CONTI'2008
The 8th INTERNATIONAL CONFERENCE ON TECHNICAL INFORMATICS, 5-6 June 2008, TIMISOARA, ROMANIA

A Framework for Mining Association Rules in Data on Perinatal Care

Roxana Dogaru*, Daniela Zaharie*, Diana Lungeanu**, Elena Bernad** and Maria Bari***

* Department of Computer Science, West University of Timisoara, Faculty of Mathematics and Computer Science, 4 V. Parvan Blvd, 300223 Timisoara, Romania
Phone: (40) 256-592150, Fax: (40) 256-592316, E-Mail: {rdogaru,dzaharie}@info.uvt.ro

** University of Medicine and Pharmacy, 2 Eftimie Murgu Sq., Timisoara, Romania

*** Department of Obstetrics-Gynaecology, University of Medicine and Pharmacy, 37 Dionisie Lupu Street, Bucharest, Romania

Abstract ***–We propose a data processing framework for mining associations between medical diagnoses in obstetrical and perinatal care. Starting from the ICD-10 codes reported in the DRG system, we focused on identifying and implementing procedures for data aggregation and filtering to be applied as preprocessing tools in mining medical data. The proposed framework was validated for real data provided by two hospitals of obstetrics-gynaecology.***

Keywords: ***medical data mining, association rules, DRG, data aggregation, data filtering, interestingness measures***

I. INTRODUCTION

Since 2005, when the Diagnosis Related Groups (DRG) system became compulsory as a reimbursement system for the hospital care in Romania, a large amount of data have been collected. This information refers to the principal and secondary diagnoses coded in the International Classification of Diseases v.10 (ICD-10) system [1] and to the applied medical procedures. Although mainly used as a payment system, the DRG datafiles contain valuable information which could be exploited to get useful medical knowledge as well. As pointed out in [2], the ICD codes can be used in medical research to study patterns of disease, patterns of care, outcomes of disease and to document the comorbidities of patients. In order to extract useful knowledge from DRG files, adequate data mining techniques should be applied. For instance, clustering techniques could be applied to extract patient profiles or association rules mining might be useful to identify frequent co-occurrences of diagnoses. Unfortunately, the DRG datafiles are not suitable to be easily mined using classical data mining techniques and tools. Therefore they have to be preprocessed in order to meet the requirements of a given data mining task. On the other hand, the results provided by a data mining method are not always easy to manage or understand. For instance, a large unstructured set of rules produced by a rule mining method might be of limited utility for end-users. This motivates the need for adequate post-processing of the results in order to make them manageable.

The aim of the work presented in this paper was to identify preprocessing and postprocessing steps and provide a data processing framework for extracting association rules from DRG datafiles. The framework was tested for obstetrical and perinatal data by analyzing several mining scenarios aiming to identify associations between diagnoses of mothers and their newborns.

II. PREVIOUS WORK

Rules mining provides knowledge which can be more easily interpreted than the results provided by other data mining techniques (e.g. neural networks). This explains the interest in using this technique for analyzing medical data. Association rules mining was successfully used in identifying potentially meaningful relationships between concepts in biomedical literature [3] and in extracting risk patterns from medical data [4].

The association rules mining in correlation with DRG data were used in finding frequent sequential patterns within the patient pathway [5] and to estimate weights in case-mixing [8]. The approach proposed in [8] is based on the idea that by extracting association rules from DRG data and by analyzing their relevance one can identify frequent combinations of diagnoses which can be used in estimating weights of different groups of diagnoses.

Most works on mining for association rules in medical data identified as the main drawback of the mining process the fact that it usually leads to a large set which may contain many redundant and uninteresting rules. Thus the mining process must be guided by an adequate filtering of initial data and the resulting set of rules should be post-processed in order to extract potentially useful/interesting rules.

III. PARTICULARITIES OF DRG DATA

DRG files contain one record for each hospitalization episode of a patient. Apart from the patient ID, the records

contain information on the patient medical status and the applied care procedures. The former consists of a list with the ICD-10 codes corresponding to the principal diagnosis and several secondary diagnoses, while the latter are specific procedural codes. The DRG data collected in perinatal care have a particular property: there exist records both for mothers and their newborns. This would allow the mining of associations between mother and newborn diagnoses. Unfortunately, the DRG records containing data about mothers and newborns are completely separated and there is no information to uniquely and unambiguously connect a mother to her baby. Therefore a preliminary aggregation of these data is necessary in order to make them usable as an information source for identifying frequent associations between mothers and newborns medical status and in identifying risk factors corresponding to abnormal situations, e.g. preterm birth.

