Advance Access publication on April 4, 2008
doi:10.1093/comjnl/bxn021

Serum Proteomic Abnormality Predating Screen Detection of Ovarian Cancer

ALEX GAMMERMAN[1,5,*], VOLODYA VOVK[1], BRIAN BURFORD[1], ILIA NOURETDINOV[1], ZHIYUAN LUO[1], ALEXEY CHERVONENKIS[1], MIKE WATERFIELD[2], RAINER CRAMER[2,3], PAUL TEMPST[4], JOSEP VILLANUEVA[4], MUSARAT KABIR[5], STEPHANE CAMUZEAUX[5], JOHN TIMMS[5], USHA MENON[5] AND IAN JACOBS[5]

[1]*Computer Learning Research Centre, Royal Holloway, University of London, UK*
[2]*Ludwig Institute for Cancer Research, University College London, UK*
[3]*BioCentre and Department of Chemistry, University of Reading, UK*
[4]*Memorial Sloan-Kettering Cancer Center, New York, USA*
[5]*Institute for Women's Health, University College London, UK*
**Corresponding author: alex@cs.rhul.ac.uk*

Ovarian cancer is characterized by vague, non-specific symptoms, advanced stage at diagnosis and poor overall survival. A nested case control study was undertaken on stored serial serum samples from women who developed ovarian cancer and healthy controls (matched for serum processing and storage conditions as well as attributes such as age) in a pilot randomized controlled trial of ovarian cancer screening. 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. Our exploratory analysis shows 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. This paper demonstrates that mass spectra from the low molecular weight serum proteome carry information useful for early detection of ovarian cancer. The next step is to identify the specific biomarkers that make early detection possible.

Keywords: bioinformatics; biomarkers; CA125; human serum proteome; mass spectra; ovarian cancer

Received 19 November 2007; revised 20 January 2008

1. BACKGROUND

Recent advances in mass spectrometry (MS)-based serum proteomics have generated high expectations for early detection of diseases. The high diagnostic accuracy for cancer reported in preliminary studies and the non-invasive character of the test have further fuelled research in the area. Using modern sample preparation and MS technologies coupled with appropriate data analysis techniques, it is now possible to detect abnormalities in the serum proteome that reflect underlying disease states.

Many publications bringing together the fields of mass spectrometry and machine learning deal with identifying peaks in mass spectra (and eventually the peptides represented by those peaks) which can be used to distinguish between symptomatic cancer patients and healthy individuals [1–3]. In this investigation, we add an extra dimension by using samples which predate the diagnosis of cancer in the cancer patients. Our objective, however, is more limited than identifying proteomic biomarkers: we only aim to demonstrate that mass spectra carry information useful for early detection of cancer; this is done by successfully (in the sense of achieving statistically significant results) distinguishing individuals with cancer from healthy controls.

Ovarian cancer is characterized by vague, non-specific symptoms, advanced stage at diagnosis and poor overall survival [4]. Survival is significantly improved if the disease is diagnosed at an early stage, suggesting that early detection might impact on the high mortality rates. A variety of biomarkers are being investigated but none so far have proven to be clinically useful [5]. It is hoped that using proteomics and other emerging technologies together with novel

bioinformatics, biomarkers with improved performance characteristics will be identified. In this early exploratory study we investigate the possibility of detecting early ovarian cancer utilizing proteomic analysis of a unique set of serial serum samples.

2. RESULTS AND DISCUSSION

Our aim is to demonstrate that the information contained in mass spectra, in combination with the level of the established tumour marker serum CA125 (see e.g. [6]), is useful for early detection of ovarian cancer. Our data set will allow us to reject the null hypothesis of no useful information at significance level 5% for detection up to 12 months in advance of the original diagnosis. The actual *P*-values obtained are given in the last column ('*P*-value 3') of Table 1, with the time to diagnosis (in months, denoted *t*) given in the first column.

We start, in the first subsection, from a brief description of the data set on which the null hypothesis of no useful information is tested. As the test statistic we use, essentially, the smallest number of errors achieved by the classification rules, described in the following subsection. In the final subsection we comment on the results presented in Table 1.

2.1. Pilot study data set

The unique feature of the sample set used in this study is that the women in the study were screened for ovarian cancer for up to 7 years (1995–2001) and therefore serial samples are available for each individual. Each of the 19 women who developed ovarian cancer (in one case, non-epithelial) has 2 to 11 serial samples taken prior to her cancer diagnosis in the course of these years. There are only single samples for most of the healthy control women.

