CLRC—TR—08—03

Early detection of breast cancer in UKCTOCS using proteomic biomarkers

February 9, 2009

Brian Burford[1], Dmitry Devetyarov[1], Ilia Nouretdinov[1], Zhiyuan Luo[1], Alexey Chervonenkis[1], Volodya Vovk[1], Mike Waterfield[2], Ali Tiss[3], Celia Smith[3], Rainer Cramer[2,3], Alex Gentry-Maharaj[4], Rachel Hallett[4], Stephane Camuzeaux[4], Jeremy Ford[4], John Timms[4], Usha Menon[4], Ian Jacobs[4] and Alex Gammerman*[1,4].

[1]Computer Learning Research, Royal Holloway, University of London
[2]Ludwig Institute for Cancer Research, University College London
[3]BioCentre and Department of Chemistry, University of Reading
[4]Institute for Womens Health, University College London

Abstract

This paper proves empirically that the information contained within serum proteome, mass spectra, can improve upon current standards for the early detection of Breast Cancer (BC). The data in this paper consists of samples collected up to 3 years before the original diagnosis of BC was made. Our statistical empirical investigation has identified a peak at mass 6638.7Da. This peak carries statistically significant information up to 11 months prior to the original BC diagnosis.

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 breast cancer.

The data set consists of breast cancer (BC) case samples each with two matched controls. The data set contains 258 samples: 86 BC cases and 172 control samples from health women. The data is divided into two sets: training and validation. The training set contains 62 case samples with all corresponding matched controls and the validation set contains 24 case samples with the corresponding matched controls. We will refer to the collection of a case sample

with two matched controls as a triplet. The training set therefore has 62 triplets and the validation has 24 triplets.

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 Pre-processing

This section briefly describes the pre-processing applied to the UCL and Reading data. The data were provided in a two-column format: the first column is a list of m/z-values, the second column is a list of corresponding intensities. Calibration had been performed prior to the data being distributed; therefore, all our further pre-processing steps were applied only to the intensities, m/z-values remained unchanged. The pre-processing steps which were applied to both data sets are described below:

1. **Re-sampling** was performed in order to decrease the number of m/z-ratios for computational optimization purposes.
2. For noise elimination, we performed **smoothing** by averaging the intensities within a moving window.
3. **Baseline subtraction** ensured that the spectra sit on the intensity = 0 axis. The algorithm applied is based on finding the lowest points between some dominant local maxima (we refer to these lowest points as *troughs* later on). We define a dominant local maximum as the point with the highest intensity within some range, the width of this range is a parameter of the algorithm. We then apply Piecewise Cubic Hermite Interpolating Polynomial to all the points marked as troughs in order to construct a baseline. Further steps are applied to correct the baseline for any points where the baseline is above the spectra.
4. We performed **normalisation** to make sure that the total amount of ions across different samples were the same. The algorithm involved dividing the intensity of each point in a spectra by the sum of all intensities.
5. The goal of the **peak identification** step was to generate a list of peaks for each sample. This was achieved by identifying all local maxima in the mass spectra above an intensity threshold and above a certain signal-to-noise ratio threshold. As a result of this algorithm we had a table for each sample with the following columns:
   (a) Unique sample ID — for any single sample this column has the same number for each entry, the reason for this column will become apparent in the next step.

(b) Number of peak in the initial array of m/z-values.

(c) M/z-value of a local maximum.

(d) Signal-to-noise ratio within a window of the certain width.

(e) Corresponding intensity.

6. Peaks obtained in the previous step do not strictly coincide because of the noise and possible presence of different isotopes. Therefore, the next step (called **peak alignment**) is to find common peaks (that is, peaks with m/z-values close to each other) among the samples. We combined all peak lists constructed for individual samples into one list and sorted in descending order relative to column 5 — peak intensity. We then worked down the list taking one of the following actions for each peak:

   (a) If the peak is within some predefined range of an existing peak group and there is no other peak from the same sample in that group, then the current peak can be added to the group; else the peak is ignored.

   (b) If there are no groups close to the current peak, then a new peak group is created containing the single peak.

7. The result of the previous step is a set of peak groups, each of which can potentially contain between 1 peak and N peaks, where N is the number of samples in the data set. Now we have to compute peak intensities for each sample for each peak group. For those peak groups that have less than N peaks, we need to estimate the intensities for the remaining samples. For this purpose, we set a mass separation parameter which we use to define a range in the m/z-vector given a single m/z-value — the maximum m/z-ratio of all the peaks in the group. Then the intensity of each missing sample is defined as the maximum intensity within this range in the spectrum of the sample.

3 Setting of the problem

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

The 186 samples extracted from the training data are divided into 62 such triplets. Each cancer sample has a time associated with it, which is the amount time before the last moment in months. We denote a cancer sample by OC_i, for the patient i. Similarly we denote the two corresponding matched controls by $M1_i$ and $M2_i$. We denote a triplet by

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

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 OC_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 colection of triplets based on t and θ as $S_{t,\theta}$

Our goal is to find decision rules that can successfully select the cancer 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 cancer sample as the element with the highest of these values.

