# A New Method of Data Preparation for Cardiological Decision Support

R Mlynarski[1], G Ilczuk[2], A Wakulicz-Deja[3], W Kargul[1]

[1]Electrocardiology Department of Medical University of Silesia, Katowice, Poland
[2] Siemens AG Medical Solutions, Erlangen, Germany
[3]Institut of Informatics University of Silesia, Sosnowiec, Poland

## Abstract

*There has been huge progress in the introduction of new digital methods, such as decision support, in cardiology. Data preparation is the most important and the most time-consuming part of the data mining process. We present a newly developed hierarchical method of text classification based on regular expressions. This method is the basis of our data mining system during the pre-processing stage to transform Latin-based free-text medical reports into a decision table. In this study we also compare the accuracy and scalability of our method with an approach based on dictionary phrases.*

## 1. Introduction

There has been huge progress in the introduction of new Artificial Intelligence (AI) based methods in medicine. A good example of this is decision support. Our earlier experiments with creation of the complex system for decision support for cardiology were presented at many international Congresses including Computers in Cardiology 2005 in Lyon. We showed the potential usefulness of these types of methods based on real data in the clinical practice [1-3].

The key elements our decision support system are:
1. Importing data from medical information systems.
2. Converting the information extracted from narrative text into data understandable by machine-learning algorithms (ICD-10 codes).
3. Selecting attributes depending on user choice for further rule generation.
4. Generating decision rules using our own implementation of the rough sets MLEM2 algorithm.
5. Visualization of the knowledge discovery in a form easily understandable by humans.
6 - Verification by experts (doctors).

The biggest challenges in this system are still the accuracy and shortening the time consumption needed to obtain results. We have done many experiments to evaluate which stage of the process is the most important from the point of view of the clinical accuracy of the results. According to the analyses, data preparation has the most influence. Also other researchers such as Pyle defined it as the most important part of a data exploration process which leads to success [4]. In most research data preparation takes about 60% of the time needed for the whole data mining process. Additionally other modules of the system directly depend on its quality: attribute selection (responsible for removing irrelevant and redundant information based on rough set theory), rule induction (calculation of decision rules) and visualization/testing module (checking the collected knowledge and presenting it in a form easily understandable by humans) [5,6]. Our focus concentrated on creating a complete solution suitable for improving medical care and clinical workflow through revealing new patterns and relations among data in cardiology [7-10]. Most of the medical information useful for data mining is still written in the form of free-text medical Latin-based reports. These reports are mostly used to extend a lapidary diagnosis written in statistical ICD-10 codes. There are some problems to solve when analyzing such reports, e.g.: different descriptions for the same disease, non-standard abbreviations, misspelled words and the floating structure of such reports. In our first solution of these problems, we used a phrase dictionary to map information from a report to an attribute. The main disadvantage of this approach was a lack of scalability and difficult maintenance. These facts led us to develop a different approach and this method and the results achieved using it are presented in this study.

## 1.1. Aim of the study

The aim of this study was:

- to prepare an algorithm for recognition of free-text Latin-based cardiological reports using hierarchically organized records,
- to implement this algorithm into the rough sets based software developed by our team to fulfil the needs of cardiology,
- to test the results in clinical work-flow using real-life data from the Electrocardiology Department.

# 2. Methods

In our research in analyzing medical data we would like to extend and complement information collected from clinical information systems in the form of ICD-10 codes with additional information stored in free-text descriptions. A typical example of a such description is the Latin diagnosis shown below:

```
Status post implantationem pacemakeri modo
VVI (1981, 1997) ppt. diss. A-V gr.III.
Exhaustio pacemakeri. Morbus  ischaemicus
cordis. Insufficientia coronaria chronica
CCS I. Myocardiopathia ischaemica in stadio
comp. circulatoriae. Fibrillatio atriorum
continua.
```

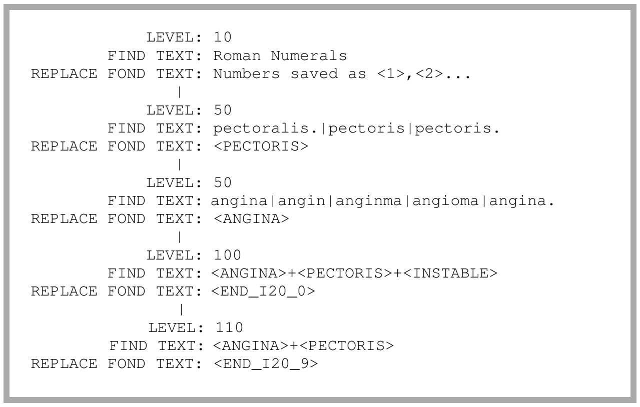
A method suitable for our needs should therefore fulfil the following requirements:

- it should recognize misspelled and abbreviated words,
- it should interpret whole sentences,
- it should provide a back tracing so that an expert can always validate an assigned mapping,
- all mappings must done based 100% on information from an original text,
- it should be easily maintainable and extendible.

