Early Detection of Ovarian Cancer in Samples Pre-Diagnosis Using CA125 and MALDI-MS Peaks

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

[1]EGA Institute for Women's Health, University College London, London, WC1E 6BT, UK;
[2]Computer Learning Research Centre, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK;
[3]BioCentre and Department of Chemistry, University of Reading, Whiteknights, Reading, Berkshire, RG6 6AD, UK;

Correspondence to: Dr John F Timms, Cancer Proteomics Laboratory, Institute for Women's Health, University College London, Cruciform Building 1.1.03, Gower Street, London, WC1E 6BT, UK. Tel: +44 2076796598, FAX: +44 2076913104, e-mail: jtimms@wibr.ucl.ac.uk

Key Words: MALDI MS serum profiling, ovarian cancer; biomarkers, early diagnosis, CA125, connective tissue-activating peptide III (CTAPIII), platelet factor 4 (PF4).

Abstract: Aim: A nested case-control discovery study was undertaken to test whether information within the serum peptidome can improve on the utility of CA125 for early ovarian cancer detection. Materials and Methods: High-throughput matrix-assisted laser desorption ionisation mass spectrometry (MALDI MS) was used to profile 295 serum samples from women pre-dating their ovarian cancer diagnosis and from 585 matched control samples. Classification rules incorporating CA125 and MS peak intensities were tested for discriminating ability. Results: Two peaks were found which in combination with CA125 discriminated cases from controls up to 15 and 11 months before diagnosis, respectively, and earlier than using CA125 alone. One peak was identified as connective tissue-activating peptide III (CTAPIII), whilst the other was putatively identified as platelet factor 4 (PF4). ELISA data supported the down-regulation of PF4 in early cancer cases. Conclusion: Serum peptide information with CA125 improves lead time for early detection of ovarian cancer. The candidate markers are platelet-derived chemokines, suggesting a link between platelet function and tumour development.

There are over 220,000 new cases of ovarian cancer worldwide each year (1). The disease has a poor prognosis mainly due to late diagnosis, with over 70% of patients exhibiting spread beyond the pelvis (2). Currently, ovarian cancer screening is not recommended due to lack of evidence of a mortality benefit, although large screening trials are underway to explore this issue in the low-risk population (3, 4). Strategies being tested combine serum cancer antigen 125 (CA125) assay with transvaginal sonography (TVS). CA125 is a coelomic epithelium-related glycoprotein of unknown function, secreted into the bloodstream by ovarian epithelial cells. However, CA125 is also produced by other mesothelium-derived tissues and may thus be elevated in benign gynaecological conditions and non-ovarian carcinomas, whilst not all early-stage tumours generate CA125 (5). Consequently, the diagnostic accuracy of CA125 using a cut-off is sub-optimal and additional biomarkers to improve accuracy for screening and early detection of ovarian cancer are required (6).

The performance characteristics required of an ovarian cancer biomarker depend upon the intended clinical use. Markers that can differentiate benign pelvic masses from malignancy in symptomatic women and guide surgical decisions need to improve upon existing tests which achieve sensitivities of 85-90% for detecting symptomatic ovarian cancer and specificities of 85-90% for benign disease (7). Indeed, one such multi-marker test OVA1 (which includes CA125) was recently approved by the FDA (8). When combined with clinical assessment by imaging and physical examination, it achieves a sensitivity of >90% and a negative predictive value of 90% in women with an ovarian tumour. A more challenging clinical use is screening for ovarian cancer. This requires high sensitivity together with specificity in excess of 98% so that at least one in ten women who undergo surgery as a result of screen positive results (positive predictive value (PPV) >10%) has ovarian cancer. In the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) this was achieved in the prevalence screen by using CA125 interpreted using the risk of ovarian cancer (ROC) algorithm followed by TVS (3). While it is encouraging that 48% of

the cases detected were early-stage cancer, it does raise the need for improving lead time even when high sensitivity and specificity are achieved.