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

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

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 cancer 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.

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.

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 and V two sets of possible weights to try for w and v respectively
  $E_i^* = 1$ (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 Each element $v_i \in V$ **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
  end for
  The total error is given by $\mathrm{Err}(m^*; S_{t,\theta})$

4 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 cancer 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.

4.1 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 $= (Q + 1)/(N + 1)$

4.2 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 be come 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 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 $v \neq 0$.

In the following experiments we will select $P = 100$ where needed. It is important to note that in Setting 3 this means that we have 1 fixed peak, p_1, and a selection of 99 peaks to assign to p_2.

5 Results

This section presents results for the various experiments described. Tables 1 and 2 show the results of the application of the BC data to the analysis method described above in Setting 1. Each of the two tables contains information for 2 different window sizes: 3 and 6 months in Table 1; 9 and 12 months in Table 2. The first column of each table 'Time' is the time, in months, to the original time of diagnosis. The column with name w is the value of the weight w as above, (in this case $v = 0$ as we are dealing with only one feature at a time). The 'Peak' 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.

Due the relatively small size of this data set it is not expected that we would be able to construct a reliable test over narrow window size as the number of examples would be too few. This is indeed the case when we look at a window size of 3 months (Table 1). The number of examples in each time slot varies from 1 to 9, and in the case where $t = 0$ there are no example in the window. As a result of these small numbers of examples it is difficult to draw any significant results from the data and the peaks used are not stable—however some of the peaks do become interesting later.

Table 1 also shows results for a 6 month window. The results here become slightly more stable and from 0 to 9 months there are always more the 10 examples—this is still extremely small but it seems to be enough to begin drawing some statistically significant results from the data albeit a little unstable. We can see that between 4 and 10 months inclusive most p-values are significant at the 5% level, only at 6 months does the p-value increase above 5% (6.37%). Also between 1 and 12 months inclusive the use of peak 19 is stable.

Table 2 shows results for a 9 and 12 month window size. It is clear from this table and our previous table that increasing the window size can make our results more stable and indeed statistically significant early on, i.e. close to $t = 0$ months. This comes with the disadvantage of reducing the number of months in advance that we can extend such results to. This can be seen by observing that for a window size of 9 months we have significant p-values up to 9 months in advance, with the exception of $t = 0$ where the p-value is above 5%. We compare this to a 12 months window where at the time $t = 0$ we now achieve a significant p-value of 0.3%, in this case we achieve significant p-values up to 8 months in advance and from 0 to 6 months the p-value are significant at the 1% level. This phenomenon is of course expected in such a setting and is one of the reasons that we test several window sizes.

We have shown that a 3 month window size is too small to produce any interesting or indeed stable results. For window sizes 6, 9 and 12 months we see that peak 19 is very stable, whenever this peak is used the weight $w = -1$ is used, this implies that this potential biomarker decreases with cancer. By taking into account all of the results it appears that this peak starts having an

	Window size							
	3 months				6 months			
Time, t (months)	w	Peak	P-value	Error	W1	Peak	P-value	Error
0	-	-	-	0/0	1	22	0.1036	1/8
1	-1	10	0.6695	0/2	-1	19	0.1326	2/10
2	-1	10	0.2270	0/3	-1	19	0.1631	3/12
3	1	22	0.1036	1/8	-1	19	0.0856	4/15
4	-1	19	0.0943	1/8	-1	19	0.0147	3/15
5	-1	19	0.2762	2/9	-1	19	0.0152	3/15
6	1	64	0.0513	0/7	-1	19	0.0637	2/11
7	-1	38	0.3350	1/7	-1	19	0.0032	2/14
8	-1	5	0.2850	1/6	-1	19	0.0047	2/14
9	-1	19	0.2681	0/4	-1	19	0.0053	1/11
10	-1	19	0.0159	0/7	-1	19	0.0347	1/9
11	-1	19	0.0958	1/8	-1	19	0.2774	2/9
12	-1	16	0.1910	1/7	-1	19	0.2689	2/9
13	-1	1	0.1113	0/2	-1	1	0.4535	1/5
14	1	1	0.0001	0/1	1	15	0.4095	0/3
15	1	2	0.2198	0/2	1	62	0.4396	1/7
16	1	15	0.4095	0/3	1	62	0.4943	2/9
17	1	4	0.3348	0/2	-1	70	0.5226	2/9
18	1	60	0.9252	1/5	1	23	0.8186	4/11
19	-1	70	0.7636	1/6	-1	22	0.5775	4/12
20	1	28	0.8447	2/7	1	62	0.5504	5/15
21	1	6	0.9046	2/6	-1	20	0.8006	5/13
22	1	3	0.9116	2/6	-1	54	0.9338	4/11
23	1	62	0.7967	2/8	-1	21	0.8369	4/11
24	1	27	0.8371	2/7	-1	9	0.7533	3/9

Table 1: Results under Setting 1, we select $P = 100$, $w \in \{-1, 1\}$ and $v = 0$, window sizes 3 and 6 months are shown.

