Transactions on
Case-Based Reasoning
Vol.2, No 1 (2008) 15 - 28
© ISSN: 1867-366X (Journal),
ISBN: 978-3-940501-08-0,
IBaI Publishing ISSN 1864-9734

www.ibai-publishing.org

Explaining Exceptional Cases Focusing on Solving the Missing Data Problem

Olga Vorobieva[1+2], Rainer Schmidt[1]

[1] Institute for Medical Informatics and Biometry, University of Rostock, Germany
rainer.schmidt@medizin.uni-rostock.de
[2] Sechenov Institute of Evolutionary Physiology and Biochemistry, St. Petersburg, Russia

Abstract. In medical practice and in knowledge-based systems too, it is necessary to consider exceptions and to deal with them appropriately. In this paper, a system is presented, which helps to explain cases that do not fit to a theoretical hypothesis. It is proposed to combine Case-Based Reasoning with a statistical model, where Case-Based Reasoning is used to explain the exceptional cases. Additionally, a method to partly solve the missing data problem was developed. This method combines general restoration techniques with domain dependent formulas provided by an expert. For the latter technique, Case-Based Reasoning is applied again.

1 Introduction

In medicine many exceptions occur. In medical practice and in knowledge-based systems too, these exceptions have to be considered and have to be dealt with appropriately. In our previous work [1], it is demonstrated how a dialogue-oriented Case-Based Reasoning (CBR) system can help in situations where a theoretically approved medical therapy does not produce the desired and usually expected results.

We have developed a system called ISOR, which helps doctors to explain exceptional cases in medical studies and in research. ISOR is a conversational case-based system. Conversational CBR systems have become rather popular in the recent years [2]. ISOR deals with situations where neither a well-developed theory nor reliable knowledge nor a proper case base is available. So, instead of theoretical knowledge and intelligent experience, just a theoretical hypothesis and a set of measurements are given. In such situations the usual question is, how do measured data fit to a theoretical hypothesis. To statistically confirm a hypothesis it is necessary that the majority of cases fit the hypothesis. Mathematical statistics determines the exact quantity of necessary confirmation [3]. However, usually a few cases do not satisfy the hypothesis. These cases shall be examined to find out why they do not satisfy the hypothesis. ISOR offers a dialogue to guide the search for possible reasons in all components of the data system. The exceptional cases belong to the case base. This approach is justified by a certain mistrust of statistical models by doctors, because modelling results are usually unspecific and “average oriented” [4], which means a lack of attention to individual "imperceptible" features of specific patients.

The usual Case-Based Reasoning assumption is that a case base with complete solutions is available (e.g. [5]), whereas this approach starts in a situation where such a case base is not available but has to be set up incrementally.

So, Case-Based Reasoning is combined with a model, in this specific situation with a statistical one. The idea to combine CBR with other methods is not new. Care-Partner, for example, resorts to a multi-modal reasoning framework for the co-operation of CBR and rule-based reasoning (RBR) [6]. Montani [7] rather uses CBR to provide evidences for a hybrid system in the domain of diabetes. Another way of combining hybrid rule bases with CBR is discussed by Prentzas and Hatzilgeroudis [8]. The combination of CBR and model-based reasoning is discussed in [9]. Statistical methods are used within CBR mainly for retrieval and retention [10]. Arshadi and Jurisica [11] propose a method that combines CBR with statistical methods like clustering and logistic regression.

The first application of ISOR is on hemodialysis and fitness. Unfortunately, the data contains many missing data, which makes the process of finding explanations for exceptional cases rather difficult. So, we decided to firstly attempt to solve the missing data problem. This is done by partly applying CBR again.

1.1 Dialysis and Fitness

Hemodialysis means stress for a patient’s organism and has significant adverse effects. Fitness is the most available and a relative cheap way of support. It is meant to improve a physiological condition of a patient and to compensate negative dialysis effects. One of the intended goals of this research is to convince the patients of the positive effects of fitness and to encourage them to make efforts and to go in actively for sports. This is important because dialysis patients usually feel sick, they are physically weak, and they do not want any additional physical load [12].

At the University clinic in St. Petersburg, a specially developed complex of physiotherapy exercises including simulators, walking, swimming and so on is offered to all dialysis patients but only some of them actively participate, whereas some others participate but are not really active. The purpose of this fitness offer is to improve the physical conditions of the patients and to increase the quality of their lives.

