CLRC—TR—08—04

Identification of proteomic biomarkers in the UKCTOCS heart disease data set

B. Burford, A. Tiss, S. Camuzeaux, J. Ford, A. Gentry-Maharaj, U. Menon, I. Jacobs, D. Devetyarov, Z. Luo, I. Nouretdinov, V. Vovk, J. Timms, R. Cramer, A. Gammerman

Abstract

Using heart disease data collected in the UKCTOCS trial we show that several peaks exist which carry some statistically significant information up to 23 months in advance of the date of death of the patients. Our study identifies two possible 'late detection' biomarkers at 4055Da and 4211Da and one one early detection biomarker at 5338Da. We show that combinations of these peaks can produce a more general rule that carries information over the whole 23 month range.

1 Data Introduction

We present here the analysis of a subset of the UKCTOCS data set. The samples in this study have been chosen for the analysis of heart disease.

The data set consists of heart disease (HD) case samples each with two matched controls. The data set contains 561 samples: 187 HD cases and 374 control samples from healthy women.

The numbers in the description above were the data set as it stood after some initial basic filtering was performed. This filtering removed triplets under the following conditions:

- If one of the samples in the triplet did not have a spectra,
- ...or, if only one matched control was provided.

2 Setting of the problem

Our goal is to select the HD sample from a collection of 3 samples which we call a triplet. In each triplet exactly one sample is from a HD patient, the other two samples are matched control samples from healthy people (controls).

The 561 samples extracted from the data are divided into 187 such triplets. Each cancer sample has a time associated with it, which is the amount time before the time of death in months. We denote a cancer sample by HD_i, for the patient i. Similarly we denote the two corresponding matched controls by $M1_i$ and $M2_i$. We denote a triplet by

$$\tau_i = \{\mathrm{HD}_i,\ M1_i,\ M2_i\}.$$

Each τ_i has a time stamp that corresponds to the amount of time before the time of diagnosis for the cancer patient with sample HD_i, we will denote this time stamp by $T_i > 0$. In our analysis we will be interested at looking at a number of *time slots*, each time slot is defined by a starting point and a window size, for example, for a window size of θ months and a starting point of $t \geq 0$ months we will be interested in all samples i, such that T_i is in between t months before the original time of diagnosis and $t + \theta$ months before the original time of diagnosis. For convenience we will often refer to a number of months before the original time of diagnosis as a number of months 'in advance'. We denote a collection of triplets based on t and θ as $S_{t,\theta}$.

Our goal is to find decision rules that can successfully select the HD sample from a triplet, based on the intensity of a single peak or a linear combination of several peaks. Our class of decision rules is defined as follows: For each sample, in a triplet, we calculate a linear combination:

$$w \log I(p_1) + v \log I(p_2) \tag{1}$$

where w and v are weights and $I(p)$ is the intensity of peak p. Our decision rule is thus: for each element in $\tau_i \in S_{t,\theta}$ calculate the value for (1) and select the HD sample as the element with the highest of these values. In the initial algorithm described in this report we will consider the case where p_2 is absent or fixed. The extension of the algorithm from the case described to the case where we have both p_1 and p_2 is trivial.

When we apply a combination of weights (w, v) and a selected peak, p, to this procedure for a single triplet, τ_i, the resulting error will be defined by

$$\mathrm{err}(w, v, p; \tau_i) = \begin{cases} 0 & \text{if the choice of HD sample was correct} \\ 1 & \text{otherwise.} \end{cases}$$

Also for a set of triplets let

$$\mathrm{Err}(w, v, p; S_{t,\theta}) = \sum_{\tau \in S_{t,\theta}} \mathrm{err}(w, v, p; \tau)$$

for weights w, v, peak selection p and set of triplets $S_{t,\theta} = \{\tau_t^1, \tau_t^2, \ldots, \tau_t^{|S_{t,\theta}|}\}$.

Having introduced the notation we now present our classification procedure in Algorithm 1. Each time we make a prediction for which sample in a triplet is the HD sample an error is incurred, either 0 or 1 (correct/incorrect); this is equivalent to sensitivity. In our results we only present this one error where

the specificity is simply half the sensitivity (in this particular setting). Due to the small size of the data set we will not be looking into any test / training set divisions or indeed applying leave-one-out. Instead we will be testing each decision rule on the whole training set and selecting the best rule based on the classification error. As this is a biased setting it makes our predictions unreliable. We will later describe a method for calculating p-values which we use to test the significance of the errors obtained under this setting.