IV. THE DATA PROCESSING FRAMEWORK

Due to the particularities of DRG files, applying a data mining method directly to the raw data would probably not lead to the expected results (i.e. easy to understand and useful medical knowledge). In order to reach such a goal, several pre-processing and post-processing steps should be followed. The data processing workflow in the case of mining for association rules in data about perinatal care is illustrated in Fig. 1. This framework contains several tools with different degrees of specificity. For instance, the aggregation tool is specific to obstetrical data, while the analytical tool can be a general one (not particularly tailored for the task of obstetrical data mining). The particularities of each tool involved in the data processing workflow are summarized in the following.

A. The Aggregation Tool

The role of the data aggregation process is to combine data from different sources. In the case of obstetrical and perinatal data the main sources are represented by the DRG files provided by the Obstetrics-Gynaecology (OG) and the Neonatology (NN) wards. In each case one has to deal with a database containing at least three separate tables with corresponding records for each patient, related through the patient ID: one containing the main diagnosis, another with the list of secondary diagnoses and a similar one with the codes of medical procedures. Additional data (e.g. socio – economic status and constitutional characteristics of the mother) collected through specific software systems may be also used in the aggregation process.

One of the main and most difficult tasks of the aggregation process is that of matching the records of mothers to those corresponding to their newborns. Since a key relating the mother's and her newborn's records does not exist, the matching should rely on existing, although less reliable information. The only pieces of information which would allow to pair a OG record with a NN one are the family

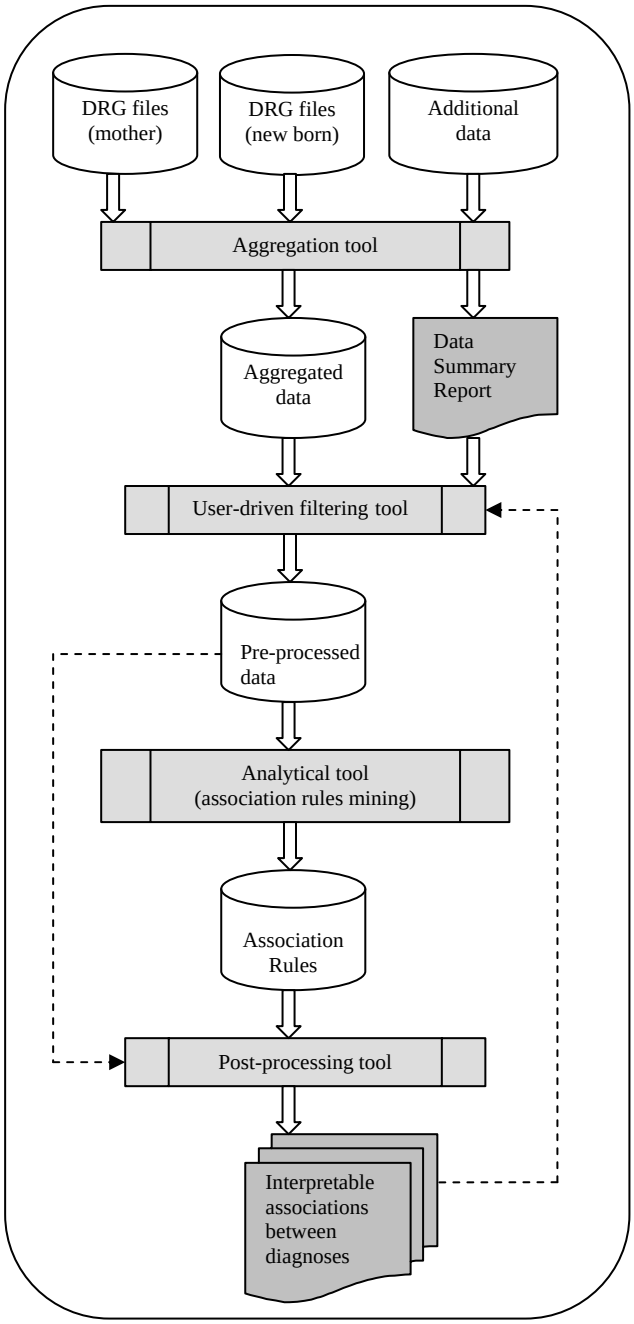

Fig. 1. The overall structure of the data processing framework

name and the hospitalization period. The matching process is based on partial information thus there can be cases when no match is found.