The theoretical hypothesis is that actively participating in the fitness program improves the physical condition of dialysis patients. Instead of reliable theoretical knowledge and intelligent experience, just this theoretical hypothesis and a set of measurements are given. In such situations the usual question is, how do measured data fit to theoretical hypotheses. To statistically confirm a hypothesis it is necessary, that the majority of cases fit the hypothesis. Mathematical statistics determines the exact quantity of necessary confirmation [3]. However, usually a few cases do not satisfy the hypothesis. They should be examined to find out why they do not satisfy the hypothesis. ISOR offers a dialogue to guide the search for possible reasons in all components of its data system.

2 Incremental Development of an Explanation Model for Exceptional Dialysis Patients

For each patient a set of physiological parameters is measured. These parameters contain information about burned calories, his oxygen uptake, his oxygen pulse (volume of oxygen consumption per heartbeat) and others. There are also biochemical parameters like haemoglobin and other laboratory measurements. More than 100 parameters were planned for every patient. But not all of them were really measured. Parameters are supposed to be measured four times during the first year of participating in the fitness program. There is an initial measurement followed by a next one after three months, then after six months and finally after a year. Unfortunately, since some measurements did not happen, many data are missing. Therefore the records of the patients often contain different sets of measured parameters.

It is necessary to note that parameter values of dialysis patients essentially differ from those of non-dialysis patients, especially of healthy people, because dialysis interferes with the natural, physiological processes in an organism. In fact, for dialysis patients all physiological processes behave abnormally. Therefore, the correlation between parameters differs too.

For statistics, this means difficulties in applying statistical methods based on correlation and it limits the usage of a knowledge base developed for normal people. Non-homogeneity of observed data, many missing data, many parameters for a relatively small sample size, all this makes the data set practically impossible for usual statistical analysis.

Since the data set is incomplete, additional information has to be found in other available data sources. These are the already existent individual base, the sequentially created case base, and the medical experts as a special source of information.

2.1 Setting up a Model

For the question "Does special fitness improve the physiological condition of dialysis patients?" physical conditions of active and non-active patients have to be compared. Patients are divided into two groups, depending on their activity, active patients and non-active ones. According to the assumption, active patients should feel better after some months of fitness, whereas non-active ones should feel rather worse. The meaning of “feeling better” and “feeling worse” has to be defined in this context. Therefor, a medical expert selects these factors:

- F1: O2PT - Oxygen pulse by training
- F2: MUO2T - Maximal Uptake of Oxygen by training
- F3: WorkJ – performed Work (Joules) during control training

Subsequently the “research time period” has to be determined. Initially, this period was planned to be twelve months, but after a while the patients tend to give up the fitness program. This means, the longer the time period, the more data are missing. Therefore, a compromise between time period and sample size had to be made. A period of six months was chosen.

The next question is whether the model shall be quantitative or qualitative? The observed data are mostly quantitative measurements. The selected factors are also of quantitative nature. On the other side, the goal of this research is to find out whether physical training improves or worsens the physical condition of dialysis patients. Each patient has to be compared with his own situation some months ago, namely just before the start of the fitness program. The success shall not be measured in absolute values, because the health statuses of patients are very different. Thus, even a modest improvement for one patient may be as important as a great improvement of another. Therefore, we simply classify the development in two categories: “better” and “worse”. Since the usual tendency for dialysis patients is to worsen in time, those few patients where no changes could be observed are added to the category ”better”.

The three main factors are supposed to describe the changes of the physical conditions of the patients. The changes are assessed depending on the number of improved factors:

- Weak version of the model: at least one factor has improved
- Medium version of the model: at least two factors have improved
- Strong version of the model: all three factors have improved

The final step means to define the type of model. Popular statistical programs offer a large variety of statistical models. Some of them deal with categorical data. The easiest one is a 2x2 frequency table. The “better/ worse” concept fits this simple model very well. So the 2x2 frequency table is accepted. The results are presented in Table 1.

According to the assumption after six months of active fitness the conditions of the patients should be better.

