First Steps towards a Frequent Pattern Mining with Nephrology Data in the Medical Domain

- Extended Abstract -

Matthias Niemann[1] , Danilo Schmidt[2], Gabriela Lindemann – von Trzebiatowski[3], Carl Hinrichs[2],

University Hospital Charité
[1]Department of Transfusion Medicine
[2]Department of Nephrology

Humboldt University of Berlin
[3]Department of Computer Science

matthias.niemann@charite.de
danilo.schmidt@charité.de
gabriela.lindemann@uv.hu-berlin.de
carl.hinrichs@charité.de

Abstract. Constantly increasing, complex and heterogeneous data collections are characteristic for our time. That occurs in nearly all spheres of our life. One of the main challenges is the handling of this huge amount of data. Data mining is a field of computer science which is dealing with this special problem. This paper describes our project which uses Data mining in the medical domain. Our medical data come from the Department of Nephrology and include data of patients which have got kidney transplantations. In this domain a lot of different kind of data were produced, because patients with organ transplantation must be in treatment for the rest of their lives. One of the big aims in the transplantation science is a long-term graft survival. In this project we use Data mining to identify new pattern in this complex data, which have an influence on a better outcome after kidney transplantation.

Keywords: data mining, electronic patient record, medical data

Introduction

The advances in transplant medicine have grown rapidly over the last 50 years. Despite a variety of new transplantation methods, surgical techniques, immuno-suppressive regimens and innovative rejection treatment it is not yet managed successfully to increase long-term survival of transplanted kidneys significantly. Because of the complex interplay between the existing physical damage caused by a long period of renal insufficiency, especially in the area of blood vessels, and complex immunological reactions, it is not trivial to make specific statements without an adequate epidemiological background.

The Charité – Universitätsmedizin Berlin has one of the most assured and valid collection of records of hundreds of kidney transplants. To analyze the wealth of data to identify new patterns that allow a useful patient outcome statement is a major challenge to identify important factors retrospectively.

Statistic analysis is the regularly way of finding correlations in complex data sets. Preliminary the researcher must have an intuition of interesting parameters. But this also means that he can only find what he a-priori searches for. A suitable way to avoid this disadvantage is data mining. Data mining is working explorative and the chance to get a statistical bias is not very high, because an artificial system has no intuitive preconditions. Data mining allows processing of a huge amount of data with the possibility of getting details of potential interests in short time [HAN et. al. 2011]. In the Charité several data sources exit with a huge amount of different highly complex data. The number of data sets is high enough to find significant rules in this data.

The aim of the project has been the implementation of an explorative data search engine including Data mining features. One of these features is the Frequent Pattern Mining which we use in our project.

Data source

Our plan is a joint project between the department of Nephrology and the Laboratory of Tissue Typing. This implies that the data are coming from both departments. The bulk of the data comes from the nephrology, which deals with illnesses of the kidney including the organ replacement (dialysis, transplantation). Illnesses of the kidney cause a considerable cost in the German health service and worldwide. Considering Germany, annually more than 2.5 billion Euros are necessary for the financiering of dialysis treatments, whereby the incidence of the dialysis-requiring kidney insufficiency raises annually around 5-8%. The transplantation of a kidney as alternative to dialysis is reduced by the small availability of donor organs.

The Charité - Universitätsmedizin Berlin has three departments for the treatment of nephrological diseases due to the historical development of the union of the clinical centers of the University Hospitals of the Humboldt University of Berlin and Freie Universität Berlin.

