CLRC—TR—08—02

Analysis of serial UKCTOCS-OC data: discriminating abilities of proteomics peaks

December 2, 2008 *

Dmitry Devetyarov[1], Ilia Nouretdinov[1], Brian Burford[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

*Correspondence should be addressed to A.G. alex@cs.rhul.ac.uk

Abstract

This study aims to prove empirically that the information contained in serum proteome mass spectra can improve the discriminating ability of CA125 in early stages of ovarian cancer (OC) development. Our serial data collected in the UKCTOCS project over the period of 7 years allow us to reject the null hypothesis at significance level 5% for detection up to 15 months in advance of the diagnosis which was confirmed by histology/cytology. Moreover, in addition to what has been previously reported we identified certain mass spectrometry (MS) peaks that, in combination with CA125 level, can provide reliable long term prediction of OC. The peaks with m/z-values 7772 Da and 9297 Da were the most informative in extending the period of significant discrimination, and in combination with CA125 they proved to be able to predict the disease up to 6 and 4 months earlier than CA125 alone, respectively.

[*] The date of version 1 - May 28, 2008

1 Introduction

Clinical symptoms in OC do not appear until the late stage of the disease, and this often leads to poor outcome in the treatment of the disease. On the other hand, discovery of the disease in early stages usually results in excellent prognosis. Screening for high-risk population and usage of antigen CA125 significantly improved the diagnosis of OC. However, the CA125 test gives a large number of false positives and the CA125 elevated level usually becomes known at late stages of OC development and very often too late to make use of it. Recently, it has been shown that the use of modern mass spectrometry techniques allow us to identify some potential biomarkers in serum proteome that can improve the diagnosis of OC [4]. The assumption is that the protein/peptide patterns are linked to clinical conditions and therefore can help to reliably diagnose patients. In our previous study—the "pilot" study[1]—we reported that mass spectra from the low molecular weight serum proteome carry information useful for early detection of ovarian cancer. It was shown that combining a single MS peak with CA125 allows statistically significant discrimination at the 5% level between cases and controls up to 12 months in advance of the original diagnosis of ovarian cancer. Such combinations work much better than a single peak or CA125 alone.

In this paper we continue our analysis of the UKCTOCS data [3]. The unique feature of this study is that the women were screened for up to 7 years. The serum samples underwent prefractionation using a reversed-phase batch extraction protocol prior to MALDI-TOF MS data acquisition. A nested case control study was undertaken on serial serum samples collected from women who developed ovarian cancer and matched controls. The serum sample data was processed using the MALDI-TOF mass-spectrometry technique [7].

The paper first describes the data, our pre-processing techniques followed by the results of our data analysis. Our analysis shows that the period of statistically significant discrimination in advance of diagnosis confirmed by histology/cytology can be extended from 9 to 15 months if CA125 is combined with certain peaks extracted using MALDI-TOF mass spectrometry data acquisition.

2 Description of the data and pre-processing

The serum samples studied in this paper are from UKCTOCS [3] and collected from patients who developed cancer (referred to as *cases*) and from healthy individuals.

Each case sample is accompanied by two samples taken from healthy individuals, we will call these *matched controls*. In addition, there are 43 unmatched controls that were also included. There are 924 samples in total: 295 cases, 586 matched controls and 43 unmatched controls. For 4 case samples, the mass spectrum for one of matched controls was excluded for quality control reason. The set of 295 cancer samples was derived from 69 serial and 90 single patients eventually diagnosed with ovarian cancer. There are up to 5 serial samples for each of the cases.

All samples are provided with their CA125 measurement. For each case the *time to diagnosis* $t > 0$ is known. This is the time interval, measured in months, between the date when the measurement was taken and the date when OC was diagnosed by histology/cytology. Other information for the samples such as the time to spinning is also available.

Each sample, case or control, was analysed using MALDI-TOF based mass spectrometry and no more than 12 replicate mass spectra were generated. There were 11048 mass spectra in total for 924 samples. For each mass spectrum, the data was 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 (down-sampling, smoothing, baseline subtraction, normalisation, peak identification and peak alignment) described in [1] were applied only to the intensities, m/z-values remained unchanged.

After the pre-processing, a set of peak groups are identified. The peak groups are ordered by their frequency (i.e., percentage of samples having a non-aligned peak belonging to this peak group). In the following, we consider 67 most frequent peaks, with frequency at least 33%. The full list of 67 peaks is provided in Appendix A. The most frequent 5 peaks are represented in Table 1.

Number	M/z-values, Da	Frequency
1	4213.39	11043 out of 11048 spectra
2	7772.06	11040 out of 11048 spectra
3	9297.84	11040 out of 11048 spectra
4	5341.04	11036 out of 11048 spectra
5	5909.15	11029 out of 11048 spectra

Table 1: The list of 5 most frequent peaks.

Examples of peak groups 2 and 3 plotted together are shown in Figure 1.

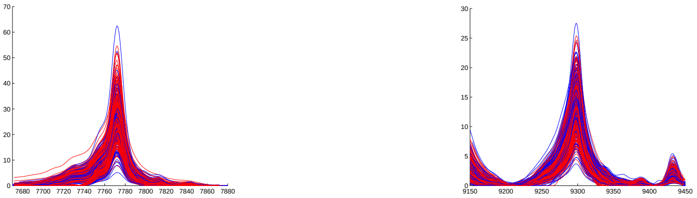

Figure 1: Peak groups 7772 Da (peak 2) and 9297 Da (peak 3)

As peak plots show, there is no clear visual separation between cases and controls that was achieved in [4]. However, the preprocessing described in this work was applied to the data described in [4] and provided visual separation as well.

Finally, for each sample, intensities of 67 peaks are averaged across the replicates of this sample. A sample, after pre-processing, includes information on averaged intensities $I(1), \ldots, I(67)$ of 67 most frequent peaks. We obtained peak intensities of 924 samples, but we are analysing only 873 samples that are divided into 291 *triplets*, each consisting of a case sample and the two matched control samples. We will say that the case sample is *labelled (as a case).*

For classification, unmatched controls and cases with only one matched control were excluded. We also had to exclude some more triplets due to missing information and triplets containing at least one sample with time to spinning more than 24 hours.

As a result, there remain 179 triplets that contain 358 controls and 179 cases taken from 104 case patients:

- 59 case patients with one measurement,
- 26 case patients with 2 measurements,
- 11 case patients with 3 measurements,
- 5 case patients with 4 measurements,
- 3 case patients with 5 measurements.

A set of triplets from the same patient is called a *triplet group*. Each triplet τ is assigned a non-negative value $T(\tau)$, the time to diagnosis for the case measurement in the group. Each sample is provided with the CA125 level C.

