Enhanced Laboratory Diagnosis of Human *Chlamydia pneumoniae* Infection through Pattern Recognition Derived from Pathology Database Analysis

Alice Richardson[1], Simon Hawkins[2], Fariba Shadabi[1], Dharmendra Sharma[1], John Fulcher[3]
and Brett A. Lidbury[4]

[1]Faculty of Information Sciences and Engineering, University of Canberra
[2]Faculty of Health, University of Canberra
[3]School of Computer Science & Software Engineering, University of Wollongong
[4] Centre for Biomolecular & Chemical Sciences, Faculty of Applied Science, University of Canberra

Abstract. This study focuses on pattern recognition in pathology data collected from patients tested for *Chlamydia pneumoniae* (Cp) infection, with co-infection by *Mycoplasma pneumoniae* (Myco) also considered. Both Cp and Myco are microbes that cause respiratory disease in some infected people. As well as the immunoassay results revealing whether the patient had been infected, or not, an extensive range of other routine pathology data was also available for each patient, allowing the analysis of associations between a positive immunoassay laboratory result for Cp or Myco, and a range of tests for biochemical and cellular markers (e.g. liver enzymes, electrolyte balance, haematological indices such as red/white cell counts). Decision trees and logistic regression were used to enhance laboratory diagnosis of these respiratory infections via the formulation of association rules derived from immunoassay results and associated pathology data.

Keywords: Data mining, decision tree, logistic regression, human pathology

1 Introduction

This study is based on the application of standard statistical methods to clinical pathology databases to recognise patterns that will identify small groups in the population that are at higher relative risk of infection by the respiratory pathogens *Chlamydia pneumoniae* (Cp) and/or *Mycoplasma pneumoniae* (Myco). Once rules have been discovered, they can be used to develop a system of electronic alerts that will identify patterns of pathology results that identify a patient at a high relative risk of a true infection through enhancing the predictive capacity of combined laboratory test results.

Chlamydia species are obligate intracellular bacteria responsible for atypical pneumonia, sexually transmitted disease or eye infection in humans. [1]. *Mycoplasma pneumoniae* is also an obligate intracellular bacteria, and unusual in comparison to other bacteria in that it lacks a cell wall and has a less sophisticated internal organisation. *Mycoplasma* (Myco) infection is associated with respiratory disease in humans, which can also lead to extrapulmonary complications and hospitalisation in some cases [2].

The aims of this study were:

- To enhance the predictive power of routinely performed diagnostic pathology laboratory results by combining multiple tests on individual patients and taking into account the reliability of each test as an individual disease marker.
- To discover patterns (grouping of pathological assays) within routinely collected diagnostic pathology tests and use these to identify the characteristics of patient groups at higher relative risk of infectious respiratory diseases.

2 The Data

The data used in this study was obtained from de-identified data extracted from ACT Pathology Laboratory databases (The Canberra Hospital). Variables with missing values were removed as the analysis methods used for this study required complete data. Patients with multiple entries (for the same lab test) were removed to ensure that the analysis related solely to single-visit patients. The selected variables used for building the models are summarised in Table 1.

M. Chetty, S. Ahmad, A. Ngom, S. W. Teng (Eds.): PRIB2008,
Supplementary Conf. Proc., pp.227-234, 2008.
(c) PRIB 2008

Table 1. Pathology laboratory variables used in model building for the study of *Chlamydia pneumoniae* infection in patients who were tested by ACT Pathology, the Canberra Hospital

Response Variables	**Description**
IgA	Antibody specific for Cp; presence of antibody means previous Cp infection
Myco	Serum IgG titre against the microbe Myco; presence means previous Myco infection
Predictor variables	**Description**
Age	Patient's age in years
Sex	Patient's gender
Na^+ (defined as Nas in section 3.1)	"Salt" in the serum. With potassium (K^+), chloride (Cl^-) etc., known as "electrolytes". Helps with assessment of physiological water balance and kidney function
Urea	Waste product of nitrogen metabolism; measures kidney function
Crea	Creatinine (also a waste product of nitrogen) that helps with the assessment of kidney function
ALT	Alanine aminotransferase (an intracellular enzyme released in high concentrations after liver damage)
GGT	Gamma-glutamyl transferase (an intracellular enzyme like ALT - also relevant to liver function)
Hb	Haemoglobin (Oxygen carrying pigment of red blood cells)
RCC	Red cell count (Red cells also called red blood cells or erythrocytes)
MCV	Mean corpuscular volume (average red cell volume - helps diagnosis of anaemia)
Hct	Haematocrit (formerly known as "packed cell volume") is a measure of cell density in a blood sample. It helps in the diagnosis of certain anaemias.
RDW	Red cell distribution width
MCH	Mean corpuscular haemoglobin (average concentration of Hb per red cell)
MCHC	Mean corpuscular haemoglobin concentration (a parameter calculated from MCH. Like MCH, MCV and RCC, assists with the diagnosis of anaemia)
Total WCC	White cell count (White cells such as lymphocytes and monocytes form the cellular component of the immune response. Elevated with infection or allergy)
Plt	Platelets (which are cell fragments involved in blood clotting)