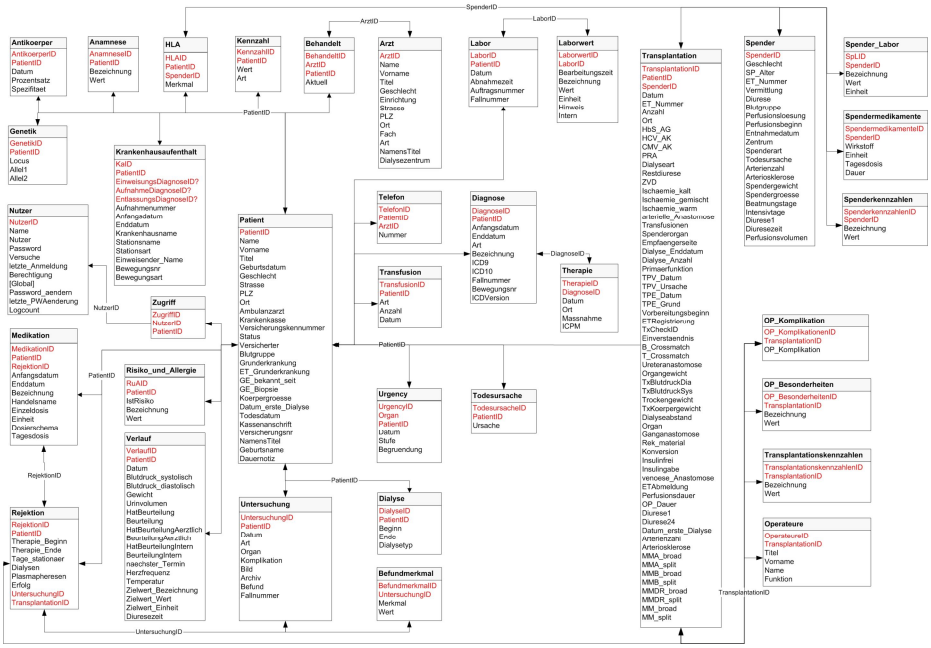

Figure 1: TBase© data structure

Each one of the three mentioned nephrology departments treats 3000 - 4000 cases per year. The availability of all cases of Berlin's university nephrology departments, the sum of about 12.000 cases, would open completely new possibilities of detection of rare disease patterns and quality management [SCHMIDT et. al. 2010].
All of the three are using the web-based Electronic Patient Record database TBase©. It is based on a "normal" relational database structure, which means, that it has fixed patterns of entities and their relations (see figure 1 only to get an impression). We used the web-based Electronic Patient Record TBase© which was implemented in a German kidney transplantation program as cooperation between the Nephrology of Charité Campus Mitte and Campus Virchow and the AI Lab of the Institute of Computer Sciences of Humboldt University of Berlin [SCHRÖTER et. al. 2000], [LINDEMANN et. al. 2000]. Currently TBase© automatically integrates essential data from the laboratory, clinical pharmacology, nuclear medicine, findings from radiology and administrative data from the SAP©-system of the Charité.

The Laboratory of Tissue Typing is the second partner of our project. For the patients which are coming from the patient record TBase© the laboratory identifies the tissue type, which is very important for transplantation. Currently we have more than 10 million data sets of laboratory values, examination and treatment data which were compiled over the last 13 years. We will use these data sets for our Data mining project, supporting the retrieval of information about the complex mechanism of an optimized outcome for the transplantation and patient survival hidden in the huge amount of data. Questions of this kind are of great interest because these complex mechanisms are not completely comprehended at present. We are sure that our Data mining project will give some hints and support in this special research area.

Pre-processing of the medical data

In our Frequent Pattern Mining two dimensions are interesting: the time and the value. For the first dimension "time" the user can define three categories of time intervals by our GUI.

1.) Static time interval with a start date and end date
2.) Event time interval with start event and an end event
3.) mixed time interval with event or date as start or end point

In our tool it is possible to generate as many time intervals as needed with respect to these categories (see figure 2).

