reported in tables 1 and 2. Table 3 is a summary table, obtained by summing tables 1 and 2 cell by cell.

From the outcome of the statistical tests $X^2$, $\tau_a$ and $\tau_b$, we can conclude that there is no evidence of a dependency between rows and columns: knowing the identity of the therapy advisor does not provide preliminary information on the score she/he obtained by peer reviewers, and vice-versa. In fact, the hypothesis
Table 1. Experts E1’s judgments. $X^2 = 5.11, \tau_a = 0.0148, \tau_b = 0.0143$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>14</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>3</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 2. Experts E2’s judgments. $X^2 = 14.56, \tau_a = 0.034, \tau_b = 0.041$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>6</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>8</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

of dependence would be acceptable only with $X^2 > 14.68$ (never verified in the three tables). Moreover, $\tau_a$ and $\tau_b$ values are very close to 0.

On a total of 238 peer review answers, A obtained 33 satisfactory judgments (i.e. scored 1 or 2), B and D obtained 32, and C 30. 65/118 answers were positively scored for the physicians (A and B), and 62/120 for the systems (C and D). Although more data should probably be collected before formulating a conclusion, these not significant differences are an encouraging premise towards the statement that the RBR and the MMR systems are able to perform at an expert level. On the other hand, no significant benefit in using MMR in comparison to RBR can be demonstrated at this stage.

The number of agreements between E1 and E2 on judging the first level advisors were the following: (A) 6/29; (B) 10/30; (C) 11/30; (D) 11/30 (Mean=0.278; SD=0.07). Therefore, a lack of consensus was particularly present when evaluating colleagues; systems evaluation came out to be more uniform.

On one patient cases, both RBR and MMR were rated unsatisfactorily (with a mean score>3), with RBR considered almost unacceptable (average score=3.33) by both experts, and MMR evaluated slightly better (average score=3.25). With such patient, the correct strategy would be a complete revision of the therapy structure. The RBR system, having been built as a very conservative system [4, 20], never proposes strong changes in insulin distribution and doses. On the other hand MMR, meant to overcome RBR limitations by suggesting more substantial changes, did not come out with a revision in the required direction; as a matter of fact, such patient’s cases belonged to a competence gap region, and retrieval did not allow a valuable tuning of rule parameters. In that situation the human (A and B) advice proved to be the best suited.
Table 3. Sum of the E1 and E2 experts’ judgments. \( X^2 = 5.68, \tau_a = 0.006, \tau_b = 0.07 \)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>22</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>9</td>
<td>11</td>
</tr>
</tbody>
</table>

3.4 How to improve MMR reliability

As observed above, although MMR can be considered sufficiently safe and reliable, it does not provide a significant benefit in comparison to RBR. MMR limitation seems to be linked to the presence of competence gaps. Our guess is that MMR would largely benefit from a richer case library, a goal that would be automatically achieved by introducing the system usage into clinical practice.

Such prediction has been confirmed through a simulation study, in which we queried the therapy suggestion for a sample case, providing the MMR system with a case library whose content was progressively enlarged. We started by applying RBR (i.e. MMR when the case library is empty). Then the number of cases was increased in 5 additional steps. Each time the quality of MMR decision support was tested by applying the prescribed therapy, and by simulating 21 days of Blood Glucose data on its basis.

We verified an enhancement in the simulated patient’s metabolic condition as the case library competence augmented (see figure 2). We resorted to the M index [21] (see formula (3)), that calculates the difference on a log scale between each measured BGL (\( BGL_i \)) and a reference value (\( BGL_0 \)).

\[
M = 10^7 \sum_i (\log \frac{BGL_i}{BGL_0})^3
\]  

4 Conclusions

Through a tight integration between CBR and RBR, we have been able to define a suitable methodology for supporting therapy planning in the context of IDDM patients management. RBR is devoted to provide information about established domain knowledge; on the other hand, CBR is able to manage and upgrade the implicit knowledge repository, during the routine clinical activities.

This methodology can be viewed in a larger KM perspective: we have been able to define an anytime continuously learning system, that exploits a naturally growing knowledge source (i.e. the case library, upgraded at each patient visit), together with formalized information. When a new case is collected, it is stored in the case library, without requiring an additional workload to physicians. At retrieval time learning takes place, as the system output is enriched
**Fig. 2.** Polynomial fitting of the M index values vs. the case library enlargement. The first M index value was calculated on BGL data simulated by applying the initial therapy. Therapy was then revised resorting to RBR (point 2), and to MMR in 5 consecutive steps (points 3 to 7), while the library competence grew. After each revision, 21 days of Blood Glucose measurements were simulated. According to the M index, the patient’s metabolic condition was improved as the case library was enlarged.

by the presence of additional examples. Retrieval provides information on physicians’ expertise, and on their approach towards new therapeutic solutions (e.g. the use of new insulin types, such as Lyspro), thus enriching the Health Care Organization domain knowledge, without requiring a formalization of the recently acquired information within a structured knowledge base. When the case library is sufficiently competent, rule parameters are tuned on the basis of Case Based retrieval to provide decision support.

The MMR system can hence be integrated in a more general KM cycle, described in figure 3, in which the physician plays the active role of defining a new therapy, and of assessing it, relying on all the available types of information (both implicit and explicit). Therapy planning is a knowledge creation activity: the new therapy would normally be stored in the case library as a piece of implicit information. On the other hand, through this KM cycle, decision making gives birth to explicit knowledge, already analyzed in the light of domain know-how, past experience and patient’s features.

This process of continuous learning, and of interoperability between the user and the system, seems to be particularly suited for the IDDM management context, in which a standard in therapy planning cannot be defined, and a lack of inter-expert consensus exists (see section 3.3).

By applying a formal evaluation methodology, we tested the classifier accuracy and the system safety and reliability, prerequisites for its introduction
into clinical practice. The system came out to be reliable, and comparable to physicians’ performances.

The MMR functionality, fully integrated in the T-IDDM project web-based environment, has been made available at the pediatric department of Policlìnico S. Matteo Hospital in Pavia. By providing decision support capabilities within a larger KM perspective, instead of building a stand-alone tool, we believe our system could be really deployed in clinical practice, overcoming the well known weaknesses shown by the majority of decision support systems developed in medicine, due to usability, accessibility and organizational factors [22, 23].

To evaluate the effectiveness of this paradigm shift, we will have to verify the effect on patients metabolic outcomes, and to address clinical impact and usability issues. The first task will be fulfilled by comparing glycosylated hemoglobin and daily insulin requirement before and after the introduction of the methodology. To assess usability and clinical impact, we plan to measure the intensity of the information flows among health care providers, the number of visits pre and post usage, and the time/costs gained.

References