With these requirements in mind, we have developed a method which is based on a fixed number of user-defined records, each containing the following three attributes:

- a level value (shown in Figure 1 as 'LEVEL'),
- a mask coded using regular expressions for searching a phrase of text ('FIND TEXT') and
- a string of text which will be used for replacing if the searched phrase is found ('REPLACE FOUND TEXT').

An example is shown in the Figure 1. From this example The lowest level value is 10, so that the algorithm begins to select a group of records having this level value. In our case only one record replaces all the found Roman numbers with the following schema '<'+number+'>' for example a number 'II' will be replaced with <2> and 'IV' results in <4>. After this processing the next higher level value is selected (50) together with a group of records having the same value of their level attribute. In the example shown, there are two such records, which are independent of each other and therefore can be processed in parallel. It is important to note that the definition of a mask used for searching (field 'FIND TEXT') contains not only the correct version of a phrase but also several misspelled combinations stored using the alteration operation of regular expressions for example (pectoralis|pectoris|...).



```
            LEVEL: 10
        FIND TEXT: Roman Numerals
REPLACE FOND TEXT: Numbers saved as <1>,<2>...
               |
            LEVEL: 50
        FIND TEXT: pectoralis.|pectoris|pectoris.
REPLACE FOND TEXT: <PECTORIS>
               |
            LEVEL: 50
        FIND TEXT: angina|angin|anginma|angioma|angina.
REPLACE FOND TEXT: <ANGINA>
               |
            LEVEL: 100
        FIND TEXT: <ANGINA>+<PECTORIS>+<INSTABLE>
REPLACE FOND TEXT: <END_I20_0>
               |
            LEVEL: 110
        FIND TEXT: <ANGINA>+<PECTORIS>
REPLACE FOND TEXT: <END_I20_9>
```

Figure 1. Example of record structures used by the presented algorithm.

If a phase is found then it will be replaced with a specified replacement string independently whether it was written correctly or incorrectly. This allows the correction of simple errors and focuses on sentence analysis. As an example records with their level value 100 and 110 can be used. These two records search for a combination of symbolic replacements previously replaced by records with the level value 50, so that these two records can correctly assign an '<END_I20_0>' code not only to a properly written 'angina pectoris' diagnosis but also to a wide range of misspelled combinations of these two words as for example 'angin pectoralis'.

The described algorithm has following advantages:

- it allows filtering of redundant and noisy information at the entry processing stage,
- it correctly recognizes misspelled diagnoses,
- the process of interpreting whole sentences is simplified because only connections between symbolic phrases must be analyzed and not all possible combinations which can be found in an input text,
- it is possible to stop the algorithm at any stage and analyze or eventually correct the replacement process,
- in our implementation we use only combinations of words already found in input text which decreases the possibility of false positive recognitions

## 2.1. Experimental environment

The data used in our research was obtained from the Electrocardiology Department of Medical University of Silesia in Katowice - the leading Electrocardiology Department in Poland specializing in the hospitalization of patients with severe heart diseases including heart rhythm disorders. For our experiments we took a data set of 4000 patients hospitalized in this Department between

2003 and 2005. This data were imported into a PostgreSQL database and then divided into eight groups (G-C1, ..., G-C8), where G-C1 contained first 500 records from the database and each subsequent group had 500 more records than the previous group, so that the last group G-C8 contained all 4000 records. Each record in a group contained a single free-text report which was then analyzed for the presence of one of the following diseases presented in Table I.