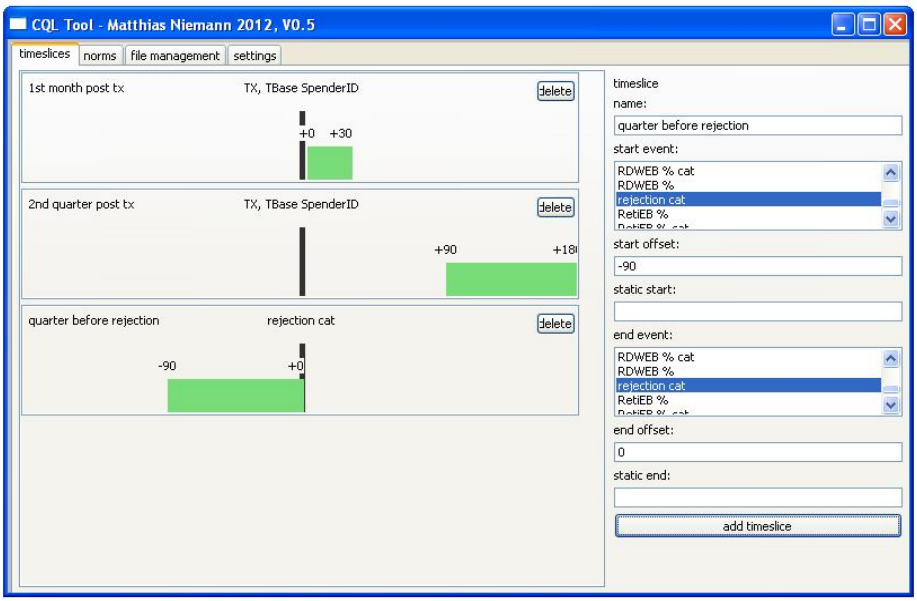

Figure 2: time slices

For the second dimension “value” the user can preprocess the data in so called norm ranges. In a first step different labeled values which are meaning the same contend can be combined under a common “norm family name”.
This is important because in retrospective data sets it is normal that parameters are not consistent labeled over the time.

In a next step the user has the possibility to define the value range e.g. for low, medium and high values or to categorize this data e.g. in positive or negative values. Here again our tool has the possibility to create as many norm definitions as needed (see figure 3).

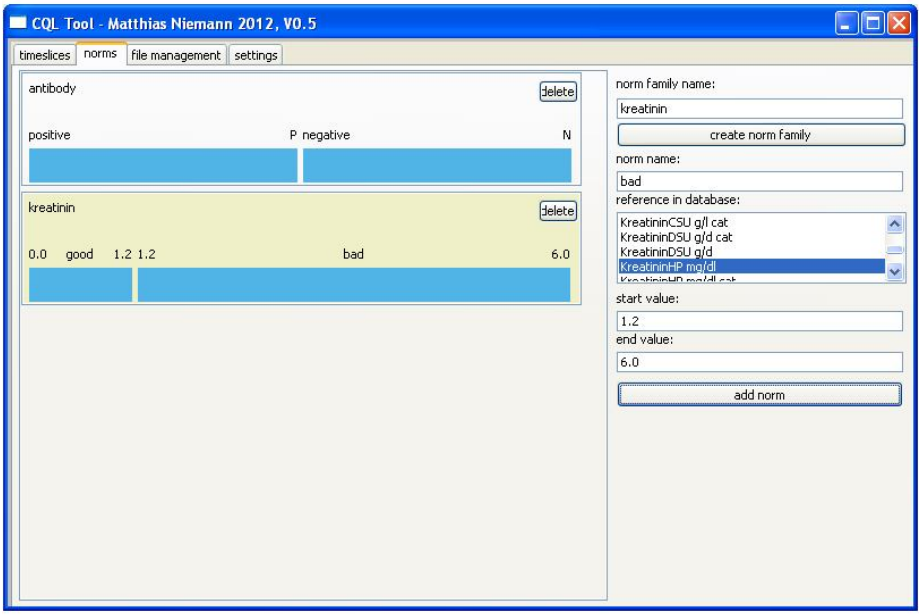

Figure 3: norms

Following these definitions, the tool creates a CQL (Charité Query Language) - file where all these information / rules are saved. The CQL file has a user identifier. So, later it can be provided to a client computer with respect to the user who defined the query (see figure 4 and 5).

