Quality Checking and Mining Nephrology Biopsy Data

Miloš Radovanović
Dept. of Mathematics and Informatics
University of Novi Sad
Novi Sad, Serbia
radacha@dmi.uns.ac.rs

Vladimir Kurbalija
Dept. of Mathematics and Informatics
University of Novi Sad
Novi Sad, Serbia
kurba@dmi.uns.ac.rs

Danilo Schmidt
Department of Nephrology
University Hospital Charité
Berlin, Germany
danilo.schmidt@charite.de

Gabriela Lindemann – von Trzebiatowski
Department of Computer Science
Humboldt University of Berlin
Berlin, Germany
gabriela.lindemann@uv.hu-berlin.de

Mirjana Ivanović
Dept. of Mathematics and Informatics
University of Novi Sad
Novi Sad, Serbia
mira@dmi.uns.ac.rs

Carl Hinrichs
Department of Nephrology
University Hospital Charité
Berlin, Germany
carl.hinrichs@charite.de

Hans-Dieter Burkhard
Department of Computer Science
Humboldt University of Berlin
Berlin, Germany
hdb@informatik.hu-berlin.de

ABSTRACT

The Charité hospital in Berlin possesses one of the most secure records of hundreds of kidney transplants, most of which originate from the electronic patient record TBase©. One of the grave problems after kidney transplantation is the recipient body's immune rejection of the transplanted organ. T-cells and antibodies attack the organ, which in the worst case can lead to graft failure. Biopsy is an important diagnostic tool to evaluate a rejection episode. TBase© includes a biopsy protocol that can easily be filled out by the physician, which helped to collect and store 1447 biopsy cases in the TBase© database. With respect to different kinds of rejections, there exist some basic rules for the entered data that are enforced during completion of the protocol. Nevertheless, because so much biopsy data was entered in by hand, it was necessary to check the quality and plausibility of existing data with respect to more complex rules that were not enforced during protocol completion. In this paper, we present the process of checking the biopsy data for consistency with complex rules provided by an expert, as well as mining of new rules using interpretable rule-based classification methods. We discovered interesting rules and relationships between features with respect to T-cell mediated rejection (TCMR), antibody-mediated rejection (AMR), interstitial fibrosis and tubular atrophy (IF/TA), and polyoma virus (BKV) nephropathy, with negative results concerning acute tubular necrosis (ATN). The discovered rules further support the quality and plausibility of the data, and open avenues for further research.

BCI'13, September 19–21, 2013, Thessaloniki, Greece.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *data mining*.

General Terms

Experimentation, Management.

Keywords

Electronic patient record, nephrology, biopsy, data mining.

1. INTRODUCTION

The University Hospital Charité in Berlin possesses one of the most secure records of hundreds of kidney transplants. Most of the data originates from the electronic patient record TBase© which was developed in cooperation between Charité and the Department of Artificial Intelligence of the Humboldt University of Berlin [1]. TBase© was introduced to the daily routine in 1999. Since then, more than 3450 patient records (with transplants or on the waiting list) and additional diagnosis and treatment data have been collected.

When starting development and implementation of the web-based electronic patient record TBase©, the initial idea was to use it only as a data collection tool which should be as complete as possible, and easy to access and handle by physicians. But, with growing complexity of medical treatments and, therefore, of patient data, the need for quality checking and use of modern analytical algorithms arises. Possible ways to obtain deeper insight into dependencies of data types in the collection is to map it onto a special ontology [2], or to reorganize the data structure itself [3]. For special purposes like case retrieval this works well, but not for rule-based analysis, which is the regular way of finding correlations in complex data sets. Preliminarily, an expert must have an intuition of interesting parameters. But this also means that he can only find what he a priori searches for. One way to avoid this disadvantage is through data mining. Data mining

functions in an explorative fashion, and the chances of getting a personal or 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 obtaining details of potential interest in a short time [4].

