Data Mining Analysis of mechanisms-based classification of musculoskeletal pain in clinical practice

The data from this study are a sample of 464 patients, each one assigned to one of the three types of pain by a group of experienced physiotherapists. The assessment was carried out according to a list of 36 binary clinical indicators ("Present"/"Absent"). These indicators are selected on an expert consensus basis regarding symptoms and features of the patients. Various data mining methods are applied to the data answer such questions as:

1. *Are there any interesting patterns in the presence/absence clinical criteria for lower back pain?*
2. *Do the patients form groups with similar presence/absence clinical criteria? Do these groups have a connection with the clinical pain types?*
3. *Can we use the presence/absence clinical criteria to accurately predict the clinical pain types?*

Data Analysis

The data was inspected to gauge the best approach in applying suitable mining methods.

	X1	X2	X3	X4	X5	X6	X7	X8	X30	X31	X32	X33	X34	X35	X36	assigned.labels
1	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	Nociceptive
2	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	Nociceptive
3	0	1	0	0	1	0	0	1	1	0	0	1	1	0	0	Nociceptive
4	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	Peripheral Neuropathic
5	1	1	0	0	1	0	0	1	0	0	0	0	0	0	0	Nociceptive
6	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	Nociceptive
7	0	0	0	1	0	1	0	0	0	1	0	0	0	0	1	Central Neuropathic
8	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	Nociceptive
9	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	Nociceptive
10	1	1	1	0	1	0	1	0	0	0	0	0	0	1	0	Peripheral Neuropathic

...

Figure 1 Sample format of data

The first 36 columns in the data matrix indicated the absence or presence of a pain type. Each code referred to a pain described in Fig. 2. The *assigned.label* column indicated the clinical assignment in one of three categories Nociceptive (NP), Peripheral Neuropathic (PN) and Central Sensitization (CN).

The 464 rows represented patients surveyed suffering from some form of lower back pain aged 18 years or older. All rows with Null values were removed leaving 425 patients. From this, some basic analysis was performed.

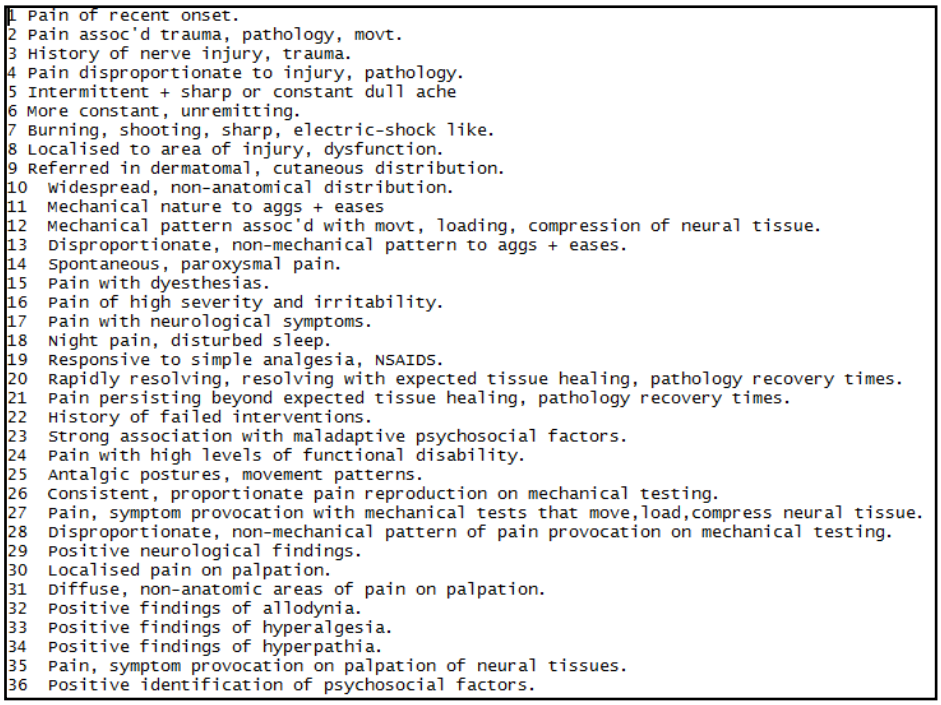


Figure 2 Pain codes

The most five common pain types were tabulated (see Table 1). It was calculated that **88.2%** of patients had at least one of these pain types. The mean number of types of pain per patient was calculated to be **13.** Clearly, the symptoms of lower pain rarely involved a small subset of the 36 pain types but rather a spread across a number of descriptors.

Table 1 Most common pain types

Pain Code	**X5**	**X11**	**X26**	**X2**	**X8**
Pain Description	Intermittent + sharp or constant dull ache	Mechanical nature to aggression + eases	Consistent, proportionate pain reproduction on mechanical testing.	Pain assoc'd trauma, pathology, movt.	Localised to area of injury, dysfunction.
Count	325	325	324	311	298

Association Rule Analysis

An apriori algorithm was used to complete an association rule analysis of the data. An initial setting of **0.65** was set for support combined with a confidence level of **0.9** to reduce the number of rules to 19.

Table 2 Association rules results sorted by lift, min support 0.65

	lhs		rhs	support	confidence	lift
16	{X5,X11}	=>	{X2}	0.6611765	0.9429530	1.288601
11	{X2,X26}	=>	{X11}	0.6847059	0.9831081	1.285603
15	{X2,X5}	=>	{X11}	0.6611765	0.9825175	1.284831
13	{X11,X26}	=>	{X2}	0.6847059	0.9387097	1.282803
18	{X5,X26}	=>	{X11}	0.6729412	0.9794521	1.280822

Although the combinations of symptoms indicated existing relationships when ordered by lift the support assignment of 0.65 only considered symptoms that occurred in **65%** of patients. This represented 7 (19%) of the 36 symptoms. Since the minimum number of symptoms in any patient was 5 (13%) and the mean was 13 (36%) then it was possible that taking a subset of 7 symptoms merely produced rules with symptoms that almost always occurred together due to their nature. To produce potential relationships that were unexpected it was likely that a wider approach might be required. To support this the top rule in Table 2 was considered in detail (with reference to the 2011 paper by Sharp et al. for full description on each pain type).

(X5 Usually intermittent and sharp with movement/mechanical provocation; may be a more constant dull ache or throb at rest)

+ (X11 Clear, proportionate mechanical/anatomic nature to aggravating and easing factors) implied:

(X2 Pain associated with and in proportion to trauma, a pathologic process or movement/postural dysfunction).

A layman's interpretation of these three descriptions would indicate that the pain was well-defined, easily produced and in proportion to the area of injury. In essence, this rule associated a set of symptoms that had clear similarities. For instance, had the symptoms been also associated with:

X30 (Disproportionate, inconsistent, nonmechanical/nonanatomic pattern of pain provocation in response to movement/mechanical testing)

To illustrate the rule relationships in detail a number of graphs were created. To reduce the number of rules further and simplify the output the support was reduced to **0.4**, the confidence remained at **0.9** but the rule size was fixed to **2**. Fig 3 illustrates rules plotted with Support vs. Confidence. The colour indicates lift and red dots have higher lift values. The isolated plot point X8 -> X11 in the bottom left corner has relatively low support (for this set of rules) and also low confidence but also low lift. All lift values greater than 1.0 indicate the relationship between two sides of the rule has more significance than if both sides were completely independent. Larger lift indicates a stronger association.