The basic idea of the matching process can be summarized by the following rule:

"IF OG.family_name=NN.family_name AND NN.birth_date is consistent with the OG.hospitalization_period THEN there is a match"

This matching rule is checked during several stages:

Stage 1. At the first stage all pairs (OG patient, NN patient) characterized by the same family name and the same date of birth (in the case of a NN record) and of child delivery (in the case of a OG record) are considered matches with a high degree of confidence. As a result of this stage we have a file containing pairs (mother, child) and files containing unmatched OG and NN records, respectively.

Stage 2. At the second stage the unmatched OG and NN records are analyzed by applying a modified version of the previous rule. More specifically, the rule still checks for the correspondence between the names but relaxes the condition on the dates: a difference of one day is accepted between the child birth date and the starting date of the delivery process (in order to catch situations like delivery starting before 12 P.M. and finishing after 12 P.M.). The rule applied to this stage also allows a difference up to three days between the date of the admission in the hospital of the child and her/his birth date (in order to catch situations when the child is born at home).

Stage 3. The still unmatched records are analyzed by taking into account only the constraints on the birth and delivery dates. The result will be a list of potential mothers and of their corresponding children which should be analyzed by a human expert in order to extract new matches. Since the OG file contains also records on patients hospitalized for other gynaecological problems only the records having main diagnoses codes from obstetrical group (e.g. O3*,O4*,O5*,O7*,O8*) are selected. On the other hand, in order to deal with situations when the difference between the two family names is caused by misspelling or typos a technique for string pattern matching, based on computing the edit distance, is used to rank the potential matches.

The rules applied at these three stages were derived based on a preliminary analysis of data collected during one year at two hospitals. The aggregation tool also gather the information corresponding to all hospitallization episodes during the pregnancy. This way, for each mother-newborn pair, there is a record containing the available information on their medical history: data about the birth, the baby medical condition at birth and as a neonate, previous hospitalization episodes of the mother (when the case, in chronological order). As an additional result of the aggregation step there are generated statistics reflecting the distribution of principal and secondary diagnoses. The frequencies of distinct diagnoses codes are further used in the filtering step and in preprocessing the data for the association rules mining.

B. The Filtering Tool

The direct application of an associations discovery tool to the entire set of data usually leads to a large set of rules. In order to guide the mining process, the data should be first filtered. The filtering tool allows to select subsets of records (e.g. those corresponding to the preterm births or to the newborn children with low birth weight etc.) and to eliminate codes corresponding to frequent diagnoses or procedures (such codes would favour rules with high support but quite trivial from the medical point of view).

The filtering process is guided by the user who, based on the statistical summary provided by the aggregation tool, decides which records (patients) and which attributes (diagnosis codes) should be taken into account.

Once obtained, the filtered data are prepared according to the requirements of the association rules mining tool, e.g. truth table format (each record contains a list of truth values corresponding to all distinct diagnoses codes).

C. Association Rules Mining

The main idea we followed in the development of the data processing framework was to ensure the interoperability with existing efficient mining tools. The efficiency issue is critical for association rule mining, which is a computationally expensive task. Therefore we selected a data mining product designed in order to deal with large amount of data, i.e. SPSS's Clementine 11.0 [http://www.spss.com/Clementine]. This data mining workbench offers the possibility of using two state-of-the-art algorithms in rules mining: Apriori [6] and GRI – Generalized Rules Induction [7]. In both cases the extracted rules can contain several antecedent terms and one consequent term. The list of attributes involved in antecedent and consequent terms is constructed by the user. For each rule some relevance measures are provided: rule support, confidence and lift value.