Algorithm 1 Optimal triplet decision rule search

Require: t—time before the last moment
Require: θ—window size in months
Require: P—the number of top ranked peaks to use
Require: W set of possible weights to try for w
  $E_i^* = 1$ (the best error found on the training set)
  $m^* = \{\{\}, \{\}, \{\}\}$ (we denote by this the best model that we find)
  for Each element $w_i \in W$ **do**
    for $p = 1, \ldots, P$ **do**
      if $\mathrm{Err}(w, v, p; S_{t,\theta}) < E^*$ **then**
        $E^* = \mathrm{Err}(w, v, p; S_{t,\theta})$
        $m^* = (w, v, p)$
      end if
    end for
  end for
  The total error is given by $\mathrm{Err}(m^*; S_{t,\theta})$

The result of the application of a set, S_t, of triplets to Algorithm 1 will be the information (E_i^*, m^*) for $i = 1, 2, \ldots, n$. Notice that the algorithm selects the higher ranked peak (lower numbered), p, in $m^* = (w, v, p)$ over peaks with lower ranking (higher number), which produce the same error, the implication of this being that ties are handled by favouring the most common peak.

3 Experimental setup

This section will introduce the method we use for calculation of the p-values for quantifying the quality of a decision rule, and hence the power of a peak for discrimination between controls and HD cases when used alone or in conjunction with another peak. We then go on to describe the experiments that will be performed on this triplet setting.

3.1 Monte-Carlo Method

We calculate valid p-values by using a Monte-Carlo method. The procedure is given in Algorithm 2. The basic process shown here models our classification algorithm repeated a large number N times (in this case $N = 10^4$). At each iteration we randomly permute the labels in each triplet and test if we can find

a decision rule such that the number of mistakes made is as good or better than the number of mistakes for the real data, under our original 'optimal' decision rule; we can call the latter the normal error. We will call it the normal rule as it relates to the normal data. In the algorithm notice that we use Q to count the number of times that the normal err is either matched or beaten by random permutations. We output our p-value at $(Q+1)/(N+1)$, the addition of one means we always count the normal case in addition to the N permutation, this prevents p-values of 0.

Algorithm 2 Monte-Carlo Method for calculating p-values

Require: t—time before the last moment.
Require: N—number of trials, should be sufficiently large.
Require: S_t—the set of triples for time t.
Require: The normal error obtained using the 'optimal' decision rule.
  $Q = 0$ - counting variable
  for $j = 1, \ldots, N$ **do**
    for $i = 1, \ldots, n$ **do**
      Randomly jumble the labels for the samples in τ_t^i.
    end for
    The new set of triples with jumbled labels is denoted by S_t'.
    Apply Triple Classification (Algorithm 1) to S_t'.
    if The total error obtained from this application is as good as or better than the normal error **then**
      $Q = Q + 1$
    end if
  end for
  p-value $= (Q+1)/(N+1)$

3.2 Experiments

Here we define 3 settings that will be used in the analysis below. We give a very general definition of each setting the importance of each will become clear in the next section. Below we refer to 'the algorithm', this in principle means calling Algorithm 2, which in turn calls Algorithm 1.

- **Setting 1** Apply the algorithm with some predefined value of P, with possible values of the weights as: $w \in \{-1, 1\}$ and $v = 0$.
- **Setting 2** Apply the algorithm with a slight modification: removing the loop over P (all peaks in Algorithm 1) and consider only 1 feature, which may be a single peak or a combination of two peaks; w and v are chosen appropriately.
- **Setting 3** The algorithm is applied by choosing some reasonable P and fixing p_1. The fixed peak, p_1, will be used in combination with a second