One of the grave problems after kidney transplantation is the recipient body's immune rejection of the transplanted organ. T-cells and antibodies attack the organ, which in the worst case can lead to graft failure. Biopsy is an important diagnostic tool to evaluate a rejection episode. TBase© includes a biopsy protocol that can easily be filled out by the physician, which helped to collect and store 1447 biopsy cases in the TBase© database. With respect to different kinds of rejections, there exist some basic rules for the entered data that are enforced during completion of the protocol. Nevertheless, because so much biopsy data was entered in by hand, it was necessary to check the quality and plausibility of existing data with respect to more complex rules that were not enforced during protocol completion (or were introduced to the protocol later).

In this paper, we present the process of checking the biopsy data for consistency with complex rules provided by an expert, as well as mining of new rules using interpretable decision tree and rule-based classification methods. Our primary goal is to establish the correctness of the data through the consistency check of the given rules and plausibility of the mined rules. As a secondary aim, we expect the data mining phase to provide interesting rules that can form a basis for deeper analysis in future research.

The rest of the paper is organized as follows. Section 2 describes the biopsy data, the process of data collection, and explains the features in the data most relevant to the subsequent analysis. Section 3 presents the analysis: the checking of rules provided by an expert (Section 3.1) and rule mining (Section 3.2), where we discovered interesting rules and relationships between features with respect to T-cell mediated rejection (TCMR), antibody-mediated rejection (AMR), interstitial fibrosis and tubular atrophy (IF/TA), and polyoma virus (BKV) nephropathy. Section 4 discusses the conclusions and future outlook of the research.

2. BIOPSY DATA

TBase© is a web-based Electronic Patient Record database for collecting and storing patient records, which enables physicians to routinely enter and access data regarding all aspects of patient illness and treatment. The kidney biopsy protocol, part of which is shown in Figure 1, supports the full range of categorical, numeric, and binary indicators relevant to all possible aspects of biopsy analysis. When expanded to a flat tabular (biopsies × features) form, the protocol currently contains 1447 biopsy entries, described by 91 features. Due to space considerations we will describe only the features most relevant to our analysis.

The parts of the kidney relevant to biopsies and the associated data are (1) the glomerulum (which basically acts as a filter), (2) the tubulum (the main tube transporting fluid), and (3) blood vessels (capillaries) attached to the tubulum. A biopsy can contain samples from any combination of the three parts; therefore features specific for the three parts can be present in the data:

- Glomerular features, with three markers taking integer values in range 0–3: *ag*, *cg*, *mm*;
- Tubular-interstitial features, with seven such markers: *ai*, *at*, *ct*, *ci*, *ptc*, *ATI*, *TTI*;
- Vascular features, with three markers *av*, *ah* and *cv*.

