A Data Mining Approach to MPGN Type II Renal Survival Analysis

Chen Yang*
Health Informatics Program
The University of Iowa
Iowa City,IA 52242
chen-yang@uiowa.edu

W. Nick Street†
Management Sciences Department
The University of Iowa
Iowa City,IA 52242
nick-street@uiowa.edu

Der-Fa Lu
College of Nursing
The University of Iowa
Iowa City,IA 52242
der-fa-lu@uiowa.edu

Lynne Lanning
College of Nursing
The University of Iowa
Iowa City,IA 52242
lynne-lanning@uiowa.edu

ABSTRACT

There are three recognized types of Membranoproliferative glomerulonephritis (MPGN). Type II or Dense Deposit Disease (DDD) has a renal survival of 50% at 10 years. The goal of this study was to better identify patients at high risk of early renal failure, and to understand the factors that lead to fast progression of the disease. We identified six diagnostic features on the 98 DDD patients who responded to a web-based survey, and examined the prognostic performance of these features in isolation and simple combinations. We then combined the features to build predictive models using both Cox proportional hazards regression (CHR), a standard statistical approach, and support vector machines (SVMs), a classification technique from the data mining literature. While the age and gender features showed some prognostic ability, the combined models – particularly the SVM – were superior in identifying cases with fast disease progression. This approach can be applied to disease survival analysis and prognosis, and might be useful to healthcare providers and patients in making healthcare decisions.

*This author thanks for the University of Iowa graduate college fellowship support, without this support, this work could not be possible.

†The corresponding author.

IHI'10, November 11–12, 2010, Arlington, Virginia, USA.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences; I.2.1 [**Computing Methodologies**]: Artificial Intelligence—*Applications and Expert Systems, Medicine and science*

General Terms

Algorithms, Experimentation

Keywords

DDD, Survival Analysis, Data Mining, SVM, CHR

1. INTRODUCTION

Membranoproliferative Glomerulonephritis (MPGN) type II or Dense Deposit Disease (DDD) cause about half of the patients to progress to end-stage renal disease (ESRD) within ten years of diagnosis [1]. MPGN affects both children and adults. Ponticelli and Glassock (1997) reported that in children DDD is typically diagnosed between ages 5 to 15 years [7]. Identifying the risk factors that potentially cause DDD to progress to end-stage renal disease (ESRD) is of interest to investigators. Little et al. reported the median time of DDD patients to ESRD was 8.3 years, revealing 5-, 10- and 20-year probabilities of DDD patients to ESRD were 32, 54, and 70%, respectively [5]. While they found age and the presence of crescents on the original biopsy to be associated with recurrence in all MPGN patients, they did not isolate the type II patients.

In this study our goal is to explore the prognostic power of six easily-available demographic and diagnostic features, as gathered from an online survey of DDD patients, and determine the best use of these features to identify the patients who will progress to ESRD the fastest. We split our dataset into two groups based on a single feature value, or a simple combination of two of them (age and gender), and compared the renal survival of the two groups. We further evaluated the features based on their coefficients from a Cox proportional hazards regression model (CHR) [4]. The prognostic

value of the Cox model was then evaluated using a leave-one-out test. Finally we transformed the data to create the classification problem of identifying cases that will advance to ESRD the fastest and solved it using support vector machines (SVMs) [9], and compared the ranking of patients in this model to that of the Cox model.

Information on 98 DDD patients was gathered from the University of Iowa MPGN Database Survey. Table 1 shows the attributes collected for each patient. To gather information, patients and their families were invited to participate in the research by completing a web-based survey. The patients were registered between 2001 and 2010. As is typical with survival datasets, the outcome is represented by a status (ESRD or not) and a time, which is either the time from diagnosis to ESRD onset or the censoring time of the observation (that is, time duration from diagnosis to March 2010).

2. METHODS

2.1 Exploratory analysis

The first stage of our analysis was to determine the prognostic power available in the collected features. Our exploratory analysis began by looking at the features in isolation. Variables were evaluated individually by splitting the dataset into two groups based on the feature values, e.g., male vs. female. We also examined the set of young females, that is, those diagnosed at age 13 or younger, and compared them to the rest of the patients. In each case we visualized the two sets using Kaplan-Meier curves and tested their differences using both Wilcoxon and Logrank statistical tests of significance. Individual features were also evaluated by examining the coefficients of a CHR model constructed from the entire dataset.