	Window size								
	6 months			9 months			12 months		
t	pk	P-value	Error	pk	P-value	Error	pk	P-value	Error
0	9	0.1207	14/33	4	0.0185	30/65	6	0.0018	35/78
1	6	0.0167	14/36	6	0.0002	23/60	6	0.0001	32/80
2	4	0.0008	13/38	4	0.0026	24/58	4	0.0008	36/81
3	4	0.0111	18/44	6	0.0012	23/57	6	0.0002	34/83
4	6	0.0008	15/43	6	0.0003	24/63	6	0.0001	37/88
5	6	0.0012	16/44	6	0.0008	28/67	6	0.0001	37/90
6	6	0.0074	18/45	6	0.0005	29/71	6	0.0002	40/93
7	6	0.0126	18/44	6	0.0069	31/69	6	0.0020	41/89
8	2	0.0593	19/43	33	0.0074	29/66	32	0.0007	38/87
9	6	0.0253	16/39	6	0.0087	27/61	6	0.0048	40/85
10	6	0.3234	22/45	38	0.0152	29/65	6	0.0413	47/93
11	38	0.0738	20/46	32	0.0282	31/67	4	0.0951	49/94
12	38	0.0229	20/48	32	0.0158	33/72	7	0.0053	54/109
13	37	0.1204	20/45	9	0.0697	36/73	7	0.0105	48/98
14	9	0.0140	18/44	7	0.0301	34/71	7	0.0011	40/89
15	32	0.0665	20/46	7	0.0010	37/83	7	0.0010	37/83
16	27	0.1308	22/48	7	0.0015	32/73	7	0.0015	32/73
17	7	0.1934	23/48	7	0.0044	29/66	7	0.0044	29/66
18	7	0.0052	26/61	7	0.0052	26/61	7	0.0052	26/61
19	7	0.0015	21/53	7	0.0015	21/53	7	0.0015	21/53
20	7	0.0072	18/45	7	0.0072	18/45	7	0.0072	18/45
21	7	0.0021	13/37	7	0.0021	13/37	7	0.0021	13/37
22	7	0.0129	8/25	7	0.0129	8/25	7	0.0129	8/25
23	7	0.0993	6/18	7	0.0993	6/18	7	0.0993	6/18

Table 1: Results under Setting 1, using $P = 41$ peaks.

peak $p_2 \in P \backslash p_1$. The weights w and v will be chosen appropriately to allow inclusion or exclusion of the fixed peak, but we will always include at least one peak. This means we allow $w = 0$ and require $v \neq 0$.

It is important to note that in Setting 3 this means that we have 1 fixed peak, p_1, and a selection of one less the total number of peaks (in this case $P = 41$) to assign to p_2 (i.e. p_2 can be one of 40 peak when $P = 41$).

4 Results

This section presents the results of various experiments under the settings explained above. We begin with the case explained in setting 1, this will identify the most significant peaks (under our triplet setting) for each time slot. Recall that a time slot is defined as a starting point and a window size. The results will include window sizes of 6,9, and 12 months. Smaller window sizes are not advisable as the number of samples becomes too small.

Table 1 shows the results under setting 1, in this case we set the number of peaks $P = 41$. This number of peak is the result of imposing a threshold on the number of samples in which a peak must appear. Peaks that fall below this threshold will not be considered, in this case we use 33% as the threshold

(this number was chosen to match the proportion of the smallest class). See the appendix for a complete listing of these 41 most common peaks.

The first column of Table 1, t, is the time, in months, to the original time of diagnosis. The pk columns indicate which peak was selected that minimises the overall error, recall we call this our normal rule: the rule achieved under normal conditions. The 'P-value' column is the p-value generated as a result of the Monte-Carlo method we introduced above. Finally we give the error of the normal rule in the 'Error' column. A column indicating the weight w has been omitted as in every case it was set to $w = -1$.

The table shows several points with interesting activity. In the 0–9 month range peak 6 is very popular and to some extent peak 4 (but much earlier on, 0–3 months). All p-values, when peak 6 is selected, indicate that the null hypothesis can be rejected at the 1% level (for window sizes of 9 and 12 months). Similarly we can reject the null hypothesis at the 5% level at points where peak 4 is superior. After 7 months in advance of the time of death the information contained in peak 6 is no longer superior to other peaks, hence in the 8–13 month period (for a 9 month window). After this, in the 14–23 month range, we see another peak, peak 7, shows statistically significant information which remains stable throughout this period (again for a 9 month window size).