Sample prefractionation and MALDI-TOF MS data acquisition were performed at the Memorial Sloan-Kettering Cancer Center (MSKCC), New York, USA, in December 2004. The serum samples were sent there 'blind' and randomized before processing.

The data set used in this paper, referred to as the *pilot study data set* later on, consists of 80 measurements taken from the 18 women eventually diagnosed with epithelial ovarian cancer (OC). For the details of how this data set was obtained, see the Methods section. The number of measurements for each of these women, referred to as the *cases*, is between 2 and 10. Each of the 80 measurements is accompanied by two measurements taken from the matched controls. Therefore, there are 240 measurements in total. Each of them includes the intensities of 402 peaks, identified at the preprocessing stage, and the CA125 level. The 402 peaks are ordered according to their frequency. For each case measurement 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 (defined as the date of the operation). See the Methods section for further details.

2.2. Classification rules used

We start by introducing some terminology and notation. Measurements will often be referred to as samples (although it should be remembered that they include not only the peak intensities but also the level of CA125).

The 240 samples are divided into 80 *triplets*, each consisting of a case sample and the two matched control samples. We will say that the case sample is *labelled* (or, more fully, *labelled as 'case'*). The 80 triplets are divided into 18 *triplet groups* corresponding to the 18 cases (with the size of a group varying between 2 and 10).

Each triplet τ in each group is assigned a non-negative value $T(\tau)$, the time to diagnosis for the case measurement in the group. For each $t = 0, 1, 2, \ldots$ let S_t be the set of triplets of measurements taken t months before the diagnosis (or as little earlier as possible). The largest S_t (for $t = 0, 1$) contain 18 triplets, whereas the smallest S_t given in Table 1 contain 14 triplets.

We will use a rather limited class of rules for *classification*, i.e. identification of the labelled sample within a triplet (according to the null hypothesis identification is impossible, for a given t). Only the first P peaks are taken into account (remember that the peaks are ordered by their frequency in the available spectra; in Table 1 we consider two cases, $P = 20$ and $P = 100$). Each classification rule

$$w \log C + v \log I(p)$$

is specified by three numbers, (p, w, v), which are a peak number $p \in \{1, \ldots, P\}$ and weights $w \geq 0$ and $v \in \{-1, 1\}$. For each triplet, the classification rule (p, w, v) predicts that the sample with the largest value of the sum w log $C + v$ log $I(p)$, where C is the CA125 level and $I(p)$ is the intensity of peak p, will be labelled as 'case'. The logarithms are taken to get rid of the arbitrary units of measurement of CA125 and the peak intensity. The coefficients w and v are weighting factors; intuitively, w determines to what degree (if at all) CA125 is taken into account in classification (we know that it tends to grow after the onset of OC, so its sign can be only positive when it is not zero), and v specifies whether we consider an elevated or lowered intensity of peak p indicative of OC.

Let

$$\mathrm{err}(\tau; p, w, v) := \begin{cases} 0 & \text{if } (p, w, v) \text{ correctly identifies} \\ & \text{the labelled sample in } \tau \\ 1 & \text{otherwise} \end{cases}$$

TABLE 1. Results for the pilot study data set.

t	$\|S_t\|$	$E1$	P-value 1	$E2$	$p2$	$w2$	P-value 2	$E3$	$p3$	$w3$	P-value 3
0	18	2	0.000001	1	−9	1	0.000001	1	−9	1	0.000006
1	18	2	0.000001	2	3	1	0.00002	1	−14	2	0.00001
2	15	4	0.0018	3	8	1	0.0094	2	18	2	0.0052
3	15	5	0.0086	3	8	1	0.0095	2	18	2	0.0056
4	15	6	0.031	3	8	1	0.0097	2	18	2	0.0056
5	15	6	0.031	2	−14	1	0.0017	2	−14	1	0.0047
6	15	7	0.088	3	−10	1	0.011	2	−10	0.5	0.0037
7	15	7	0.088	3	−10	1	0.011	2	−10	0.5	0.0037
8	14	8	0.31	3	−10	1	0.025	2	−39	1	0.021
9	14	7	0.15	3	−10	1	0.025	2	18	2	0.012
10	14	7	0.15	3	−10	1	0.025	2	18	2	0.012
11	14	7	0.15	3	−14	1	0.032	2	18	2	0.012
12	14	7	0.15	4	−10	1	0.13	2	18	2	0.012
13	14	8	0.31	3	−10	1	0.024	3	−10	1	0.063
14	14	9	0.52	4	−10	0	0.12	4	−10	0	0.29
15	14	9	0.52	5	−10	0	0.39	4	−60	0	0.51
18	14	10	0.74	4	−10	0	0.12	3	−60	0	0.15