2.2 Survival analysis

Our second goal was to combine the features to achieve the best identification of fast-progressing cases. As a baseline we constructed CHR models in a leave-one-out fashion to determine the prognostic power of this approach (see evaluation methodology in following subsection). However, previous research has shown that data mining models may perform well on survival analysis problems [8, 3]. So, we repeated the test with a transformed dataset using SVMs, a state-of-the-art classification method.

In order to transform the survival problem into a binary classification problem suitable for classification methods such as SVMs, we first choose a cutoff value for time to ESRD. Based on initial data analysis and clinical relevance, we chose a time of three years from diagnosis. Cases that progressed to ESRD in three years or less were labeled as positives. The negatives consisted both of censored and ESRD cases with observation time greater than three years. For censored cases with time less than three years, no label can be applied, so these cases were not used for training. They were however still used as test cases for validation.

We used the LIBSVM implementation of SVMs provided by Chang and Lin (2001) [2] to classify our DDD dataset. Based on our initial experiments we chose a radial basis function (RBF) kernel with default values for all parameters, including 0.167 for gamma and 1.0 for C.

2.3 Evaluation methodology

We compare the performance of the Cox model with the SVM model in a fashion similar to our exploratory analysis, that is, by dividing the dataset into two groups and performing a statistical comparison on their actual outcomes. In this case, the two groups represent those predicted to progress quickly to ESRD, and those predicted to progress more slowly.

We employed a leave-one-out test methodology as follows. One case was left aside for testing, and a model was built using the remaining cases. The model was then applied to the test case, and the result was recorded. Note that "result" means the predicted hazard ratio for CHR and a probability of three-year onset of ESRD for SVMs. The process was repeated, using each case as a test point.

Next we sorted the dataset in descending order by hazard ratio/probability. We divided the dataset into two parts by labeling the first 50% "poor prognosis" cases and the remainder as "good prognosis." The results were again visualized with Kaplan-Meier curves and tested for significant differences. The rationale is that a significant difference indicates good performance of the predictive model itself.

3. RESULTS

Of all the 98 patients with DDD, the percentage of hematuria, proteinuria, fever, renal survival time, and high blood pressure are summarized in Table 2. During the period of follow-up, 34 patients (34.7%) progressed into ESRD. Mean age at diagnosis was 14.2 years (range 1.9-38.9) and 52 (53.61%) were female. Columns 2 and 3 show counts for the various feature values; however, note that the statistical tests utilize time, as well as outcome, to evaluate the differences between groups. The female patients who were diagnosed at age 13 and younger have lower survival probabilities ($P < 0.05$). There were no significant differences between the 2 groups for any of the features used individually.

Figure 1 shows the Kaplan-Meier Dense Deposit Disease renal survival curves for the entire dataset. The median time to ESRD from diagnosis was 11.15 years and 5, 10, 20-year probabilities of renal survival were around 67%, 60%, and 50%, respectively.

Individual features were also evaluated by examining the significance of their coefficient in the Cox regression model. The results shown in Table 3 validate the conclusion that none of the features are individually significant.

Figure 2 shows Kaplan-Meier survival curves for the group of females diagnosed as age 13 or younger vs. everyone else. The 5-year renal survival probability was around 48% for the young female group (dashed line) and around 77%($p < 0.05$) for the others.

Figures 3 and 4 show the results for the Cox regression and SVM models, respectively. Although both methods can identify good prognosis and poor prognosis groups, the SVM shows a more clear separation.

4. DISCUSSION

This study obtained information directly from individuals with DDD, which was also termed MPGN type 2, as described in [6]. This is different from other research data that is obtained by healthcare professionals. The quality of data

Table 1: The MPGN dataset

Attribute	*Values*	*Description*
Gender	Female,Male	
Diagnosed age	Numeric	
Proteinuria	Yes,No	Protein in urine or not
Hypertension	Yes,No	High blood pressure or not
Hematuria	Yes,No	Blood in urine or not
Fever	Yes,No	Fever or not
ESRD	Yes, No	End-stage renal disease or not
Time	Numeric	Time from diagnosis to either ESRD or last observation

Table 2: Analysis of potential risk factors associated with the development of ESRD.