Fig 4 shows how pain types are connected with circle size indicating support value and colour indicating lift. Again rule X8 -> X11 shows how pain type X8 (Localised to area of injury, dysfunction) is unique in only connecting to pain type X11 (Mechanical nature to aggs + eases).

Fig 5 shows a matrix of circles again with circle size indicating support value and colour indicating lift.

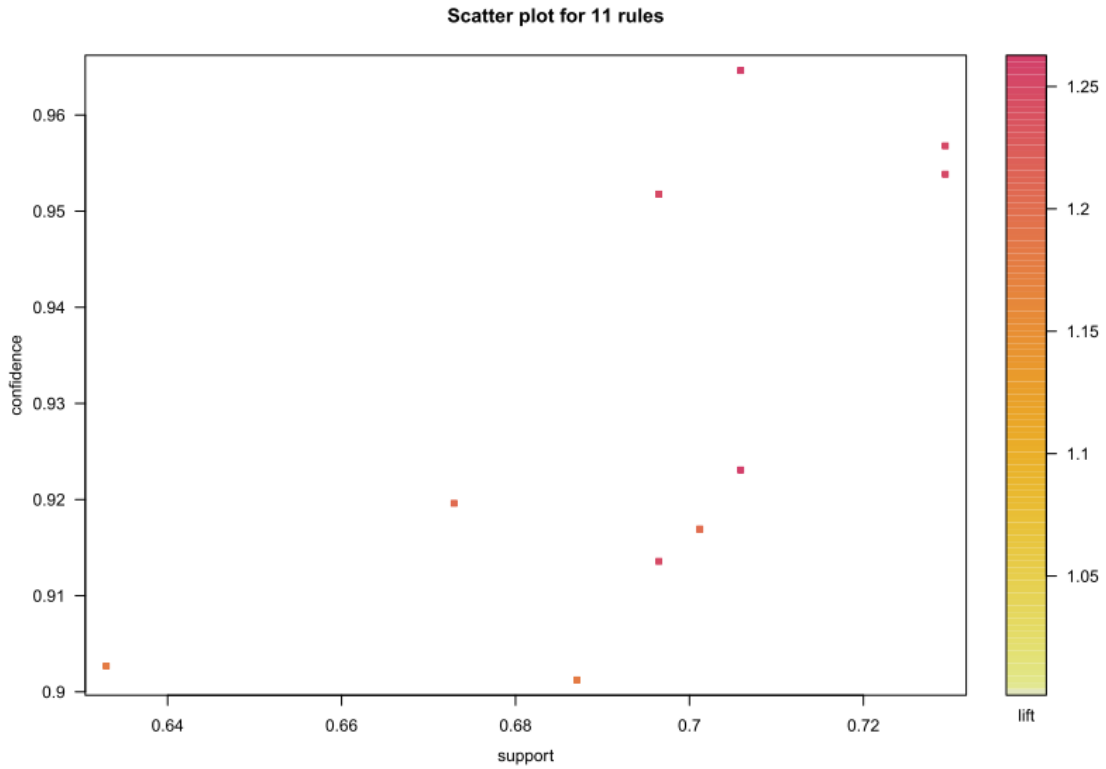


Figure 3 Support vs. Confidence for top 11 rules

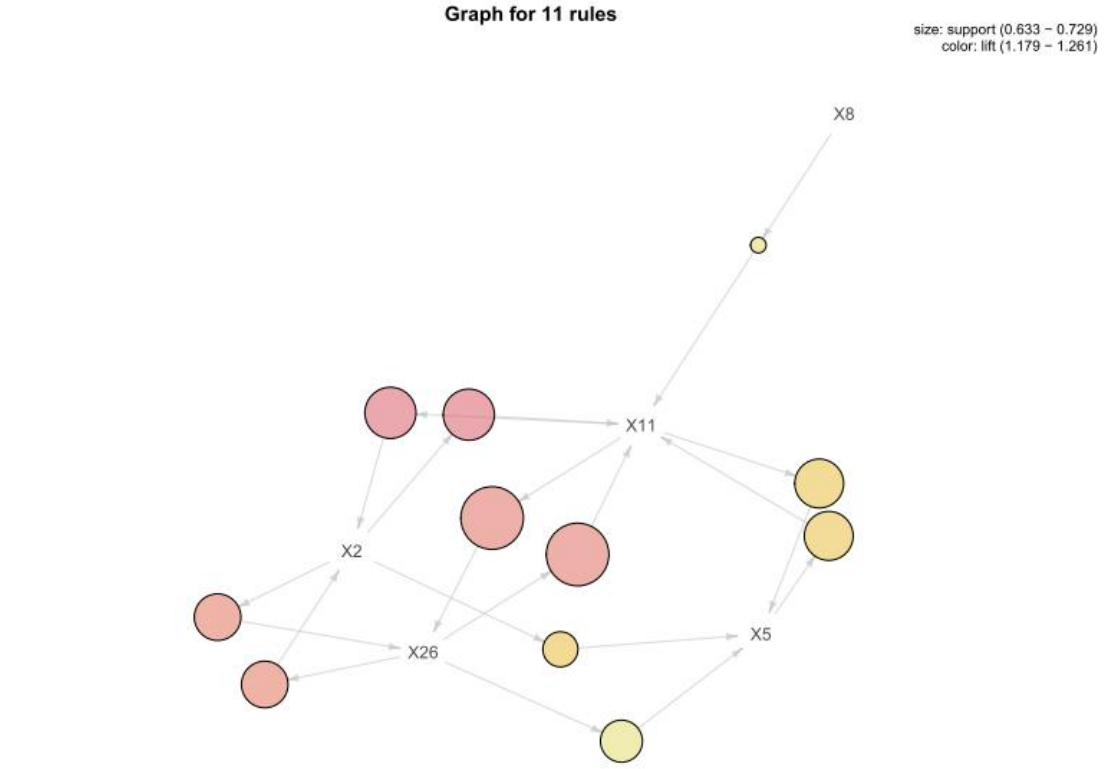
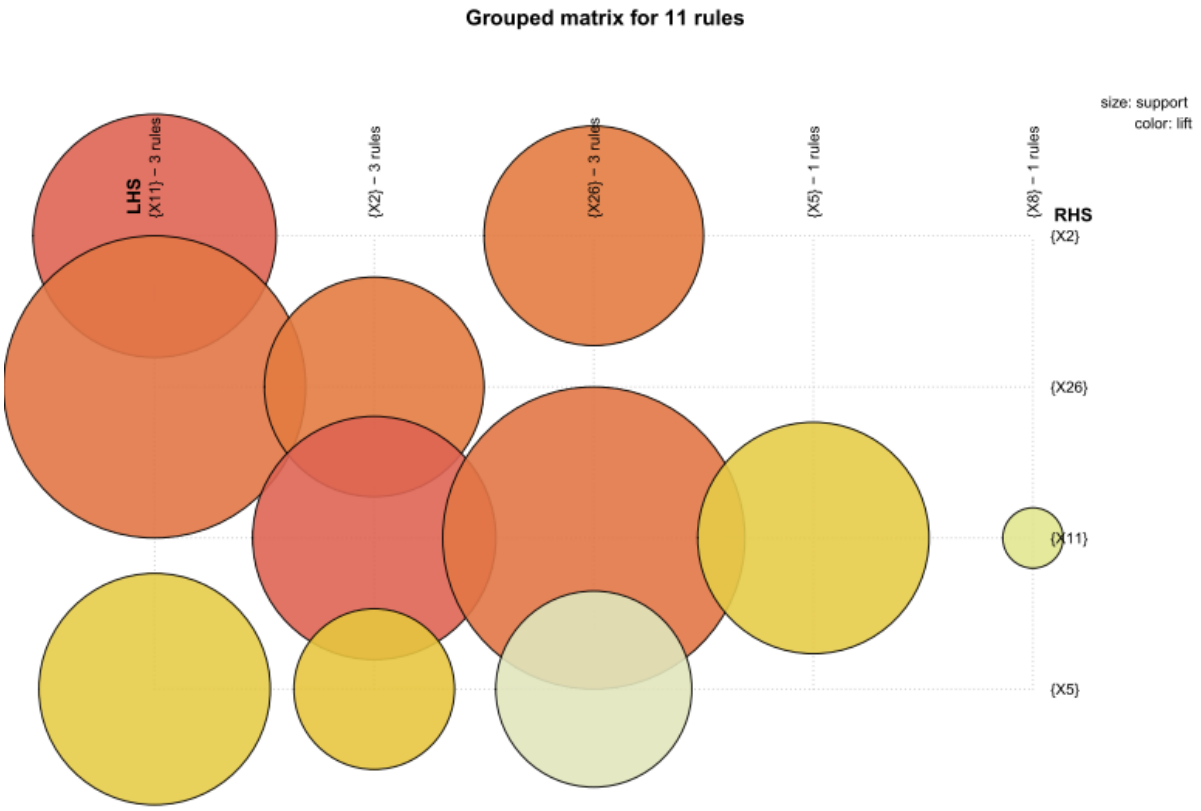


