Anti-phospholipid Antibodies and Thrombosis

Katsuhiko Takabayashi

Chiba University, School of Medicine, Chiba, Japan
takaba@ho.chiba-u.ac.jp

Abstract. The paper gives some preliminary comments on the submissions to Discovery Challenge on Thrombosis data.

1 What is APS ?

Collagen diseases were defined by Dr Klemperer, as characterized clinically by rheumatism features, or pathologically belonged to connective tissue diseases. They were also classified autoimmune diseases because of immunological abnormalities. Collagen disease is not a sole disease, but one category like pulmonary disease which is composed of several diseases. Anti-phospholipid syndrome (APS) is probably the newest one which is categorized as a collagen disease. This disease causes thrombosis (vessel stasis by blood clots), such as myocardial infarction or stroke. However, even if patients have this antibody in their serum, not all patients show thrombosis. Therefore we would like to know the mechanism of how the thrombosis occurs in these patients. Finally it is very important to predict for whom and when thrombosis may occur.

In general, anti-cardiolipin antibody (aCL) (one kind and representative antibody of anti-phospholipid antibodies) has a significant relation to thrombosis. Especially IgG type aCL (aCL IgG) is believed to be strong indicator of thrombosis. In addition, other antibodies such as aCL IgM, aCL IgA have also shown some relation to thrombosis. LAC, PT and APTT are methods to detect the abnormality of blood coagulation, especially significant in these patients. Thrombocytopenia is often accompanied with APS, however the mechanism is unknown. In fact, the phenomenon is not clear yet, and the discrepancy between the test results often confuse clinicians.

2 Data mining technique on APS

2.1 Estimation concept

Thus from current medical knowledge, we can estimate a data mining technique works well or not by observing if it can point out important key factors (aCL, LAC, PT, APTT) related to thrombosis correctly from many variants we provided. Secondly, if they can specify some combinations of several antibodies and/or other variants which reflect thrombotic events, we could comment from a medical view point. On the contrary if it shows some irregular relations from the medical point of view, we can deny them immediately. The hidden truth and rules are lying between significance and nonsense, and it might be difficult for clinicians to decide if they might be true or not. As the time going, it became clear that a few data mining techniques can detect change of some variants in the course of thrombosis from the temporal lab data, and can identify high risk patients who have no history of thrombosis so far.

Now I have just read them and I will make some comments on the results obtained by four applicants only from a medical standpoint.

As I mentioned above, evaluation of the results can be classified as follows:

1. medical sense results (positive control)
2. probable results
3. possible results from the current medical viewpoint
4. uncommentable results
5. non sense results (negative control)

2.2 Medical data set

Medical data set for 1241 patients with collagen diseases and 7 basic laboratory data for aCL for 806 cases were provided. As for the temporal laboratory data, 41 items in 57,543 laboratory tests in 17 years were prepared. Seventy-six patients had some thrombotic events in their clinical course.

2.3 Comments on each case

Ivan Coursac et al mentioned that he could predict the health state from spe-exams and lab-exams in 99.28%. If this were true, it would be very sensational matter in medical field, thus we would like to know further details of the formula of spe-exams and lab-exams. He also mentioned that CNS lupus has a relation to DNA level and IgM type anti-cardiolipin antibody. This fact also sounds interesting to us because

IgM antibodies sometimes causes these kinds of situations due to its general characteristics.

James Cunha Werner and Terence C Fogarty analyzed data by using genetic programming. They also defined the laboratory data as enough to determine the discriminate function with the cost of too much processing evaluation. Their results and predictions could be applied back to individual patients.

Jan Zytkow and Shishir Gupta identified the patterns in a dataset by residing in a Relational Database using contingency tables. They compared their results with those obtained by InfoZoom which showed us very sensible and reasonable results. They mentioned that they obtained the same results and in addition they obtained further results which are reasonable in medical aspects, but not surprising. Alveolar hemorrhage and CNS attacks are not associated with milder attacks, but I designated them to be in the same category of symptoms. I should have made the categorizations clear to avoid such misunderstandings. These kinds of upside down relationships between the causes and results are often observed in the KDD, however in this session they became rare and improved compared to before. On the contrary, ANA pattern analysis has become more interesting compared to before even if other antibodies are often deeply studied. And the estimation that the patients who had severe attacks have more possibilities of other attacks in the future may be proven true. InfoZoom was a very impressive tool for us, the clinicians. If Zytkow's method is similar to InfoZoom from the aspect of pattern visualization, it could be useful to doctors in the near future.

Susan Jensen et al analyzed by using the cross-industry standard process for data mining and reported by explaining the procedure step by step with color pictures. In summary she mentioned that LAC, ANA, U-pro, centromere-type, SSA, SSB,RNP, SM,SCl-70 were strong contributors to predicting the presence or absence of thrombosis. APTT has a high relationship to thrombosis. However, while LAC and APTT are of course related to thrombosis because of the initial definition of APS, other antibodies like SSA, SSB, scl-70 might be a higher rate as positive than healthy persons have even if they have had no relation to APS thrombosis. If she can directly demonstrate that patients who have those positive antibodies can show higher rates of thrombosis than those who lack those antibodies in APS patients, it implies an important message to the clinicians, which she also mentioned other possibilities of thrombosis without aCL antibodies later. Sequential analysis did not show interesting results including her report. It might be difficult to predict the time of thrombosis not after but prior data.

Boulicaut et al applied delta-strong classification rules for predicting collagen diseases. Though they extracted a lot of rules with 100% confidence, most of them are just common sense or nonsense rules from a medical viewpoint. For example, there is a rule that aCL >2.4 and range of aCL IgM from 1.9 to 2.7 and KCT (-) is SLE. In fact lots of persons with high aCL and KCT (-) patients exists not only in SLE but also in other patients, and to decide the range of aCL IgM level dose not make sense. The rule that sex is male and ANA is 0 is Behcet might be not wrong but it is a common sense and is not a specific rule. We would like to know the other rules not written in their report to evaluate.

3 Conclusions

In conclusion, as a data provider, I felt the responsibility to mislead some applicants in the wrong directions by misinterpreting the mutual data relationships. Medical provider's attitude is of course very important if you start data mining. Medical provider's interest but less stubbornness to their medicine might be very important to succeed in this field. In medicine, EBM and prospective studies are strongly emphasized currently. Even this trial is a retrospective approach and the data are not arranged well with many noises, I have confidence that we will obtain a lot of significant evidence from the data mining techniques.