The rule support expresses the probability that both the antecedent and the consequent are satisfied: P(A,C). It measures the quantitative importance of a rule and in the Apriori algorithm is used as a main criterion in constructing the so-called frequent itemsets. The confidence corresponds to the conditional probability of the consequent given the antecedent: P(A,C)/P(A). As emphasized in [8] the confidence can be interpreted as a measure of the qualitative importance of a rule. In the Apriori algorithm it is used as a criteria to construct the association rules starting from the frequent itemsets. The lift measure is computed as P(A,C)/(P(A)P(C)) and can be used to evaluate the significance of an association rule [8]. If the value of the lift is 1 it is considered that the association rule has no relevance. If it is smaller than one then both sides of the association rule inhibit each other with no positive effect. If the value of the lift is greater than 1 then both sides of the association rules have positive impacts on each other. Thus an association rule can be considered as relevant as the lift value is higher.
The disadvantage of using a general tool for association rules mining is that it does not deal well with the particularity of mining medical data. When analyzing medical data, low support but high lift rules can be of

interest because they provide interesting information, i.e. novel while meaningful from the clinical point of view. This drawback of using a general tool can be partially compensated in the post-processing step by ranking the rules based on different rules interestingness measures.

D. The Post-processing Tool

The simplest procedure for post-processing the set of extracted rules consists in ranking the rules based on the relevance measures provided by the association mining tool. Another approach consists in computing for each rule in the set some supplementary interestingness measures as, for instance, those analyzed in [9]. In this case, for each rule provided by the analytical tool, the pre-processed data are scanned in order to compute statistical measures used in the computation of the interestingness measures.

V. A CASE STUDY

In order to test the data processing framework, the tools specific to the obstetrical data (aggregation and filtering) were implemented in Java. The input data in the Java application would consist of Excel files exported from the DRG system and the output would be a file of records prepared according to the requirements of Clementine data mining tool. The implemented aggregation tool would firstly ensure the reorganization of the input data in a easy-to-access relational database. Secondly, following the steps described in section IV.A, it would construct the tables containing both paired (mother, newborn) records and unmatched records. In order to ensure the access to a relational database, the JDBC Java API was used. The data processing framework was tested for real datasets containing records collected during 2006. The data were provided by two hospitals of obstetrics - gynaecology.

A. Summary of analyzed data

From the data collected in the first hospital (H1) we identified (during the aggregation step) 1908 pairs of matching mother and newborn records, based on which 108 distinct ICD-10 codes were identified for mothers and 201 for newborns. In the case of the second hospital the number of identified pairs is 2341 and the number of distinct codes is 379 and 120 for mother and newborn records, respectively. Tables 1 and 2 present the most frequent main and secondary diagnoses for both hospitals.

These frequencies can be used in order to identify codes which should be filtered out. On the other hand the codes frequencies can be used to compare the coding styles in different hospitals. The differences in Tables 1 and 2 between the frequencies suggest the existence of different patterns of coding in different hospitals, as have been already identified in [10] by applying clustering methods.

TABLE 1. Top 5 frequent diagnoses codes (main and secondary associated to mothers)

Main diagnoses				**Secondary diagnoses**			
Hospital 1		**Hospital 2**		**Hospital 1**		**Hospital 2**	
Code	%	Code	%	Code	%	Code	%
O80.0	61.16	O83.8	59.27	Z39.2	82.39	Z37.0	97.18
O82.0	36.84	O82.1	18.53	O34.2	1.41	Z39.1	97.13
O81.4	0.83	O80.0	13.54	Z35.3	0.78	Z39.2	96.58
O80.1	0.31	O60	4.27	Z35.5	0.78	Z39.0	96.24
O84.2	0.31	Z39.0	1.027	O13	0.73	Z30.0	88.89

TABLE 2. Top 5 frequent diagnoses codes (main and secondary associated to newborns)

Main diagnoses				**Secondary diagnoses**			
Hospital 1		**Hospital 2**		**Hospital 1**		**Hospital 2**	
Code	%	Code	%	Code	%	Code	%
Z38.0	71	Z38.0	27.35	Z24.6	34	P59.9	86.5
P05.0	5.2	P03.4	11.14	Z23.2	30	Z38.0	60.1
P07.3	4.6	P12.1	10.08	P59.9	19	P83.1	31.73
P07.1	2.7	P15.4	9.14	P92.8	10	P12.1	11.53
P02.5	2.6	P55.0	5.38	P21.9	4.4	P15.4	9.22

B. Mining scenarios

In order to extract rules reflecting associations between the mother health status and the status of the newborn several mining scenario were conducted, based on different filtering criteria.