	Window size							
	9 months				12 months			
Time, t (months)	w	Peak	P-value	Error	W1	Peak	P-value	Error
0	-1	19	0.0856	4/15	-1	19	0.0034	4/19
1	-1	19	0.0190	4/17	-1	19	0.0001	4/24
2	-1	19	0.0070	4/18	-1	19	0.0001	5/26
3	-1	19	0.0034	4/19	-1	19	0.0001	5/26
4	-1	19	0.0001	3/22	-1	19	0.0001	4/24
5	-1	19	0.0005	4/23	-1	19	0.0003	5/24
6	-1	19	0.0010	3/18	-1	19	0.0023	4/20
7	-1	19	0.0063	3/16	-1	19	0.0229	5/19
8	-1	19	0.0141	3/15	-1	19	0.0170	4/17
9	-1	19	0.0085	2/13	-1	19	0.1921	6/18
10	-1	19	0.1570	3/12	-1	19	0.5299	7/18
11	-1	19	0.3112	3/11	-1	89	0.8133	7/18
12	-1	19	0.9320	6/14	-1	22	0.8788	9/20
13	1	62	0.9378	4/11	1	62	0.5254	6/17
14	1	62	0.7860	3/10	1	62	0.0880	5/18
15	1	62	0.5314	4/13	1	62	0.3133	7/20
16	1	62	0.1702	4/15	1	62	0.3133	7/20
17	1	62	0.1793	5/17	1	62	0.6867	8/20
18	1	62	0.7229	7/18	1	92	0.7779	8/20
19	1	94	0.6394	6/17	1	28	0.5427	9/22
20	1	92	0.8137	7/18	1	28	0.3574	9/23
21	-1	20	0.7881	6/15	-1	73	0.5709	7/19
22	1	59	0.7646	6/16	-1	73	0.3779	6/18
23	-1	73	0.8216	6/16	-1	73	0.5941	6/17
24	1	69	0.9495	5/13	-1	95	0.6888	5/15

Table 2: Results under Setting 1, we select $P = 100$, $w \in \{-1, 1\}$ and $v = 0$, window sizes 9 and 12 months are shown.

effect between 8 and 10 months in advance of the original diagnosis. Peak 19 breaks down for a windows size of 12 months at $t = 8$ months due to the larger window size. We can extend to $t = 10$ months for a window size of 6 months. Our general 'conservative' conclusion is that our tests have shown that we can extract statistically significant information up to 8 months in advance of the original diagnosis and the decision rule used is stable: $-\log I(19)$. The m/z value of peak 19 is 6638.7Da.

5.1 Testing peak 19

Now that we have seen that peak 19 may be important in early detection of breast cancer we will repeat our analysis on this peak as a single feature (Setting 2). The results above clearly show that a window size of 3 months is too small to produce stable results, we will therefore restrict our analysis to 6, 9 and 12 month window sizes.

Table 3 shows the results of testing peak 19 alone. It is clear that regardless of the window size (in this case) we can observe significant p-values up to $t = 11$ months; and in the case of a 6 months window size we can even see p-values less than 1% at 11 and 12 months in advance. Taking into account all window sizes we can draw the conclusion that up to 11 months in advance peak 19 contains statistically significant information regarding the separation of healthy and BC classes.

Figure 1 shows the distribution of the controls in each triplet relative to the cases for $-\log I(19)$. Each cross on the plot represents a triplet. The horizontal axis is the time before diagnosis of that triplet, the vertical axis is the value of $-\log I(19)$ for the case sample minus the maximum value of $-\log I(19)$ for the two matched controls. This means that the maximum of the two controls would sit on the horizontal dashed line, we call this line the 'zeroed control line'. Examples above this line are correctly classified, examples below are mistakes. Note that not all examples below the line represent triplets where the case is less than the minimum of the 2 controls, some of these triplets will have the case sample between the two controls, this is important to note as turning the picture 180 degrees will not show the distribution of examples for $+\log I(19)$. The figure shows that there is a large cluster of triplets above the zeroed control line up to approximately 12 months in advance but situated mostly before 8 months. Between 8 and 12 months there are only a small number of examples, therefore not much can be said about the information contained in this peak in this time slot.

