Guide to Medical Data on Collagen Disease and Thrombosis

Jan Zytkow, Shishir Gupta

Department of Computer Science, UNC Charlotte, Charlotte
North Carolina, USA
shishir@shishir.net

Abstract. Collagen diseases are often dangerous and can be lethal. A severe complication common to those diseases of auto-immune system is called thrombosis. It occurs when blood vessels are clogged by coagulation of blood. Data relevant to the analysis of patients with collagen diseases and thrombosis have been donated to the discovery challenge in the hope that the discovered knowledge will illuminate the mechanisms responsible for collagen diseases and will help to diagnose and predict attacks of thrombosis. Two previous discovery challenges, at PKDD-99 in Prague and PKDD-2000 in Lyon, brought preliminary results available to all participants in the new challenge. It seems that the thrombosis data offer a potential for much more knowledge. We made a number of improvements to the original, raw thrombosis data during the two years since they became available. The data are now available in the format of seven files, each representing one relational table that can be directly loaded into a relational database.

1 Thrombosis and collagen diseases data: medical problems

Collagen diseases are disorders of auto-immune system. Patients generate anti-bodies which attack their own bodies. That may result in a loss of life, when anti-bodies paralyze the organ where they develop. For example, if a patient generates anti-bodies in lungs, (s)he will chronically lose the respiratory function and finally will lose life. Little is known about the mechanisms responsible for those diseases and their classification is still fuzzy. Some patients may generate many kinds of anti-bodies and they can manifest in all the characteristics of collagen diseases.

Thrombosis is one of the most important and severe complications in collagen diseases, and one of the major causes of death. Thrombosis is an increased coagulation of blood which clogs blood vessels. Usually it lasts several hours and can repeat. It has been found that this complication is closely related to anti-cardiolipin antibodies. This was discovered by physicians, one of whom donated the dataset to discovery challenge.

Thrombosis is an emergency. It is important to predict the possibility that it will occur. It is also important to detect that it occurred and to capture temporal patterns specific and sensitive to attacks of thrombosis. Thrombosis can arise from different collagen diseases which in turn can help predict thrombosis. Doctors are moreover interested in classifying collagen diseases and in temporal patterns specific and sensitive to each collagen disease.

Many other problems may be solved with the Thrombosis Data and the Challenge is open to all of them.

2 The Thrombosis Data tables

The patients belong to three categories:

- *First type*: a patient followed at outpatient clinic in University hospital, but no special examinations are made for this patient. Those patients do not suffer from thrombosis. *[These patients have their ID's listed in PATIENT_INFO but not in ANTIBODY_EXAM and related tables]*

- *Second type*: a patient followed at University Hospital and special examinations are made for this patient. *[These patients have their ID's listed both in PATIENT_INFO, in ANTIBODY_EXAM and related tables and in LAB_EXAM]*

- *Third type*: a patient is not followed at University Hospital, but special examinations are made for this patient. About 400 patients in ANTIBODY_EXAM and related tables are belonging to third type. But they are not followed at University Hospital, they do not have temporal data. *[These patients have entries in PATIENT_INFO and ATNIBODY_EXAM and related tables but not in table LAB_EXAM]*

The Thrombosis Data consist of seven tables. Note that the attributes that belong to the key are emphasized in **bold**face and underlined.

By joining several tables useful information can be extracted. The effect of various Collagen Disease can be studied on both sexes and their corresponding ages. The effect of various anti-body concentrations and degree of Coagulation of blood with respect to them can be studied and useful patterns can be found which can help us in classifying collagen diseases and thrombosis. Further more the concentrations of various anti-bodies with respect to various other parameters from the normal laboratory examination can be studied and temporal pattern specific/sensitive to thrombosis and collagen diseases can be found.

2.1 Table PATIENT_INFO

Table PATIENT_INFO includes all the data of patients who are followed by doctors at the outpatient clinic in University Hospital at least for several months. The primary data of the patient are recorded when the patient first comes to the Outpatient clinic. This table consists of 1239 records.