3 Classification of triplets—Problem statement

For each $t = 0, 1, 2, \ldots$ let $S_{t,u}$ be the set of triplets of measurements taken t months before the diagnosis, or as little earlier as possible and not exceeding $t + u$ months. Formally, $S_{t,u}$ is defined to be the following set of triplets: start with $S_{t,u} := \emptyset$ and for each triplet group put in $S_{t,u}$ the triplet τ in the group with the smallest $T(\tau)$ satisfying $t \leq T(\tau) \leq t + u$ if it exists. Two triplets from the same group are not allowed to be used in the same set $S_{t,u}$ because they can be dependent on each other. In this paper we use a 6-month window size: $u = 6$. For example, $S_{0,6}$ contains 68 triplets, $S_{12,6}$ contains 28 triplets.

We will use a rather limited class of rules for *triplet classification*, i.e. identification of the labelled sample within a triplet.

Each classification rule is specified by three numbers (p, w, v), which are a peak number $p \in P$ (P is a set of peak numbers and can be equal to $\{1, \ldots, 67\}$ or one certain peak number) and weights w, v, $w \in W = \{0, 0.5, 1, 2\}$, $v \in V = \{-1, 1\}$. For each triplet, the classification rule (p, w, v) assigns a 'case' label to the sample with the largest value of $w \log C + v \log I(p)$, where C and $I(p)$ are the CA125 level and the intensity of peak p, respectively. The logarithms are taken to remove the arbitrary units of measurements.

If $w \neq 0$ (CA125 is taken into account), the rule has an equivalent form: $\log C + \frac{v}{w} \log I(p)$. Thus, we are usually looking for a term additional to CA125, v is responsible for the possible sign of its influence (whether it is expected to be larger at cases or at controls), and $1/w$ determines to what degree it is taken into account.

Let $err(\tau; p, w, v)$ be 0 if (p, w, v) correctly identifies the labelled sample in a triplet τ and 1 otherwise (in the case of a wrong classification), and

$$Err(S; p, w, v) := \sum_{\tau \in S} err(\tau; p, w, v)$$

stand for the number of errors made by (p, w, v) on a set S of triples. As our test statistic we take the pair (E_0, p_0) of the least number of errors and the number of the most frequent peak where it is achieved. Formally,

$$E_0 := \min_{p \in P, w \in W, v \in V} Err(S_{t,u}; p, w, v),$$

$$p_0 := \min\{p : \exists w \exists v Err(S_{t,u}; p, w, v) = E_0\}.$$

Pairs (E_0, p_0) are ordered lexicographically, and p_0 is added to break ties: if two rules lead to the same error rate, the rule involving a more frequent peak has priority when we calculate p-values.

3.1 Summary of the main findings

We consider 6-month time slots starting with months $t = 0, 1, 2, \ldots, 16$. For each set of samples $S_{t,6}$ the best pair (E_0, p_0) was selected for certain sets of parameters P, W and V as described in section above. Tests were designed in order to check the significance of our findings. The p-values given by these tests for the set of all 67 peaks ($P = \{1, \ldots, 67\}$) and separately for peak 2 ($P = \{2\}$) and peak 3 ($P = \{3\}$) are represented in Table 2. The tests check the null hypothesis that the assignment of labels within triplets is independent of the information contained in CA125 levels and intensities of certain peaks: all peaks, peak 2 only or peak 3 only. The methodology of p-value calculation is described in detail in Section 3.2.

t	$\lvert S(t,6)\rvert$	p-values for all peaks	Adjusted p-values for peak 2	Adjusted p-values for peak 3
0	68	0.0001	0.001	0.001
1	56	0.0001	0.001	0.001
2	47	0.0001	0.001	0.001
3	36	0.0001	0.001	0.001
4	27	0.0001	0.001	0.001
5	23	0.0006	0.004	0.007
6	20	0.0028	0.01	0.046
7	17	0.0141	0.017	0.098
8	17	0.0019	0.02	0.02
9	20	0.0076	0.01	0.009
10	28	0.0003	0.001	0.001
11	28	0.0042	0.008	0.004
12	28	0.0585	0.033	0.049
13	30	0.0168	0.007	0.015
14	25	0.0304	0.015	0.301
15	20	0.0464	0.022	0.577
16	10	0.4101	5.165	5.979

Table 2: Summary of p-values for triplet classification.

The first column of Table 2 shows the time to diagnosis t, which is the most recent end-point of the time window. The second column shows the size of $S_{t,6}$ — the number of samples in the considered 6-month time slot. The other columns represent p-values for the set of all peaks (these p-values do not require adjustment) and adjusted p-values for peak 2 and peak 3 separately.

The table demonstrates that CA125 and 67 peak intensities allow us to reject the null hypothesis at significance level of 5% for up to 15 months with a single exception of month 12, which still has a p-value less than 6% , while the information contained in CA125 and peak 2 or CA125 and peak 3 provide significant p-values for detection up to 15 and 13 months before diagnosis, respectively.

3.2 Statistical analysis of all peaks

In order to check the robustness of our triplet classification using MS peaks, we designed three types of statistical tests which reject the following null hypotheses about classification of $S_{t,u}$:

1. The null hypothesis that assignment of labels within triplets is independent of CA125 and peak intensities.
2. The null hypothesis that assignment of labels within triplets is independent of CA125 levels.

3. The null hypothesis that when assigning labels within triplets, pairs (label, CA125) are independent of peak intensities, that is, peak intensities do not contain any information useful to improve predictive ability of CA125.

The detailed explanation of corresponding p-value calculation and their meaning is represented below.

- **The main p-value** (or just 'p-value') is based on the null hypothesis that the assignment of the case label within each triplet in $S_{t,u}$ is independent of the information contained in CA125 and all MS peaks.

  We calculated these and all other p-values experimentally, using the Monte-Carlo method. Suppose $E_0 := \min_{p \in P, w \in W, v \in V} Err(S_{t,u}; p, w, v)$, that is the minimum number of errors occurring in prediction by CA125 and one of the peak intensities, and $p_0 := \min\{p : \exists w \exists v Err(S_{t,u}; p, w, v) = E_0\}$, the peak with the highest commonality that can provide this minimum error rate.

  The statistic equal to the pair (E', p') calculated the same way as (E_0, p_0) is collected for a large number N (we used $N = 10^4$) of times on the same data set with randomly reassigned case labels.

  We then calculate the number Q of times the statistic is as good as or better than the statistic (E_0, p_0) computed from the true labels, where 'as good as or better than' is understood in the sense of the lexicographical order: $(E', p') \leq (E_0, p_0)$ means that either $E' < E_0$ or $E' = E_0$ and $p' \leq p_0$. The p-value is then estimated as $(Q + 1)/(N + 1)$.