5.2 Adding more information to peak 19

We have, up to this point, a very interesting result regarding the information contained within peak 19. This peak however was a clear dominant peak and it may be the case that other peaks also carry significant information for the early detection of breast cancer—we would of course expect these results to be worse than those including peak 19. Rerunning the experiments with out including

	Window size					
	6 months		9 months		12 months	
Time, t (months)	P-value	Error	P-value	Error	P-value	Error
0	0.0206	2/8	0.0013	4/15	0.0002	4/19
1	0.0039	2/10	0.0003	4/17	0.0001	4/24
2	0.0046	3/12	0.0005	4/18	0.0001	5/26
3	0.0013	4/15	0.0002	4/19	0.0001	5/26
4	0.0004	3/15	0.0001	3/22	0.0001	4/24
5	0.0002	3/15	0.0001	4/23	0.0001	5/24
6	0.0013	2/11	0.0002	3/18	0.0001	4/20
7	0.0002	2/14	0.0002	3/16	0.0005	5/19
8	0.0006	2/14	0.0003	3/15	0.0004	4/17
9	0.0003	1/11	0.0004	2/13	0.0032	6/18
10	0.0016	1/9	0.0029	3/12	0.0132	7/18
11	0.0081	2/9	0.0111	3/11	0.0436	8/18
12	0.0083	2/9	0.0548	6/14	0.1887	11/20
13	0.5411	3/5	0.5345	7/11	0.7205	12/17
14	0.7055	2/3	0.7032	7/10	0.7706	13/18
15	0.7386	5/7	0.8646	10/13	0.8522	15/20
16	0.6255	6/9	0.7909	11/15	0.8522	15/20
17	0.6285	6/9	0.7174	12/17	0.7109	14/20
18	0.9216	9/11	0.8987	14/18	0.8583	15/20
19	0.8267	9/12	0.8644	13/17	0.7932	16/22
20	0.7898	11/15	0.7721	13/18	0.6934	16/23
21	0.8622	10/13	0.7913	11/15	0.6494	13/19
22	0.9220	9/11	0.8299	12/16	0.5916	12/18
23	0.7664	8/11	0.6623	11/16	0.5231	11/17
24	0.6286	6/9	0.4499	8/13	0.3822	9/15

Table 3: Results under Setting 2, we select $p_1 = 19$, $w = -1$ and $v = 0$, window sizes 6, 9 and 12 months are shown.

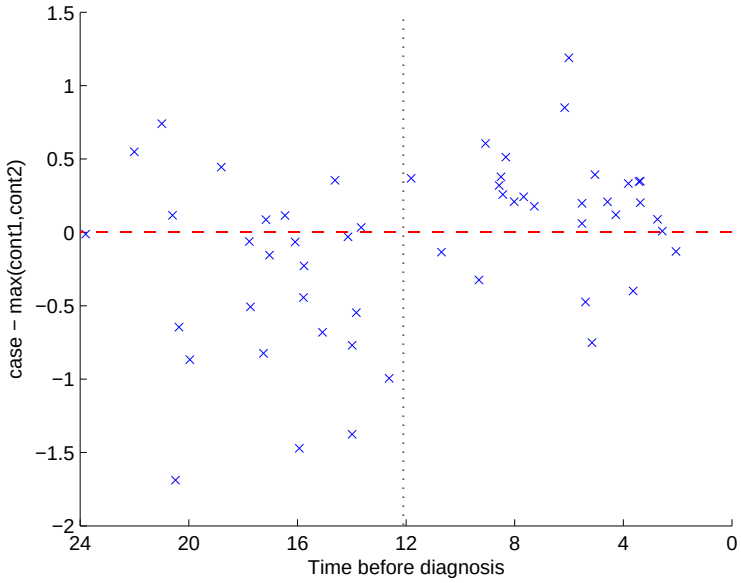

Figure 1: The figure shows the distribution of data over time and the relationship between the intensity in the cases compared to the matched controls for the decision rule: $-\log I(19)$. Each cross represents a triplet.

	Window size									
	6 months					9 months				
t	w	v	Peak	P-value	Error	w	v	Peak	P-value	Error
0	-1	1	5	0.0686	1/8	-1	1	5	0.0921	4/15
1	-1	1	5	0.0112	1/10	-1	1	5	0.0216	4/17
2	-1	1	5	0.0208	2/12	-1	1	5	0.0081	4/18
3	-1	1	5	0.0921	4/15	-1	1	5	0.0033	4/19
4	-1	1	5	0.0141	3/15	-1	1	5	0.0001	3/22
5	-1	1	5	0.0131	3/15	-1	1	5	0.0001	3/23
6	-1	-1	21	0.0999	2/11	-1	1	5	0.0011	3/18
7	-1	-1	48	0.0098	2/14	-1	1	5	0.0065	3/16
8	-1	-1	5	0.0035	2/14	-1	-1	5	0.0017	2/15
9	-1	1	5	0.0001	0/11	-1	-1	5	0.0008	1/13
10	-1	1	5	0.0010	0/9	-1	-1	5	0.0199	2/12
11	-1	-1	1	0.0187	1/9	-1	-1	5	0.0501	2/11
12	-1	-1	5	0.0298	1/9	-1	-1	5	0.5515	5/14
13	0	-1	1	0.5100	1/5	-1	1	16	0.8167	4/11
14	0	1	15	0.4074	0/3	0	1	61	0.8184	3/10
15	0	1	61	0.4860	1/7	0	1	61	0.5927	4/13
16	0	1	61	0.5568	2/9	0	1	61	0.2143	4/15
17	0	-1	69	0.5855	2/9	0	1	61	0.2246	5/17
18	0	1	22	0.8388	4/11	0	1	61	0.7761	7/18
19	0	-1	21	0.6300	4/12	0	1	93	0.7124	6/17
20	0	1	61	0.6164	5/15	0	1	91	0.8671	7/18
21	-1	1	3	0.7902	5/13	-1	1	3	0.7757	6/15
22	0	-1	53	0.9424	4/11	0	1	58	0.8107	6/16
23	-1	1	80	0.7127	3/11	0	-1	72	0.8631	6/16
24	0	-1	9	0.8149	3/9	-1	1	5	0.8120	5/13