Table 1. Table PATIENT_INFO

ITEM	EXPLANATION	ORACLE TYPE
ID	Patient's unique identification	varchar(32)
Sex	Patient's sex	char(1)
Birthday	Patients date of birth	date
Description Date	First date when the patient was recorded	date
First Date	Day the patient came to the hospital	date
Admission	Admitted after diagnosis: '+' => Admitted to hospital '-' => Followed at the outpatient clinic	char(1)

It may be useful to study the effect of antibodies on males and females separately as their effect and prediction of thrombosis could differ with sex. Age could be an important factor, too.

Normally the data analysis starts from description-date. If, however, the patient had Laboratory Examinations done before this date, description-date can be used as the date the disease(s) was first detect.

2.2 Table DIAGNOSIS

Table DIAGNOSIS lists all collagen diseases the patient is suffering from and recognized by the doctor, but also other diseases, observations and symptoms. This table contains 1942 records.

Table 2. Table DIAGNOSIS

ITEM	EXPLANATION	ORACLE TYPE
ID	Patient's unique identification	varchar(32)
Diagnosis	Single value of Diagnosis	varchar(32)
Exam Date	The date of the exam. First Date in case of table TSUM_A Exam Date in case of table TSUM_B	date
Status	Disease Status: '~' => Suspected. '+' => Confirmed.	char(1)

From Table	From which table the data was taken: 'DT' => Table TSUM_A 'ST' => Table TSUM_B	char(2)

Link the DIAGNOSIS table to ANTIBODY-EXAM(s). In case of conflicting diagnoses, the diagnoses from table PATIENT_INFO are more recent than diagnoses from ANTIBODY_EXAM.

Closer analysis of the Diagnosis field reveals that string "susp" is appended to some attribute values of Diagnosis and per documentation it means that there diagnosis have not been confirmed and the doctors only suspect that the patient could be suffering from the disease. To accommodate this information a new column needs to be added to the "DIAGNOSIS" table. Hence we have the table definition as below. The "diag" column indicates if the diagnosis is confirmed '+' or is suspected '~'. Also observed "diagnosis" coming out of TSUM_A and here essentially are the same and hence can be combined by adding an extra column to indicate which table this is coming from.

2.3 Table ANTIBODY_EXAM

This table reports on special laboratory examination performed on some patients. The data consists of various anti-body levels, blood coagulation levels and degree of thrombosis at the time when thrombosis happened. The table also includes the tests when thrombosis was suspected but it did not occur. This table includes 801 records.

Attributes ACL IgG, aCL IgM, ANA, aCL IgA refer to the anti-body concentrations present when thrombosis struck. As anti-cardiolipin antibodies are linked to thrombosis, their concentrations at the time of thrombosis occurrence will be very useful in determining the temporal patterns specific/sensitive to thrombosis.

Table 3. Table ANTIBODY_EXAM

ITEM	EXPLANATION	ORACLE TYPE
ID	Patient's unique identification	varchar(32)
Exam Date	Also the date when thrombosis struck	date
aCL IgG	Anti-Cardiolipin antibody IgG concentration	number
aCL IgM	Anti-Cardiolipin antibody IgM concentration	number
ANA	Anti-Nucleus antibody concentration	number
aCL IgA	Anti-Cardiolipin antibody IgA concentration	number
KCT	Test of coagulation '+' => Above Normal '-' => Normal	varchar(1)
RVVT	Test of coagulation '+' => Above Normal '-' => Normal	varchar(1)
LAC	Test of coagulation '+' => Above Normal '-' => Normal	varchar(1)

Thrombosis	Degree of Thrombosis: 0 => Negative (None) 1 => Positive (Most Severe) 2 => Positive (Severe) 3 => Positive (Mild)	number

2.4 Table ANA_PATTERN

Table ANA_PATTERN has 656 records.