Table 1. Analyzed diseases and ICD codes.

| **Disease** | | **ICD** |
|---|---|---|
| Essential (primary) hypertension | → | I10 |
| Past (old) myocardial infarction | → | I25.2 |
| Atrioventricular block, first degree | → | I44.0 |
| Atrioventricular block, second degree | → | I44.1 |
| Atrioventricular block, complete | → | I44.2 |
| Sinus node disfunction | → | I49.5 |

We implemented the presented algorithm in Java version 1.5. and used the Java implementation of regular expressions from the 'java.util.regex.Pattern' class.

## 3. Results

The results presented in Table II show the absolute number of cases recognized by the method described in this paper within each of the tested group.

Table 2. Number of recognized cases by the proposed method.

| **Group** | **Number of records** | **I10** | **I25.2** | **I44.0** | **I44.1** | **I44.2** | **I49.5** |
|---|---|---|---|---|---|---|---|
| G-C1 | 500 | 325<br>**1** | 91<br>**15** | 45<br>**0** | 82<br>**3** | 135<br>**-4** | 220<br>**72** |
| G-C2 | 1000 | 636<br>**3** | 190<br>**-13** | 95<br>**12** | 161<br>**31** | 245<br>**17** | 414<br>**150** |
| G-C3 | 1500 | 978<br>**12** | 276<br>**3** | 134<br>**36** | 248<br>**99** | 354<br>**68** | 618<br>**241** |
| G-C4 | 2000 | 1314<br>**21** | 368<br>**25** | 183<br>**61** | 329<br>**158** | 483<br>**131** | 824<br>**338** |
| G-C5 | 2500 | 1645<br>**30** | 442<br>**30** | 236<br>**96** | 410<br>**220** | 604<br>**192** | 1029<br>**441** |
| G-C6 | 3000 | 1959<br>**38** | 524<br>**44** | 305<br>**120** | 502<br>**258** | 728<br>**238** | 1265<br>**550** |
| G-C7 | 3500 | 2275<br>**41** | 600<br>**40** | 373<br>**125** | 590<br>**267** | 863<br>**237** | 1493<br>**640** |
| G-C8 | 4000 | 2616<br>**41** | 704<br>**43** | 427<br>**131** | 678<br>**276** | 1007<br>**233** | 1693<br>**714** |

These results are additionally compared with the dictionary method and this comparison is shown as bold numbers in each column. A positive number means the number of cases additionally recognized by the method based on regular expressions. Visualization of these numbers is shown at Figure 2, where it can be seen that the proposed method recognized more cases than the dictionary method but with a different, depending on a selected disease, characteristic. For hypertension and old myocardial infarction the number of additionally recognized cases is rather low which may be attributable to the fact that most diagnosis variants are already covered by the dictionary method. Recognition of atrioventricular block poses a bigger challenge, so that a difference in the number of recognized cases for all three types of this disease varies between 20-40% additional cases identified by the proposed method. The most spectacular results were achieved for recognizing Sick node disfunction which can be assignable with a huge number of possible combinations used to specify this diagnosis. These combinations were better covered by regular expressions and the difference to the dictionary method was almost 42%.

What was very interesting is that the number of identified cases recognized by the new method increased for all tested diseases almost linearly.

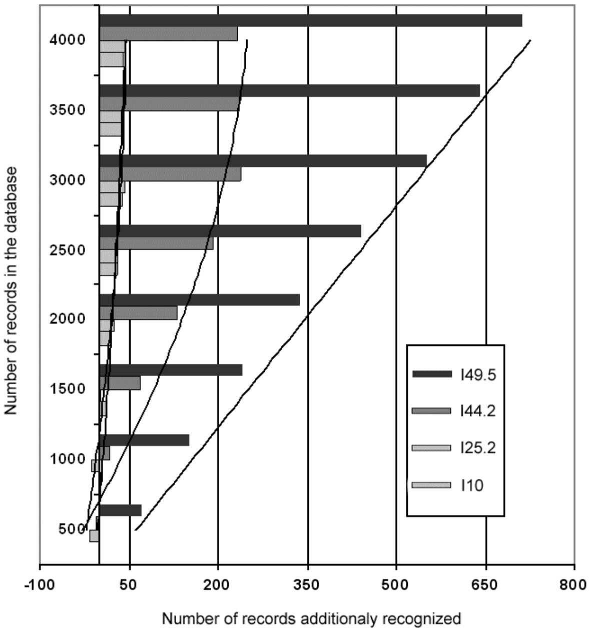


Figure 2. Additionally recognized cases by the new method based on regular expressions.