3 Enhanced Predictors for Diagnosis of *Chlamydia Pneumoniae*

Two standard statistical methods were applied to the data: decision trees [3] and logistic regression. These were chosen on the basis of their widely-studied statistical properties and the clear indications they give of which variables are important in the finals decision-making process. Both these methods require a complete data set with no missing observations for accurate analysis and pattern recognition. This requirement, along with the data pre-processing mentioned in Section 2, reduced the data set to 1495 individuals, all with the complete set of response and predictor variables defined in Table 1. On the basis of the value of IgA, there were 649 individuals with a clear positive serological test for Cp, 82 indeterminate results and 764 clear negatives. The basis of deciding on whether a result was clear positive, negative or indeterminate, was guided by internal ACT Pathology quality control protocols and reference ranges for Cp serology.

3.1 Decision tree

A decision tree model was fitted using the binary recursive-partitioning (RPART) algorithm in R 2.6.1. A pattern recognition model in the healthcare sector ideally needs to be able to deal with ever increasing amounts of data, especially in the complex field of bioinformatics. Decision tree models are able to accept numerous input variables and provide good explanation capability on how to recognise a pattern. Figure 1 shows the tree produced for all 1495 patients: no pruning was required. The numbers in the ovals and rectangles are the predicted outcome for patients at that node (1 = negative, 2 = indeterminate, 3 = positive). Table 2 shows a summary of the main path through the tree, with associated probabilities.

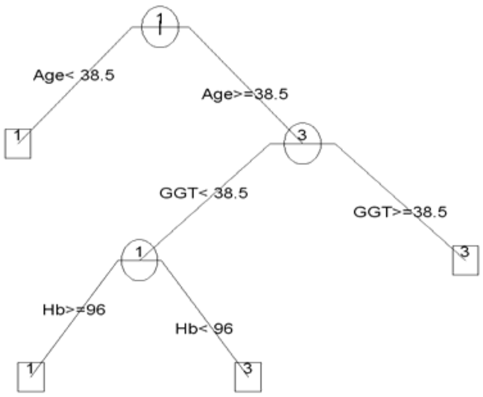

Fig. 1. Decision tree for all patients (n = 1495) tested for Cp infection

The main path through the decision tree is to the right. It is interpreted as follows. There is a 43% chance of a patient testing positive for Cp. If we also learn that the patient is ≥ 38.5 years old, the chance that the patient will test positive is now 50%. If we also learn that the patient has GGT ≥ 38.5U/L, the chance that the patient will test positive increases again to 55%. Table 2 shows these increases in percentage. Other paths in the tree can be interpreted in a similar fashion.

Table 2. Positive tests in the decision tree for all Cp patients

Rule	Percentage testing positive under this accumulation of rules
Initial group	649/1495 = 43%
Age ≥ 38.5	523/1044 = 50%
GGT ≥ 38.5	341/618 = 55%

Given the importance of age as a determining variable in Cp infection emerging from this decision tree, we next fitted two decision trees for older and younger patients separately. Figure 2 and Table 3 show the results for patients older than 38 years. The tree was pruned using a complexity parameter of 0.01.