Table 4. Table ANA_PATTERN

ITEM	EXPLANATION
ID	Patient's unique identification
Exam Date	The date of the exam.
Pattern	Pattern observed in the sheet of ANA Exam: 'P' => Peripheral Pattern [DNA] {SLE, PSS, MCTD, DLE, SjS, DILE} 'H' => Homogeneous Pattern [DNP-Histon, Histon] {SLE,DLE,DILE,SjS,PSS,RA} 'S' => Speckled Pattern [ENA,NAPA] {SLE,MCTD,SjS,PSS,PM/DM,DEL,RA} 'N' => Nucleolar Pattern [Nucleolar body] {PSS,SLE,SjS} 'D' => Discrete Speckled [Centromere] {CREST,PSS} Where: [.....] => Specific Location of Antibody {.....} => Corresponding Diseases.

2. 5 Table THROMBOSIS

Table THROMBOSIS stores data on all thrombosis attacks, 195 records.

- Symptoms: Since this multi-valued attribute refers to symptoms observed in a patient and has no bearing on any Collagen disease, (or there is no sufficient data to link them to a Collagen disease in particular) this attribute may not have a great deal of significance in our analysis process.
- Thrombosis: This attribute states the degree of thrombosis at the time of the attack and should be useful in finding temporal patterns a specific or sensitive to the disease. This attribute can also help us in searching for patterns which detect & predict thrombosis.

Table 5. Table THROMBOSIS

ITEM	EXPLANATION	ORACLE TYPE
ID	Patient's unique identification	varchar(32)
Attack Date	Date of Thrombosis attack. For 'Symptom' this is the Exam Date.	date
Symptom	Symptom observed during attack.	varchar(32)
Attack number	Number for 1-5; up to 5 attacks per patient. 0 => Symptom 1 => Symptom1 2 => Symptom2 3 => Symptom3 4 => Symptom4 5 => Symptom5	number

2. 6 Table LAB_EXAM

This table contains laboratory examinations stored in Hospital Information System (57542 records). All data include ordinary Laboratory exams and have temporal stamps. These tests are not necessarily connected to thrombosis.

{-,+-,+} is a usual notation in medical "qualitative" tests. "-" is negative (in normal range), "+-" is not negative but at the border of normal range, "+" means positive, or abnormal. {-,+-,+} can be observed in a simple test: usually, each symbol corresponds to a range of "quantitative" values. "-" is equivalent to quantitative statements, such as <8, if N<8 is the normal range.

Although the majority of test results are numbers, other values cause problems. Consider values such as ``$>107$" for the predominantly numerical attribute PLT in TSUM_C. Many attributes in TSUM_C include such values. They are allowed since data types are strings rather than numbers. We can understand the convenience of the value ``>107" when the test is not exact. But this value is hard to compare with numerical values. String values allow neither the use of number ordering, nor other numerical relations. Unfortunately, there is no quick solution. The normal values of PLT are between 100 and 400, so we can include ``>107" into the normal range, but any detailed number assignment may cause significant error. On the other hand a combination of numerical and non numerical values impedes the use of many knowledge discovery tools.

The test attributes listed in table have normal ranges specified as metadata: N<20 ; 40 <N< 400 ; N=- ; N= +- The value TR of U-PRO means "error in measurement due to a problem with the submitted blood serum."