Table 4: Results under Setting 3, we select $P = 100$, $p_1 = 19$, $w \in \{-1, 0\}$ and $v = \{-1, 1\}$, window sizes 6 and 9 months are shown.

peak 19 does not produce any significant results, a table showing the results will therefore not be included.

As the other peaks do not carry very stable information alone we will investigate the possibility of using peak 19 as a fixed feature and using all other peaks as possible second features (Setting 3). Results have shown that the rule $-\log I(19)$ was important, we will therefore either allow $w = -1$ or $w = 0$, (also $v \in \{-1, 1\}$). We will again look at 6, 9 and 12 month window sizes.

Tables 4 and 5 show the results of this experiment. In comparison to previous results using only one peak there is no major change. In some cases there is a small improvement in error however the changes are not large and more importantly do not extend further than the current 11 months that can already be archived by peak 19 alone. Despite these observations there is still a peak (peak 5) that seems to be stable when used in conjunction with peak 19. This is

	Window size				
	12 months				
t	w	v	Peak	P-value	Error
0	-1	1	5	0.0033	4/19
1	-1	1	5	0.0001	4/24
2	-1	1	5	0.0001	4/26
3	-1	1	5	0.0001	4/26
4	-1	1	5	0.0001	3/24
5	-1	1	5	0.0001	4/24
6	-1	-1	48	0.0011	3/20
7	-1	-1	48	0.0124	4/19
8	-1	-1	5	0.0024	3/17
9	-1	-1	5	0.0457	5/18
10	-1	-1	5	0.0468	5/18
11	-1	-1	5	0.1996	6/18
12	-1	-1	5	0.5202	8/20
13	0	1	61	0.6003	6/17
14	0	1	61	0.1239	5/18
15	0	1	61	0.3847	7/20
16	0	1	61	0.3847	7/20
17	0	1	61	0.7465	8/20
18	0	1	91	0.8347	8/20
19	0	1	27	0.6214	9/22
20	0	1	27	0.4264	9/23
21	0	-1	72	0.6512	7/19
22	0	-1	72	0.4549	6/18
23	0	-1	72	0.6757	6/17
24	0	-1	94	0.7705	5/15

Table 5: Results under Setting 2, we select $P = 100$, $p_1 = 19$, $w \in \{-1, 0\}$ and $v = \{-1, 1\}$, window size 12 months is shown.

especially stable for a 9 month window size where for $t = 0$ up to 7 months the rule $-\log I(19) + \log I(5)$ (i.e. $v = 1$) is favoured. Although peak 5 is present up to 12 months in advance the weight changes to $v = -1$ when $t = 8$, this is looks very strange, however this rule does not remain stable and the p-values are no longer significant after 10 months. Despite the peculiar pattern peak 5 may still hold some important information when used in conjunction with peak 19, our further analysis will therefore include this peak. The m/z value of peak 5 is 5342.4Da.

5.3 Further analysis of peak 19 and peak 5

In order to compare the use of peak 19 alone and the combination of peak 19 and peak 5 we apply our experiment in Setting 2 to the rule: $-\log I(19) + \log I(5)$. The results are shown in Table 6. It is clear that the addition of peak 5 does not produce results that are very superior to peak 19 alone, however in the case of a 12 month window the error achieved with the combination of the 2 peaks is always as good or better then peak 19 alone—the improvements represent one more correct classification.

We plot these results in a similar fashion as before (see Figure 1). Figure 2 shows the a similar cluster above the zeroed control line. However in this case despite there being less examples below the zero line (incorrect classification) there are also a larger number sitting almost on the line, implying that this rule may not be as stable with respect to fluctuations in the controls as the single peak 19 rule.

The errors in each of the experiments above can give an overall picture of the dynamic power of each rule. Here we will compare the two rules: $-\log I(19)$ and $-\log I(19) + \log I(5)$ using plots of these dynamics to give a more general picture. Figures 3, 4, 5 and 6 show the plots of the errors for window sizes 3, 6, 9 and 12 respectively. The vertical axis has been labeled '1-error' to highlight the relation to the error column in the results above, however we will also refer to it as percentage performance. We have included 3 months in order to demonstrate the instability observed in the first set of experiments. Each plot shows the dynamics of peak 19 alone and the combination of peak 19 and peak 5.