Algorithm 1 Main p-value calculation

```
Input: $t$, time to diagnosis.
Input: $N = 10^4$, number of trials.
$E_0 := \min_{p,w,v} Err(S_{t,u}; p, w, v)$
$p_0 := \min\{p : \exists w \exists v Err(S_{t,u}; p, w, v) = E_0\}$
$Q := 0$
for $j := 1, \ldots, N$ do
    Assign the case label to a randomly chosen sample in each triplet in $S_{t,u}$
    Recalculate $E' := \min_{p \in P, w \in W, v \in V} Err(S_{t,u}; p, w, v)$
    Recalculate $p' := \min\{p : \exists w \exists v Err(S_{t,u}; p, w, v) = E'\}$
    if $(E', p') \leq (E_0, p_0)$ then
        $Q := Q + 1$
    end if
end for
Output: $(Q + 1)/(N + 1)$ as the $p$-value.
```

- The previous p-value demonstrates significance of predictions provided by CA125 levels in combination with peak intensities. To compare discriminating ability of the combination of CA125 and peaks with the well-known benchmark, CA125 on its own, we need to calculate **CA125 p-values**. This test checks the null hypothesis that labels at $S_{t,u}$ are independent of CA125 levels.

These p-values were also calculated using the Monte-Carlo method. However, similar p-values can be calculated theoretically as described in Appendix B.

Suppose $E_0 = Err(S_{t,u}; 1, 1, 0)$, that is the number of errors occurring in prediction by CA125 only. The statistic equal to number of errors is collected for $N = 10^4$ times on the same data set with randomly reassigned case labels, and then the algorithm counts the number Q of times the statistic is as good as or better than the statistic E_0 computed from the true labels. The p-value is then estimated as $(Q + 1)/(N + 1)$.

Algorithm 2 CA125 p-value calculation

```
Input: $t$, time to diagnosis.
Input: $N = 10^4$, number of trials.
$E_0 := Err(S_{t,u}; 1, 1, 0)$
$Q := 0$
for $j := 1, \dots, N$ do
  Assign the case label to a randomly chosen sample in each triplet in $S_{t,u}$
  Recalculate $E' := Err(S_{t,u}; 1, 1, 0)$
  if $E' \leq E_0$ then
    $Q := Q + 1$
  end if
end for
Output: $(Q + 1)/(N + 1)$ as the $p$-value.
```

- In our current research we also introduced **conditional p-values** — in addition to main p-values and CA125 p-values that were considered in our previous work [1].

  Suppose that the main p-value is significant. That means that CA125 with MS peaks are able to discriminate between controls and cases. In this case we wish to separate contributions of CA125 and MS peaks. For this reason, we consider conditional p-value.

  Our null hypothesis is the independence between a pair (label, C) and $(I(1), \dots, I(67))$ as opposed to the main p-value hypothesis that can be interpreted as the independence between a label and $(C, I(1), \dots, I(67))$.

  The only difference from the computation of the main p-values is the following. At each loop step instead of assigning the case label randomly, we generate a random permutation of 3 elements for each triplet and apply the permutation to both labels (case, control, control) and CA125 levels of triplet measurements.

Algorithm 3 Conditional p-value calculation

```
**Input:** $t$, time to diagnosis.
**Input:** $N = 10^4$, number of trials.
$E_0 := \min_{p,w,v} Err(S_{t,u}; p, w, v)$
$p_0 := \min\{p : \exists w \exists v Err(S_{t,u}; p, w, v) = E_0\}$
$Q := 0$
for $j := 1, \ldots, N$ do
  for each triplet in $S_{t,u}$ do
    Consider a random permutation $s : \{1, 2, 3\} \to \{1, 2, 3\}$
    Apply $s$ to the labels (case, control, control) of this triple
    Apply the same $s$ to the CA125 values $(C_1, C_2, C_3)$ of this triple
  end for
  Recalculate $E' := \min_{p \in P, w \in W, v \in V} Err(S_{t,u}; p, w, v)$
  Recalculate $p' := \min\{p : \exists w \exists v Err(S_{t,u}; p, w, v) = E'\}$
  if $(E', p') \leq (E_0, p_0)$ then
    $Q := Q + 1$
  end if
end for
**Output:** $(Q + 1)/(N + 1)$ as the $p$-value.
```

Neither of p-values described above require adjustment. Main p-values as well as conditional p-values do not require any adjustment as the numbers of errors are calibrated by the Monte-Carlo procedure taking into account all the rules at the same stage. CA125 p-value does not require any adjustment, since in this case there is no set of rules to select from.

Experimental results

Given that we have only a limited number of samples, we considered 6-month time slots starting in different months ($S_{t,6}$, where t = 0, 1, ..., 16). For each of these slots we checked hypotheses of random label distribution, calculating p-values described above, and looked for certain peaks that carry the most useful information for discrimination between cancer and healthy samples. The initial results of the analysis are represented in Table 3.

Table 3: Initial statistical analysis. The columns are: time t to diagnosis in months; the number $|S_{t,6}|$ of cases with measurement taken between t and $t+6$ months before diagnosis; the number E_C of errors when classifying the triplets in $S_{t,6}$ with CA125 alone; CA125 p-value; the minimal number $E_{C,p}$ of errors when classifying with CA125 (taken with weight $w \in \{0, 1/2, 1, 2\}$) and intensity $I(p)$ of a peak (taken with weight $v \in \{-1, +1\}$); number p of this peak; v/w; p-value for overall significance of this result; conditional p-value for significance of non-CA125 component.

t	$\|S_{t,6}\|$	CA125 only		CA125 and all peaks				
		E_C	CA125	$E_{C,p}$	p	v/w	Main	Conditional
			p-value				p-value	p-value
0	68	2	0.0001	1	01	$+1$	0.0001	0.3164
1	56	4	0.0001	2	07	$+1$	0.0001	0.2446
2	47	6	0.0001	3	15	-2	0.0001	0.1795
3	36	8	0.0001	4	15	-2	0.0001	0.0746
4	27	7	0.0001	4	15	-2	0.0001	0.6734
5	23	7	0.0008	4	15	-1	0.0006	0.4885
6	20	6	0.0010	4	15	-1	0.0028	0.6973
7	17	6	0.0071	4	01	-1	0.0141	0.6034
8	17	5	0.0021	3	01	-1	0.0019	0.1523
9	20	7	0.0042	5	02	-1	0.0076	0.4497
10	28	14	0.0503	6	03	-1	0.0003	0.0013
11	28	15	0.1028	8	03	-1	0.0042	0.0078
12	28	17	0.3164	10	02	-2	0.0585	0.0658
13	30	16	0.0895	10	02	-2	0.0168	0.0428
14	25	16	0.4661	8	02	-2	0.0304	0.0206
15	20	13	0.5211	6	02	-2	0.0464	0.0609
16	10	6	0.4406	2	67	$+1/0$	0.4101	0.5066

The first and second columns of Table 3 are the same as in Table 2. Columns 3 and 4 represent analysis results for prediction with CA125 only. Columns 5–9 show results for prediction with CA125 in combination with the set of all peaks.