Note: t, the time to diagnosis in months; $|S_t|$, the number of cases with measurements taken $\geq t$ months before diagnosis; $E1$, the number of errors when classifying the triplets in S_t with CA125 alone; P-value 1, the corresponding P-value; $E2$, the minimal number of errors when classifying with CA125 (taken with weight 0 or 1) and one of the peaks 1–20 (weight −1 or 1); $p2$ and $w2$, the peak number (smallest, if there are several such peaks) and the value of the weight for CA125, respectively, attaining $E2$ errors (the – sign in front of the peak number in the $p2$ column indicates that $E2$ errors are attained when the peak is assigned weight −1); P-value 2, the corresponding P-value; $E3$, the minimal number of errors when classifying with CA125 (weights 0, 1/16, 1/8, . . . , 8 ,16) and one of the peaks 1–100 (weight −1 or 1); $p3$ and $w3$, the peak number and the value of the weight for CA125, respectively, attaining $E3$ errors; P-value 3, the corresponding P-value.

be the indicator of whether (p, w, v) is successful on a triplet τ, and let

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

stand for the number of errors made by (p, w, v) on a set S of triplets. As our test statistic we take the pair (E, p), where E is the best number of errors achieved by the classification rules in the class and p is the smallest peak number used by the classification rules achieving E errors. Formally,

$$\begin{aligned} E &:= \min_{(p,w,v)} \mathrm{Err}(S_t; p, w, v), \\ p &:= \min\{\, p : \mathrm{Err}(S_t; p, w, v) = E \text{ for some } w \text{ and } v \,\}. \end{aligned} \tag{1}$$

Pairs (E, p) are ordered lexicographically, with small values deemed significant; the idea behind adding p is that otherwise there would be too many ties and the P-values would deteriorate. (In principle, there might be another kind of ties, when a classification rule attains its largest value on more than one sample in one triplet, but such ties never happen in this study.) The P-values are calculated using the Monte Carlo procedure; for details, see the Methods section.

2.3. Results: further details

The first column of Table 1 shows the time to diagnosis t , and the second column shows the size of S_t . The third column gives the number $E1$ of errors made on the triplets in S_t by the classification rule $w \log C + v \log I(p)$ with $w = 1$ and $v = 0$; this is the baseline classification rule taking into account CA125 alone. For example, the entry 7 corresponding to $t = 6$ means that out of the 15 cases with samples taken at least 6 months before the diagnosis, the baseline classification rule made 7 errors on the most recent of those samples (confused them with one of the matched control samples). For $t = 6$ the expected number of errors when classification is done at random is 10 (in general, 2/3 of the number of triplets in S_t), so 7 is better than random, but it is not statistically significant (with the P-value of $\simeq$9%).

Column $E3$ shows the smallest number of errors made on the triplets in S_t by the classification rules $w \log C + v \log I(p)$ with w ranging over the set of weights {0, 1/16, 1/8, . . . , 8,16}, v ranging over $\{-1, 1\}$, and $p = 1, \ldots, 100$ ranging over the 100 most frequent peaks. For example, the entry