Mass spectrometry (MS)-based proteomic profiling of polypeptides in serum (or plasma) holds promise for the identification of novel cancer biomarkers (9-11). However, serum proteome profiling is challenging due to the complexity and large dynamic range of abundance of its component proteins. Together with significant intra- and inter-individual heterogeneity, this has meant that few, if any, robust cancer biomarkers have been identified to date using proteomic methods. Indeed, most candidates reported have been abundant, non-specific acute-phase proteins (12, 13) that are likely to be secondary host responses to any diseased state rather than specific markers useful for accurate diagnosis (14). In addition, numerous studies have highlighted alterations in serum and plasma proteins that are attributable to sample handling that is largely driven by differential proteolysis (15-21) (22). Finally, concerns have been raised over assay reproducibility and the robustness of class-discriminating algorithms used for proteomic profiling biomarker discovery (14, 23, 24).

Matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) MS is a technique that can be used for high-throughput profiling of clinical samples, particularly when linked to automated handling. MALDI-TOF MS profiling has revealed the complexity of the low-molecular weight proteome (peptidome) of serum and plasma (25-27). Most peptides observed in MALDI-TOF profiles are derived from relatively few abundant proteins as the result of extensive proteolysis, and it has been reported that these cleavage patterns can discriminate cancer cases from healthy controls (28). It was hypothesized that these peptide patterns are generated by tumour-specific exopeptidases during coagulation, and as such, represent *ex vivo* surrogate markers of cancer (28).

In the present discovery study, we have used semi-automated peptide extraction linked to MALDI-TOF MS to profile the peptidomes of serum samples taken from the UKCTOCS trial. Samples were obtained from women at various timepoints prior to diagnosis of primary invasive

ovarian cancer and from healthy controls which were matched based on clotting time, collection centre and age. Given that these samples pre-date symptomatic cancer, they provide a unique source of early markers in the absence of late-stage confounding markers. We tested the power of peptide peaks alone and in combination with serum CA125 for discriminating cases from controls at different timepoints prior to diagnosis building on previous work from a pilot study where we reported that combining a single MS peak with CA125 allowed significant discrimination between cases and controls up to 12 months in advance of diagnosis (29).

Materials and Methods

Serum samples: A nested case-control study was undertaken on serum samples collected from 159 women who developed primary invasive ovarian cancer and 585 healthy women recruited to the UKCTOCS (30) (http://www.instituteforwomenshealth.ucl.ac.uk/academic_research/gynaecologicalcancer/gcrc/ukctocs/design). This secondary study was approved by the Joint UCL/UCLH Research Ethics Committees on the Ethics of Human Research, Committee A (Ref. No. 05/Q0505/57) and written informed consent was obtained from all donors. No data allowing identification of patients was provided. There were 880 samples in total collected between April 2001 and January 2007 at 13 trial centres, of which 295 samples were from the 159 women who went on to develop ovarian cancer (referred to as cases). Of these there were single samples from 90 cases and up to 5 serial samples from 69 cases. According to a standard operating procedure used at all centres, blood samples were taken by venepuncture into gel tubes (8 ml gel separation serum tubes; Greiner Bio-One, Stonehouse, UK) and transported by courier at ambient temperature to the central laboratory. All samples received more than 56 h after venepuncture were discarded and repeat samples requested. The blood was centrifuged at 1500 x *g* for 10 min, the serum separated and aliquoted into bar-coded straws which were frozen at –80℃ and then transferred the next day

to liquid nitrogen for long-term storage. For each case the time to diagnosis was known, and is the time interval (measured in months) between the date of collection and the date of diagnosis (defined as the date of surgery/biopsy). Matched controls were 585 samples from healthy women with two samples matched per case (only 5 cases had 1 matched control). These were taken as close as possible in time to the case sample on the same day and at the same collection centre, and hence were clotted for the same time and stored identically as matched cases. Clotting times ranged between 1 and 49 h (average 23 h ± 4 h). Samples were also matched by age at sample draw as a secondary criterion if time and centre matching allowed. Baseline characteristics of cases and controls and statistical assessment of differences are presented in Table I.