The third column provides the number E_C of errors made on the triplets $S_{t,u}$ by the classification rule $\log C$, i.e., classification by CA125 only.

The fourth column (CA125 p-values) is the measure of significance for this result, the probability to obtain an equal or even more extreme result, by chance doing classification at random. Recall that the expected probability of error in triple classification is $2/3$, so misclassifying 16 of 30 triplets (as for $t = 13$) is not significant (p-value $= 9\%$) but still better than random. The results for classification using CA125 only are significant for up to 9 months.

The fifth column $E_{C,p}$ gives the best quality (the smallest number of errors) which may be achieved by a classification rule from our list. The following rules are considered: classification by $w \log C + v \log I(p)$ where $p \in \{1, \dots, 67\}$, w is any of $0, 1/2, 1, 2$, v is either $+1$ or -1. Equivalent form for the rule is $\log C + (v/w) \log I(p)$ where v/w is any of $-1/0, -2, -1, -1/2; +1/2, +1, +2, +1/0$. Notation $+1/0$ (or $-1/0$) means that the peak $I(p)$ is used alone without CA125. The sixth and seventh columns of the table are p and v/w for the best classification rule.

Experiments show that any result can be more or less improved in such a way ($E_{C,p}$ is less than E_C), but we need to check that this is not obtained by chance, due to large choice of classification rules. Moreover, we need to check both the probability to obtain low error rate $E_{C,p}$ by chance and the probability to decrease it from E_C to $E_{C,p}$ by chance. Two types of p-values answer these two questions. Main p-value measures the chance to obtain the error rate not worse than $E_{C,p}$ at random, and conditional p-value — the chance to get the error equal to or less than $E_{C,p}$ if peak intensities are reshuffled, but the CA125 value is real. Actually, conditional p-values are much more important for us, as we know in advance that CA125 is a useful biomarker and, therefore, we are mainly interested in whether addition of anything else to it may lead to statistically significant improvement.

Main and conditional p-values are represented in the last 2 columns of Table 3. Main p-values are also shown in the summary table — Table 2. From 'Main p-value' column, we can see that the null hypothesis can be rejected at level 5% for up to 15 months (with the only exception being 12 months where it is about 6%). This contrasts with using CA125 alone ('CA125 p-value' column), which produces significant results for up to 9 months. Conditional p-values show that contribution of the added MS peak becomes essential only from 10 to 15 months, so we should pay more attention to the best quality rules for these months, that is, peaks 2 and 3 with corresponding coefficients v and w.

3.3 Statistical analysis of peaks 2 and 3

In this section we check significance of peaks 2 and 3 identified as the most informative. Thus, we check whether a specific peak (2 or 3) contains some information useful to improve triplet classification in comparison with the use of CA125. These peaks are plotted in Figure 1.

Main p-values and conditional p-values presented before are adjusted to this problem. Main p-values will be given by the test checking the null hypothesis that assignment of labels within triplets is independent of CA125 and peak 2 (3). Conditional p-values will be calculated by the test checking the null hypothesis that when assigning labels within triplets, pairs (label, CA125) are independent of peak 2 (3).

The only difference in calculation of these p-values from Algorithms 1 and 3 is the set of rules we are selecting from. In this case, we consider the following sets of parameters: P is peak 2 (or peak 3), W and V are the same ($\{0, 0.5, 1, 2\}$ and $\{-1, 1\}$, respectively).

Experimental results

Table 4 represents the results for prediction by CA125 and peak 2 (columns 5–8) or CA125 and peak 3 (columns 11–14). Error rate and p-value for prediction using CA125 only are included in the beginning of the table for comparison (columns 3 and 4). ‘Rule’ columns show the best rules selected: CA in this columns means $\log C$, p2 means $\log I(2)$, p3 means $\log I(3)$. Columns 6(12), 7(13), 8(14) represent error rates for the best rule, main p-values and conditional p-values, respectively.

Conditional p-value shows that improvement achieved by adding information from peak 2 or 3 is not significant up to 9 months, as CA125 itself is good for separation, and it can’t be improved significantly by adding a peak-dependent term.

The table demonstrates that in time slots starting with 10–16 months, main p-values and conditional p-values are similar, hence significant or insignificant at the same time. This confirms that possible improvement in predictive ability of CA125 with a peak is achieved due to information contained in this peak.

Main and conditional p-values shown in the table require adjustment.

When adjusting by 10 peaks the threshold for significance is $0.05/10 = 0.005$ and thus the results are statistically significant for up to 15 months (with peak 2) and 13 months (with peak 3), respectively. Main p-values for peak 2 and peak 3 represented in Table 2 are adjusted.

Table 4: Experimental results for triplet classification with a fixed peak (peak 2 or 3).

t	$\lvert S_{t,6}\rvert$	CA125 only		CA125 and peak 2			
		E_C	p-value	Rule	Err.	p-value	Cond. p-value
0	68	2	0.0001	2CA−p02	2	0.0001	1
1	56	4	0.0001	2CA−p02	4	0.0001	1
2	47	6	0.0001	2CA−p02	5	0.0001	0.5295
3	36	8	0.0001	2CA−p02	7	0.0001	0.834
4	27	7	0.0001	2CA−p02	6	0.0001	0.9297
5	23	7	0.0008	CA−p02	6	0.0004	0.4103
6	20	6	0.001	CA−p02	5	0.0010	0.1618
7	17	6	0.0071	CA−p02	4	0.0017	0.1612
8	17	5	0.0021	CA−p02	4	0.0020	0.3303
9	20	7	0.0042	CA−p02	5	0.001	0.083
10	28	14	0.0503	CA/2−p02	7	0.0001	0.0007
11	28	15	0.1028	CA/2−p02	9	0.0008	0.002
12	28	17	0.0895	CA/2−p02	10	0.0033	0.0045
13	30	16	0.3164	CA/2−p02	10	0.0007	0.0014
14	25	16	0.4661	CA/2−p02	8	0.0015	0.0011
15	20	13	0.5211	CA/2−p02	6	0.0022	0.0011
16	10	6	0.4406	CA/2+p02	5	0.5165	0.4836

t	$\lvert S_{t,6}\rvert$	CA125 and peak 3			
		Rule	Err.	p-value	Cond. p-value
0	68	2CA−p03	2	0.0001	0.9114
1	56	CA+p03	4	0.0001	0.8815
2	47	2CA−p03	5	0.0001	0.5279
3	36	2CA−p03	7	0.0001	0.6777
4	27	CA+p03	6	0.0001	0.8735
5	23	2CA−p03	6	0.0007	0.6316
6	20	CA−p03	6	0.0046	0.6872
7	17	CA−p03	5	0.0098	0.5735
8	17	CA−p03	4	0.0020	0.2781
9	20	CA−p03	5	0.0009	0.0693
10	28	CA−p03	6	0.0001	0.0002
11	28	CA−p03	8	0.0004	0.0005
12	28	CA−p03	10	0.0049	0.0049
13	30	CA−p03	10	0.0015	0.0016
14	25	CA−p03	10	0.0301	0.0181
15	20	CA−p03	8	0.0577	0.0683
16	10	CA/2+p03	5	0.5979	0.7643