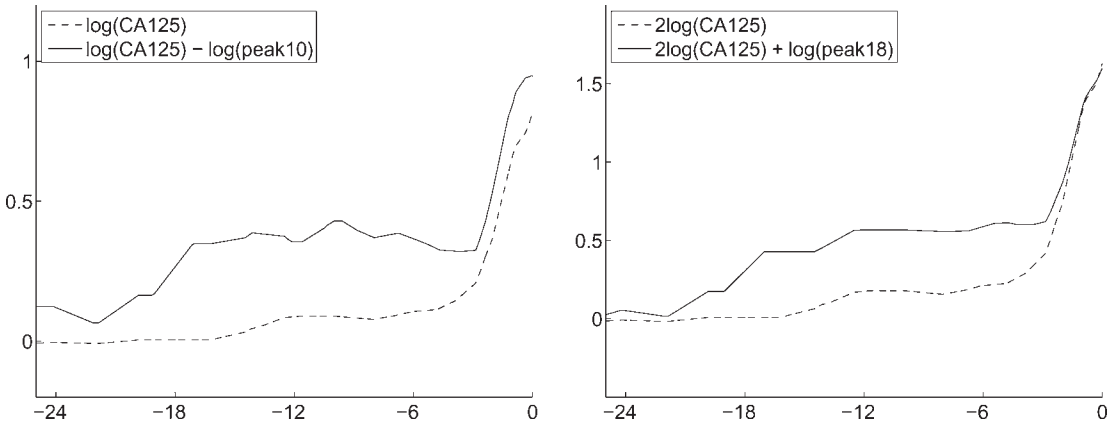

FIGURE 1. The performance of the two classication rules on the pilot study data set (log standing for log10): each plot essentially shows the deviation of the value taken by the corresponding classication rule on the cases from the values taken on the controls (see the text for full details); the horizontal axis shows the time to diagnosis in months.

2 corresponding to $t = 6$ means that out of the 15 cases with samples taken at least 6 months before the diagnosis, the best classification rule made two errors on the most recent of those samples. This is a very small number compared to 10, but it should be remembered that the numbers of errors are heavily biased: the training set S_t includes the test samples. However, the P-value, $\simeq 0.4\%$, is still highly significant.

In general, from the last column of Table 1, which gives the P-values calculated from (E, p) (see (1)) using the Monte Carlo method, we can see that the null hypothesis can be rejected at significance level 5% for t up to 12 months. This contrasts with using CA125 alone, which does not produce significant results even for $t = 6$. It is important that our P-values do not require adjustment: the numbers of errors are calibrated with the Monte Carlo procedure described in the Methods section. Columns $p3$ and $w3$ give the peak and the weights at which $E3$ errors are attained for various values of t; interestingly, there is no consistency: often different peaks and weights are chosen for different t.

Columns $E2$, $p2$, $w2$ and 'P-value 2' demonstrate the robustness of our approach. It might be argued that 100 peaks is too many, so now we only use the peaks present in at least 40% of all samples; these are peaks 1–20. We also decide in advance that CA125 and the peak intensity should be taken with equal weights (when CA125 is taken into account at all). As column 'P-value 2' shows, this does not significantly affect the P-values.

Figure 1 illustrates the performance of the classification rules $\log C - \log I(10)$ and $2\log C + \log I(18)$ that occur most frequently in Table 1. Informally, each of the four plots in the figure shows the deviation of the value taken by the corresponding classification rule on the cases from the values taken on the controls. For a fuller description, see the Methods section. If at time $-t$ (as shown on the horizontal axis) the plot is significantly above zero, we can say that the cases differ significantly from controls t months in advance of the diagnosis. The solid lines in both plots are significantly raised above zero one year before the diagnosis. (The statistical significance of the raise follows from the P-values given in Table 1.) For comparison, the plot for CA125 (dashed line) is significantly raised only within the last six months before the diagnosis. For the reader wishing to interpret the marks on the vertical axis, all logarithms are to the base 10. (It is important, however, that the heights of the graphs should not be compared directly across the panels, since the log-level of CA125 is taken with different weights.)

3. METHODS

In this section we describe the unique set of serum samples from which the pilot study data set was derived, introduce the prefractionation, MALDI-TOF MS data acquisition and pre-processing techniques used to obtain peaks for classification, and give the details of statistical analysis.

3.1. Sample collection and handling

The serum samples used in this study were collected as part of a pilot randomized controlled trial of ovarian cancer screening involving 13 460 post-menopausal women, aged over 50 [6]. Information on reproductive history, family history of cancer and HRT use was obtained on all women at baseline. Following recruitment, 6682 were randomized to the screen group. These women underwent annual screening for 2 to 6 years between 1995 and 2001. The women were posted venepuncture equipment and asked to visit their general practitioners for venepuncture. From 1995 to 2000, samples were collected using 10 ml serum sample tubes (BD red top). After 2001, samples were collected in Greiner-Bio-One serum gel tubes.