4.1 Testing single peaks

In the previous section we analysed the set of 41 peaks, meaning at any time slot only the best peak (in terms of error) is observed. It may however be the case that other peaks are almost as good as the best in other time slots. We will therefore take some of the more interesting peaks from above and analyse them individually. Here we analyse peaks 7,6 and 4 respectively.

In this section we will be analysing single peaks selected from the original 41. In order to remove the possibility that significant p-values are generated by chance in any of the 1 out of 41 possible experiments we are required to perform some adjustment to the p-values, or in fact an adjustment of the level of significance. For example, in this case, if we were to require a 5% significant level then the adjusted significance level would be 0.122% (so numbers below 0.00122 in our tables). This is known as the Bonferroni adjustment.

4.1.1 Peak 7

In the original analysis peak 7 showed some very early diagnostic information in the 14-22 month period. Ideally we are looking for markers (peaks) that continue to grow or fall consistently all the way to the time of death. Table 2 shows the result of the single peak analysis, clearly we have statistically significant information from 12-22 months for a 12 month window. Lower window sizes however do not show such a trend however there are some significant p-values. After 14 months, as expected, the p-values become highly significant.

	Window size					
	6 months		9 months		12 months	
Time, t (months)	P-value	Error	P-value	Error	P-value	Error
0	0.5455	20/33	0.0375	35/65	0.0134	41/78
1	0.0092	16/36	0.0011	27/60	0.0020	39/80
2	0.0032	16/38	0.0011	26/58	0.0083	42/81
3	0.0026	19/44	0.0010	25/57	0.0030	42/83
4	0.0002	16/43	0.0006	28/63	0.0017	44/88
5	0.0021	19/44	0.0195	35/67	0.0045	46/90
6	0.0093	21/45	0.0291	38/71	0.0059	49/93
7	0.0703	23/44	0.1024	39/69	0.0510	50/89
8	0.4710	26/43	0.0881	37/66	0.0370	48/87
9	0.3830	23/39	0.1005	34/61	0.0375	47/85
10	0.5984	28/45	0.3161	39/65	0.0395	52/93
11	0.3191	27/46	0.1124	38/67	0.0080	50/94
12	0.2889	28/48	0.1110	41/72	0.0004	54/109
13	0.4212	27/45	0.0457	40/73	0.0005	48/98
14	0.0324	22/44	0.0015	34/71	0.0001	40/89
15	0.0619	24/46	0.0001	37/83	0.0001	37/83
16	0.0237	24/48	0.0001	32/73	0.0001	32/73
17	0.0119	23/48	0.0004	29/66	0.0004	29/66
18	0.0007	26/61	0.0007	26/61	0.0007	26/61
19	0.0002	21/53	0.0002	21/53	0.0002	21/53
20	0.0012	18/45	0.0012	18/45	0.0012	18/45
21	0.0002	13/37	0.0002	13/37	0.0002	13/37
22	0.0013	8/25	0.0013	8/25	0.0013	8/25
23	0.0087	6/18	0.0087	6/18	0.0087	6/18

Table 2: Results under Setting 2, where $p1 = 7$.

	Window size					
	6 months		9 months		12 months	
Time, t (months)	P-value	Error	P-value	Error	P-value	Error
0	0.1028	17/33	0.0014	30/65	0.0002	35/78
1	0.0014	14/36	0.0001	23/60	0.0001	32/80
2	0.0012	15/38	0.0002	24/58	0.0001	36/81
3	0.0007	18/44	0.0001	23/57	0.0001	34/83
4	0.0001	15/43	0.0001	24/63	0.0001	37/88
5	0.0001	16/44	0.0003	28/67	0.0001	37/90
6	0.0007	18/45	0.0001	29/71	0.0001	40/93
7	0.0012	18/44	0.0004	31/69	0.0003	41/89
8	0.0241	21/43	0.0007	30/66	0.0005	42/87
9	0.0020	16/39	0.0004	27/61	0.0005	40/85
10	0.0211	22/45	0.0051	32/65	0.0033	47/93
11	0.0052	21/46	0.0041	33/67	0.0051	49/94
12	0.0043	22/48	0.0028	35/72	0.0023	57/109
13	0.0440	23/45	0.0126	38/73	0.0128	53/98
14	0.0136	21/44	0.0164	37/71	0.0104	47/89
15	0.0546	24/46	0.0440	46/83	0.0440	46/83
16	0.0526	25/48	0.0492	40/73	0.0492	40/73
17	0.2856	28/48	0.1568	38/66	0.1568	38/66
18	0.1653	35/61	0.1653	35/61	0.1653	35/61
19	0.1629	30/53	0.1629	30/53	0.1629	30/53
20	0.2732	26/45	0.2732	26/45	0.2732	26/45
21	0.4322	22/37	0.4322	22/37	0.4322	22/37
22	0.5744	15/25	0.5744	15/25	0.5744	15/25
23	0.4465	10/18	0.4465	10/18	0.4465	10/18