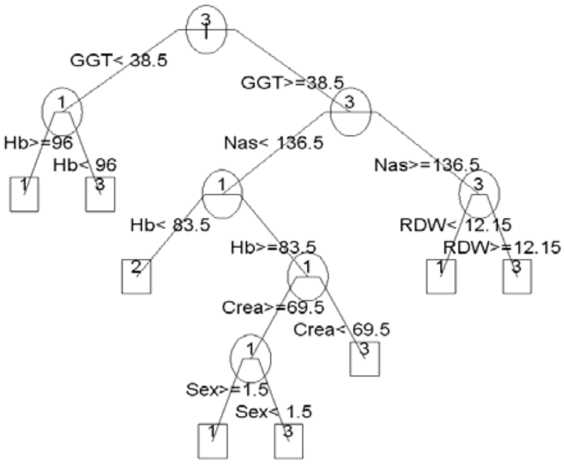

Fig. 2. Decision tree for patients older than 38 years (n = 1044) tested for Cp infection.

The main path through the decision tree is to the right once again. It is interpreted in a similar fashion to Figure 1, and the increases in percentage is shown in Table 3.

Table 3. Positive tests in the decision tree for Cp patients over 38 years of age

Rule	Percentage testing positive under this accumulation of rules
Initial group	523/1044 = 50%
GGT ≥ 38.5	341/618 = 55%
Na+ ≥ 136.5	262/436 = 60%
RDW ≥ 12.15	261/426 = 61%

Figure 3 and Table 4 show the results for patients 38 years of age and under. The tree was pruned using a complexity parameter of 0.02.

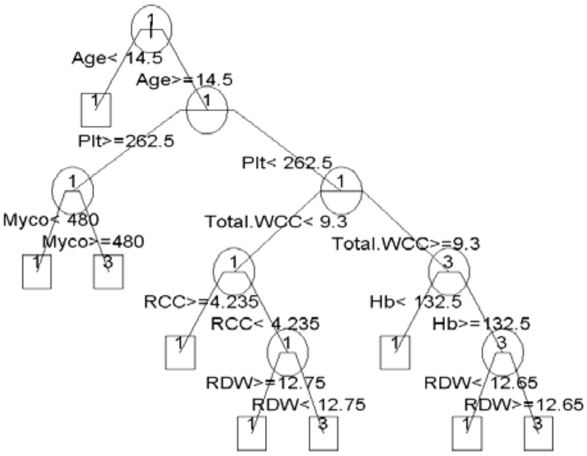

Fig. 3. Decision tree for patients 38 years of age and under (n = 451) tested for Cp infection

Table 4. Positive tests in the decision tree for Cp patients 38 years of age and under

Rule	Percentage testing positive under this accumulation of rules
Initial group	126/451 = 28%
Age ≤ 14.5	114/360 = 32%
Plt < 262.5	72/179 = 40%
Total WCC ≥ 9.3	39/73 = 53%
Hb ≥ 132.5	35/55 = 64%
RDW ≥ 12.65	31/41 = 76%

The main path through the decision tree is to the right once again. It is interpreted in a similar fashion as Figures 1 and 2.

3.2 Logistic regression in the Chlamydia data set

Logistic regression was used to predict the likelihood (%) of a clear positive for Cp infection (as informed by serology/immunoassay - regular quality control by ACT Pathology ensures accurate Cp immunoassay data, within two standard deviations). Since logistic regression requires two outcomes (positive or negative), indeterminate IgA was combined with negative IgA.

The logistic regression modelling was done in R 2.6.1. The final model, in terms of log odds of being IgA positive, is:

$$\log(P(\text{IgA positive})/1 - P(\text{IgA positive})) =$$
$$-2.72 + 0.02\ \text{Age} - 0.42\ (\text{Sex} = \text{M}) + 0.001\ \text{ALT} + 0.017\ \text{MCV}.$$

In general, the coefficients in the logistic model estimate the change in the log odds of IgA positive when a variable is increased by 1 unit, holding all the other variables in the model fixed. The exponentials of the coefficients estimate the change in the odds-ratio of IgA positive when a variable is increased by 1 unit, holding all the other variables in the model fixed.

This means that;

a. When all variables are set to 0, and holding all other variables constant, the odds of an IgA positive result is 0.066
b. For every extra year of age, there is a 2% increase in the odds of Cp IgA positive
c. For males compared to females, there is a 34% decrease in the odds of Cp IgA positive
d. For each extra unit of ALT, there is a 0.1% increase in the odds of Cp IgA positive
e. For each extra unit of MCV, there is a 2% increase in the odds of Cp IgA positive

The *p*-values for the coefficients are shown in Table 5.