Statistical analysis shows a significant dependence between the patients activity and improvement of their physical condition. Unfortunately, the most popular Pearson Chi-square test is not applicable here because of the small values “2” and “3” in Table 1. But Fisher’s exact test [3] can be used. In the three versions shown in Table 1 a very strong significance can be observed. The smaller the value p is, the more significant the dependency.

Table 1. Results of Fisher’s Exact Test. The cases printed in bold have to be explained.

Improve-ment mode	Patient's physical condition	Active	Non-active	Fisher Exact p
Strong	Better	28	2	< 0.0001
	Worse	**22**	21	
Medium	Better	40	10	< 0.005
	Worse	**10**	12	
Weak	Better	47	16	< 0.02
	Worse	**3**	6	

Exceptions. Though the performed Fisher test confirms the hypothesis, there are exceptions, namely active patients whose health conditions did not improve. Exceptions should be explained. Explained exceptions build the case base. According to Table 1, the stronger the model, the more exceptions can be observed and have to be explained.

In the following section the set-up of a case base on the strongest model version is explained.

2.2 Setting up a Case Base

We begin to set up the case base up sequentially. That means, as soon as an exception is explained, it is incorporated into the case base and can be used to help explaining further exceptional cases. A random order for the exceptional cases was chosen. In fact, they were taken in alphabetical order.

The retrieval of already explained cases is performed by keywords. The main keywords are "problem code", "diagnosis", and "therapy". In the situation of explaining exceptions for dialysis patients the instantiations of these keywords are "adverse effects of dialysis" (diagnosis), "fitness" (therapy), and two specific problem codes. Besides the main keywords additional problem specific ones are used. Here the additional keyword is the number of worsened factors. Further keywords are optional.

However, ISOR does not only use the case base as knowledge source but further sources are involved, namely the patient's individual base (his medical history) and observed data (partly gained by dialogue with medical experts). Since in the domain of kidney disease and dialysis medical knowledge is very detailed and much investigated but still incomplete, it is unreasonable to attempt to create an adequate knowledge base. Therefore, a medical expert, observed data, and just a few rules serve as medical knowledge sources.

2.2.1 Expert Knowledge and Artificial Cases

Expert's knowledge can be used in many different ways. Firstly, it is used to acquire rules, and secondly, it can be used to select appropriate items from the list of retrieved solutions, to propose new solutions and last but not least – to create artificial cases.

Initially, artificial cases are created by an expert, afterwards they can be used in the same way as real cases. They are created in the following situation. An expert points out a factor F as a possible solution for a query patient. Since many data are missing, it can happen that just for the query patient values of factor F are missing. The doctor's knowledge in this case can not be applied, but it is sensible to save it anyway. Principally, there are two different ways to do this.

The first one means to generate a correspondent rule and to insert it into ISOR's algorithms. Unfortunately, this is very complicated, especially to find an appropriate way for inserting such a rule. The alternative is to create an artificial case. Instead of a patient's name an artificial case number is generated. The other attributes are either inherited from the query case or declared as missing. The retrieval attributes are inherited. This can be done by a short dialogue and ISOR's algorithms remain intact. Artificial cases can be treated in the same way as real cases, they can be revised, deleted, generalised and so on.

2.2.2 Why Did Some Patients Conditions Became Worse?

As results a set of solutions of different origin and different nature is obtained. There are three categories of solutions: additional factor, model failure, and wrong data.

Additional factor. The most important and most frequent solution is the influence of an additional factor. Only three main factors are obviously not enough to describe all medical cases. Unfortunately, for different patients different additional factors are important. When ISOR has discovered an additional factor as explanation for an exceptional case, the factor has to be confirmed by the medical expert before it can be accepted as a solution. One of these factors is Parathyroid Hormone (PTH). An increased PTH level sometimes can explain a worsened condition of a patient [12]. PTH is a significant factor, but unfortunately it was measured only for some patients.

Other additional factors that sometimes could explain exceptional cases are a high phosphorus level and a very long time of dialysis (more than 60 months) before a patient began with the fitness program.

Model failure. We regard two types of model failures. One of them is deliberately neglected data. As a compromise we just considered data of six months, whereas further data of a patient might be important. In fact, three of the patients did not show an improvement in the considered six months but in the following six months. So, they were wrongly classified and should really belong to the "better" category. The second type of model failure is based on the fact that the two-category model was not precise enough. Some exceptions could be explained by a tiny and not really significant change in one of the main factors.