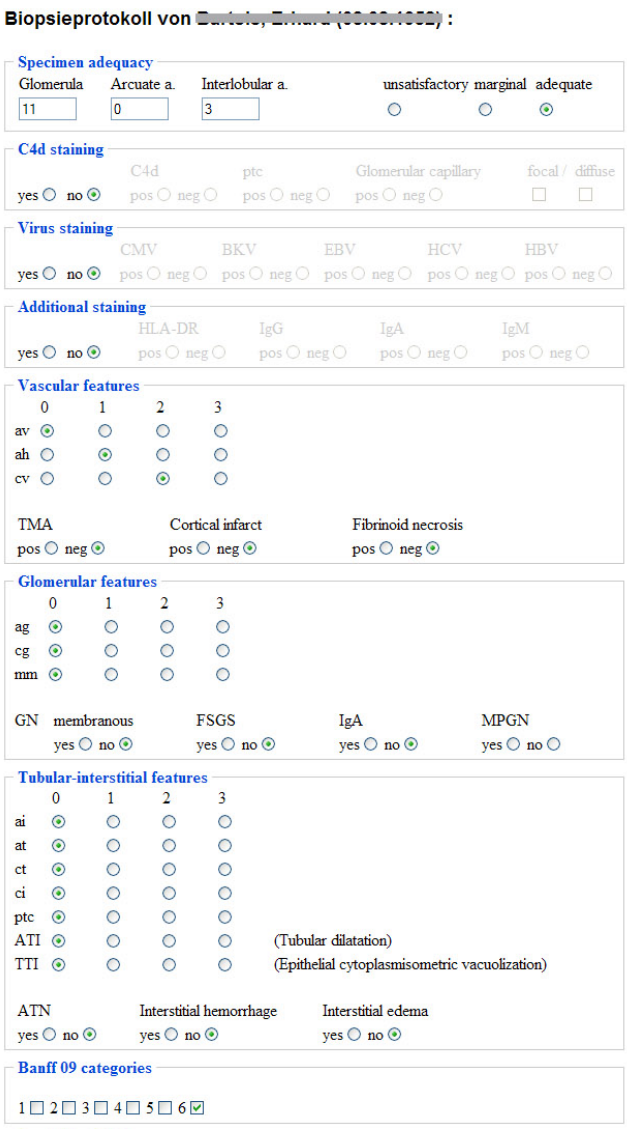

Figure 1. The TBase© biopsy protocol (portion).

The binary *Virus staining* feature indicates whether the physician ordered checks for certain types of viruses, with more specific binary features *CMV*, *BKV*, *EBV, HCV* and *HBV* denoting presence/absence of a specific kind of virus. Staining can also be performed for the presence of the C4d protein, with associated positive/negative valued markers for the presence of *C4d* in general, and in particular its presence in a peritubular capillary (*ptc*), and *Glomerular capillary*. Detection of C4d can be focal or diffuse. *Additional staining* can be ordered, which includes markers *HLA-DR*, *IgG*, *IgA* and *IgM*.

The Banff Conference on Allograft Pathology regularly held in Banff, Canada, defines categories 1–6 concerning changes detected in renal biopsies. The Banff 09 category set [5] determines the following binary features in our data:

1. Normal;
2. Antibody-mediated changes, including antibody-mediated rejection (AMR). An additional feature *Banff 09 ABM changes* provides the details, with possible values (grades) I, II and III for acute AMR, or alternatively the determination of chronic AMR;

3. Borderline changes;
4. T-cell mediated rejection (TCMR). An additional feature *Banff 09 TCMR Type/Grade* provides the details, with possible grades Ia, Ib, IIa, IIb and III for acute TCMR, or alternatively the determination of chronic TCMR;
5. Interstitial fibrosis and tubular atrophy (IF/TA), with additional feature IF/TA providing grades I, II and III;
6. Other.

In our analysis we will pay special attention to Banff categories 2, 3 and 5. We will also examine one of the binary chronic-transplant nephropathy features – *BKV Nephropathy* – which indicates nephropathy caused by the polyoma virus (BKV). Finally, we will consider the 20 binary features that describe various graft conditions not related to rejection, under the common heading *non-rejection diagnosis*.

3. DATA ANALYSIS

3.1 Checking of Expert Rules

In the first phase of our analysis, we checked the data for compliance with rules provided by an expert in the nephrology domain. The rules express relationships between different features in the data, and are as follows:

1) If *C4d staining* = **no** then there are no *C4d*, *ptc*, *Glomerular capillary* and *focal/diffuse* data.

2) If *C4d staining* = **yes** then *C4d*, *ptc* and *Glomerular capillary* should have **positive** or **negative** values.

3) If *Virus staining* = **no** then there are no *CMV*, *BKV*, *EBV, HCV* and *HBV* data.

4) If *Virus staining* = **yes** then *CMV*, *BKV*, *EBV, HCV* and *HBV* should have **positive** or **negative** values.

5) If *Additional staining* = **no** then there are no *HLA-DR*, *IgG*, *IgA* and *IgM* data.

6) If *Additional staining* = **yes** then *HLA-DR*, *IgG*, *IgA* and *IgM* should have **positive** or **negative** values.

7) If *Banff category 3* = **true** then *Banff 09 TCMR Type/Grade (acute rejection)* = **borderline** and *Banff category 4* = **false**.

8) If *Banff category 4* = **true** then *Banff category 3* = **false**.

We found no violations of rules 1, 3, 5 and 8 in the data. As for rules 2, 4 and 6, there were up to several hundred violations of these rules, however only in the sense of some staining markers having missing values instead of the required positive/negative. This is understandable since the nephrologist need not order all types of staining to be performed, for various reasons, thus the rules actually need to be updated to allow missing values. Similarly, rule 7 was violated by having missing values of the feature *Banff 09 TCMR Type/Grade* when it should have had the value borderline. However, this was a more serious violation, since the TCMR feature did not have the correct expected values. We fixed this by altering the data, setting the TCMR value to borderline whenever *Banff category 3* was true.