Table 3: Results under Setting 2, where $p1 = 6$.

4.1.2 Peak 6

Also in the original analysis peak 6 showed some very early information in the 0-9 month period. As above for peak 7 we will analyse the information contained in peak 6 separately. Table 3 shows the result of the single peak analysis, clearly we have statistically significant information from 0-9 for larger window sizes. After this there is no statistically significant information—recall that we are comparing the p-values against the adjusted significance level of 0.00122.

4.1.3 Peak 4

Peak 4 also showed a small amount of information very early on in months 0 and 2. Despite the occurrence being only sporadic there may be some information contained in the peak which is only slightly worse that peak 6. We therefore apply the setting as above to peak 4. Table 4 shows the result of the single peak analysis, clearly we have statistically significant information from 0-4 for all window sizes. Despite this the p-values become larger much earlier than for peak 6 and are not be significant for many months under the p-value adjustment.

	Window size					
	6 months		9 months		12 months	
Time, t (months)	P-value	Error	P-value	Error	P-value	Error
0	0.0200	15/33	0.0009	30/65	0.0004	36/78
1	0.0032	15/36	0.0002	24/60	0.0001	34/80
2	0.0001	13/38	0.0002	24/58	0.0002	36/81
3	0.0005	18/44	0.0003	24/57	0.0001	36/83
4	0.0011	17/43	0.0006	27/63	0.0002	41/88
5	0.0058	20/44	0.0017	32/67	0.0004	43/90
6	0.0093	21/45	0.0009	33/71	0.0012	47/93
7	0.0022	19/44	0.0017	33/69	0.0026	45/89
8	0.0974	23/43	0.0155	34/66	0.0182	47/87
9	0.0131	18/39	0.0288	32/61	0.0136	45/85
10	0.0852	24/45	0.0764	36/65	0.0134	50/93
11	0.0283	23/46	0.0382	36/67	0.0060	49/94
12	0.0940	26/48	0.0364	39/72	0.0048	58/109
13	0.2744	26/45	0.0473	40/73	0.0125	53/98
14	0.1229	24/44	0.0146	37/71	0.0100	47/89
15	0.3239	27/46	0.0427	46/83	0.0427	46/83
16	0.0923	26/48	0.0258	39/73	0.0258	39/73
17	0.1010	26/48	0.0541	36/66	0.0541	36/66
18	0.0288	32/61	0.0288	32/61	0.0288	32/61
19	0.0251	27/53	0.0251	27/53	0.0251	27/53
20	0.0438	23/45	0.0438	23/45	0.0438	23/45
21	0.0769	19/37	0.0769	19/37	0.0769	19/37
22	0.1894	13/25	0.1894	13/25	0.1894	13/25
23	0.4359	10/18	0.4359	10/18	0.4359	10/18

Table 4: Results under Setting 2, where $p1 = 4$.