Wrong data are usually due to a technical mistake or to not really proved data. One patient, for example, was reported as actively participating in the fitness program but really was not.

3 Missing Data

Databases with many variables have specific problems. Since it is very difficult to overview their content, usually a priori a user does not know how complete the data set is. Are there any data missing? How many of them and where are they located?

In the dialyse data set many data are missing randomly and without any regularity. It can be assumed that the data set contains groups of interdependent variables but a priori it is not known how many such groups there are, what kind of variables are dependent, and in which way they are dependent. However, we intend to make use of all possible forms of dependency to restore missing data, because the more complete the observed data base is, the easier it should be to find explanations for exceptional cases and furthermore the better the explanations should be. Even for setting up the model the expert user should select those parameters as main factors with only few missing data. So, the more data are restored, the better the choice for setting up the model can be.

A data analysis method is often assessed according to its tolerance to missing data as in [13]. In principle, there are two main approaches to the missing data problem. The first approach is a statistical restoration of missing data. Usually it is based on non-missing data from other records.

The second approach suggests methods that accept the absence of some data. The methods of this approach can be differently advanced, from simply excluding cases with missing values up to rather sophisticated statistical models [14].

Gediga and Düntsch [15] propose to use CBR to restore missing data. Since their approach does not require any external information, they call it a non-invasive imputation method. Missing data are supposed to be replaced by their correspondent values of the most similar retrieved cases.

So, why don't we just apply statistical methods? Statistical methods require homogeneity of the sample. However, there are no reasons to expect the set of dialyse patients to be a homogenous sample. Since the data consists of many parameters, sometimes missing values can be calculated or estimated from other parameter values. Furthermore, the number of cases in the data set is rather small, whereas usually statistical methods are the more appropriate the bigger the number of cases.

3.1 The Data Set

For each patient a set of physiological parameters is measured. These parameters are supposed to be measured four times during the first year of participating in the fitness program. Since some parameters, e.g. the height of a patient, are supposed to remain constant within a year, they were measured just once. The other ones are regarded as factors with four grades, they are denoted as F0 – the initial measurement of factor F, and F3, F6, and F12 – the measurements of factor F after 3, 6, and 12 months.

All performed measurements are stored in the observed database, which contains 150 records (one patient – one record) and 460 variables. 12 variables are constants, the other 448 variables represent 112 different parameters.

The factors can not be considered as completely independent from each other, but there are different types of dependency among specific factors. Even a strict mathematical dependency can occur, for example in this triple: time of controlled training, work performed during this time, and average achieved power, expressed as Power = Work/Time. Less strict are relations between factors of biochemical nature. An increase of parathyroid hormone, for example, implies an increase of blood phosphorus. Such relations between factors enable us to fill some missing data in the data set.

3.2 Restoration of Missing Data

CBR is applied to restore missing data, the calculated values are filled in the observed database. The whole knowledge is contained in the case base, namely in form of solutions of former cases.

There are three types of numerical solutions: exact, estimated, and binary ones. Some examples and restoration formulas are shown in table 3. All types of solutions are demonstrated by examples.

Table 3. Some examples of solutions and of restoration formulas. Abbreviations:
BC = Breath consumption, BF = Breath frequency, BV = Breath volume,
HT = Hematocrit, P = Phosphorus, PTH = Parathyroid hormone, PV =plasma volume
Y indicates a general formula, of course in the second hematocrit line it is HAT in the concrete application

Missing parameter	Type of solution	Numeric solution (examples)	Description	Parameters
PTH	Binary	1	If P(T) >= P(t) then PTH(T) >= PTH(t) Else PTH(T) < PTH(t)	P, PTH
HT	Exact	36,2	HT = 100 * (1–PV/0.065 * Weight)	PV, Weight
HT	Estimated	29,1	Y(6) = Y(3)*0.66 + Y(12) * 0.33	HT
WorkJ	Exact	30447,1	WorkJ = MaxPower * Time * 0.5	Time, MaxPower
BC	Exact	15,6	BC = BF * BV	BF, BV
Oxygen pulse	Estimated	10,29	Linear regression	O2plus

When a missing value can be completely restored, it is called exact solution. Exact solutions are based on other parameters. A medical expert has defined them as specific relations between parameters. He has done it during the use of ISOR. When they have been used once, they are stored in the case base and can be retrieved for further cases.