3.2 Rule Mining

To investigate further the plausibility of the data, potentially discovering interesting patterns and rules, we executed several standard rule-based and tree-based classification algorithms on the data, starting with the 0-Rule and 1-Rule baselines [6], and going on to Quinlan's C4.5 decision tree classifier [7], and Cohen's RIPPER rule learner [8]. The choice of classifiers was driven by their easy interpretability by a human expert. All experiments were done using the Weka machine-learning workbench [6], with the classifiers trained with their default parameters. We also attempted generation of rules featuring arbitrary sets of features using the standard Apriori algorithm [9], however the result contained too many meaningless and noisy rules to facilitate feasible analysis and interpretation.

We selected several nominal features from the data and treated each of them separately as the class feature in the process of building classifier models. The choice of class features was driven by the meaningfulness of attempting to express the relationship of the class feature with other features, and the lack of a trivial mechanism for assigning values to the class feature. The selected class features (classification problems) are therefore: *Banff 09 TCMR Type/Grade*, *Banff 09 Category 2* (AMR), *Banff 09 Category 5* (IF/TA), and *BKV Nephropathy*. Furthermore, in each classification problem we removed from the data the features that have trivial dependencies with the modeled class feature (e.g. when modeling on one Banff category, all others are removed).

With all classification problems we found the RIPPER classifier to produce results that are most accurate (or at least competitive), at the same time producing models that depend on fewer features and seem to overfit the data less. Therefore, we report the rule models built by RIPPER on the whole data set, as well as 10-fold cross-validation accuracies and confusion matrices.

Banff 09 TCMR Type/Grade. In this classification problem, RIPPPER revealed that the class depends on two features: the vascular feature *av*, and tubular-interstitial feature *at*. The rules, which should be read in a top-down manner, with the left sides of the arrows in each row representing conditions on a particular feature (or features), and the right sides denoting class values, are given below:

av ≥ 3 → III
av ≥ 2 → IIb
av ≥ 1 → IIa
at ≥ 3 → Ib
at ≥ 2 → Ia
→ borderline

The generated rules actually closely mimic the most important part of the hands-on ruled nephrologists use for determining T-cell mediated rejection. The 10-fold cross-validation accuracy of the rules is 94.86%, while the confusion matrix is shown in Table 1. The matrix indicates that errors are spread fairy evenly across the class values.

Table 1. Confusion matrix for Banff 09 TCMR Type/Grade

a	b	c	d	e	f	<-- classified as
245	3	2	0	2	0	a = borderline
5	183	1	1	1	1	b = Ia
1	1	83	2	4	0	c = IIa
1	3	1	83	2	1	d = Ib
2	0	1	0	45	0	e = IIb
0	0	0	0	0	7	f = III

Banff 09 Category 2 (AMR). The RIPPER model for determining the presence of antibody-mediated rejection depends predominantly on the *C4d* marker, as well as glomerular feature *ag*, tubular-interstitial feature *ATI*, and the *Glomerula* indicator. The rules are thus:

(C4d = positive) → true
(ag ≥ 2) and (ATI ≤ 0) and (Glomerula ≤ 18) → true
→ false

Although the accuracy of the above rule model is high, 97.72%, the confusion matrix in Table 2 reveals that for this imbalanced problem there exists a fairly large number of false positives (24), compared to the true positives, i.e. correct classifications (114). Nevertheless, the main diagnostic parameters from the biopsies were correctly identified by the classifier.

Table 2. Confusion matrix for Banff 09 Category 2 (AMR)

a	b	<-- classified as
114	9	a = true
24	1300	b = false

Banff 09 Category 5 (IF/TA). The problem of determining the occurrence of interstitial fibrosis and tubular atrophy was modeled simply and effectively using tubular-interstitial features *ci* and *ct*:

ci ≥ 1 → true
ct ≥ 1 → true
→ false

The accuracy of the above model is 96.75%, with the confusion matrix (Table 3) showing an error-preference towards false positives in this slightly imbalanced classification problem.