Sample preparation, MALDI-TOF MS profiling and data processing: Blinded serum samples were subjected to pre-fractionation using robotic C18 ZipTip® (Billerica, MA, USA) reversed-phase extraction prior to automated MALDI-TOF MS data acquisition as described previously (31)-(32). Briefly, polypeptides were enriched from 5 µl of serum using a semi-automated protocol based on reversed phase pre-packed tips (C18 ZipTips®). A CyBi®-Disk robot (CyBio AG, Jena, Germany) equipped with a 96-piston head for 25 µl tips was adapted and used for this purpose. After C18 ZipTip® purification, enriched polypeptides were eluted from the ZipTips®, and 2 µl from the eluate were mixed with α-cyano-4-hydroxycinnamic acid (CHCA) matrix and spotted onto a 600 µm-AnchorChip target plate (Bruker Daltonics, Bremen, Germany) for analysis on an Ultraflex II MALDI-TOF/TOF mass spectrometer (Bruker Daltonics) using FlexAnalysis v3.0 software (Bruker Daltonics) for further data handling. Up to 12 replicate mass spectra were generated for each sample giving 11,048 mass spectra in total. These were mass calibrated against peptide and protein standards run at the same time and converted to two column ASCII files of *m/z* values and corresponding intensities for further processing and analysis. Data processing steps (data reduction, smoothing, baseline subtraction, normalization,

peak defining and peak alignment) were applied to the data, as previously described (29). Documentation and Matlab system code for these steps are available at http://www.clrc.rhul.ac.uk/projects/proteomic3.htm. Following processing, a set of peak groups were defined, ordered and numbered according to their frequency of appearance in the dataset. The 67 most common peaks appearing in >33% of all spectra (cases and controls) were used for subsequent data analysis (Table II). Finally, the intensities of the 67 peaks were averaged across replicates for each sample.

Classification and statistical testing: For classification, data from cases with only one matched control, and samples with a clotting time greater than 24 h were excluded, resulting in 179 samples (from 104 cases) and 358 matched control samples organized into 179 triplets. A total of 59 cases had one measurement, 26 had 2 measurements, 11 had 3 measurements, 5 had 4 measurements, and 3 had 5 measurements. Each triplet was assigned to a time to diagnosis group (in months) using a 6-month time window. Log-linear models using peak intensity were then tested for triplet classification, *i.e.* identification of the case sample within a triplet. Models were tested with and without CA125 values. For each triplet, the classification rule assigns a 'case' label to the sample with the largest value of $w \log C + v \log I(\mathrm{p})$, where C and $I(\mathrm{p})$ are the CA125 level and the intensity of peak p, respectively, and w and v are various weights (w={0, 0.5, 1, 2}; v={–1, 1}), the latter taking into account the direction of regulation of the peak in controls *versus* cases, logarithms were taken to remove the arbitrary units of measurements. Error rates were then calculated for various classification rules taking the peak giving the least number of errors. If two rules led to the same error rate, the rule involving the most frequently occurring peak across the dataset had priority when calculating *P*-values.

In order to check the robustness of the triplet classification, three types of statistical tests were used which reject the following null hypotheses about the classification. (i) Assignment of the label 'case' within each triplet of each time group is random. We apply this hypothesis to

compare the performance of the classification rules on the actual data set with a randomly permuted data set. (ii) Assignment of labels within triplets is independent of CA125 levels – we wanted to check here if using CA125 alone is good enough to separate cases from controls. (iii) Peak intensities do not contain information useful to improve the predictive ability of CA125. Rejection of this third hypothesis means that separation between cases and controls is significantly improved after adding peak intensities to the information given by CA125. Based on the Monte-Carlo method, we developed a procedure to calculate *P*-values (see (29)) for the first null hypothesis. Here, the test statistic is calculated a large number of times (N) for the data set with a randomly assigned label 'case' within each triplet, and counting the number of times (Q) this statistic is equal to or less than the actual statistic computed from the observed data set. The *P*-value is then estimated as Q/N. For the second null hypothesis, the error rate was computed using CA125 only (*i.e.* v=0). For the third null hypothesis, the same set of classification rules is used as for the first null hypothesis, but instead of randomly assigning the 'case' label, we randomly permute the three peak intensities within each triplet (6 possible permutations), leaving the labels and CA125 levels as they are. All three procedures produce valid *P*-values that do not need adjustment in the sense that for each threshold δ the probability that the computed *P*-value does not exceed δ will be at most δ (under the null hypothesis). A detailed account of the corresponding *P*-value calculations using the Monte-Carlo method, their meanings and validity is described in Supplementary Information 1 (see also (29).