myNewQuery.automator.xml | myNewQuery.cql.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<query xmlns="http://www.mniemann.de/charite/mining/ChariteQueryLanguage"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.mniemann.de/charite/mining/ChariteQueryLanguage"
queryName="krea and antibodies"
userMail="matthias.niemann@charite.de"
id="6"
>
    <timeslice id="1st month post tx">
        <start value="TX, TBase SpenderID" type="event" offset="0"/>
        <end value="TX, TBase SpenderID" type="event" offset="30"/>
    </timeslice>
    <timeslice id="2nd quarter post tx">
        <start value="TX, TBase SpenderID" type="event" offset="90"/>
        <end value="TX, TBase SpenderID" type="event" offset="180"/>
    </timeslice>
    <timeslice id="quarter before rejection">
        <start value="rejection cat" type="event" offset="-90"/>
        <end value="rejection cat" type="event" offset="0"/>
    </timeslice>
    <normFamily id="antibody">
        <norm id="positive" reference="Antibody ELISA class 1 qualitative cat">
            <qualiNorm value="P"/>
        </norm>
```

Figure 4: CQL file

Additionally it creates another automated file. This file regulates the next steps in the mining process and allows a parallel processing of the data. The files include the schedule of the necessary tasks (see figure 5 and 6).

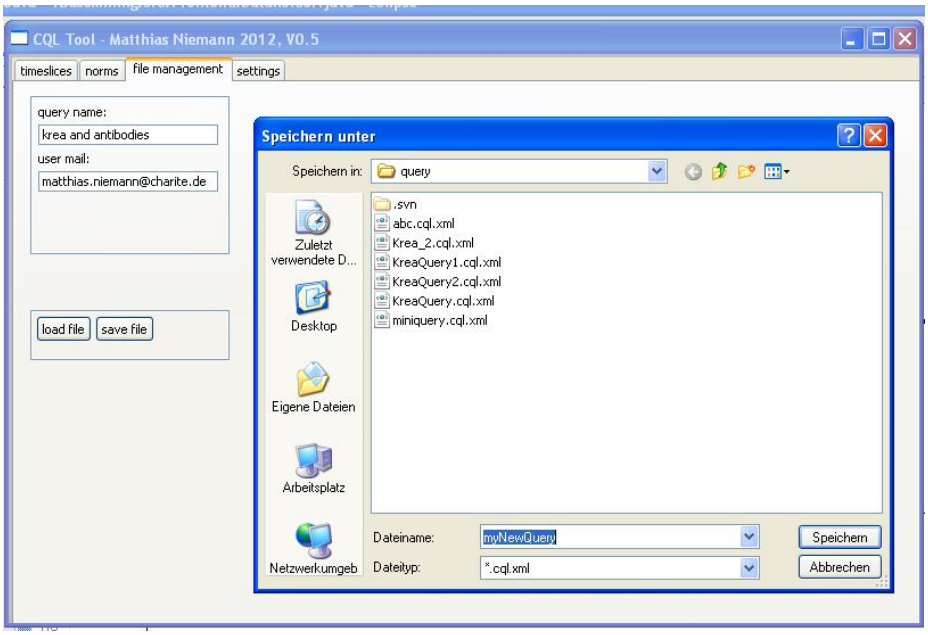

Figure 5: CQL file is user based

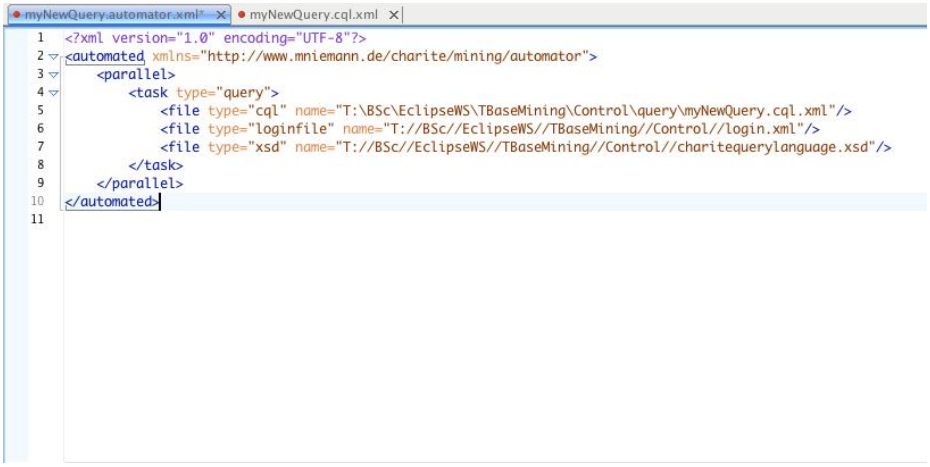

Figure 6: Automated file (XML view)

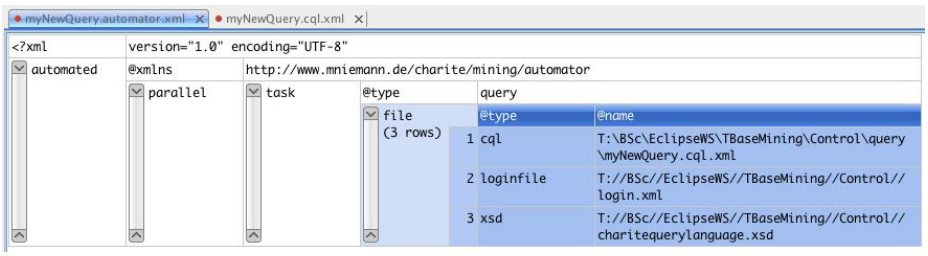

Figure 7: Automated file (structured view)

The preprocessing algorithm generates the mining item via the CQL rules out from the data source. The first preliminary result of our project is a table with all created items.

Outlook

These structured items are suitable for the Frequent Pattern Mining. Frequent Pattern Mining is a well-known Data mining technique. At the beginning of the 90's years *Agrawal* published the first efficient algorithm for the generation of association rules. The name of this algorithm is "Apriori". [HAN et. al. 2011] A further approach is the "FP-Growth" algorithm published by Han et al.. The "FP-Growth" algorithm is based on a tree structure. In contrast to the "Apriori" algorithm this allows a better performance in huge amount of patterns or very long patterns. We except to find such patterns in the medical data and therefore we think that the FP-growth algorithm is the most useful one. In a next step we will implement such an algorithm.