Figure 4 Rule connection graph



Clustering Analysis

Although analysing the data with association rules provides insight into how pain types relate to each other it does not connect the rules to the clinical pain type assignment Nociceptive (NP), Peripheral Neuropathic (PN) and Central Sensitization (CN). This is the ground truth to how a subset of the pain types are classed as one of three overall pain categories. An attempt was made to analyse the 425 patients and group them into clusters based on common characteristics related the pain types experienced. The advantage of this approach is that the results can be compared against the clinical assignment (NP, PN and CN).

A K-medoids analysis was carried on out the dataset with the *assigned.label* column removed. The k-value (number of cluster centres) was varied between 2 and 10 to create an array of average silhouette widths. A high average silhouette width indicates the points within a cluster are tightly associated and similar. A low silhouette width (relative to the rest of the data) indicates the cluster points are dissimilar.

The Sum of Squares calculation indicates the average distance between all points in all clusters and their assigned centre. The lower this value, the tighter each set of points are to each other and the centre. A K-means analysis was performed on the data with the assigned.label column removed. Again, the k-value (number of cluster centres) was varied between 2 and 10 to create an array of sum of squares values.

The Sum of Squares graph indicated that there was significant drop off in average distance when the K went from 2 to 3. This indicated that K = 3 would be an optimal choice. However, the average silhouette width indicated that the optimal K value was 2. To examine this further the silhouette profile for both was produced. It could be seen from both silhouette plots that clustering seemed consistent for cluster 2 (fig 7) and cluster 3 (fig 8). It appeared the algorithm had selected the same cluster for 2 and 3 but when K=3 it had split cluster 1 into two separate clusters. To compare this further a plot of clusters was produced for K=2 and K=3 (see fig 9 and fig 10). Clearly cluster 1 (when K=2) had been split into two when K=3. Although the average silhouette width was higher for K=2 it appeared a potential cluster of points was visible in the top left corner (which fig 10 illustrated). So while the points in cluster 1 were deemed more "similar" when k=2, their average distance from the centroid was higher (fig 5). Combined with this information and the consideration that the assigned.labels assignment was made up of three categories it was decided a measure of accuracy would be made against clusters when K=3.

Using the assigned cluster numbers the assigned.labels were extracted and stored as the ground truth. From this a cross-tabulation was produced. This correlated well with the clinical pain type assignment (NP, PN and CN). Overall accuracy was calculated to **92.5%.**

```
Overall Statistics

               Accuracy : 0.9247
                 95% CI : (0.8954, 0.9479)
    No Information Rate : 0.5529
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.8733
 Mcnemar's Test P-Value : 0.01895

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.9404   0.8544   0.9655
Specificity            0.9263   0.9783   0.9675
Pos Pred Value         0.9404   0.9263   0.8842
Neg Pred Value         0.9263   0.9545   0.9909
Prevalence             0.5529   0.2424   0.2047
Detection Rate         0.5200   0.2071   0.1976
Detection Prevalence   0.5529   0.2235   0.2235
Balanced Accuracy      0.9334   0.9163   0.9665
```

Figure 5 Clustering confusion table results. NP=Class1, PN=Class2, CN=Class3

```
          clust_assign
act_vals   1   2   3
       1 221  13   1
       2   5  88   2
       3   9   2  84
```