Table 6. TAble LAB_EXAM

ITEM	EXPLANATION	ORACLE TYPE
ID	Patient's unique identification	varchar(32)
Date	Examination Date	date
GOT	Normal Range : N < 60	number
GPT	Normal Range : N < 60	number
LDH	Normal Range : N < 500	number (Exceptions!!)
ALP	Normal Range : N < 300	number
TP	Normal Range : 6.0 < N < 8.5	number
ALB	Normal Range : 3.5 < N < 5.5	number
UA	Normal Range : N > 8.0 (Male) N > 6.5 (Female)	number
UN	Normal Range : N < 30	number
CRE	Normal Range : N < 1.5	number
T-BIL	Normal Range : N < 2.0	number
T-CHO	Normal Range : N < 250	number
TG	Normal Range : N < 200	number
CPK	Normal Range : N < 250	number (Exceptions!!)
GLU	Normal Range : N < 180	number
WBC	Normal Range : 3.5 < N < 9.0	number (Exceptions!!)
RBC	Normal Range : 3.5 < N < 6.0	number
HGB	Normal Range : 10 < N < 17	number
HCT	Normal Range : 29 < N < 52	number
PLT	Normal Range : 100 < N < 400	number (Exceptions!!)
PT	Normal Range : N < 14	number
Note	Note	varchar(10)
APTT	Normal Range : N < 45	number
FG	Normal Range : 150 < N < 450	number
AT3	Normal Range : 70 < N < 130	number
A2PI	Normal Range : 0 < N < 30 or TR	number
U-PRO	Normal Range :	number (Exceptions!!)
IGG	Normal Range : 900 < N < 2000	number (Exceptions!!)
IGA	Normal Range : 80 < N < 500	number (Exceptions!!)
IGM	Normal Range : 40 < N < 400	number (Exceptions!!)
CRP	Normal Range : N = +, -, +- or N < 1.0	varchar(4) (Exceptions!!)
RA*	Normal Range : N = +, - or +-	varchar(4)
RF	Normal Range : N < 20	number (Exceptions!!)
C3	Normal Range : N > 35	number (Exceptions!!)
C4	Normal Range : N > 10	number (Exceptions!!)
RNP*	Normal Range : N = +, - or +-	varchar(4)
SM*	Normal Range : N = +, - or +-	varchar(4)
SC170*	Normal Range : N = +, - or +-	varchar(4)
SSA*	Normal Range : N = +, - or +-	varchar(4)

SSB*	Normal Range : N = +, - or +-	varchar(4)
CENTROMIA*	Normal Range : N = +, - or +-	varchar(4)
DNA	Normal Range : N < 8	number
DNA-II	Normal Range : N < 8	number

*Some values of this attribute are listed as 2, 4, 8, 16, 32, 64 ... They are interpreted as: Less than 16 -> '-'
Equal to 16 -> '+-'
Greater than 16 -> '+'

2. 7 Table DISEASE

This table lists all the diseases as diagnosed by the doctors. It also contains remarks against the diagnosis if one is available. Many values of diagnosis name collagen diseases, but some others refer to non-collagen diseases, observations and symptoms. The table includes 65 records.

Table 7. Table DISEASE

ITEM	REMARKS	ORACLE TYPE
Disease Name	The name of the disease. All the values occur in the DIAGNOSIS Table.	varchar(64)
Disease Type	The type of the disease 'C' => Collagen disease 'D' => Non-collagen disease 'N' => No diagnosis 'O' => Observation	varchar(1)
Comments	Doctor's comments if any	varchar(128)

3 Brief history of the challenge on Thrombosis Data

Three contributions were made to the September 1999 Prague Challenge chaired by Petr Berka (Beilken & Spenke, 1999; Levin et al. 1999; Taylor, 1999). Also in September 1999, four contributions were made to a workshop in Japan chaired by Shusaku Tsumoto (Ichise & Numao, 2000; Nakamoto, Yoshida & Suzuki, 2000; Negishi, Suyama & Yamaguchi, 2000; Tsukada, Inokuchi, Washio & Motoda, 2000). Two contributions were submitted to PKDD-2000 Challenge in Lyon (September 2000; Meidan et al. 2000; Tawfik & Strickland, 2000). Most of the contributions are interesting but the results are preliminary. The most interesting results were obtained by Beilken and Spenke's Infozoom, which captures not only reasonable rules from antibody exams but also very interesting temporal patterns of laboratory tests before the thrombosis episode. Other rule induction methods obtain reasonable results but did not induce interesting temporal patterns.

4 The raw Thrombosis data

The raw Thrombosis Data for the 1999 Discovery Challenge were organized into three tables, TSUM_A.CSV, TSUM_B.CSV, TSUM_C.CSV (for simplicity we will skip the extensions .CSV). The tables can be connected by the ID number unique for each patient.