Scenario 1. In the first scenario all diagnoses codes of mothers corresponding to the birth moment were considered as potential antecedent terms and all diagnoses codes of the newborns were considered as potential consequent terms.

Scenario 2. The aim of this second scenario was to identify possible associations between the diagnoses of the mother at the birth moment and the diagnoses of preterm newborns and of newborns with small weight. In order to do this the newborn records containing one of the codes P05* (*Slow fetal growth and fetal malnutrition*) or P07* (*Disorders related to short gestation and low birth weight, not elsewhere classified*) as main or secondary diagnosis were selected. In order to avoid the generation of uninteresting rules frequent codes carrying little novel information were eliminated (e.g. P59.9-*neonatal jaundice*, Z23.2-*need for immunization against tuberculosis*, Z39.2-*routine postpartum follow-up*).

Scenario 3. In this scenario we looked for associations between diagnoses of the mother in previous hospitalization episodes and her diagnoses at the birth moment. Therefore only the records of mothers having previous hospitalization periods were selected. All diagnoses codes corresponding to previous hospitalization episodes (except for Z34* - *supervision of normal pregnancy*) were considered as potential antecedent terms while all diagnoses corresponding to the birth moment (except for O80.0 – *spontaneous vertex delivery* and Z39.2

- *routine postpartum follow-up*) were considered as potential consequent terms.

Scenario 4. This scenario aimed to identify associations between diagnoses of mothers which had hospitalization episodes before the birth moment and the newborn characteristics. Thus the same records as in the previous scenario were used and the potential antecedent terms were all the available mother diagnoses (corresponding to the previous hospitalization periods and to the birth moment). The consequent terms were from the newborn diagnoses. In order to avoid the generation of high support but less interesting rules, frequent secondary codes with trivial clinical information were ignored (e.g. P59.9 -*neonatal jaundice*, Z23.2 - *need for immunization against tuberculosis*, Z24.6 - *need for immunization against viral hepatitis* and Z38.0 - *singleton born in hospital*).

C. Results

Scenario 1. The rules with the highest confidence obtained by applying the Apriori and GRI algorithms are presented in Tables 3 and 4, respectively. As results in Tables 3 and 4 illustrate, this scenario, especially in the case of the Apriori algorithm, leads to rules which are trivial from a medical point of view. They were extracted by the mining algorithm mainly due to the fact that the codes involved in the antecedent and consequent parts are from the list of frequent codes presented in Tables 1 and 2. These results suggested the necessity of designing different scenarios in order to guide the mining process toward more interesting rules.

TABLE 3. Highest confidence rules extracted by the Apriori algorithm. Significance of codes: O80.0 – spontaneous vertex delivery, P59.9 – neonatal jaundice, Z24.6 – need for immunization against viral hepatitis,Z39 - postpartum care and examination, Z48.0 – attention to surgical dressings and sutures.*

Antecedent	Consequent	Support %	Confidence %
Hospital 1			
O80.0	Z24.6	61.16	34.7
O82.0	Z24.6	36.79	33.04
Hospital 2			
Z48.0, Z39.0 - Z39.2	P59.9	81.67	87.60

TABLE 4. Highest confidence rules extracted by the GRI algorithm. Significance of codes: O66.4 - failed trial of labor, O84.2-multiple deliveries, all by caesarean section, P07.1 – other low birth weight,Z29.8 – other specified prophylactic measures, Z30.0 – general counseling and advice in contraception, Z38.0 – singleton born in hospital, Z92.2 – personal history of long term use of medicaments.

Antecedent	Consequent	Support %	Confidence%
Hospital 1			
O84.2, Z92.2	P07.1	0.31	83.33
Hospital 2			
O66.4, Z29.8, Z30.0	Z38.0	10.68	82.4

Scenario 2. From the set of 35 rules extracted when applying the Apriori algorithm on the data from the first hospital, a subset of six were selected during the post-processing step: the best two rules with respect to each relevance measure (Table 5). As expected, the high support rules did not provide surprising associations.

Scenario 3. For this scenario the Apriori algorithm generated 15 rules with support larger than 1% and confidence larger than 30%. This scenario was aimed at finding meaningful association between the hospitalization episodes prior to the birth moment and the diagnoses at the birth moment. The hospitalization episodes were specified as follows: E0 – birth moment, E1 – last hospitalization before the birth moment, E2 – the previous hospitalization episode etc. Six rules corresponding to the highest values of support, confidence and lift are listed in Table 6.