Figures 2 and 3 illustrate the performance of classification rules $\log C - 2\log I(2)$ and $\log C - \log I(3)$, respectively, in comparison with the performance

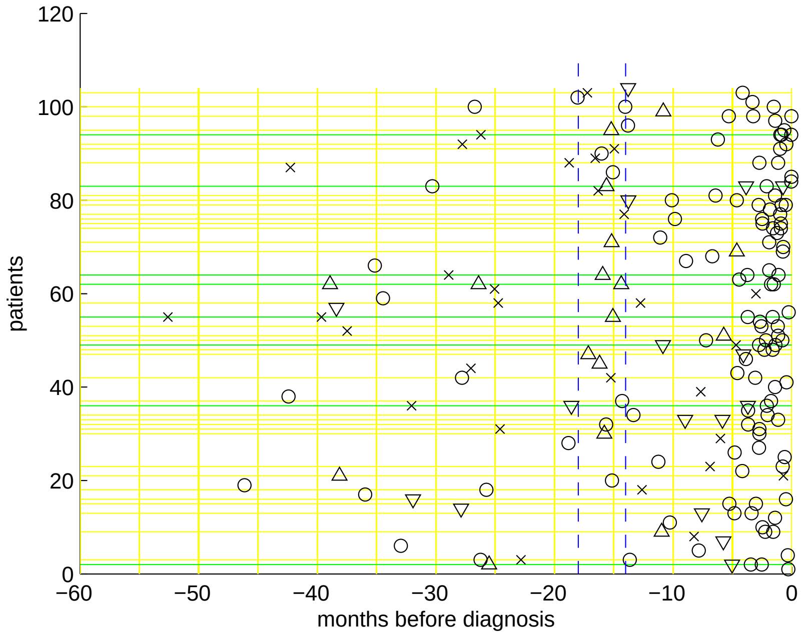

Figure 2: Comparison of $\log C$ and $\log C - 2\log I(2)$ rules on time/patient scale. A circle means that a triple was correctly classified by both rules. A cross means misclassification in both cases. A triangle shows either improvement (up-directed) or deterioration (down-directed) after addition of $-2\log I(2)$ component. The figure demonstrates that most samples where addition of a peak to CA125 works (marked as up-directed triangles) are in the interval of 13–16 months before the diagnosis.

of $\log C$. The horizontal axis shows time to diagnosis, the vertical one — triplet groups in this time interval. The figures demonstrate that most triplets with the measurement date close to diagnosis date are predicted correctly even by the $\log C$ rule, most samples where addition of a peak to CA125 allows correct classification (marked as up-directed triangles) are in the interval of 13–16 months before the diagnosis.

Figures 4 and 5 show the median dynamics of $\log C$ versus $\log C - \log I(3)$ and $\log C - 2\log I(2)$ for case measurements. For each time moment, the latest available case measurement for each triplet group is taken into account. These measurements are averaged by median through all triplet groups. The figures illustrate that rules combining CA125 with peak intensity start to grow earlier than $\log C$. However, the CA125 growth at the moments close to diagnosis is quicker due to the exponential growth of CA125.

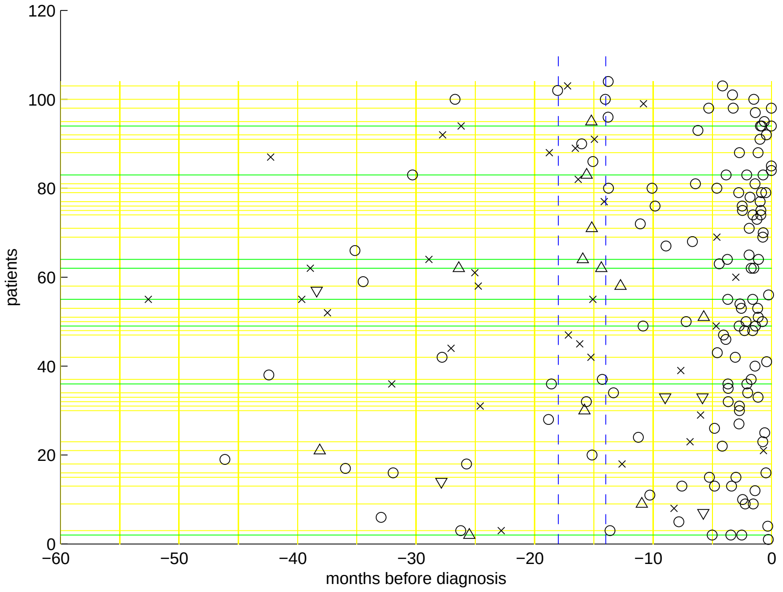

Figure 3: Comparison of $\log C$ and $\log C - \log I(3)$ rules on time/patient scale. A circle means that a triple was correctly classified by both rules. A cross means misclassification in both cases. A triangle shows either improvement (up-directed) or deterioration (down-directed) after addition of $-\log I(3)$ component. The figure demonstrates that most samples where addition of a peak to CA125 works (marked as up-directed triangles) are in the interval of 13–16 months before the diagnosis

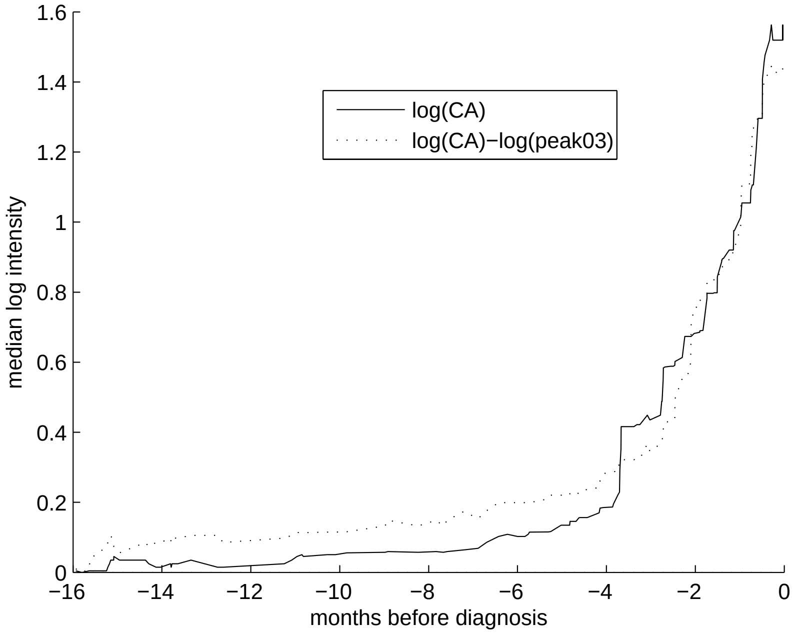

Figure 4: Median dynamics of rules $\log C$ and $\log C - \log I(3)$ (for cases only)