	Window size									
	9 months					12 months				
t	w	v	Peak	P-value	Error	w	v	Peak	P-value	Error
0	-1	-1	6	0.0039	28/65	-1	-1	36	0.0011	33/78
1	-1	-1	6	0.0001	20/60	-1	-1	6	0.0001	31/80
2	-1	-1	6	0.0001	20/58	-1	-1	6	0.0003	34/81
3	-1	-1	6	0.0002	20/57	-1	-1	6	0.0001	34/83
4	-1	-1	6	0.0001	22/63	-1	-1	6	0.0002	36/88
5	-1	-1	6	0.0040	29/67	-1	-1	36	0.0002	37/90
6	-1	-1	6	0.0040	31/71	-1	-1	36	0.0001	38/93
7	-1	-1	36	0.0034	29/69	-1	-1	14	0.0003	38/89
8	-1	-1	36	0.0010	26/66	-1	-1	14	0.0003	36/87
9	-1	-1	36	0.0009	24/61	-1	-1	14	0.0003	35/85
10	-1	-1	14	0.0077	28/65	-1	-1	14	0.0013	42/93
11	-1	-1	14	0.0029	28/67	-1	-1	14	0.0019	43/94
12	-1	-1	14	0.0033	31/72	-1	-1	36	0.0002	48/109
13	-1	-1	14	0.0036	32/73	-1	-1	36	0.0007	43/98
14	-1	-1	14	0.0016	30/71	-1	-1	36	0.0003	37/89
15	-1	-1	36	0.0006	35/83	-1	-1	36	0.0006	35/83
16	-1	1	1	0.0004	30/73	-1	1	1	0.0004	30/73
17	-1	1	1	0.0069	29/66	-1	1	1	0.0069	29/66
18	-1	0		0.0058	26/61	-1	0		0.0058	26/61
19	-1	0		0.0033	21/53	-1	0		0.0033	21/53
20	-1	0		0.0088	18/45	-1	0		0.0088	18/45
21	-1	0		0.0031	13/37	-1	0		0.0031	13/37
22	-1	0		0.0150	8/25	-1	0		0.0150	8/25
23	-1	1	1	0.1071	6/18	-1	1	1	0.1071	6/18

Table 5: Results under Setting 3 p1 = 7, the 6 month window has been omitted as the results are similar to the 9 month window with respect to p-values and decision rules selected.

4.2 Peak combinations

Given that some peaks perform well alone in small time intervals, we further our analysis to find if there are peaks that can improve p-values when used in combination with one of the best performing peaks. This experiment will be run under setting 3 described above. First we will look at the performance of other peaks in combination with peak 7. We will of course allow w to be taken from $\{-1, 1\}$ as before. v will be taken from $\{-1, 1, 0\}$ to allow the choice not to add anything to peak 7. Table 5 shows the results of this analysis. During the period 18–22 months peak 7 works best when used alone, for some months before this the peak selection is rather unstable. In the range 0–6 months we see that $p_2 = 6, v = -1$ is always selected (for a 9 month window for example). This shows that the information contained in peak 6 is also useful when used as a combination.

4.3 Further analysis of peaks 6 and 7.

The most stable peak in the previous experiment, where we analysed the performance of all peaks in combination with peak 7 was peak 6. This also happens

	Window size					
	6 months		9 months		12 months	
t	P-value	Error	P-value	Error	P-value	Error
0	0.0543	17/33	0.00007	28/65	0.00002	34/78
1	0.000005	12/36	0.000001	20/60	0.000001	31/80
2	0.000003	10/38	0.000002	20/58	0.000005	34/81
3	0.000002	14/44	0.000001	20/57	0.000001	34/83
4	0.000001	11/43	0.000001	22/63	0.000003	36/88
5	0.00002	15/44	0.000007	29/67	0.000007	39/90
6	0.00007	17/45	0.00002	31/71	0.00002	42/93
7	0.0007	19/44	0.0006	33/69	0.0006	44/89
8	0.0927	24/43	0.0026	34/66	0.0026	45/87
9	0.0358	20/39	0.0015	31/61	0.0015	43/85
10	0.0791	25/45	0.0368	36/65	0.0036	49/93
11	0.0296	24/46	0.0101	35/67	0.0033	49/94
12	0.0277	25/48	0.0051	37/72	0.00008	53/109
13	0.0803	25/45	0.0055	38/73	0.0002	48/98
14	0.0061	21/44	0.0024	35/71	0.000061	41/89
15	0.0146	23/46	0.0005	39/83	0.0005	39/83
16	0.0125	24/48	0.0001	34/73	0.0001	34/73
17	0.0270	25/48	0.0011	31/66	0.0011	31/66
18	0.0008	28/61	0.0008	28/61	0.0008	28/61
19	0.0003	23/53	0.0003	23/53	0.0003	23/53
20	0.0011	20/45	0.0011	20/45	0.0011	20/45
21	0.0023	16/37	0.0023	16/37	0.0023	16/37
22	0.0079	10/25	0.0079	10/25	0.0079	10/25
23	0.0041	6/18	0.0041	6/18	0.0041	6/18