Figure 6 Cluster Assignment

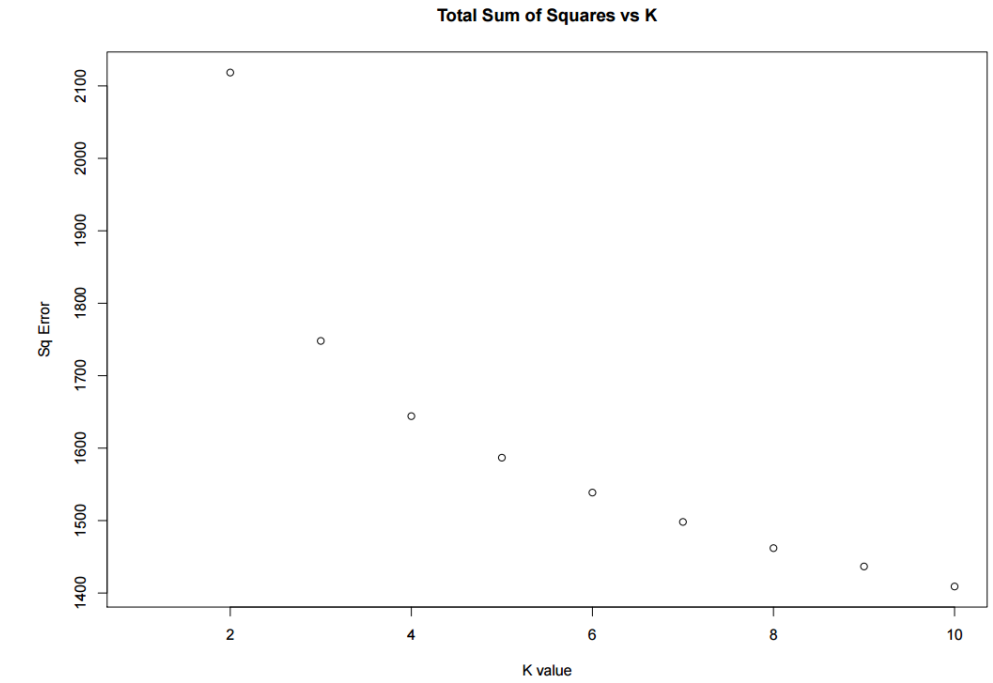


Figure 7 Total Sum of Squares vs K

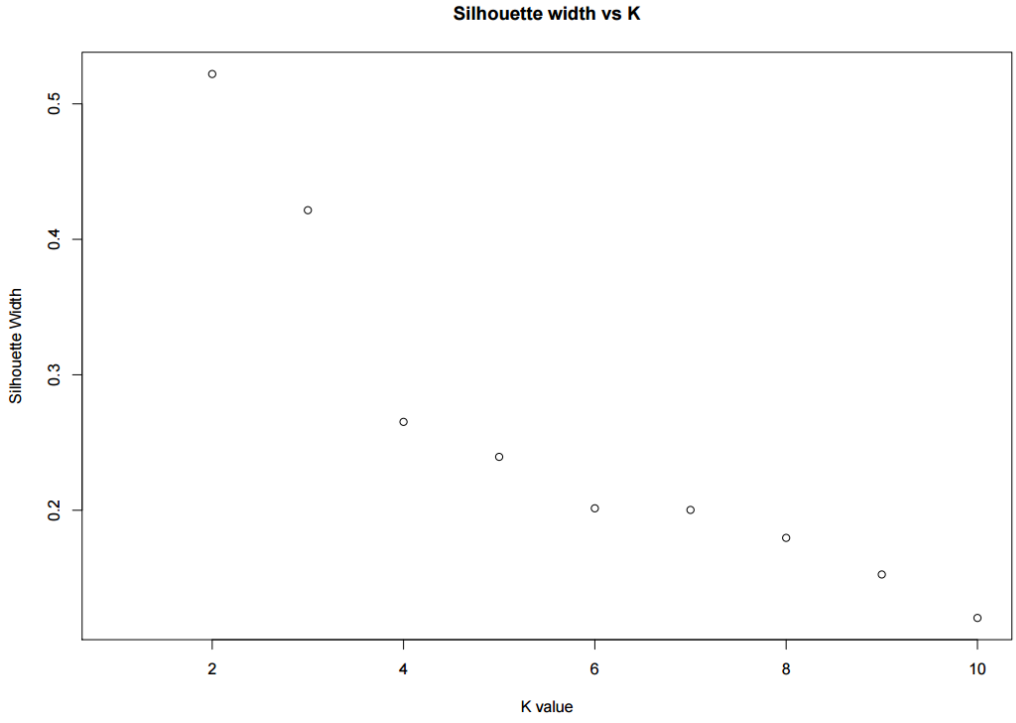


Figure 8 Silhouette width vs K

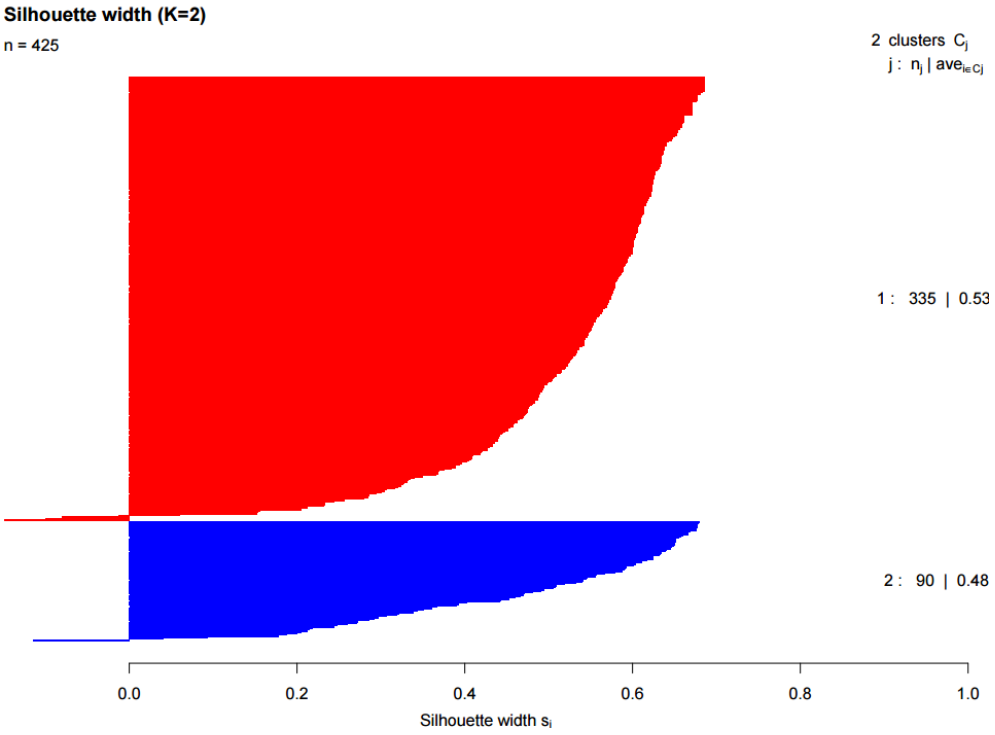


Figure 9 Silhouette plot K=2

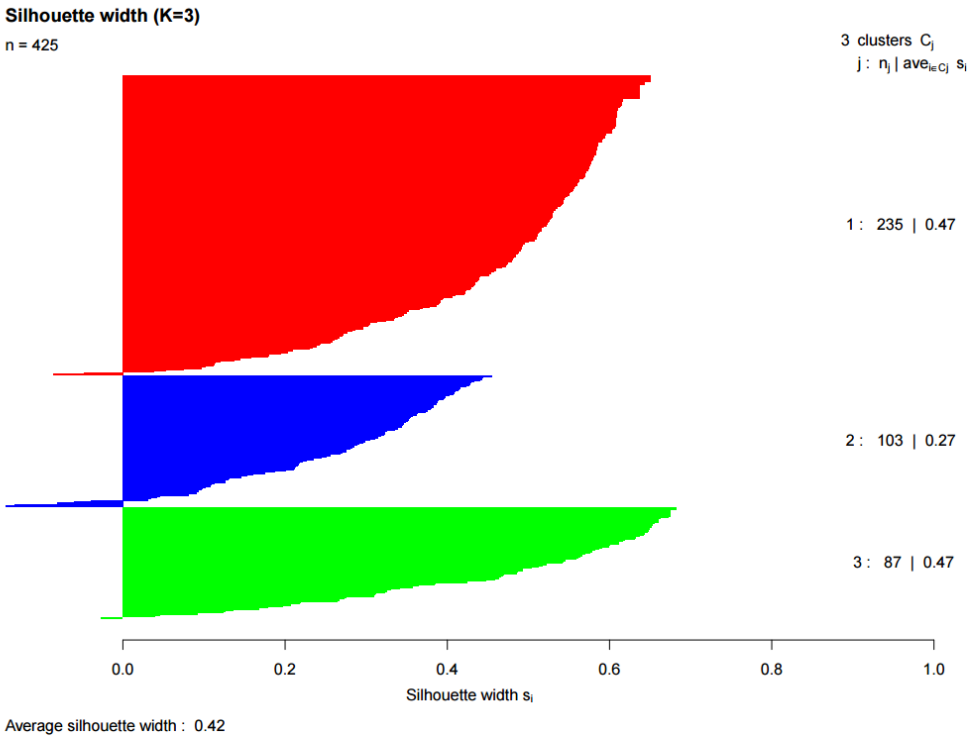


Figure 10 Silhouette plot K=2

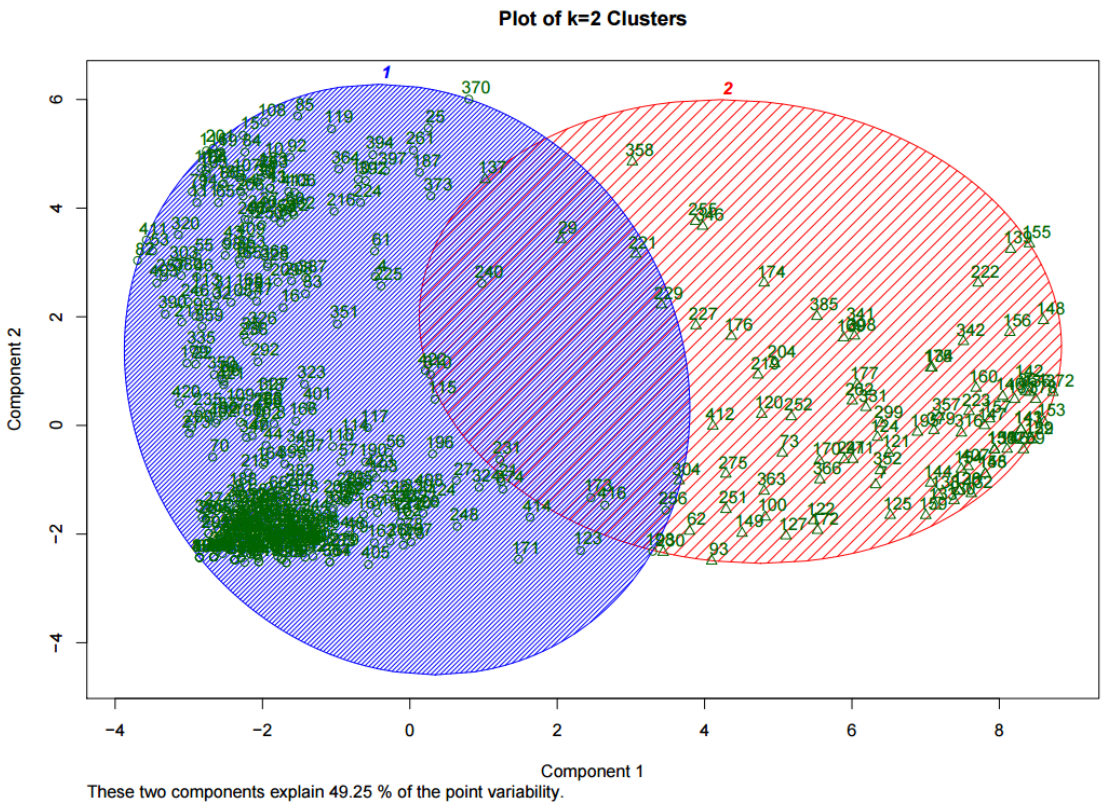


Figure 11 Cluster plot k=2

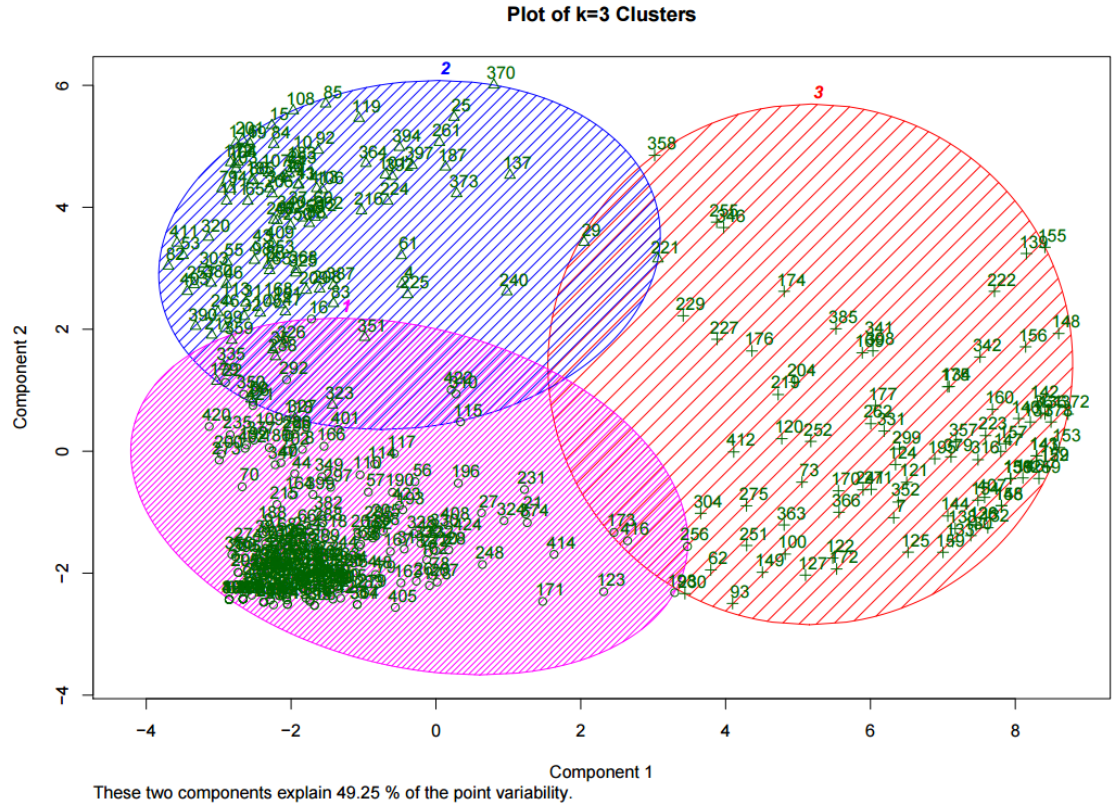


Figure 12 Cluster plot K=3 NP=Cluster1, PN=Cluster2, CN=Cluster3

Logistic Regression and classification tree Analysis

To provide context to previous analysis the paper by Smart et al. (2011) was reviewed to guide further data mining and improve on methods used. The approach in the Smart Paper treated each type of pain assignment Nociceptive (NP), Peripheral Neuropathic (PN) and Central Sensitization (CN) independently. Each classification was taken in turn and assigned '1' for presence of the trait and '0' for absence. A model was created for each classification and optimized by removing the least important variables (determined by posterior probability) until the final model represents variables considered to have the highest relevance to the model fit.

To reproduce this approach the assigned label for each pain classification (NP, PN and CN) was renamed to 1 or 0 in turn for each classification and a model fit was created using the **Random Forest** method. The Random Forest method indicated an OOB estimate error rate of **6.35%.** Using this fit the variable importance was extracted and ordered. The importance was assigned based on the mean decrease in the Gini Coefficient when that variable was removed from the model. A higher mean decrease in Gini Coefficient indicates a higher probability of the importance of that variable. From this ordered list the top ten variables were used to fit a model. (The figure of ten was found to be a value at which adding more variables did not decrease the **AIC** significantly.) This model was used as the basis to perform a logistic regression on the full dataset. Figs 13, 14 and 15 show the logistic regression output. Note, the X rules have an extra 1 added for an unknown reason after the label, eg X81 actually refers to X8.