Further we intend to develop an adjustment of confidence and support values, because we are sure it is important and comfortable for the users. Moreover we will create a feature to filter found items. All in all we want to reduce the complexity of control of the tool for the convenience for users.
At the same time we will implement an exporter so that the user has the possibility to investigate the items with other data mining /statistic tools e.g. WEKA or SPSS.
At the end of our development we will split the medical data in a training bulk and a control bulk. After the Frequent Pattern Mining processing we will evaluate the results with statistic methods.

Conclusion

We made a data integration of transplantation data, laboratory values and clinical reports which come from the electronic patient record TBase and laboratory values, which comes from the Laboratory of Tissue Typing. We have built a tool for preprocessing for the Frequent Pattern Mining. Via a GUI the users have the possibility to define time scales with respect to defined events. Furthermore the users can configure values with special characteristics. These defined rules are stored in an especial file structure named CQL (Charité Query Language). These files can be referred to the user. The file will sent to a server were the items are generated according to the defined rules. The final items set can be used for the Frequent Pattern Mining. The plan is to use a FP-Growth Algorithm. Furthermore an exporter should be implemented for other mining or statistic tools.

References:

[HAN ET. AL. 2011] J. Han, M. Kamber, J. Pei: Data Mining: Concepts and Techniques, Morgan Kaufmann Series in Data Management Systems, 3rd revised edition, ISBN: 978-0123814791, New York, 2011.

[LINDEMANN ET. AL. 2000] G. Lindemann, L. Fritsche: Web-Based Patient Records - The Design of TBase2. In Bruch, Köckerling, Bouchard, Schug-Paß (Eds.) New Aspects of High Technology in Medicine; Seiten 409-414; 2000.

[SCHMIDT ET. AL. 2010] D. Schmidt , G. Lindemann,: Precision and Recall of the Intelligent Catalogue in the OpEN.SC - Extended Abstract -. In: Proceedings of "Concurrency, Specification & Programming" – CS&P2010; Vol. I, II, III. Informatik Berichte, Institute of Computer Science, Humboldt University Berlin, ISSN: 0863-95X, Germany, 2010.

[SCHRÖTER ET.AL. 2000] K. Schröter, G. Lindemann, L. Fritsche: TBase2 – A web-based Electronic Patient Record. Fundamenta Informaticae 43, 343-353, IOS Press, Amsterdam, 2000