Smaller window sizes, particularly 3 months and to some extent 6 months look more unstable, however this is to be expected due to the smaller number of examples and more rapid change in examples from window to window—for example at a 3 month window the examples at $t = 0$ ar all different from the examples at $t = 3$. However for a 12 month window one must increase to $t = 12$ in order to produce a distinct set of examples to $t = 0$. This accounts for the more gradual looking trends in the higher window sizes.

It is clear from the plots that there is a definite increase approximately 12 months before the original time of diagnosis. Before this 12 month point (i.e. at time point more than 12 month till diagnosis) the general trend (for window sizes 6, 9 and 12) is for the $-\log I(19)$ rule to fluctuate on approximately 30% performance—this is natural for the triplet setting, a trivial algorithm based on

	Window size					
	6 months		9 months		12 months	
t	P-value	Error	P-value	Error	P-value	Error
0	0.0027	1/8	0.0020	4/15	0.0001	4/19
1	0.0011	1/10	0.0004	4/17	0.0001	4/24
2	0.0005	2/12	0.0002	4/18	0.0001	4/26
3	0.0020	4/15	0.0001	4/19	0.0001	4/26
4	0.0002	3/15	0.0001	3/22	0.0001	3/24
5	0.0002	3/15	0.0001	3/23	0.0001	4/24
6	0.0095	3/11	0.0001	3/18	0.0001	4/20
7	0.0009	3/14	0.0005	3/16	0.0006	5/19
8	0.0002	2/14	0.0004	3/15	0.0005	4/17
9	0.0001	0/11	0.0001	1/13	0.0009	5/18
10	0.0001	0/9	0.0006	2/12	0.0032	6/18
11	0.0008	1/9	0.0019	2/11	0.0148	7/18
12	0.0012	1/9	0.0162	5/14	0.0884	10/20
13	0.2099	2/5	0.2895	6/11	0.3285	10/17
14	0.7055	2/3	0.7032	7/10	0.3944	11/18
15	0.7386	5/7	0.8646	10/13	0.5218	13/20
16	0.6255	6/9	0.5949	10/15	0.5218	13/20
17	0.6285	6/9	0.3223	10/17	0.3381	12/20
18	0.9216	9/11	0.5957	12/18	0.5287	13/20
19	0.6098	8/12	0.5184	11/17	0.2940	13/22
20	0.3731	9/15	0.4059	11/18	0.1985	13/23
21	0.4591	8/13	0.3763	9/15	0.1473	10/19
22	0.5206	7/11	0.2631	9/16	0.1160	9/18
23	0.2914	6/11	0.1215	8/16	0.0744	8/17
24	0.1454	4/9	0.0348	5/13	0.0316	6/15

Table 6: Results under Setting 2, we select $p_1 = 19$ $w = -1$, $p_2 = 5$ and $v = 1$, window sizes 6, 9 and 12 months are shown.

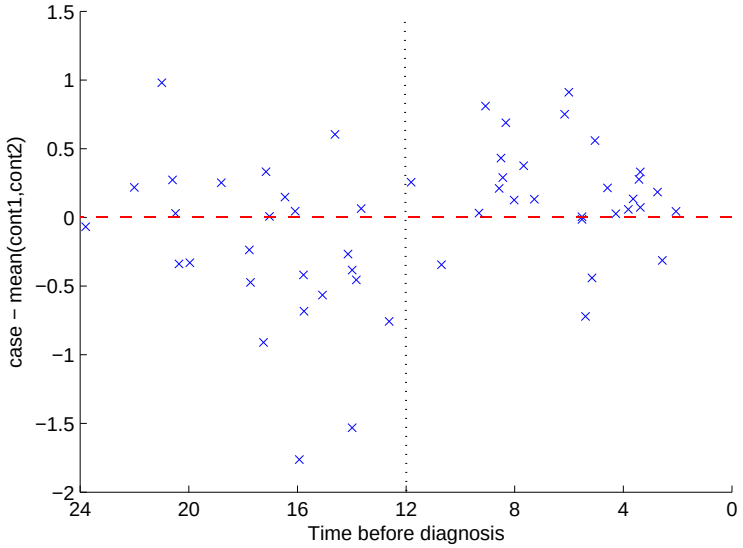

Figure 2: The figure shows the distribution of data over time and the relationship between the intensity in the cases compared to the matched controls for the decision rule: $-\log I(19) + \log I(5)$. Each cross represents a triplet.