Since estimated solutions are usually based on domain independent interpolation, extrapolation, or regression methods, a medical expert is not involved. An estimated solution is not considered as full reconstruction but just as estimation.

A binary solution is a partly reconstruction of a missing value. Sometimes ISOR is not able to construct neither an exact nor an estimated solution, but the expert may draw a conclusion about increasing/decreasing of the missing value. So, a binary solution expresses just the assumed trend. “1” means that the missing value should have increased since the last measurement, whereas “0” means that it should have decreased. Binary solutions are used in the qualitative models of ISOR.

First example: Exact solution.
The value of hematocrit (HT) after six months is missing. Hematocrit is the proportion of the blood volume that consists of red blood cells. So, the hematocrit measurements are expressed in percentage. The retrieved solution (third line of table 3) requires two additional parameters, namely plasma volume (PV) and the weight of the patient. For the query patient these values (measured after six months) are “weight = 74 kg and PV = 3,367”. These values are inserted in the formula and the result is a hematocrit value of 30%.

This restoration is domain dependent, it combines three parameters in such a specific way that it can not be applied to any other parameters. However, the formula can of course be transformed in two other ways and so it can be applied to restore values of PV and the weight of the patient. The formula contains specific medical knowledge that was once given as a case solution by the expert.

Second example: Estimated solution.
It is the same situation as in the first example. The value of hematocrit that should have been measured after six months is missing. Unlike the first example, now the PV value that is required to apply the domain dependent formula is also missing. Since no other solution for exact calculation can be retrieved, ISOR attempts to generate an estimated solution. A domain independent formula ist retrieved (fourth line of table 3). It states that a missing value after six months should be calculated as the sum of two-thirds of the value measured after three months and one-third of the value measured after twelve months. This general calculation can be used for many parameters.

Third example: Binary solution.
The value of parathyroid hormone (PTH) after six months is missing and shall be restored. The retrieved solution involves the initial PTH measurement and the additional parameter phosphorus (P), namely the measurement after six months, P(6), and the initial measurement, P(0). Informally, the solution states that with an increase of phosphorus goes along an increase of PTH too. More formal the retrieved solution states:

```
If P(6) >= P(0)
        then PTH(6) >= PTH(0)
        else PTH(6) < (PTH(0)
```

So, here a complete restoration of the missing PTH value is not possible but just a binary solution that indicates the trend, where “1” stands for an increase and “0” for a decrease.

3.3 Case-based Reasoning

In ISOR, cases are mainly used to explain further exceptional cases that do not fit the initial model. Just a sort of secondary application is the restoration of missing data. The solutions given by the medical expert are stored in form of cases so that they can be retrieved for solving further missing data cases.

Since the number of stored cases is rather small, retrieval is not crucial. The retrieval is performed by keywords. The four main keywords are: Problem code (here: “missing value”), diagnosis, therapy, and time period. As an additional keyword the parameter where the value is missing can be used. Solutions that are retrieved by using the additional keyword are domain dependent. They contain medical knowledge that has been provided by the medical expert. The domain independent solutions are retrieved by using just the four main keywords.

What happens when the retrieval provides more than one solution? Though only very few solutions are expected to be retrieved at the same time, only one solution should be selected. At first ISOR checks whether the required parameter values of the retrieved solutions are available. A solution is accepted if all required values are available. If more than one solution is accepted, the expert selects one of them. If no solution is accepted, ISOR attempts to apply the one with the fewest required parameter values.

Each sort of solution has its specific adaptation. A numerical solution is just a result of a calculation according to a formula. This kind of adaptation is performed automatically. If all required parameter values are available, the calculation is carried out and the query case receives its numerical solution.