Peak identification: MALDI-TOF MS peaks were identified using 96 pooled UKCTOCS sample eluates from C18 ZipTip® extraction as a source of material and a combination of MALDI MS/MS, ESI MS/MS and off-line LC fractionation and ESI MS/MS with or without tryptic digestion The methodology (see (27)) was designed to ensure that identifications could be directly linked to the relevant peak by confirming the presence of that peak in fractions at each stage of the procedure by MALDI-TOF MS.

For peak identification using MALDI MS/MS, the pooled eluates were mixed in a ratio of 1:1 (v:v) with either 5 mg/ml CHCA or 50 mg/ml 2,5-dihydroxybenzoic acid (DHB) matrix prepared in 50% acetonitrile (ACN)/0.1% trifluoroacetic acid (TFA) and spotted onto a MALDI target plate for analysis on a Q-TOF Premier (Waters, Manchester, UK) equipped with a 337 nm-UV-MALDI source. For peak identification using ESI MS/MS, the pooled eluates were first dried, then resuspended in 50% ACN/0.1% formic acid and sprayed using a Nanomate HD (Advion Biosciences, Ithaca, NY, USA) as nanoESI ion source in front of the Q-TOF Premier. A positive voltage of 1.7 kV was applied to the chip and the flow rate was kept constant at ~80 nl/min. At this flow rate, a sample volume of 10 µl provided stable static electrospray for at least 2 h.

For peak identification using off-line LC fractionation and ESI MS/MS, 100 µl of the pooled eluate was loaded onto a HiChrom ACE C18 column (2.1mm ID, 150 mm length) (HiChrom Ltd., Reading, UK), pre-equilibrated with 5% ACN/0.1% formic acid (FA) and using an Agilent 1100 LC system (Agilent Technologies inc., Santa Clara, CA, USA). Elution was achieved using a binary solvent gradient of 5-70% ACN in 0.1% FA at a flow rate of 100 µl/min over 90 min, and finally up to 95% ACN/0.1% FA in 10 min. UV detection was set at 214 nm with fractions collected every minute. Fractions were dried and resuspended in 40 µl of 50% ACN/0.1% FA, then 15 µl was taken for chip-based nanoESI analysis using the Nanomate and Q-TOF Premier, 2 µl were used for MALDI-TOF analysis (Ultraflex), and the remaining was dried before resuspension in 15 µl of trypsin solution (100 ng trypsin in 100 mM ammonium bicarbonate, pH 8.5) and incubation overnight at 37°C. Finally, digested fractions were subjected to C18 ZipTips® clean-up as detailed above for MALDI-TOF MS profiling, except that the elution was performed using 10 µl of 50% ACN/0.1% FA. The purified digested fractions were then loaded on the Nanomate system and nanoESI MS/MS data were collected using data dependent acquisition.

Identification of peptides was performed by searching against the human protein sequence library in NCBInr (v20081128, 216937 sequences) using the Mascot 2.2 search engine (Matrix

Science, London, UK). Searches were performed choosing “None” for enzyme (“Trypsin” for the digested fractions) and with a mass tolerance of 0.1 Da for parent ions and 0.2 Da for fragment ions. “Ammonia-loss (N-term C)”, “Deamidated (NQ)”, “Dehydrated (N-term C)”, “Oxidation (M)”, “Phospho (ST)”, and “Phospho (Y)” were set as variable modifications. GPMAW software v7.10 (Lighthouse data, Odense, Denmark) was used to match accurate mass data obtained from undigested samples with theoretical peptide masses of protein hits obtained from the MASCOT analysis of the MS/MS data of the digested fractions. Isotope pattern software v3.0 (Bruker Daltonics) was used for comparison of experimental isotopomer distributions with the theoretical distribution from putative peptide identifications and the UniProt and NCBI protein databases were used to extract additional information for substantiating or rejecting putative identities, *e.g.* information with regard to alternative splicing, pro-sequence cleavage, disulfide bridges and post-translational modifications.