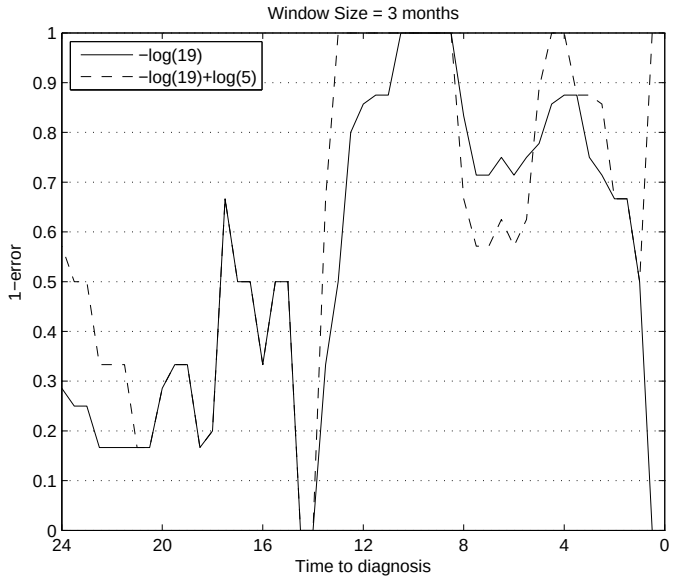

Figure 3: This plot shows the dynamic behaviour of the error yield by the two important decision rules for the window size of 3 months.

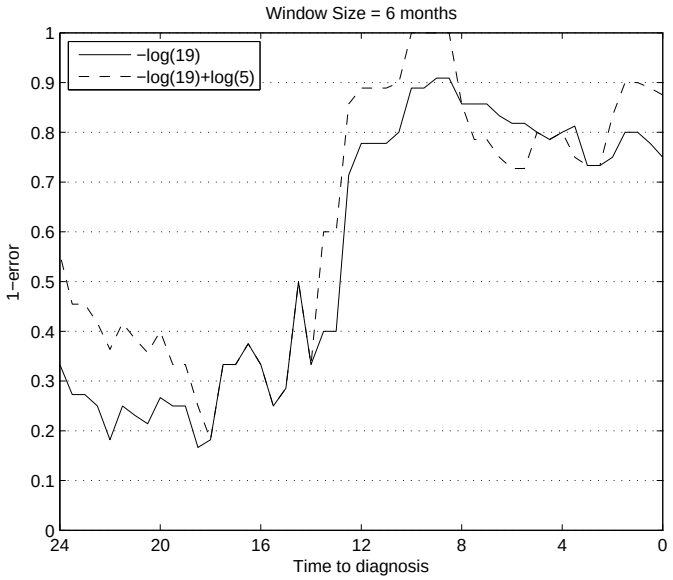

Figure 4: This plot shows the dynamic behaviour of the error yield by the two important decision rules for the window size of 6 months.

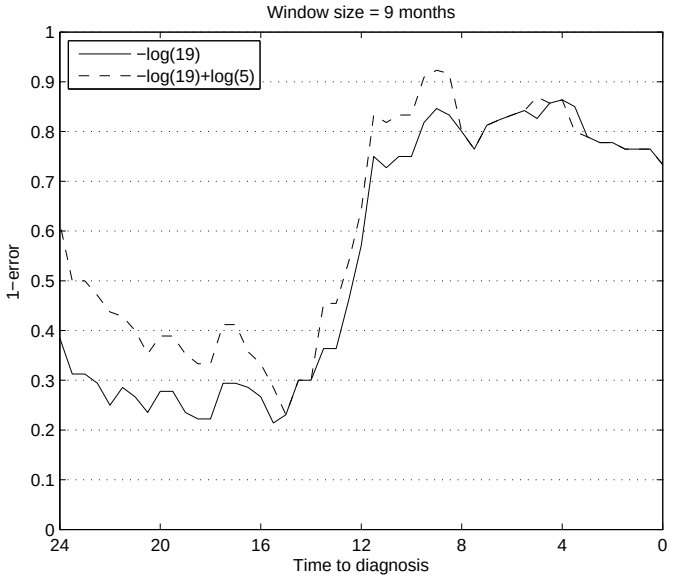

Figure 5: This plot shows the dynamic behaviour of the error yield by the two important decision rules for the window size of 9 months.

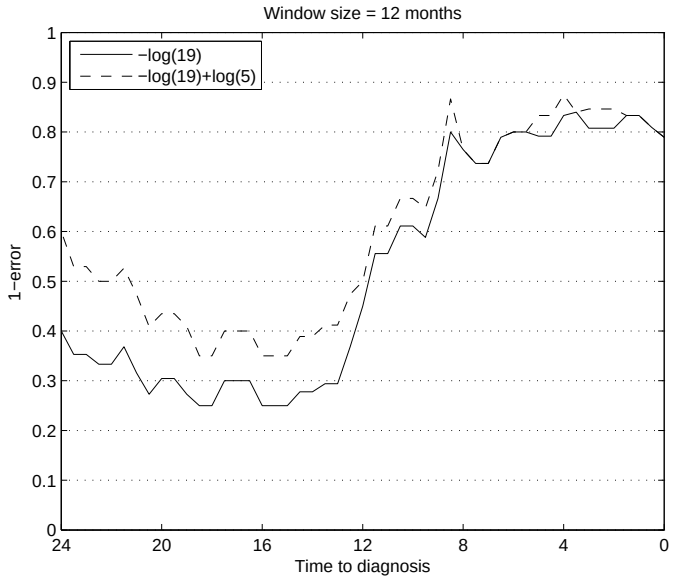

Figure 6: This plot shows the dynamic behaviour of the error yield by the two important decision rules for the window size of 12 months.