Table 5. Coefficient *p*-values for variables included in logistic regression model of Cp infection

Variable	*p*-value
Intercept	0.0008
Age	0.0000
Sex = M	0.0000
MCV	0.0524
ALT	0.0684

The order of significance of the variables is Age, Sex, MCV and ALT. This correlates broadly with the decision tree results, where Age was the first variable on which decision-making was based, and a liver function variable was also important (GGT in two of the decision trees and ALT in the logistic regression.)

The area under the ROC curve (Figure 4) is 0.6424. This suggests that the model is able to explain approximately two-thirds of the variation in the data. This is not a high value but is nonetheless an improvement over the baseline value of 50%.

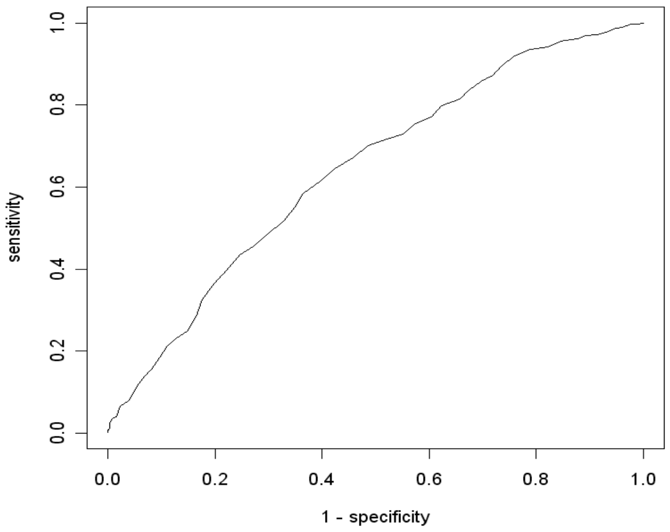

Fig. 4. Receiver Operating Characteristic (ROC) curve for the logistic regression model of Cp infection

4 Discussion

The logistic regression shows Age, Sex, ALT and MCV as key markers with which to predict a positive serology result for Cp. A limitation of this study was that apart from Myco serology, there were no results to indicate whether the patients were infected, or not, with other respiratory pathogens, for example influenza virus, rhinoviruses, coronaviruses, *Streptococcus sp.* and so on. For the purpose of this report, it is assumed that the patients are not infected with other pathogens (patients who were positive for both Cp and Myco were excluded from the above analyses).

Age was divided as either ≥ or < 15 years of age. Access to larger data sets will allow an enhanced age analysis. In general, as the immune system matures (which should happen around 12-15 years of age), the majority of infectious agents are likely to have been encountered and immunological memory established. For the purposes of this study, this is a generalisation that does not take into account the large genetic diversity in immune genes and immune gene expression for human populations, and hence, variable immune responses between individuals. However, age effect and susceptibility to disease post infection has been recognised, for example, with the arthrogenic Australian alphavirus, Ross River (RRV). Disease post RRV infection is generally only seen in people between 20 and 60 years of age [4].

The liver function test marker, ALT, was implicated in logistic regression analysis as associated with a greater probability of a positive serology result for Cp. Further analysis requires access to additional liver function data to decide whether liver involvement is physiologically associated with Cp infection, and not simply a mathematical association. Analysis of the biomedical literature has not found any comprehensive clinical or pre-clinical studies of Cp infection indicating liver sequelae. However, a case study on a patient suffering from severe Cp-associated disease has reported abnormal liver enzymes and other markers of liver function in routine pathology analysis of this patient's serum [5]. Another study has isolated *Chlamydia trachomatis* from a liver biopsy of a patient who had a 10-month history of chills, fever and abdominal pain. The biopsy result showed inflammatory infiltrates in the portal tracts, and serum analysis detected elevated alkaline phosphatase (ALP) and 5'-nucleotidase [6]. Case study reports have also linked *Mycoplasma* infection to abnormal liver function test results [7, 8]. In the context of this study, further investigations are required with markers of true liver function (e.g. serum albumin, bilirubin) to determine a pathological connection of Cp and Myco infection with liver function.