The second kind of adaptation modifies a restoration formula. This kind of adaptation can not be done entirely automatically but the expert is involved. When a (usually short) list of solutions is retrieved, ISOR at first checks whether all required values of the exact calculation formulae are available. If required parameter values are not available, there are three alternatives to proceed. First, to find an exact solution formula where all required parameter values are available, second to find an estimation formula, and third to attempt to restore the required values too. Since for the third alternative there is the danger that this might lead to an endless loop, this process can be manually stopped by pressing a button in the dialogue menu. When for an estimated solution required values are missing, ISOR asks the expert. The expert can suggest an exact or an estimated solution. Of course, such an expert solution has also to be checked for the availability of the required data. However, the expert can even provide just a numerical solution, a value to replace the missing data – with or without an explanation of this suggested value. Furthermore, adaptation can be differentiated according to its domain dependency. Domain dependent adaptation rules have to be provided by the expert and they are only applicable to specific parameters. Domain independent adaptation uses general mathematical formulae that can be applied to many parameters. Two or more adaptation methods can be combined.

In ISOR a revision occurs. However, it is just the attempt to find better solutions. An exact solution is obviously better than an estimated one. So, if a value has been restored by estimation and later on (for a later case) the expert has provided an appropriate exact formula, this formula should be applied to the former case too. Some estimation rules are better than other. So it may happen that later on a more appropriate rule is incorporated in ISOR. In principle holds, the more new solution methods are included in ISOR, the more former already restored values are attempted to revise.

Since every piece of knowledge provided by a medical expert is supposed to be valuable, ISOR saves it for future use. If an expert solution cannot be used for adaptation for the query case (required values might be missing too), the expert user can generate an artificial case by using a special dialogue menu. Artificial cases have the same structure as real ones. They are also stored in the case base.

4 Results

At first, we undertook some experiments to assess the quality of our restoration method, subsequently we attempted to restore real missing data and then we set up a new model for the original hypothesis that actively participating in the fitness program improves the conditions of the patients.

4.1 Experimental Restoration

Since ISOR is a dialogue system and the solutions are generated within a conversation process with the user, the quality of the solutions does not only depend on ISOR but also on the expert user. To test the method a random set of parameter values was deleted from the observed data set. Subsequently, the method was applied and it was attempted to restore the deleted values - but not the really missing ones!

In table 4 it is summarised how and how many deleted values could be restored. More than half of the deleted values could be at least partly restored and nearly a third of them could be completely restored. However, 39% of the deleted values could not be restored at all. The main reasons are that for some parameters no proper method is available and that specific additional parameter values are required that sometimes are also missing.

Table 4. Summary of the numbers of randomly deleted and restored values. Only the deleted values were attempted to restore, not the really missing ones.

Number of Parameters	112
Number of values	448
Number of really missing values	104
Number of randomly deleted values	97
Number of completely restored values	29
Number of estimated values	17
Number of partly restored values (binary)	13
Number of automatically restored values	34
Number of expert assistance	25
Number of values that could not be restored	38

Another question concerns the quality of the restoration. That means how close are the restored values to the real values. It has to be differentiated between exact, estimated and binary restored values. Just one of the 13 binary restored values was wrong. However, this shows the “quality” of the expert user. The deviation (percentage) between the restored values and the real ones is shown in Table 5. Concerning the two exactly restored values with more than 5% deviation, we consulted the expert user, who consequently altered one formula, which had been applied for both values. For the estimated values, it is conspicuous that for few values the deviation is rather big. The probable reason is that the general estimation methods have problems with varying parameter courses.

Table 5. Closeness of restored values. The numbers in brackets show the average deviations in %.

Deviation	Number of exactly restored values	Number of estimated values
< 3 %	14 (2.2)	9 (1.8)
< 5 %	13 (5.7)	5 (6.1)
< 10 %	2 (8.5)	2 (7.4)
> 10 %	0	1 (12.3)

4.2 Restoration of Real Missing Data and Setting up a New Model

As a consequence of the experimental results, we assumed that our method is not perfect but at least applicable. So, it was attempted to restore real missing data. It is no surprise that more missing values could be restored (table 6) than randomly deleted ones (see table 4), because all restoration methods rely on other parameter values and generally holds that with an increased number of given values the chance of restoring missing values increases too. In the experiment (table 4) not just the randomly deleted values were missing but also the real missing ones.

Table 6. Summary of the numbers of missing and restored values.