random selection of the cancer sample in each triplet would yield a $\sim 33.3\%$ accuracy. Between 12 and 8 months till diagnosis the percentage performance rises to $\sim 80\%$ and fluctuates on this value up till the point of diagnosis ($t = 0$).

6 Validation

We apply here the two interesting decision rules to the validation set. The procedure is outlined as follows:

1. For each triplet in the validation set apply the decision rule to find the predicted cancer sample.

2. The true labels for the validation set are blind, i.e. they are not known. Information provided will be percentage performance for a number of time thresholds . Given a time threshold, $t^* > 0$, we will test all triplets with time stamp $T_i \in [0, t^*]$.

Table 7 shows results for different value of t^* for the two decision rules. Note that these values can be interpreted as the sensitivity in this context. Specificity is directly related to sensitivity, it can be calculated using the following formula:

$$\text{specificity} = 50 + \frac{\text{sensitivity}}{2},$$

	Decision rule			
	$-\log I(19)$		$-\log I(19) + \log I(5)$	
t^*	Sensitivity	Specificity	Sensitivity	Specificity
3	100%(3/3)	100%	33.3%(1/3)	66.7%
4	83.3%(5/6)	91.7%	50.0%(3/6)	75.0%
5	71.4%(5/7)	85.7%	42.9%(3/7)	71.5%
6	62.5%(5/8)	81.3%	37.5%(3/8)	68.8%
7	60.0%(6/10)	80.0%	40.0%(4/10)	70.0%
8	57.1%(8/14)	78.6%	42.9%(6/14)	71.5%
9	50.0%(8/16)	75.0%	37.5%(6/16)	68.8%
10	55.6%(10/18)	77.8%	44.4%(8/18)	72.2%
11	55.6%(10/18)	77.8%	44.4%(8/18)	72.2%
12	55.0%(11/20)	77.5%	45.0%(9/20)	72.5%

Table 7: Percentage performance on the validation set; for each row we consider all triplets in the range $[t^*, 0]$. The performance is given as a percentage performance and a in brackets: (number correct classifications / total number of triplets)

as the impact on specificity (in this context) is always half that of sensitivity.

Despite the two decision rules being close in performance on the training set this is not reflected in the validation set. The results for the single peak (peak 19) are much better, particularly a points less than 6 months in advance, than the two peaks (19 and 5). This shows a classic case of overfitting, even with as little as 2 features (peaks) we have over fitted the training data. As the validation data represents only 1/3 of this already limited sized data set there are only few example to work with. However, despite this it is clear that peak 19 can still perform well particularly before 6 months in advance of the last moment where we see accuracies of 71.4% for $t^* = 5$, 83.3% for $t^* = 4$ and 100% for $t^* = 3$. Values for $t^* < 3$ are omitted as the number of triplets falls to 0.

7 Conclusions and further work

Analysis of the data has shown that there is a very interesting peak at 6638.7Da that carries statistically significant information up to 11 months in advance of the original time of diagnosis. Small improvements can be made to this peak through combination with a second peak at 5342.4Da. The improvement is with respect to the overall error, however it does not extend predictions past 11 months. It is clear from the plots of the dynamics of the errors that there is some interesting behaviour in peak 19 and in the combination of peak 19 and peak 5 at approximately 12 months before diagnosis. As mentioned results have shown that this information becomes statistically significant at 11 months. Testing the two rules on the validation data has confirmed that peak 19 is indeed useful

for early prediction, due to the small number of examples is is difficult to get extremely relatable validation, however all the results are better than a random selection, which would yield $\sim 33.3\%$. The results on the validation set for the two peak rule were much closer to this random value, it is clear that adding this feature may have performed well on the training set but is a clear case of overfitting.

It is clear from the analysis that there is are two major problems with the data provided. The first is a problem of numbers, there are to few examples to allow small enough window sizes which would enable us to be more precise with our analysis. The second problem regards the distribution of the data with respect to the time stamps (the time before diagnosis). It is of course difficult to regulate such data however in this investigation there is relatively dense data in the 0–8 month range when compared to the more sparse data in the 8–12 month range. This is also unfortunately the range in which our important peaks begin to hold statistically significant information, it would therefore be advantageous to have more data in the 8-12 month range in order to confirm the results presented here.