TABLE 5. Rules expressing associations between mothers diagnoses at birth moment and diagnoses of newborns with low weight. Significance of codes: O80.0 – spontaneous vertex delivery, O82.0 – delivery by elective caesarean section, P07.3 – other preterm infants, P05.0 – light for gestational age, P74.4 – other transitory electrolyte disturbances of newborn, P91.3 – neonatal cerebral irritability, Z35.0 – supervision of pregnancy with history of infertility, Z34.9 – supervision of normal pregnancy (unspecified), Z34.8 – supervision of other normal pregnancy, Z35.9 – supervision of high risk pregnancy

Antecedent	Consequent	Support %	Confidence %	Lift %
Highest support rules				
O80.0	P07.3	21.16	35.41	0.97
O80.0	P05.0	19.08	31.94	1.20
Highest confidence rules				
Z35.0	P07.3	4.97	54.54	1.49
Z34.9	P07.3	3.32	53.53	1.46
Highest lift rules				
Z34.8,O82.0	P91.3	1.66	30.76	3.70
Z35.9	P74.4	2.90	35	3.37

TABLE 6. Rules expressing associations between mothers diagnoses at previous hospitalization episodes and at the birth moment. Significance of codes: O82.0 – delivery by elective caesarean section, O20.0 – threatened abortion, O47.9 – false labor, Z34.0 – supervision of normal first pregnancy, Z35.9 – supervision of high risk pregnancy, Z35.0 – supervision of pregnancy with history of infertility.

Antecedent	Consequent	Support %	Confidence %	Lift %
Highest support rules				
Z34.0 (E1)	O82.0	8.55	34.78	0.97
O20.0(E1)	O82.0	8.02	38.46	1.07
Highest confidence rules				
Z35.9 (E1)	O82.0	7.48	50	1.99
O20.0(E2), O47.9(E1)	Z35.0	2.67	45.45	5.31
Highest lift rules				
O20.0(E2), O47.9(E1)	Z35.0	2.67	45.45	5.31
Z35.9(E1), O47.9(E1)	Z35.0	3.2	42.85	5.009

Scenario 4. In the case of scenario 4, two variants were analyzed: one when the antecedent were represented by diagnoses corresponding to hospitalization prior the birth moment (E1,E2,…) and a second one when the diagnoses corresponding to the birth moment (E0) were also taken into account. In the first case, by applying the Apriori method, we obtained 72 rules with support higher than 1% and confidence higher than 30%. The best three rules for each criterion are presented in Table 7. In the second case we obtained nine rules with support higher than 1% and confidence over 30%. The rules corresponding to the highest values for the relevance measures are presented in Table 8.

TABLE 7. Rules expressing associations between mothers diagnoses at previous hospitalization episodes and the child diagnoses. Significance of codes: O20.0 – threatened abortion, O47.9 – false labor, P05.0 – light for gestational age, P07.1 – other low birth weight, P07.3 – other preterm infants, P74.9 – transitory metabolic disturbance of newborn, P92.8 – other feeding problems of newborn, Z34.0 – supervision of normal first pregnancy, Z38.6 – other multiple born in hospital.

Antecedent	Consequent	Support %	Confidence %	Lift %
Highest support rules				
Z34.0 (E1)	P92.8	4.81	19.56	1.74
Z34.0 (E1)	P05.0	3.74	15.21	1.67
O20.0(E1)	P74.9	3.20	15.38	2.21
Highest confidence rules				
Z34.0 (E1) O47.9(E1)	P92.8	2.13	28.57	2.54
O20.0(E2), O47.9(E1)	P07.1	1.60	27.27	7.28
O34.3(E1)	P07.3	2.13	26.66	2.93
Highest lift rules				
O20.0(E2), O47.9(E1)	Z38.6	1.07	18.18	17
O20.0(E3), O20.0(E2)	Z38.6	1.07	15.38	14.38
O20.0(E3)	Z38.6	1.07	14.28	13.35

TABLE 8. Rules expressing associations between mothers diagnoses at all hospitalization episodes (including the birth moment) and the child diagnoses. Significance of codes: O20.0 – threatened abortion, O34.3- maternal care for cervical incompetence, O47.9 – false labour, P07.3 – other preterm infants, P07.1 – other low birth weight, P92.8 – other feeding problems of newborn, Z34.0 – supervision of normal first pregnancy, Z39.2 – routine postpartum follow-up, Z37.0 – single live birth, Z34.9 –supervision of normal pregnancy.