Attributes	*ESRD (n=34)*	*Non-ESRD (n=64)*	*Wilcoxon Test (p-value)*	*Logrank test (p-value)*
Age(years)(mean,range)	23.65(10.09-52.26)	20.90(3.18-45.68)	0.1646	(t-test)
Diagnosed age(years)(mean,s.d)	12.956(8.80)	14.01(8.607)	0.3116	(t-test)
Female $\leq$ 13	17(17.34%)	16(16.33%)	0.0061	0.0105
All others	17(17.34%)	48(48.98%)		
Diagnosed age $>$ 13			0.1044	0.1080
Yes	10(10.20%)	27(27.55%)		
No	24(24.49%)	37(37.55%)		
Gender(n,%)			0.0656	0.0843
Female	22(22.45%)	30(30.61%)		
Male	12(12.25%)	34(34.69%)		
Proteinuria(n,%)			0.0894	0.2013
Yes	33(33.67%)	55(56.13%)		
No	1(1.02%)	9(9.18%)		
Hypertension(n,%)			0.1448	0.1080
Yes	20(20.41%)	32(32.65%)		
No	14(14.29%)	32(32.65%)		
Hematuria(n,%)			0.7965	0.4997
Yes	28(28.57%)	54(55.10%)		
No	6(6.12%)	10(10.21%)		
Fever(n,%)			0.1756	0.3557
Yes	8(8.16%)	11(11.22%)		
No	26(26.54%)	53(54.08%)		

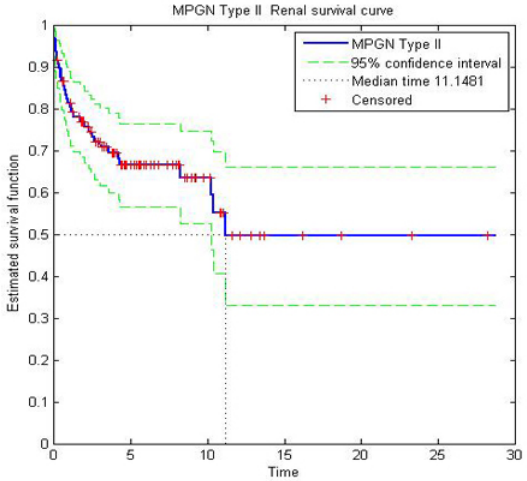

Figure 1: Kaplan-Meier MPGN II renal survival curves. The median time to ESRD from diagnosis was 11.15 years and 5,10,20-year probabilities of renal survival were around 67%, 60% and 50%, respectively. Error bars show standard errors.

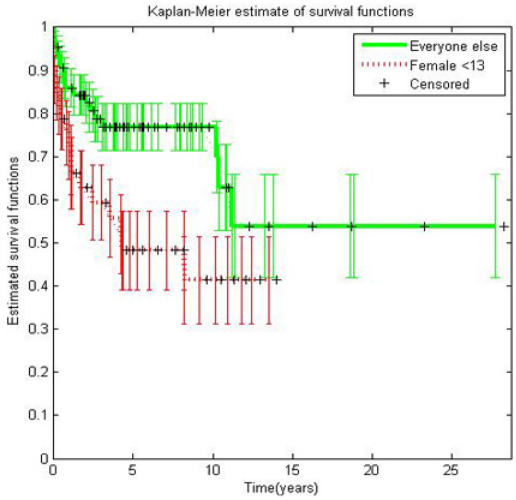

Figure 2: Kaplan-Meier MPGN II renal survival curves for female patients diagnosed at age 13 or younger vs. everyone else.

Table 3: Multivariate Cox regression analysis of variables potentially associated with ESRD.

Attributes	*Regression Coefficient*	*p-value*	*Standard Error*	*Exp(coefficient)*
Gender	-0.6320	0.0878	0.3703	0.5315
Diagnosed age(years)	-0.0320	0.2221	0.0262	0.9685
Proteinuria	1.4938	0.1495	1.0364	4.4541
Hypertension	0.2670	0.4731	0.3722	1.3061
Hematuria	0.7249	0.1285	0.4769	0.4844
Fever	0.1902	0.6634	0.4371	1.2095

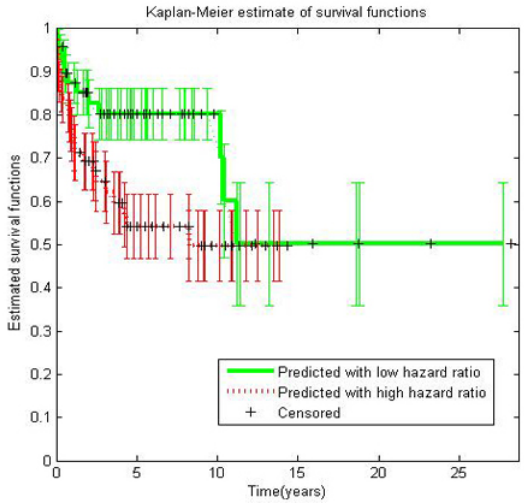

Figure 3: Cox Regression results. The group predicted by CHR to have a higher hazard ratio have significantly worse survival probabilities than those predicted to have a lower hazard ratio ($p < 0.05$).