Table 6: Results under Setting 2, we select $p_1 = 7$ $w = -1$, $p_2 = 6$ and $v = -1$, window sizes 6, 9 and 12 months are shown.

to be the best performing peak when used alone in the 0–6 month range. The two peaks may both be important: one for showing risk of heart disease in the months close to death and the other showing risk of heart disease at 14–23 months before death. Despite these findings it is still more practical to have a single test for the identification of such a risk. We will therefore attempt to find a decision rule that produces statistically significant information for the whole time range.

Table 6 shows the results of our analysis under setting 2, where $p_1 = 7$ $w = -1$, $p_2 = 6$ and $v = -1$. Clearly there are very small p-values and most are significant if we set a level of 0.05. However, as above we must also consider p-value adjustment.

In this case we are pre-selecting pairs of peaks from a list of 41 peaks. As there are 1681 ways of doing this, significant at the 5% level needs to be adjusted to a 0.003% level. This would mean that the p-values need to be less than 0.00003 in order to be significant. Even though this adjustment is very restrictive, there are still some significant p-values. The fact that we have chosen only one rule means we have no knowledge of how other rules perform.

5 Conclusions and further work

This work has identified 3 peaks which may be useful in the early detection of heart disease. Two of the peaks, at 4055Da and 4211Da, carry statistical significant information up to 9 and 4 months respectively in advance of the original time of death. Of these 2 peaks the most dominant is 4055Da in this time frame. We have also identified a third peak, at 5338Da, which carries statistically significant information in the 14–22 month range and some limited information in the 0–13 month range. By limited we mean that not all months showed significant p-values in our test due to the p-value adjustment made.

Given that we have a peak that provides information in the months close to diagnosis and another peak providing information much later it stands to reason that some combination of the two peaks may create a more general rule with respect to time. We have shown that the combination $-\log I(4405) - \log I(5338)$ can provide some information in the whole 0-23 month range however it is not clear how significant this information is as the p-values would require huge adjustments.

This work has only shown that some peaks exist that can well outperform other peaks under out triplet analysis. Despite the encouraging p-values it should be noted that the rate of error observed was not as low as one would hope in this situation. It is therefore required that these findings be validated using more data and other technologies, which would enable us to move away from the triplet setting under which the analysis in this paper is set.

Appendix A: Peak list

Table 7 is a complete peak listing given the 1/3 commonality threshold.

Peak number	peak m/z	commonality
1	7767.8	100.0 %
2	9293.2	99.9 %
3	5905.9	99.9 %
4	4211.1	99.8 %
5	3242.3	99.0 %
6	4055	98.5 %
7	5338.3	98.4 %
9	1946.1	97.8 %
10	4645.5	96.3 %
11	2661.8	96.3 %
12	4965.2	96.0 %
13	1547.2	95.1 %
14	6632.3	91.4 %
15	2355.3	91.4 %
16	1467.5	90.9 %
17	3508.8	90.6 %
18	3957.3	84.7 %
19	3524.6	83.0 %
20	2025.5	79.7 %
21	1450.2	75.9 %
22	1898.2	72.2 %
23	2115.4	71.1 %
24	2380.4	70.4 %
25	1262.2	67.8 %
26	8605	64.4 %
27	2934.5	61.6 %
28	1017	60.1 %
29	1521.4	54.9 %
30	1617.5	48.8 %
31	8132	48.2 %
32	3770.1	47.3 %
33	5005.5	47.2 %
34	991.87	46.9 %
35	1789.2	43.3 %
36	2083.7	39.5 %
37	2273.6	38.7 %
38	4284	38.5 %
39	3159.7	38.5 %
40	3193.4	36.4 %
41	2605.7	35.4 %

Table 7: Top 41 peaks in the Reading data set.