Blood samples were sent to the Ovarian Cancer Screening Unit by first class mail at room temperature, centrifuged at 4000 rpm for 10 min and the serum separated, aliquoted and stored at −20°C. All blood samples received more than 56 hours after venepuncture were discarded and repeat samples requested. The date blood was taken and received was recorded. Serum CA125 levels were determined by commercial radioimmunoassay (CA 125II kit, Centocor, Malvern, PA) within two weeks of receipt of the blood sample. In 2004, all samples were moved to −80°C freezers.

The serum bank from the study currently includes a set of 92 serum samples from 19 women dating from less than 1 to more than 6 years prior to the development of primary epithelial ovarian/fallopian tube cancer (broadly referred in the rest of this paper as ovarian cancer, or OC) or non-epithelial ovarian cancer (in the case of one woman, later removed from the data set for the purpose of this study). For the current pilot study, one aliquot of the sample was thawed and aliquoted into 100 μL aliquots, which were packaged into dry ice containers for transport to the analysis laboratory.

Each case serum sample was matched to two samples closest in the freezer to it in order to match for storage conditions. The chosen control sample also fulfilled age (within 5 years) and use of hormone replacement therapy matching criteria. Usually two control samples were selected for each case sample. All cases and controls are women aged over 50, post-menopausal with no high-risk family history (all have only one or no relatives with ovarian cancer).

For all cases, the date of diagnosis is the date when the woman underwent surgery and a pathological diagnosis of cancer was confirmed.

Approval was obtained from the local ethics committee for this pilot study using proteomic analysis.

3.2. Serum prefractionation

The serum samples were subjected to prefractionation using a previously reported reversed-phase (RP) batch extraction protocol [12, 13]. The procedure was carried out in 96-well Template III PCR non-skirted and half-skirted plates (USA Scientific, Ocala, FL; catalogue numbers 1402-9600 and 1402-9700) and all steps were automated on a 'Genesis Freedom 100' liquid handling work station (Tecan; Research Triangle Park, NC).

The protocol utilizes paramagnetic beads with C8 RP binding ligands (ChemiCell GmbH, Berlin, Germany) and was performed as follows. Magnetic bead pellets were re-suspended by pipetting up and down 10 times. Bead suspension (25 μL) was transferred to a well containing an aliquot of serum (50 μL), magnetic beads and serum were then mixed for 30 seconds, beads were then pulled to the side of the well using magnets and allowed to settle and the supernatant was removed and discarded. Washing solution (200 μL) (0.1% TFA; Pierce, Rockford, IL) was added to the bead pellet, beads were pulled five times from left to right and back on a magnetic holder, beads were allowed to settle on one side and the washing solution was removed. This washing step was repeated. During the second washing step only 120 μL of washing solution was removed. The sample plate was then moved to a sample holder where the beads were further re-suspended, beads were then pulled to the bottom of the well by magnets and the remaining supernatant was removed. Elution solvent (7 μL) consisting of 50% acetonitrile (Burdick and Jackson, Muskegon, WI) was added to the bead pellet and mixed by pipetting. A fraction of the eluate (5 μL) was transferred to another well and 5 μL of pre-made α-cyano-4-hydroxycinnamic acid CHCA matrix solution (Agilent, Palo Alto, CA) was added to the eluate and mixed. Then 1 μL of each mixture was spotted in duplicate onto a ground steel MALDI target (Bruker Daltonics, Bremen, Germany) in every other column of the 384-spot layout and left to dry. Finally, samples were allowed to dry at room temperature. In addition, 30 fmols (per peptide) and 500 fmols (per protein) of commercially available calibration standards (Bruker Daltonics, Bremen, Germany; catalogue numbers 206195 (≤4 kD) and 206355 (≥4 kD)) were also mixed with CHCA matrix and separately deposited onto the target plates adjacent to each spotting duplicate (one standard/sample), in the alternating columns.