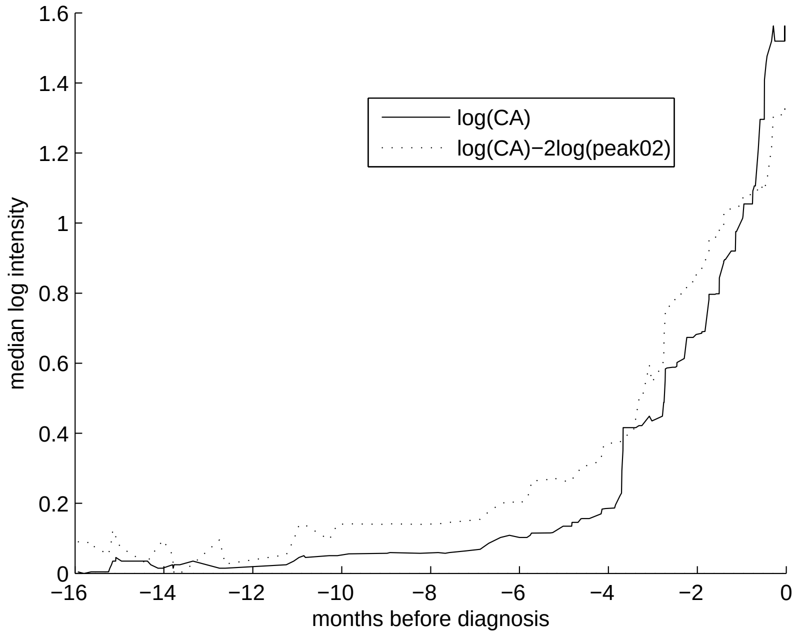

Figure 5: Median dynamics of rules $\log C$ and $\log C - 2 \log I(2)$ (for cases only)

The median dynamics of $\log C$ versus $\log C - \log I(3)$ and $\log C - 2\log I(2)$ for case measurements complemented with statistical characteristics of controls are shown in Figures 6 and 7 in Appendix C.

Dynamics of peak 2 and peak 3 are represented in Figures 8 and 9 of Appendix C.

4 Conclusion

The purpose of this study has been to demonstrate that mass spectra carry significant information to make an early diagnosis of OC and to identify peaks that allow making this early diagnosis. Intensities of certain peaks combined with the level of CA125 provide statistically reliable information for cancer prediction. The predictive power of CA125 alone is more limited.

We can pinpoint two peaks that can make early detection of OC possible and deserve much attention:

- peak 2 (m/z-value = 7772 Da);
- peak 3 (m/z-value = 9297 Da).

Peak 3 is similar to UKOPS peak CTAPIII with m/z-value 9288 Da, which has lower intensities for diseased samples as well.

In our previous work (the "pilot" study[1]) we identified peaks 10 and 18 as ones that occurred the most frequently in rules significantly discriminating OC cancer patients and healthy patients. Among 67 most frequent peaks identified in UKCTOCS OC data we found peaks with m/z-ranges closest to "pilot" study peaks 10 and 18. They are the following:

- UKCTOCS OC peak 40 (m/z-range 4277.9-4288.7 Da, mean value 4285.7 Da) is the closest to "pilot" study peak 10 (m/z-range 4288.0-4300.8 Da, mean value 4292.5 Da);
- UKCTOCS OC peak 49 (m/z-range 3136.1-3165.1 Da, mean value 3161.7 Da) is the closest to "pilot" study peak 18 (m/z-range 3169.1-3176.4 Da, mean value 3172.1 Da).

One could speculate that peaks 40 and 49 represent the same peaks as 10 and 18 in the "pilot" study since m/z-value ranges are close enough, and different data and different machines could cause a shift in m/z-values. However, peaks 40 and 49 do not occur in selected classification rules.

The detailed results of our current research are shown in Tables 3 and 4. The main findings are accumulated in Table 2. The difference between Tables 3 and 4 is in the hypothesis being checked:

- Hypothesis 1 checked in Table 3 is that all the peaks contain some information useful to improve triplet classification in comparison with the use of CA125.

- Hypothesis 2 checked in Table 4 is that a specific peak (peak 2 or peak 3) contains such useful information.

In our previous work [1] we found out that only Hypothesis 1 could be confirmed statistically; here we are also checking Hypothesis 2.

In general, we see that CA125 itself is enough for satisfactory prediction for up to 9 months before the diagnosis. It follows from conditional p-values that, in this range, other peaks cannot add significant improvement to this, and the quality of non-conditional p-values is caused by CA125 itself.

For more than 9 months before the diagnosis, CA125 produces less information, and this can be significantly completed by other peaks. The time where combination with MS peaks works varies for different hypotheses. This time is 15 months for Hypothesis 1 and 15 or 13 months for Hypothesis 2 for peaks 2 or 3, respectively. These results are summarised in Table 5.

	Period of significant discrimination
CA125	9 months
CA125 + all peaks	15 months
CA125 + peak 2	15 months
CA125 + peak 3	13 months

Table 5: The predictive ability of CA125 on its own, with all peaks and certain peaks.

Thus, we can come to the following conclusions:

1. Mass spectra contain information extending the period of statistically significant discrimination between controls and cases provided by CA125 to up to 15 months in advance of the diagnosis by histology/cytology.

2. Peak 2 (m/z-value = 7772 Da) and peak 3 (m/z-value = 9297 Da) separately contain such information for up to 15 and 13 months in advance of the diagnosis by histology/cytology.

An interesting point in this discussion is that the moment $T = 0$ is the moment of the diagnosis confirmed by histology/cytology, but not the clinical diagnosis. The women had no clinical symptoms and were picked up by the screening either by the CA125/Risk of Ovarian Cancer Algorithm (ROC) strategy [3] or the ultrasound strategy. The time of clinical diagnosis is not known for any of these women since the doctors involved in the study did not wait for the women to become symptomatic and have a clinical diagnosis. They operated on them and diagnosed the cancer earlier. Previous studies suggested that screening may be able to pick up ovarian cancer 18 months before they would have presented clinically as ovarian disease. This means that CA125 and MS peaks are able to discriminate between healthy samples and samples with OC up to 33 months in advance of the clinical diagnosis.