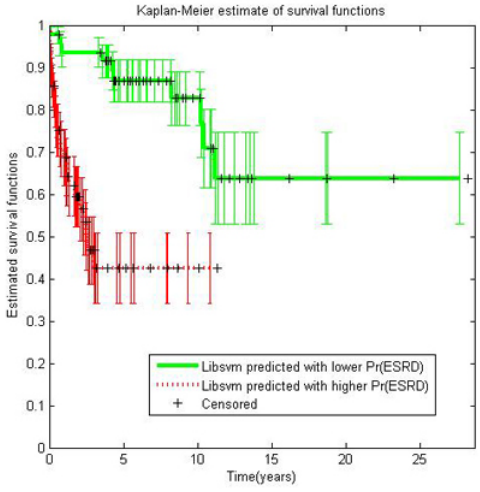

Figure 4: SVM results. Three-year renal survival in the group predicted by the SVM classifier to have a higher probability of fast-onset ESRD have renal survival probability around 40% (dashed line), compared to around 90% for the group predicted to have a lower probability of ESRD ($p < 0.0001$).

depends on the users of a web-based questionnaire. We have 98 patients in our data set; this is a small dataset on which to train and test classifier models. However, it is to our knowledge the largest existing dataset for this rare disease. Future work would include increasing the number of participating patients. This would likely improve the accuracy and significance of the results. However, our results show that based on clinical symptoms the classification approach can determine which patients are more likely to progress to ESRD in a short time. It can guide healthcare individuals to develop more individualized treatment models based on knowledge of patients who are at increased risk of more rapid ESRD. The findings from this study can be used to identify high-risk patients for ESRD. Since current medical prognostic prediction for this disease is largely based on the practitioner's personal experience, this study provides a step toward data-driven clinical decision support. The research team is also incorporating genotype information into the analysis in order to find genetic patterns for high-risk patients.

5. CONCLUSION

Dense Deposit Disease (DDD) is a rare disease with poorly understood prognosis. We used a data mining approach to find important patterns that could be useful to develop clinical interventions and specific individual care plans. These data mining approaches are better than the conventional survival analysis methods for this application. The SVM classifier can be used to predict outcomes and can be applied to identify high-risk patients more reliably than individual features or Cox regression. The results are also generalizable to other diseases.

6. REFERENCES

[1] M. A. Abrera-Abeleda, C. Nishimura, J. L. H. Smith, S. Sethi, J. L. McRae, B. F. Murphy, G. Silvestri, C. Skerka, M. Józsi, P. F. Zipfel, G. S. Hageman, and R. J. H. Smith. Variations in the complement regulatory genes factor H (CFH) and factor H related 5 (CFHR5) are associated with membranoproliferative glomerulonephritis type II (dense deposit disease). *J Med Genet*, 43(7):582–589, Jul 2006.

[2] C. Chang and C. Lin. LIBSVM: A library for support vector machines, 2001.

[3] C.-L. Chi, W. N. Street, and W. H. Wolberg. Application of artificial neural network-based survival analysis on two breast cancer datasets. In *Proc. American Medical Informatics Association Annual Symposium*, pages 130–134, November 2007.

[4] D. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

[5] M. A. Little, P. Dupont, E. Campbell, A. Dorman, and J. J. Walshe. Severity of primary MPGN, rather than MPGN type, determines renal survival and post-transplantation recurrence risk. *Kidney Int*, 69(3):504–511, Feb 2006.

[6] D.-F. Lu, A. M. McCarthy, L. D. Lanning, C. Delaney, and C. Porter. A descriptive study of individuals with membranoproliferative glomerulonephritis. *Nephrol Nurs J*, 34(3):295–302; quiz 303, 2007.

[7] C. Ponticelli and R. Glassock. *Treatment of Primary Glomerulonephritis*. Oxford University Press, 1997.

[8] W. N. Street. A neural network model for prognostic prediction. In *Proc. 15th Int. Conf. on Machine Learning*, pages 540–546, 1998.

[9] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.