3.3. Mass spectrometry

Peptide profiles were analysed using an Autoflex MALDI-TOF mass spectrometer (Bruker Daltonics, Bremen, Germany) equipped with a 337-nm nitrogen laser, a gridless ion source, delayed-extraction (DE) electronics, and a 2-GHz digitizer. Separate spectra were obtained for two restricted mass-to-charge (m/z) ranges, corresponding to polypeptides with molecular mass of 0.7–4 kDa ('≤4 kD') and 4–15 kDa ('≥4 kD') assuming $z = 1$, under specifically optimized instrument settings. Each spectrum was the result of 400 laser shots per m/z segment per sample, delivered in four sets of 100 shots (at 50-Hz frequency) to each of four different locations on the surface of the matrix spot. The effective laser energy delivered to the target was carefully controlled to be 16 μJ (±10%) per shot. The entire irradiation program was automated using the instrument's 'AutoXecute' function. Spectra were acquired in linear mode geometry under 20 kV (18.6 kV during DE) of ion acceleration and −1.3 kV multiplier potentials, and with gating of mass ions ≤400 m/z (≤4 kD segment) or ≤3000 m/z (≥4 kD segment). DE was maintained for 80 (≤4 kD) or 50 nanoseconds (≥4 kD) to give appropriate time-lag focusing after each laser shot. In total, 125 206 data points were generated for the total mass range (0.7–15 kDa) with the corresponding intensities. In total, 265 samples were successfully analysed, of which 91 were case samples taken from 19 women with ovarian cancer, and 175 control samples. Peaks were identified

from all 265 samples, but in the main part of the study we used 240 samples; we removed the only case with non-epithelial OC (this decision had been taken before data analysis started), and we removed the measurements in which the case sample did not have both of the matched control samples.

3.4. Pre-processing

Mass spectrometry instruments are very sensitive, and artefacts can be introduced into spectra from physical, electrical or chemical sources in experiments. Pre-processing is an important step to attempt to remove these systematic artefacts and isolate the true protein signal. The goals of pre-processing are to reduce noise, normalize the spectra from different samples and reduce the dimensionality of MS data. The implicit assumption is that each spectrum can be considered as composed of three components: true peak signal, baseline and random noise.

Our pre-processing of the raw data included the following standard steps:

calibration using 13 calibrants associated with each sample and performed separately on the low mass range data ($\leq$4 kD) and high mass range data ($\geq$4 kD);

smoothing by averaging the intensities within a moving window;

baseline subtraction involving the steps of baseline estimation and subtraction of the estimated baseline from the mass spectrum;

normalization to make sure that the total amounts of ions across different samples are the same;

peak identification by finding local maxima in the mass spectra with a signal-to-noise ratio exceeding a certain threshold;

peak alignment, i.e. relating the peaks identified in one spectrum to the peaks found in other spectra.

The step of peak identification led to 7216 'non-aligned peaks', each non-aligned peak pertaining to one sample. After peak alignment these non-aligned peaks were clustered into 402 groups, called *peaks* in the rest of the paper, each group having no more than one non-aligned peak in each sample. The frequency of a peak is defined as the percentage of samples having this peak (i.e. having a non-aligned peak belonging to this peak). The 402 peaks are ordered by their frequency.

Table 2 shows the 20 peaks whose frequency exceeds 40%. The first column of this table gives the ordinal number for each of these peaks; in columns 2 and 3 we show the mean m/z value of the peak and the m/z range, respectively. Columns 4 and 5 provide information on the number of samples that contain the peak.

TABLE 2. Top 20 peaks and peak 39.

Peak number	mean m/z	m/z range	Number of samples having the peak	% of total
1	3188.9	3185.3–3191.0	246	92.8
2	6646.1	6636.5–6652.8	245	92.5
3	3330.5	3325.2–3333.4	198	74.7
4	2004.1	2001.2–2005.3	190	71.7
5	1764.6	1762.0–1766.6	184	69.4
6	818.5	817.4–818.9	165	62.3
7	9307.0	9294.7–9319.7	154	58.1
8	2982.3	2978.8–2985.2	146	55.1
9	2020.9	2019.7–2021.9	144	54.3
10	4292.5	4288.0–4300.8	140	52.8
11	3280.0	3276.2–3282.1	139	52.5
12	2548.2	2543.5–2550.2	137	51.7
13	2562.8	2561.4–2564.2	123	46.4
14	8942.2	8930.8–8955.3	118	44.5
15	3296.9	3294.3–3299.1	115	43.4
16	1888.8	1887.7–1889.7	114	43.0
17	899.9	898.6–901.0	113	42.6
18	3172.1	3169.1–3176.4	109	41.1
19	3229.8	3226.7–3234.2	109	41.1
20	5010.4	5004.0–5017.0	109	41.1
39	2016.8	2013.6–2019.6	56	21.1

The top 20 peaks are those present in more than 40% of the samples, and peak 39 is the only peak that occurs more rarely but is still useful for discrimination according to Table 1.