To sum up, different techniques allow us to provide statistically significant predictions of OC up to 33 months in advance of the date of clinical diagnosis. This period can be broken down by the cause of discrimination:

- 18 months - due to screening methods;
- 10 months - due to information contained CA125 levels;
- 5 months - due to information contained in proteomics peaks.

Acknowledgements

This work was undertaken in UCL/UCLH, Computer Learning Research Centre, Royal Holloway, University of London and Reading University. UCLH/UCL received a proportion of funding from the Department of Health's NIHR Biomedical Research Centres funding scheme. The work was also supported by EPSRC grants EP/E00053/1 (Machine Learning for Resource Management in Next-Generation Optical Networks), EP/F002998/1 (Practical competitive prediction), MRC grant G0301107 (Proteomic analysis of the human serum proteome), VLA of DEFRA grant on Development and Application of Machine Learning Algorithms for the Analysis of Complex Veterinary Data Sets, and EU FP7 grant O-PTM-Biomarkers (2008–2011).

References

[1] Alex Gammerman, Volodya Vovk, Brian Burford, Ilia Nouretdinov, Zhiyuan Luo, Alexey Chervonenkis, Mike Waterfield, Rainer Cramer, Paul Tempst, Josep Villanueva, Musarat Kabir, Stephane Camuzeaux, John Timms, Usha Menon and Ian Jacobs. **Serum Proteomic Abnormality Predating Screen Detection of Ovarian Cancer.** The Computer Journal, 2008. Published by Oxford University Press on behalf of The British Computer Society.

[2] T.P.Conrad, V.A.Fusaro, S.Ross, D.Johan, V.Rajapakse, B.A.Hitt, S.M.Steinberg, E.C.Kohn, D.A.Fishman, G.Whiteley, J.C.Barrett, L.A.Liotta, E.F.Petricoin, and T.D.Veenstra. **High-resolution serum proteomic features for ovarian cancer detection.** Endocrine-Related Cancer (2004), v.11, 163-178.

[3] Menon U, Skates SJ, Lewis S, Rosenthal AN, Rufford B, Sibley K, Macdonald N, Dawnay A, Jeyarajah A, Bast RC Jr, Oram D, Jacobs IJ: **Prospective study using the risk of ovarian cancer algorithm to screen for ovarian cancer**. *J Clin Oncol* 2005, **23**:7919–7926.

[4] Villanueva, J., Shaffer, D. R., Philip, J., Chaparro, C. A., Erdjument-Bromage, H., Olshen, A. B., Fleisher, M., Lilja, H., Brogi, E., Boyd, J., Sanchez-Carbayo, M., Holland, E. C., Cordon-Cardo, C., Scher, H. I. &

Tempst, P. **Differential exoprotease activities confer tumor-specific serum peptidome patterns**. *J Clin Invest* 2006, **116**:271–284.

[5] Villanueva J, Lawlor K, Toledo-Crow R, Tempst P: **Automated serum peptide profiling**. *Nature Protocols* 2006, **1**:880–891.

[6] Villanueva J, Shaffer DR, Philip J, Chaparro CA, Erdjument-Bromage H, Olshen AB, Fleisher M, Lilja H, Brogi E, Boyd J, Sanchez-Carbayo M, Holland EC, Cordon-Cardo C, Scher HI, Tempst P: **Differential exoprotease activities confer tumor-specific serum peptidome patterns**. *J Clin Invest* 2006, **116**:271–284.

[7] Timms JF, Arslan-Low E, Gentry-Maharaj A, Luo Z, T'Jampens D, Podust VN, Ford J, Fung ET, Gammerman A, Jacobs IJ, Menon U: **Preanalytic influence of sample handling on SELDI-TOF serum protein profiles**. *Clin Chem 2007*, **53**:645–656.

Appendix A: Full peak list

Number	M/z-values, Da	Frequency (out of 11048)	Number	M/z-values, Da	Frequency (out of 11048)
1	4213.4	11043	35	1704.5	7292
2	7772.1	11040	36	992.2	7109
3	9297.8	11040	37	2084.5	7048
4	5341.0	11036	38	9724.2	6658
5	5909.2	11029	39	5583.3	6570
6	6636.0	10994	40	4285.7	6517
7	4647.8	10904	41	2993.5	6460
8	4057.2	10878	42	2606.0	6422
9	3243.9	10739	43	2274.3	6358
10	3959.0	10705	44	740.9	6349
11	4967.9	10682	45	2512.8	6345
12	3772.6	10609	46	4759.2	6193
13	2757.3	10596	47	3527.0	6005
14	8610.9	10537	48	1742.6	5873
15	8937.3	10437	49	3161.7	5591
16	4478.8	10393	50	3886.1	5586
17	2662.9	10367	51	6385.4	5361
18	1547.6	10222	52	1262.6	5188
19	5007.8	10174	53	853.1	5001
20	2381.1	10088	54	1790.0	4910
21	3510.1	9870	55	6810.0	4909
22	6437.3	9819	56	7476.7	4859
23	2356.4	9749	57	905.9	4733
24	2025.7	9274	58	1041.2	4627
25	8135.7	9195	59	2191.2	4583
26	2116.6	9140	60	1618.2	4304
27	1946.9	9042	61	870.1	4298
28	1467.7	9024	62	2871.2	4294
29	2935.4	8984	63	3266.5	4191
30	1450.7	8545	64	6231.3	3841
31	1016.7	8432	65	1278.4	3724
32	2212.2	8034	66	3616.4	3683
33	1898.9	7797	67	1115.1	3650
34	3194.8	7599			

Table 6: The list of 67 most frequent peaks.

Appendix B: CA125 p-value alternative calculation

Suppose $E_0 = Err(S_{t,u}; p, w, v)$. In this case, random prediction leads to 1/3 probability to guess the correct result in each of $|S_{t,u}|$ independent cases. Probability to make at most E_0 errors by chance is

$$\text{CA125 } p\text{-value} = \sum_{i=0}^{E_0} (2/3)^i (1/3)^{|S_{t,u}|-i} C^i_{|S_{t,u}|},$$

where $C^k_n = \frac{n!}{k!(n-k)!}$ is the number of combinations of k elements without repetitions out of n elements. This formula provides theoretical calculation of CA125 p-values.