Antecedent	Consequent	Support %	Confidence %	Lift %
Highest support and confidence rules				
Z34.0 (E1) O47.9(E1) Z39.2(E0)	P92.8	2.14	33.33	2.96
O34.3 (E1) Z39.2(E0)	P07.3	2.14	30.76	3.38
Highest lift rules				
O47.9(E1) Z37.0(E0)	P07.1	1.60	30	8.01
O20.0(E3), O20.0(E2) Z39.2(E0)	P07.1	1.60	30	8.01
Z34.9(E0) Z39.2(E0)	P07.3	1.60	30	3.74

These results illustrate the fact that even a small change in the filtering criteria can lead to significantly different sets of rules.

VI. CONCLUSIONS

Association rule mining allows the extraction of meaningful information from the DRG files, while the mining process is carefully guided and the discovered rules are validated by medical experts to assess their clinical interestingness. The results we obtained illustrate the importance of appropriately approaching the pre- and post-processing steps. The steps with the highest impact on the final results proved to be those related with data aggregation and filtering. The use of heuristic criteria based on the edit distance between the names of mothers and newborns and on the relationships between several dates (date of newborn birth, date of hospital admission, date of surgical procedure on the mother etc.) allowed us to reduce the number of unmatched pairs during the aggregation stage. The proposed data processing framework proved to be a viable solution in mining DRG data, an important source of information not fully exploited yet in our healthcare system.

ACKNOWLEDGMENTS

This work is supported by grant 99-II CEEX 03 - INFOSOC 4091/31.07.2006.

REFERENCES

[1] WHO. ICD-10 Online. (Last access Feb 28th 2007) *http://www.who.int/classifications/apps/icd/icd10online*

[2] K.J. O'Malley, K.F. Cook, M.D. Price, K.R. Wildes, J.F. Hurdle and C.M. Ashton, "Measuring diagnosis: ICD code accuracy", HSR: Health Services Research 40:5, Part II, pp. 1620-1639, 2005.

[3] M. Berardi, M. Lapi, P. Leo, C. Loglisci, "Mining generalized association rules on biomedical literature", Proc. of the 18th International Conference on Innovations in Applied Artificial Intelligence, LNCS 3533, pp. 500-509, 2005.

[4] J. Li, A.W. Fu, H. He, J. Chen, H. Jin, D. McAullay, G. Williams, R. Sparks, C. Kelman, "Mining risk patterns in medical data" , Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 770-775, 2005.

[5] .N. Jay, G. Herengt, E. Albuisson, F. Kohler, A. Napoli, "Sequential pattern mining and classification of patient path", in M. Fieschi et al. (Eds), Proc. of MEDINFO 2004, pp. 1667, 2004.

[6] R. Agrawal and R. Srikant, ″Fast algorithms for mining association rules″, Proc. of the 20th Int'l Conference on Very Large Databases, pp. 487-499, 1994.

[7] P. Smyth, R.M. Goodman, "An information theoretic approach to rule induction from databases", IEEE Trans. On Knowledge and Data Engineering, vol. 4, no. 2, pp. 301-316, 1992.

[8] C. Baragoin, C.M. Andersen, S. Bayerl, G. Bent, J. Lee, C. Schommer, Mining Your Own Bussiness in Health Care, IBM RedBook, 2001.

[9] D.R. Carvalho, A.A. Freitas, N..Ebecken, "Evaluating the correlation between objective rule interestingness measures and real human interest", Proc. of PKDD 2005, LNAI 3721, pp.453-461, 2005.

[10] D. Lungeanu, S. Holban, D. Zaharie, M. Bari, E. Bernad, D. Navolan, "A clustering approach in characterizing ICD-10 code usage patterns in obstetrics and perinatal care", Rev Med Chir Soc Med Nat Iasi (ISSN 0048-7848), Vol 111, Nr. 2 Supl.2: 132-136, 2007.