For the purpose of this paper, each peak p needs to have a well-defined intensity $I(p)$ in each sample (see e.g. the definition of classification rules in the Results section). If peak p is not present in the sample, $I(p)$ is defined as the average intensity in the m/z range associated with the peak.

3.5. Statistical analysis

To calculate the P-values in Table 1 we use the Monte Carlo method. For each $t = 0, 1, \ldots, 15, 18$, our null hypothesis is that the assignment of the label 'case' within each triplet in S_t is random (given the samples). Therefore, in calculating the P-values we compare the performance of our classification rules on the actual data set with their performance on the data set with the labels randomly permuted.

The formal definition of S_t is as follows: start from $S_t := \emptyset$ and for each triplet group put in S_t the triplet τ in the group with the smallest $T(\tau)$ satisfying $T(\tau) \geq t$ (if this group contains triplets with $T(\tau) \geq t$).

Our procedure for computing P-values, Algorithm 1, calculates the statistic (E, p) according to (1) a large number N (we used $N = 10^6$) of times on the pilot study data set with randomly reassigned 'case' labels and counts the number Q of times the statistic is as good as or better than

the statistic computed from the true labels. 'As good as or better than' is understood in the sense of the lexicographic order: $(E, p) \leq (E^{\text{true}}, p^{\text{true}})$ means that either $E \leq E^{\text{true}}$ or $E = E^{\text{true}}$ and $p \leq p^{\text{true}}$. The P-value is then estimated as Q/N.

Algorithm 1. P-value calculation

Input: t, time to diagnosis.
Input: N, number of trials.
Input: $(E^{\text{true}}, p^{\text{true}})$, the output (E, p) of Equation (1) for the true labels.
Output: The corresponding P-value.
$Q := 0$.
for $j = 1, \ldots, N$ **do**
Assign label 'case' to a randomly chosen sample in each triplet in S_t.
Set (E, p) according to Equation (1) applied to the new data set.
if $(E, p) \leq (E^{\text{true}}, p^{\text{true}})$ **then**
$Q := Q + 1$.
end if
end for
Output Q/N as the P-value.

The validity of the P-values produced by Algorithm 1 is a simple mathematical fact; the only difference from the standard procedure is that we use the lexicographic order rather than the usual order on the real line.

Formally, the plots in Fig. 1 are obtained as result of the following sequence of steps:

- for each case, the points $(-t, D(t))$ are plotted, where $t \in [0, 25]$,

$$D(t) := w \log C_{\text{case},t} + v \log I_{\text{case},t}(p) - \frac{1}{2}(w \log C_{\text{contr1},t} + v \log I_{\text{contr1},t}(p) + w \log C_{\text{contr2},t} + v \log I_{\text{contr2},t}(p)),$$

  $w \log C + v \log I(p)$ is the classification rule being considered ($\log C - \log I(10)$, $2 \log C + \log I(18)$, $\log C$ or $2 \log C$), $C_{\text{case},t}$ is the latest level of CA125 for the given case measured at least t months in advance of the diagnosis, $I_{\text{case},t}(p)$ is the intensity of peak p for the case found from the latest sample taken at least t months in advance of the diagnosis, and $C_{\text{contr1},t}$, $I_{\text{contr1},t}(p)$, $C_{\text{contr2},t}$, $I_{\text{contr2},t}(p)$ are the analogous values for the two matched controls;
- for each t, the $|S(t)|$ values of $D(t)$ (corresponding to the $|S(t)|$ cases, with $|S(t)|$ given in Table 1) are replaced by their median;
- the resulting plot is smoothed with a moving window of width 2 months.

4. CONCLUSIONS

The purpose of this study has been to demonstrate that mass spectra carry information about diagnosing with OC in the relatively remote future. It turned out that this is true when they are combined with CA125; the predictive power of CA125 alone is more limited. One of the pressing directions of further research is to pinpoint and identify the peaks that make early detection of OC possible. For example, the peak that occurs most frequently in Table 1 is peak 10, which has the mean mass-to-charge (m/z) value of 4292.5 Da. One can speculate that, within the margin of experimental error, this peak may be identified as a fragment of inter-α-trypsin inhibitor heavy chain H4 (ITIH4) with the m/z value given as 4286 Da in [7] and 4281.78 Da in [8]. This is reinforced by the fact that other authors using different technology have identified this fragment of ITIH4 to discriminate patients with ovarian cancer from healthy controls and patients with benign pelvic masses [9]. This putative acute-phase protein ITIH4 was also shown to be associated with other diseases [10, 11].