Each patient first came to the Hospital's Outpatient Clinic on collagen diseases, as recommended by a home doctor or a general physician in a local hospital. The primary data on the patient were recorded at that time. TSUM_A table consisted of approximately 1240 records and contained that information. The table was defined in detail by Tsumoto (1999). Besides ID the attributes included sex birthday, the first date when patient's data were recorded, the date when the patient came to the hospital, whether the patient was admitted to the hospital or followed in the outpatient clinic. The last attribute was DIAGNOSIS. This was a multi-valued attribute and upon closed examination the values turned out to belong to several categories, only some directly related to collagen diseases.

The table TSUM_B included special results obtained in the Laboratory on Collagen Diseases. The data were input by doctors. They only include the patients who underwent those special tests. The data include patient ID, examination date, concentrations of three anti-cardiolipin antibodies (IGG, IGM, IGA), anti-nucleus antibody concentration (ANA), ANA patterns (a multi-valued attribute), three measures of degree of coagulation (KTC, RVVT, LAC). One attributes described degree of thrombosis, while two other multi-valued attributes described diagnosis and symptoms. The problems with multi-valued attributes are similar to diagnosis in TSUM_A. One can assume that the examination date was frequently close to the date of thrombosis.

The third table, TSUM_C, included ordinary laboratory examinations, one record per one date of the tests. Distinct attributes permit storage of values of 42 specific tests recorded. ID is a foreign key to TSUM_A and TSUM_B. Many records with dates that stretch over a long time are available on some patients, raising a possibility of time-series analysis. Background knowledge available on attributes in TSUM_C included the range of normal values of each test and the meaning of each test described in one or a few words.

5 Enhancements prior to PKDD Challenges 2000 and 2001

The past challenges demonstrated that multi-relational and multi-valued data are difficult for knowledge miners. Tools are not available and problems go beyond traditional tasks of PKDD. Further problems are presented by string-valued attributes and by missing critical information on the dates of thrombosis attacks. Upon closer inquiry it turned out that some of the patients suffered multiple attacks of thrombosis, and many of relevant data were not included in TSUM_B.

5.1 String values

String format caused problems. It is vulnerable to misspelled values, different spacing in disease names, and other non-essential changes. While some values could be easily identified by commas ("SLE, PM, PSS"), many cases required help from database provider, for instance:

"ANA□□□□"and "Spleen infarction+R[-784]C, PH,throm bophlebitis".

The same diagnosis occurred under different names, such as

CHRONIC EB
CHRONIC EB VIRUS INFECTION
CHRONIC EBV
CHR EB

Value identification is a case-by-case effort that is especially helpful when the number of records with a particular string value is small, so that by recognizing the same values records can be ungrouped and significance of findings can improve.

5.2 Temporal information missing on thrombosis attacks

The guide to 1999 TSUM_B says that the tests in the Laboratory on Collagen Diseases were related to thrombosis attacks, so that the date of attack and date of the exam were similar. But this can be true only to a degree. For instance, the thrombosis attribute indicates that many patients did not suffer from thrombosis, so the exam recorded in TSUM_B may not follow a thrombosis attack. Second, doctors know that some patients suffered more than one attack. Upon closer investigation it turned out that the dates of multiple attacks can be retrieved from hospital database. Symptoms observed during each attack are also available and may be useful. The new TSUM_B includes up to four attacks per patient, which is the maximum number registered for any single patient. Each attack is described by date and symptoms observed during the attack.

The data on up to four attacks, each on a specified date enable a better use of tests in TSUM_C. Now we can distinguish data relevant to prediction of thrombosis: those are test results prior to the onset. We can also distinguish the data that can lead to detection of a past attack of thrombosis: they include test that follow the attack.

Now, TSUM_C can be JOINed with the new TSUM_B on records selected by their relevance to prediction or to diagnosis of an attack. Tests before an attack can be compared to tests after the attack. It is always important to compare such tests with a control group of patients who did not suffer thrombosis.