Taking Figure 15 as an example, this shows the **CN** variable importance, it can be seen that pain type **X13** had a significant positive relationship, ie., if pain type X13 *(Disproportionate, nonmechanical, unpredictable pattern of pain provocation in response to multiple/nonspecific aggravating/easing factors)* was present it was likely the clinical classification would be *Central Neuropathic - Pain initiated or caused by a primary lesion or dysfunction in the central nervous system*. This seemed to correlate with a basic understanding of the descriptions – a central nerve issue would result in multiple, non-specific unpredictable behaviour. The opposite seemed true with a negative correlation was examined. The *absence* of pain type X2 indicated evidence that the clinical classification would be CN. *X2 Pain associated with and in proportion to trauma, a pathologic process or movement/postural dysfunction.* This seemed to describe a pain that was more direct and easier to reproduce mechanically. This was supported by the information from fig 13 for NP that the *presence* of this pain was likely to result in an NP classification. In essence, **pain types that seemed quite dissimilar in their description showed positive and negative importance in one clinical classification and vice-versa in another.**

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1781     1.8857   0.094  0.92474
X81           2.7753     0.8740   3.176  0.00150 **
X121         -2.4785     1.1704  -2.118  0.03421 *
X91          -2.6781     0.8679  -3.086  0.00203 **
X111          1.1630     1.7276   0.673  0.50084
X271         -1.9348     0.7763  -2.492  0.01269 *
X31          -1.4636     0.8836  -1.656  0.09762 .
X21           3.2418     1.0212   3.175  0.00150 **
X131         -2.5468     2.0117  -1.266  0.20553
X61          -5.0244     1.1457  -4.386 1.16e-05 ***
X241         -1.4134     0.6742  -2.096  0.03605 *
```

Figure 13 Important variables NP

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.0560     1.1855  -4.265    2e-05 ***
X121           1.7999     0.9854   1.827  0.06776 .
X91            1.2778     0.8942   1.429  0.15300
X31            2.4493     0.9850   2.487  0.01290 *
X271           2.3459     0.9258   2.534  0.01128 *
X291           0.7198     0.7032   1.024  0.30600
X171           0.0300     0.7889   0.038  0.96967
X81           -2.3991     0.8049  -2.981  0.00288 **
X351           0.1070     0.7027   0.152  0.87896
X151           2.0956     0.8022   2.612  0.00899 **
X281         -19.2819  1578.0096  -0.012  0.99025
```