Another direction of research is to identify the classification rules that can discriminate between cases and controls in the general population. In this study we have used classification rules for testing our null hypothesis, but those rules are only applicable to groups of samples (such as our triplets) in which one and only one sample is known to be a case. To make them applicable to individual samples, they need to be supplemented with a threshold that separates cases from controls. It is likely that more sophisticated rules, perhaps involving several peaks, will be more efficient, but finding such rules would require larger data sets.

ACKNOWLEDGEMENTS

We are grateful to Professor A. Philip Dawid for a useful discussion of the statistical aspects of this work.

FUNDING

This work was undertaken at UCLH/UCL, who received a proportion of funding from the Department of Health's NIHR Biomedical Research Centres funding scheme. It was also supported by MRC grant G0301107 (Proteomic analysis of the human serum proteome), the Helen Feather Trust, and the Cyprus Research Promotion Foundation.

REFERENCES

[1] Diamandis, E.P. (2006) Serum proteomic profiling by matrix-assisted laser desorption-ionization time-of-flight mass spectrometry for cancer diagnosis: next steps. *Cancer Res.*, **66**, 5540–5541.

[2] Engwegen, J.Y., Gast, M.C., Schellens, J.H. and Beijnen, J.H. (2006) Clinical proteomics: searching for better tumour markers with SELDI-TOF mass spectrometry. *Trends Pharmacol. Sci.*, **27**, 251–259.

[3] Xiao, Z., Prieto, D., Conrads, T.P., Veenstra, T.D. and Issaq, H.J. (2005) Proteomic patterns: their potential for disease diagnosis. *Mol. Cell. Endocrinol.*, **230**, 95–106.

[4] Ozols, R.F., Rubin, S.C., Thomas, G.M. and Robboy, S.J. (2005) Epithelial ovarian cancer. In Hoskins, W.J., Perez, C.A. and Young, R.C. (eds), *Principles and Practice of Gynecologic Oncology* (4th edn), pp. 823–872. Lippincott Williams & Wilkins, Philadelphia.

[5] Menon, U. (2007) Ovarian cancer: challenges of early detection. *Nat. Clin. Pract. Oncol.*, **4**, 498–499.

[6] Menon, U. *et al*. (2005) Prospective study using the risk of ovarian cancer algorithm to screen for ovarian cancer. *J. Clin. Oncol.*, **23**, 7919–7926.

[7] Timms, J.F. *et al*. (2007) Preanalytic influence of sample handling on SELDI-TOF serum protein profiles. *Clin. Chem.*, **53**, 645–656.

[8] Koomen, J.M., Shih, L.N., Coombes, K.R., Li, D., Xiao, L.-c., Fidler, I.J., Abbruzzese, J.L. and Kobayashi, R. (2005) Plasma protein profiling for diagnosis of pancreatic cancer reveals the presence of host response proteins. *Clin. Cancer. Res.*, **11**, 1110–1118.

[9] Zhang, Z. *et al*. (2004) Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. *Cancer Res.*, **64**, 5882–5890.

[10] Song, J., Patel, M., Rosenzweig, C.N., Chan-Li, Y., Sokoll, L.J., Fung, E.T., Choi-Miura, N.-H., Goggins, M., Chan, D.W. and Zhang, Z. (2006) Quantification of fragments of human serum inter-α-trypsin inhibitor heavy chain 4 by a surface-enhanced laser desorption/ionization-based immunoassay. *Clin. Chem.,* **52**, 1045–1053.

[11] Villanueva, J. *et al*. (2006) Differential exoprotease activities confer tumor-specific serum peptidome patterns. *J. Clin. Invest.*, **116**, 271–284.

[12] Villanueva, J., Philip, J., Entenberg, D., Chaparro, C.A., Tanwar, M.K., Holland, E.C. and Tempst, P. (2004) Serum peptide profiling by magnetic particle-assisted, automated sample processing and MALDI-TOF mass spectrometry. *Anal. Chem.*, **76**, 1560–1570.

[13] Villanueva, J., Lawlor, K., Toledo-Crow, R. and Tempst, P. (2006) Automated serum peptide profiling. *Nat. Protoc.*, **1**, 880–891.