Appendix C: Dynamics plots

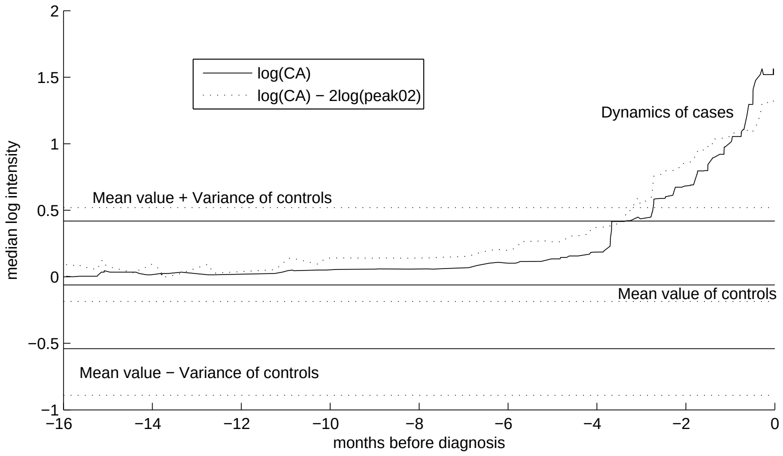

Figure 6: Median logarithm dynamics of rules $\log C$ and $\log C - 2 \log I(2)$ (for both controls and cases)

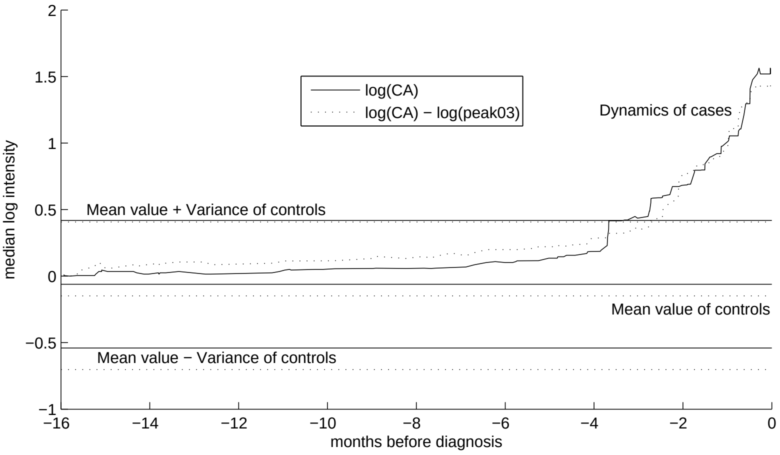

Figure 7: Median logarithm dynamics of rules $\log C$ and $\log C - \log I(3)$ (for both controls and cases)

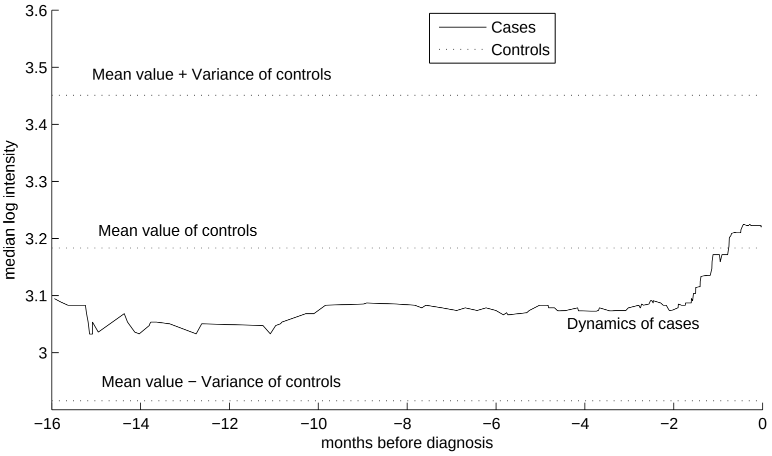

Figure 8: Median logarithm dynamics of Peak 2

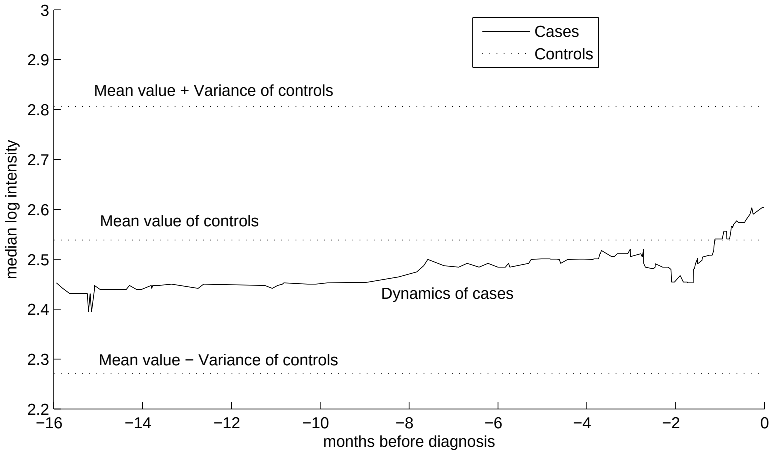

Figure 9: Median logarithm dynamics of Peak 3

subsec:summary of findingsfig:dyn CA vs CA-2p2

Appendix D: Duplicate classification for UCL data

The data provided by UCL consisted of duplicates: pairs of matched cases and controls. The pre-processing applied to this data was the same as described in Section 2. The statistical analysis of all peaks analogous to the one in Section 3.2 was performed. The results are represented in Table 7 (the table is similar to Table 3 for Reading data).

t	$\lvert S_{t,6}\rvert$	E_C	p-value 1	$E_{C,p}$	p	v/w	p-value 2	p-value 3
0	56	2	0.0001	0	04	-2	0.0001	0.2788
1	46	2	0.0001	0	50	-1	0.0001	0.3521
2	38	2	0.0001	1	09	-1	0.0001	0.6973
3	26	3	0.0001	2	34	$+1$	0.0006	0.9821
4	19	2	0.0004	1	01	$+2$	0.0007	0.1118
5	17	3	0.0066	1	38	$+2$	0.0175	0.5354
6	15	3	0.0198	1	49	$+1$	0.0718	0.7111
7	13	3	0.0441	1	06	$-1/2$	0.0657	0.1623
8	14	3	0.0262	0	36	-2	0.0082	0.0253
9	17	5	0.0689	3	13	$+2$	0.3532	0.8845
10	23	5	0.0053	3	18	$+2$	0.0255	0.5392
11	24	7	0.0307	5	18	$+2$	0.2738	0.6075
12	24	9	0.1526	6	18	$+2$	0.6297	0.6897
13	26	9	0.0830	6	18	$+2$	0.3558	0.4663
14	23	9	0.2045	6	18	$+2$	0.7494	0.7926
15	18	6	0.1150	4	18	$+2$	0.6764	0.8294
16	11	6	0.7218	1	42	$+1/0$	0.4646	0.3812

Table 7: Duplicate classification results for UCL data.

The table demonstrates that CA125 on its own predicts well up to 8 months and for months 10 and 11. But according to p-values 2 and 3, the addition of MS information does not allow us to obtain statistically significant predictions on any other months. For example, the only significant p-value 3 was achieved for month 8 with peak 36 (m/z = 3195.1 Da). Worse results can be explained by the fact that it is more probable to label the case by chance in a duplicate than in a triplet.