Figure 14 Important variables PN

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.0060     2.7071  -2.957  0.00310 **
X131          7.4402     2.3801   3.126  0.00177 **
X281          0.7000     1.2382   0.565  0.57187
X21          -3.1124     1.4171  -2.196  0.02807 *
X111          5.9417     2.4433   2.432  0.01503 *
X261         -1.4649     1.1565  -1.267  0.20527
X41           1.1848     1.3346   0.888  0.37467
X311          4.0950     1.4035   2.918  0.00353 **
X101          0.7527     1.1238   0.670  0.50300
X61           2.4718     0.8987   2.751  0.00595 **
X231          0.8653     0.9091   0.952  0.34121
```

Figure 15 Important variables CN

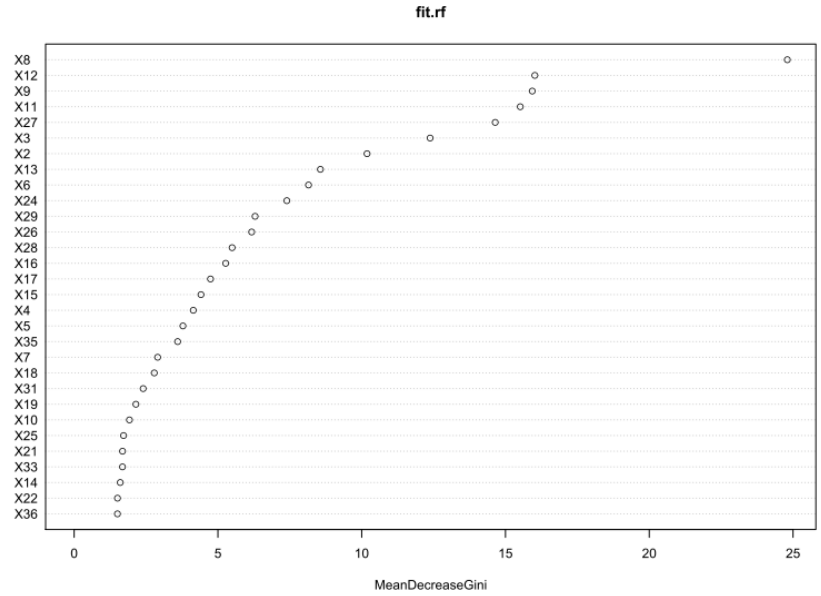


Figure 16 NP Variable importance

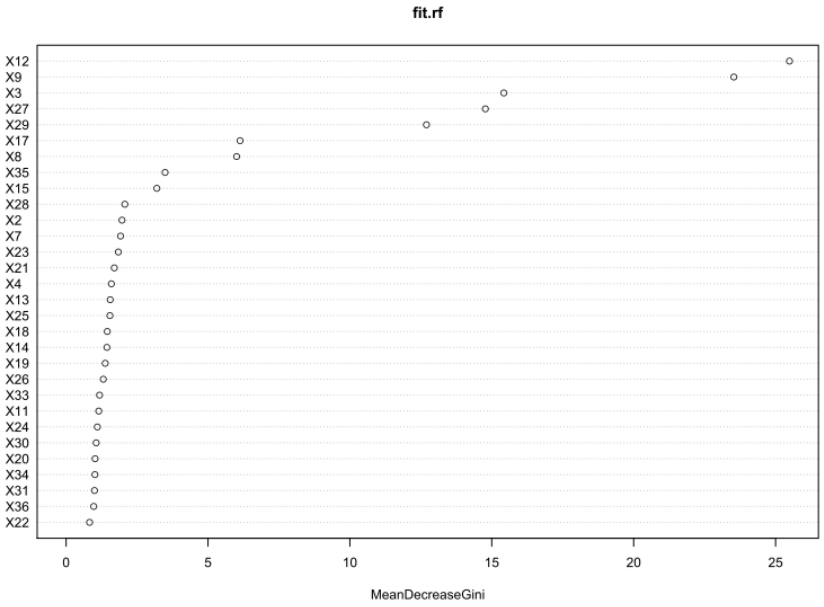


Figure 17 PN Variable importance

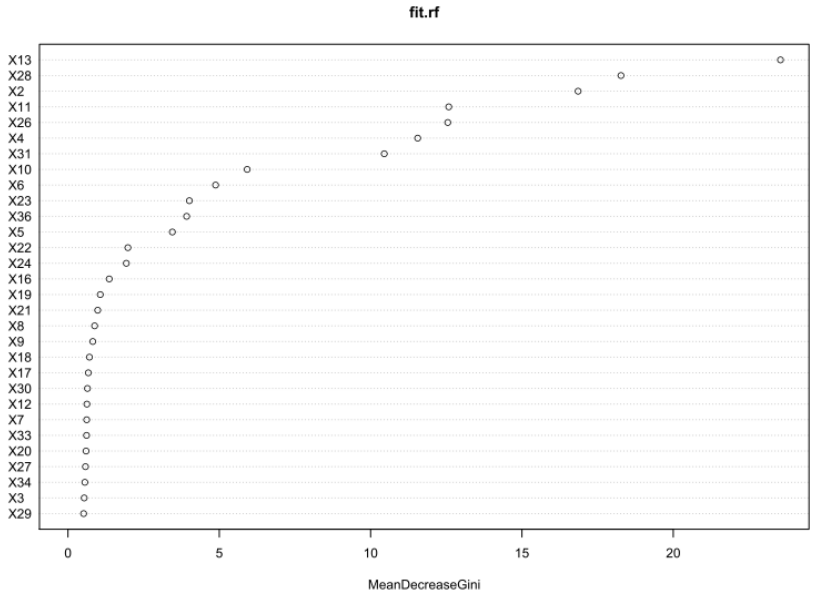


Figure 18 CN Variable importance

Using this model and comparing against the true assignment an estimate on accuracy of the model was then assessed. The possible values of the assignment were '1' for pain classification presence and '0' for absence. A threshold (tau) was assigned across a range between 0 and 1 in 0.001 increments and applied to the predictions from the selected model and counts the number of correct and incorrect predictions. An optimal value of tau was obtained for each model per pain classification (see Table 3). Using the values for Sensitivity, Specificity, Accuracy, PPV (positive predictive value), NPV (negative predictive value) the data provided a graphical representation of the model fit ROC (Receiver Operator Curve), (See Fig) plotting False Positive Rate (FPR) against True Positive Rate (TPR). A model is considered "good" when it plots points closer to the top left corner, ie., the True Positive Rate is high when the False Positive Rate is low. Each of the models for each pain classification indicated strong fit characteristics based on the ROC graphs.

Also, to illustrate the classification tree process carried out (similar to the Random Forest process) a sample tree was created using the partition tree package in R. This shows how decisions on how values are binned are made at each level. In fig 19 it showed that if pain type X13 was present the classification was almost always CN. Similar classification binning was then performed based on X12 and X9 for PN and NP.

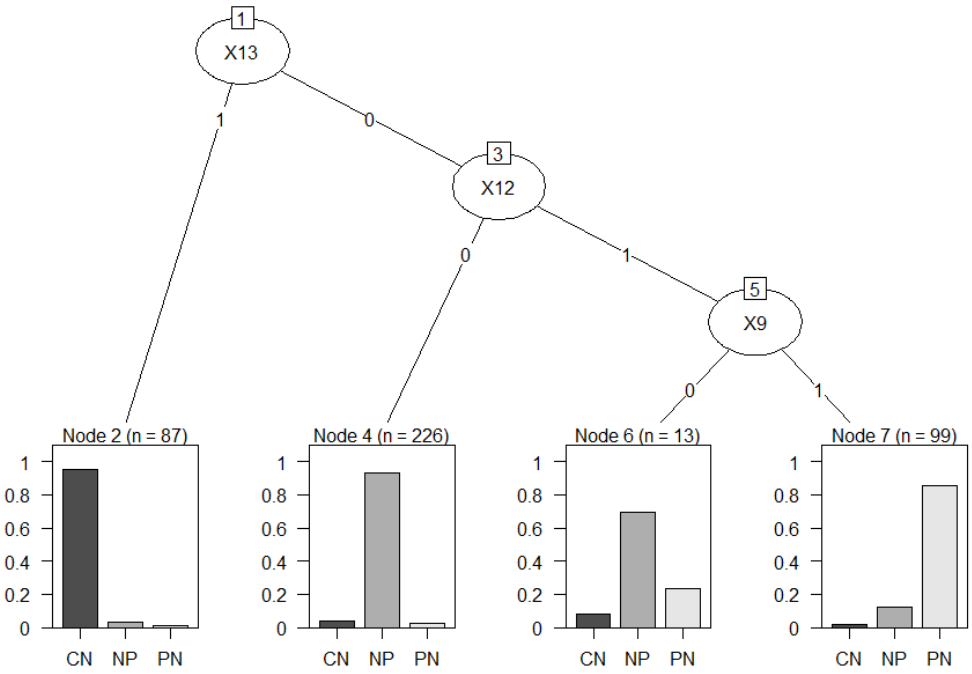


Figure 19 Partitioning sample tree

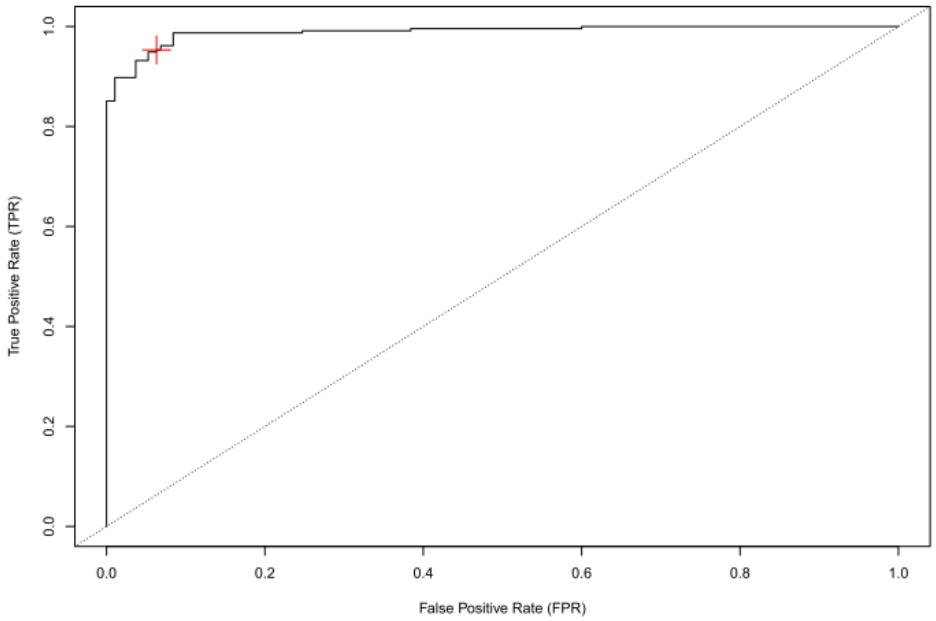


Figure 20 ROC NP

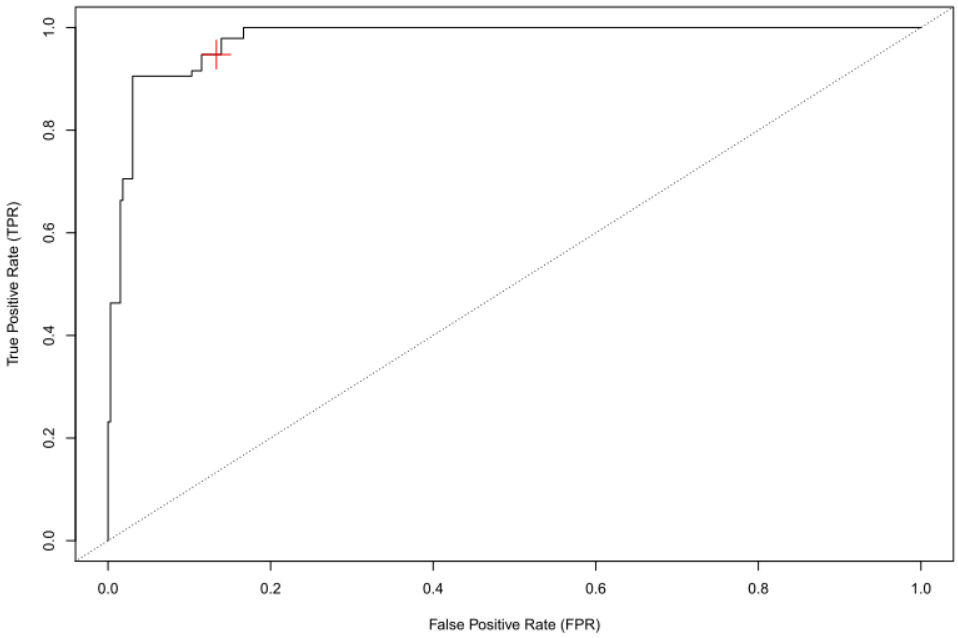


Figure 21 ROC PN

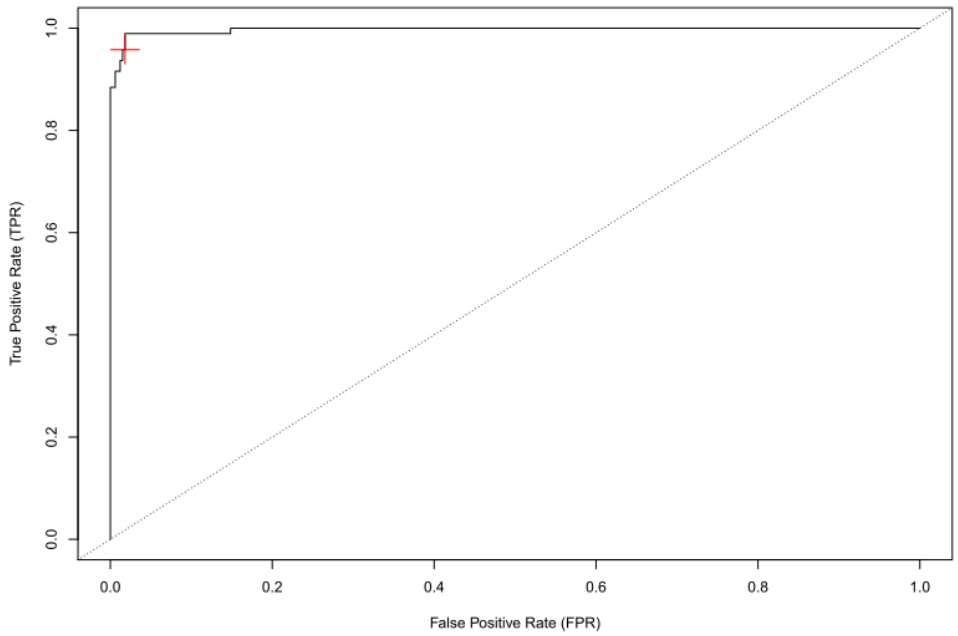


Figure 22 ROC CN

Table 3 Logisitic regression results

	tau	Sensitivity	Specificity	PPV	NPV	Accuracy	FDR	FPR
NP	0.373	0.9532	0.9368	0.9492	0.9418	0.9459	0.0508	0.0632
PN	0.144	0.9474	0.8667	0.6716	0.9828	0.8847	0.3284	0.1333
CN	0.139	0.9579	0.9818	0.9381	0.9878	0.9765	0.0619	0.0182

The results on accuracy from all three models supported the indication that the performance shown from the ROC graphs were in agreement (Table 3). The overall accuracy exceeded the values obtained from the clustering methods. One weakness to the logistic regression approach was the possible bias due to the fit and prediction being extracted from the same set of data. A possible optimization here would be to split the data set into test and training for each model and re-evaluate.