Table 3. Confusion matrix for Banff 09 Category 5 (IF/TA)

a	b	<-- classified as
512	12	a = true
35	888	b = false

BKV Nephropathy. The model for diagnosis of BKV virus induced nephropathy was expectedly linked to the value of the BKV marker and the tubular-interstitial feature ai (but surprisingly not at). In addition, a plausible relationship is evident with the *non-rejection diagnosis 18* (viral infection), and the *Glomerulosclerosis%* feature, with the somewhat surprising inclusion of *Arcuate a.* (specimen adequacy feature) in the rules:

(BKV = positive) and (ai >= 2) → true
(BKV = positive) and (Glomerulosclerosis% <= 33) → true
(non-rejection diagnosis 18 = true) and (Virus-staining = no) → true
(BKV = positive) and (Arcuate a. >= 1) → true
→ false

The accuracy of the rules is extremely high, 99.42%, however the problem is also extremely imbalanced, as shown by the confusion matrix in Table 4. Nevertheless, the model produces no false negatives.

Table 4. Confusion matrix for BKV Nephropathy

a	b	<-- classified as
42	0	a = true
8	1325	b = false

Non-rejection diagnosis. Finally, we report that the binary features pertaining to non-rejection diagnosis can not be adequately modeled from our data. For many of the binary classification problems there are simply too few examples in the data to train usable classifiers. Even when sufficient examples are present, the obtained models are very inaccurate, indicating that features other than those present in the biopsy data need to be considered. The most extreme example is diagnosis 5, *acute tubular necrosis* (ATN), where the number of false positives and/or negatives is an order of magnitude higher than the number of true positives for all classifiers we tested (besides RIPPER and C4.5, we considered SVMs and Bayesian networks as well).

4. CONCLUSIONS AND OUTLOOK

The primary aim of this work, which was to establish the correctness and plausibility of the collected kidney biopsy data, was fulfilled through conformance checks with provided expert rules, as well as generation of plausible rules using data mining rule-based classification techniques. However, the mined rules mostly mimic the hands-on knowledge already possessed by nephrology experts. In order to obtain radically novel insight into patterns and relationships underlying kidney transplants, the biopsy data will need to be combined with other sources of data, such as patients' general medical records, and cohort data with time series of various measurements taken from the blood.

Acknowledgments. Above all, we would like to thank Dr. Kaiyin Wu of University Hospital Charité, Berlin, for providing the data, expert rules, and thoughtful comments. Also, we gratefully acknowledge the support of this work by DAAD bilateral project "Intelligent Techniques for Data Integration and Decision Support in the Medical Domain" (2011–2012). M. Radovanović, V. Kurbalija and M. Ivanović thank the Serbian Ministry of Education, Science and Technological Development for support through project no. OI174023, "Intelligent Techniques and their Integration into Wide-Spectrum Decision Support."

5. REFERENCES

[1] Schröter, K., Lindemann, G., and Fritsche, L. 2000. TBase2 – A web-based electronic patient record. *Fundamenta Informaticae* 43, 343–353.

[2] Paslaru Bontas, E., Tietz, S., Tolksdorf, R., and Schrader, T. 2004. Generation and management of a medical ontology in a semantic web retrieval system. *Lecture Notes in Computer Science* 3290.

[3] Lindemann, G., Schmidt, D., Schrader, T., and Keune, D. 2007. The resource description framework (RDF) as a modern structure for medical data. *Proceedings of World Academy of Science, Engineering and Technology, CESSE*, 422–427.

[4] Han, J., Kamber, M., and Pei, J. 2011. *Data Mining: Concepts and Techniques*. Third Edition, Morgan Kaufmann Publishers.

[5] Sis, B. et al. 2010. Banff '09 meeting report: Antibody mediated graft deterioration and implementation of Banff working groups. *American Journal of Transplantation* 10, 464–471.

[6] Witten, I. H., Frank, E., and Hall, M. A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Third Edition, Morgan Kaufmann Publishers.

[7] Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

[8] Cohen, W. W. 1995. Learning to classify English text with ILP methods. In De Raedt, L., editor, *Advances in Inductive Logic Programming*, 124–143, IOS Press.

[9] Agrawal, R., and Srikant, R. 1994. Fast algorithms for mining association rules in large databases. *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB*, 478–499.