An unexpected result of the analyses showed MCV and RDW as factors contributing to the enhanced likelihood of a positive Cp serology test result. Explanation for this result can broadly refer to possible anaemia as a symptom of Cp infection or pathology, although it would be expected that other red cell markers would also feature as important in the mathematical results, as MCV results are calculated from other measured red cell parameters (Table 1). The involvement of MCV as a marker for Cp infection can be explained by the results of a study from Al Younes *et al.* [9]. This study reported that Cp can sequester iron (Fe^{+++}) from the infected patient, and that iron availability modulates the course of Cp infection. Iron is found in abundance in the red cell pigment haemoglobin. With the Cp microbe capable of utilising patient iron, it is likely that variability in red blood cell parameters will occur and hence be a positive factor in the enhanced prediction of Cp infection, in addition to Cp antibody testing by immunoassay in the laboratory.

The results of this study provide good indications of biomedical associations for the laboratory diagnosis of Cp infection, leading to enhanced prediction of true positive results for Cp via pattern recognition in routine pathology data. Extended analysis of the data could include comparison of the rules from the methods presented here with results from techniques such as support vector machines [10, 11]. It is possible that such methods may yield stronger results in terms of ROC, for example, but these may come at the cost of transparency of which markers are important in decision-making. Future uses of analyses could also include:

- Use of the discovered clinical rules to develop a prototype expert system to flag patients at higher relative risk of viral diseases.
- Link database analyses to patient gene transcript profiles to further augment the diagnostic power of this strategy.

Future work will also benefit from the provision of extra data for each anonymous patient in the original data set. For example, the analyses thus far have indicated a consistent association of a positive Cp serology result with the routine liver (enzyme) markers alanine aminotransferase (ALT) and gamma-glutamyl transferase (γGT). Whether this association is incidental, or truly reflects an impact of infection on liver function, requires data for additional laboratory markers of true liver function. Given the inflammatory nature of respiratory infection by Cp, the inclusion of data for C-Reactive Protein (CRP) will also be beneficial.

References

1. Valdivia R.H. Chlamydia effector proteins and new insights into chlamydial cellular microbiology. *Curr Opin Microbiol.* 11:53-59. (2008)
2. Waites K.B., Talkington D.F. Mycoplasma pneumoniae and its role as a human pathogen. *Clin Microbiol Rev.* 17:697-728 (2004)
3. Gu L., Li J., He H., Williams, G., Hawkins S., Kelman C. Lecture Notes in Artificial Intelligence 2903/2003 : Advances in Artificial Intelligence: 16th Australian Joint Conference on Artificial Intelligence, Perth, Australia 3-5, 2003. Proceedings. Tamás D. Gedeon, Lance Chun Che Fung (Eds.), Springer-Verlag (2003)
4. Liu C., Johansen C., Kurucz N., Whelan P. (National Arbovirus and Malaria Advisory Committee). Communicable Diseases Network Australia National Arbovirus and Malaria Advisory Committee annual report, 2005-06. *Commun Dis Intell.*, 30:411-429 (2006).
5. Sundelöf B., Gnarpe H., Gnarpe J. An unusual manifestation of *Chlamydia pneumoniae* infection: meningitis, hepatitis, iritis and atypical erythema nodosum. *Scand J Infect Dis.*, 25:259-261 (1993)
6. Dan M., Tyrrell L.D., Goldsand G. Isolation of *Chlamydia trachomatis* from the liver of a patient with prolonged fever. *Gut*, 28:1514-1516 (1987)
7. Grullich C., Baumert T.F., Blum H.E. Acute *Mycoplasma pneumoniae* infection presenting as cholestatic hepatitis. *J.Clin.Micro.*, 41:514-515 (2003)
8. Cunha B. Liver involvement with *Mycoplasma pneumoniae* community-acquired pneumonia. *J.Clin.Micro.*, 41: 3456–3457 (2003)
9. Al Younes H.M., Rudel T., Brinkmann V., Szczepek A.J., Meyer T.F. Low iron availability modulates the course of *Chlamydia pneumoniae* infection. *Cell Microbiol.,* 3: 427–437 (2001)
10. Karatzoglou A., Meyer D., Hornik K. Support vecvctor machines in R. *J. Stat.Software,* 15: [online] http://www.jstatsoft.org/. (2008)

11. Debnath R., Muramatsu M. and Takahashi H. An efficient support vector machine learning method with second-order cone programming for large-scale problems. *Applied Intelligence*, 23: 219-239 (2005).

Acknowledgements. This project was funded by a University of Canberra Interdisciplinary Grant in 2007. The authors would like to thank Gus Koerbin and Carolyn Hawkins of ACT Pathology for access to the data; and the Research Centre for Nursing and Midwifery Practice, The Canberra Hospital for research space.