Using the variable importance gained from the Random Forest analysis the association rule output was re-examined. Each pain classification NP, PN and CN had relationships to the presence or absence of pain types (shown in fig 13, 14 and 15). Each pain type had a level of dominance in each clinical classification. These important variables were compared against in a matrix (Table 4) to determine if any rules conflicted. A green box indicated a variable important to NP, a red box indicated a variable important to CN. No colour indicated the variable was not significantly important to any classification. No obvious conflicts could be observed.

Table 4 Association rules compared with variable importance

	lhs	rhs		support	confidence	lift	
6	{X12}	=>		{X27}	0.2494118	0.913793	2.814218
1	{X4}	=>		{X21}	0.2188235	0.939394	1.839827
2	{X20}	=>		{X2}	0.3317647	0.979167	1.338089
3	{X20}	=>		{X26}	0.3317647	0.979167	1.284401
4	{X20}	=>		{X11}	0.3294118	0.972222	1.271368
5	{X20}	=>		{X5}	0.3270588	0.965278	1.262286
14	{X11}	=>		{X2}	0.7058824	0.923077	1.26144
13	{X2}	=>		{X11}	0.7058824	0.96463	1.26144

Conclusions

Data mining analysis largely agreed with the clinical assignment of each pain grouping, NP, PN and CN with a high degree of accuracy. Returning to the original questions posed:

- *Do the patients form groups with similar presence/absence clinical criteria? Do these groups have a connection with the clinical pain types?*

It was shown from clustering methods that groups of patients with similar pain types could be found to have common profiles of presence and absence of the same symptoms. This was supported with strong agreement when the number of target clusters was set to three and then compared against the three clinical pain assignments NP, PN and CN. The accuracy indicated **92.4%** in prediction ability supporting the suggestion that these clusters related to the clinical pain types they were assigned by medical staff. This grouping assignment was further supported by the low Out-of-Bag error rate from the Random Forest of just **6.35%.**

- *Can we use the presence/absence clinical criteria to accurately predict the clinical pain types?*

It was shown in clustering and logistic regression that it is possible to predict clinical pain types with a high degree of accuracy given the list of symptoms provided. Logistic regression provided an

overall accuracy rate of **93.57%**. It was notable that both approaches provided high accuracy despite very different analysis methods. Clustering methods treated the classifications with 3 potential assignments, NP, PN and CN while logistic regression converted the assignments to either 0 or 1 depending on the model assessed NP, PN and CN and then combined the average accuracies at the end.

Finally, the methods applied showed favourable results when compared against the original 2011 paper by Smart et al. The authors of the paper showed that **93.73%** accuracy could be achieved using a modified logistic regression approach.

One weakness discussed in the paper also applied to methods used here. It was observed that potential existed to improve the clinical pain type assignment:

"In the absence of a "diagnostic" gold standard by which to determine mechanisms-based classifications of pain, patients were necessarily classified on the basis of a "reference standard" of "experienced" clinical judgement. The robustness of the reference standard may have been improved if patients' pain had been classified on the basis of a unanimous agreement after independent assessments by 2 (or more clinicians)."