Number of Parameters	112
Number of values	448
Number of missing values	104
Number of completely restored values	37
Number of estimated values	24
Number of partly restored values (binary)	19
Number of automatically restored values	49
Number of expert assistance	31
Number of values that could not be restored	24

After this restoration we returned to the original problem, namely to set up a model for the hypothesis that actively participating in the fitness program improves the conditions of the patients (see section 2.1). Since many missing values have been restored, the expert user can select other main factors to set up the model, namely also those ones where many data had been missing before. In fact, he chose a different third factor, PTH instead of WorkJ. The resulting strongest model is shown in Table 7. The result is obviously much better than the model before (see table 1 in section 2.1). However, since the missing data problem is not responsible for all exceptional cases, also for this model some (eleven) exceptional cases still have to be explained.

Table 7. Results of Fisher's Exact Test, for $p < 0.0001$.

Patient's physical condition	Active	Non-active	Fisher Exact p
Better	39	1	< 0.0001
Worse	**11**	21	

6 Conclusion

In this paper, we propose to use CBR to explain cases that do not fit a statistical model. Here one of the simplest models is presented. However, it is relatively effective, because it demonstrates statistically significant dependencies. In the example between fitness activity and health improvement of dialysis patients the model covers about two thirds of the patients, whereas the other third can be explained by applying CBR. Since just qualitative assessments (better or worse) were chosen, very small changes appear to be the same as very large ones. As a future step, it is intended to define these concepts more precisely, especially to introduce more assessments. The presented method makes use of different sources of knowledge and information, including medical experts. This approach seems to be a very promising method to deal with a poorly structured database, with many missing data, and with situations where cases contain different sets of attributes.

Additionally, a method to restore missing values was developed. This method combines expert knowledge, which is delivered as formulae for specific situations and can be used for later similar situations too, with domain independent techniques.

References

1. Schmidt R, Vorobieva O: Case-Based Reasoning Investigation of Therapy Inefficacy. Knowledge-Based Systems 19 (5); (2006) 333-340
2. Aha, D.W., McSherry, D., Yang, Q.: Advances in conversational Case-Based Reasoning. Knowledge Engineering Review 20 (2005) 247-254.
3. Kendall MG, Stuart A: The advanced theory of statistics. 4 ed. New York: Macmillan publishing, New York (1979)
4. Hai GA: Logic of diagnostic and decision making in clinical medicine. Politheknica publishing, St. Petersburg (2002)
5. Perner, P. (ed.): Case-Based Reasoning on Images and Signals. Springer-Verlag, Berlin Heidelberg (2007)
6. Bichindaritz, I., Kansu, E., Sullivan, K.M.: Case-based Reasoning in Care-Partner. In: Proc EWCBR-98. Springer, Berlin Heidelberg (1998) 334-345.
7. Montani, S., Magni, P., Bellazzi, R., Larizza, C., Roudsari, A.V., Carsson, E.R.: Integration model-based decision support in multi-modal reasoning system for managing type 1 diabetic patients. Artificial Intelligence in Medicine 29 (1-2) (2003) 131-151
8. Prentzas, J., Hatzilgeroudis, I.: Integrating Hybrid Rule-Based with Case-Based Reasoning. In: Craw, S., Preeece, A. (eds.): Proc ECCBR 2002, Springer, Berlin Heidelberg (2002) 336-349
9. Shuguang, L., Qing, J., George, C.: Combining case-based and model-based reasoning: a formal specification. Proc APSEC'00 (2000) 416
10. Corchado, J.M., Corchado, E.S., Aiken, J. et al.: Maximum likelihood Hebbian learning based retrieval method for CBR systems. In: Ashley, K.D., Bridge, D.G. (eds.): Proc ICCBR 2003, Springer, Berlin Heidelberg (2003) 107-121
11. Arshadi, N., Jurisica, I.: Data Mining for Case-based Reasoning in high-dimensional biological domains. IEEE Transactions on Knowledge and Data Engineering 17 (8); (2005) 1127-1137
12. Davidson, A.M., Cameron, J.S., Grünfeld. J.-P. et al. (eds.): Oxford Textbook of Nephrology, Volume 3. Oxford University Press (2005)
13. McSherry, D.: Interactive Case-Based Reasoning in sequential diagnosis. Applied Intelligence 14 (1) (2001) 65-76
14. Little, R., Rubin, D.: Statistical analysis with missing data. John Wiley & Sons, New York (1987)
15. Gediga, G., Düntsch, I.: Maximum Consistency of Incomplete Data via Non-Invasive Imputation. Artificial Intelligence Review 19 (1) (2003) 93-107