5.3 Multi-valued attributes

Together, the three multi-valued attributes that occurred in TSUM_A and TSUM_B were replaced by two relational tables DIAGNOSIS and ANA_PATTERN. Actually, the values were single strings, but many strings included multiple values. In the process of separation of individual values, many were determined identical, and were represented by the same name.

PATIENT_INFO is the remainder of TSUM_A, after DIAGNOSIS was put into a separate DIAGNOSIS table. DIAGNOSIS includes values of diagnosis from TSUM_A and TSUM_B, and separates them to a single value per record. It includes 1942 records.

Information from TSUM_B was distributed in four tables, including DIAGNOSIS (see above, for a multivalued attribute), ANA_PATTERN (multivalued attribute, 656 records), THROMBOSIS (195 records) and the remainder was left in ANTIBODY_EXAM (801 records).

5.4 Meaning of diagnosis

Upon inspection, different values of diagnosis turned out to belong to different categories. The values indicated not only collagen diseases but also other diseases and various observations and symptoms. It was important to create a table that identifies the category of each value of DIAGNOSIS, since the main focus of the data was collagen diseases.

6 References

Tsumoto, S. 1999. Guide to the Medical Data Set. In: Berka P. ed. {*Workshop Notes on Discovery Challenge, PKDD-1999, Prague, Sep.15-18*}, Univ.of Economics, Prague, p.45-47.

Beilken, C. & Spenke, M. 1999. Visual, Interactive Data Mining with InfoZoom -- the Medical Data Set. In: Berka P. ed. {*Workshop Notes on Discovery Challenge, PKDD-1999, Prague, Sep.15-18*}, Univ.of Economics, Prague, p.49-54.

Ichise Ryutaro & Numao Masayuki. 2000. Knowledge Discovery from Medical Database with Multistrategy Approach (in Japanese). {*Proceedings of SIG-FAI/KBS-9902, Japan Assoc. of Artificial Intelligence*}, p.1-4.

Levin, B., Meidan, A., Cheskis, A., Gefen, O. & Vorobyov, I. 1999. PKDD99 Discovery Challenge -- Medical Domain. In: Berka P. ed. {*Workshop Notes on Discovery Challenge, PKDD-1999, Prague, Sep.15-18*}, Univ.of Economics, Prague, p.55-57.

Meidan, A., Cheskis, A., Gefen, O., Levin, B. & Vorobyov, I. 2000. The WizWhy analysis of the PKDD 2000 Discovery Challenge Medical Domain. In Siebes A. & Berka P. eds. Discovery Challenge, PKDD, Lyon, France, September 12-16, 2000.

Nakamoto Kazuki, Yoshida Mieko & Suzuki Einoshin, 2000. Analysis of Collagen-Disease Data Set Based on KDD Process Model (in Japanese), {*Proceedings of SIG-FAI/KBS-9902, Japan Assoc. of Artificial Intelligence*}, p.9-153.

Negishi Naoya, Suyama Akihiro \& Yamaguchi Takahira. 2000. Automatic Composition of Inductive Applications to Collagen Diseases Database Using Inductive Learning Method Ontologies. {*Proceedings of SIG-FAI/KBS-9902, Japan Assoc. of Artificial Intelligence*}, p.5-8.

Taylor, C. 1999. PKDD'99 Discovery Challenge:Medical Data Set. In: Berka P. ed. {*Workshop Notes on Discovery Challenge, PKDD-1999, Prague, Sep.15-18*}, Univ.of Economics, Prague, p.59-64.

Tawfik, A. & Strickland K. 2000. Mining Medical Data for Causal and Temporal Patterns. In Siebes A. & Berka P. eds. Discovery Challenge, PKDD, Lyon, France, September 12-16, 2000.

Tsukada Makoto, Inokuchi Akihiro, Washio Takashi & Motoda Hiroshi. 2000. Discretization of Numerical Attributes on Structured Data for Basket Analysis, (in Japanese) {*Proceedings of SIG-FAI/KBS-9902, Japan Assoc. of Artificial Intelligence*}, p.17-24.