This satisfactory result shows the ability of the presented method to recognize new records with a relatively small number of definitions (500 in regular expressions compared to more then 4800 in dictionary phrases).

## 3.1. Clinical tests

We had also randomly selected a set of 100 records and with the help of domain experts from the Electrocardiology Department manually identified them for three diseases. These numbers were then compared with the results achieved by both, the regular expression and the dictionary method. Based on these results it can be seen that for a relatively small group of records the method based on regular expression recognized all hypertension and atrioventricular block (first degree) cases. Of course, it will be a matter of additional time and effort needed to extend the recognition accuracy of the dictionary method but this is precisely the advantage of the proposed algorithm, which with a significantly smaller number of records presents better scalability and in the case of new data also a better update ability.

## 4. Discussion

In this paper we presented an algorithm for the recognition of free-text Latin-based medical reports which is based on hierarchically organized records. These records use regular expressions to find a specified phrase in an input text and replace it with a user-defined text. The hierarchically organized records convert an input text step by step replacing simple words into symbolic phrases first then these symbolic phrases into more complicated expressions and finally whole sentences are mapped to user-defined codes. Such codes can then be easily used to construct a decision table used by the next data mining algorithms.

Our experiments have shown that the presented method achieves better recognition accuracy than the method based on fixed dictionary phrases and this result can be achieved with a significantly smaller number of records used for the definition. This small number of easily modifiable and very flexible records is truly an advantage of the described method.

Our idea to reduce the complexity of recognizing Latin-based diagnosis through defining short parts of the whole sentence using regular expressions and then to join such pieces of information together hierarchically allowed us to cover, with a finite, small number of records, a huge number of possible combinations. This advantage and the fact that the presented method fulfils all the specified requirements shows that it can be used in our data exploration system during a pre-processing stage for processing also laboratory, ECG and Echo descriptions.

## 4.1. Conclusions

1. A high accuracy of the newly developed method for recognition of free-text Latin-based medical reports based on hierarchically organized records was achieved.
2. Implementation of this method in automated decision support software can improve the accuracy and cut down the time needed for calculation in comparison to full syntax analysis.
3. Additional research is necessary to improve the recognition accuracy.

## Acknowledgements

We would like to thank Iwona Grzegorczyk MS from the statistical department of the Upper Silesia Medical Center for her assistance in exporting the databases.

## References

[1] Mlynarski R, Ilczuk G, Pilat E, Wakulicz-Deja A, Kargul W. Automated Decision Support and Guideline Verification in Clinical Practice. Computers in Cardiology 2005;32:375-378

[2] Ilczuk G, Mlynarski R, Wakulicz-Deja A, Drzewiecka A, Kargul W. Rough Set Techniques for Medical Diagnosis Systems. Computers in Cardiology 2005;32:837-840

[3] Ilczuk G, Mlynarski R, Wakulicz-Deja A, Drzewiecka A, Gardas R, Kargul W. Attribute selection techniques for medical diagnosis systems. Kardiologia Polska 2005; 63: 3 (supl.1): 218-219

[4] Pyle D. Data preparation for data mining. San Francisco Morgan Kaufmann.; 1999.

[5] Pawlak Z. Knowledge and Uncertainty: A Rough Set Approach. In: SOFTEKS Workshop on Incompleteness and Uncertainty in Information Systems; 1993. p. 34-42.

[6] Pawlak Z, Grzymala-Busse JW, Slowinski R, Ziarko W. Rough Sets. Commun ACM. 1995;38(11):88-95.

[7] Ilczuk G, Wakulicz-Deja A. Rough Sets Approach to Medical Diagnosis System. Proceedings Lecture Notes in Computer Science. 2005;3528:204-210.

[8] Ilczuk G, Wakulicz-Deja A. Attribute selection and rule generation techniques for medical diagnosis systems. Proceedings Lecture Notes in Computer Science. 2005;3642:352361.

[9] Wakulicz-Deja A, Paszek P. Applying Rough Set Theory to Multi Stage Medical Diagnosing. Fundam Inform. 2003;54(4):387-408.

Address for correspondence

**Cardiological section**
Rafal Mlynarski
Klinika Elektrokardiologii
ul. Ziolowa 45/47
Katowice 40-635, Poland
email: joker@mp.pl

**Computer section**
Grzegorz Ilczuk
Heuweg 12A
91334 Hemhofen, Germany
email:
Grzegorz.Ilczuk@Ilczuk.com