CA125 and PF4 immunoassays: CA125 analysis was performed using an electro-chemiluminescence immunoassay on a Roche Elecsys 2010 analyser (Roche Diagnostics, Burgess Hill, UK). The assay uses monoclonal antibodies OC125 as the detection antibody and M11 as the capture antibody (Fujirebio Diagnostics; Oxford Biosystems, Oxford, UK). PF4 assays were performed on 173 of the case samples and 173 matched controls using an Asserachrom PF4 ELISA kit (Diagnostica Stago UK Ltd). Serum samples were diluted 1:2,100 v/v with dilution buffer and assayed according to the manufacturer's instructions.

Results

Peak discovery: MALDI MS profiling of a set of blinded ovarian cancer pre-diagnosis sera and matched controls from UKCTOCS was conducted to identify possible candidate markers for early detection. Raw MS spectral data were processed (Figure 1A) and classification rules applied to peak intensity information and CA125 values and these tested for significance. Given

the limited number of samples, 6-month time slots starting in different months with cases and two matched controls grouped into triplets were considered. For each time slot, hypotheses of random label distribution were tested, calculating *P*-values and looking for peaks carrying information for discriminating pre-diagnosis cancer cases from matched controls. Using single peak information alone, several of the 67 peaks analysed (Table II) were able to discriminate cases from controls at a confidence level of $P<0.05$, and mostly at the early time slots (Table III). However, none provided significant discrimination after adjustment for multiple testing.

Peak information was next combined with CA125 values to examine if any peaks could improve on CA125 in terms of early detection. The expected probability of error in triple classification is 2/3, so misclassifying 16 out of 30 triplets (as for the 13 month time group) was not significant ($P=0.09$), but still better than random (Table IV). The results for classification using CA125 alone were significant up to 9 months before diagnosis, an important finding. When peak information was added, the classification was improved, with the lead time of detection significant ($P<0.05$) up to 15 months prior to diagnosis, with the only exception being the 12 months time group with a *P*-value of 0.059. Peaks with processed m/z values of 7772 and 9297 (referred to as peaks 2 and 3 based on their frequency of occurrence) were used most often for discrimination of cases and controls at these earlier time slots. The spectral profiles of these peaks are shown in Figure 1B. None of these *P*-values require adjustment as they satisfy the property of validity described in the Materials and Methods. However, they do not demonstrate the significance of individual peaks; they only show that the peaks are significant en masse. A further *P* value was thus calculated. Whereas the overall 'main' *P*-value tests the hypothesis that CA125 and peak intensity do not help to discriminate cases from controls, the 'conditional' *P*-values presented in Table 5 test the hypothesis that given CA125, the peaks do not carry additional information useful for discrimination. For this analysis, the conditional *P*-values are more important, since it is known that CA125 is a useful biomarker and therefore the interest is in whether addition of other data leads to significant improvement. The

conditional *P*-values show that the contribution of adding MS peaks becomes essential only from 10 to 15 months.

The significance of peaks 2 and 3 were next checked directly by adjusting the main and conditional *P*-values using the sets of rules with these peak numbers built in. Table V presents the results for prediction by CA125 and peak 2 (columns 5–8), and by CA125 and peak 3 (columns 9–12). Error rate and *P*-value for prediction using CA125 alone are included in the table for comparison. The 'rule' columns show the best rules selected. As above, the conditional P-values show that improvement on CA125 cannot be achieved by adding information from either peak 2 or 3 up to 9 months before diagnosis, but that these peaks do improve its predictive ability beyond 9 months. Indeed, the results are significant for up to 15 months using peak 2 and 11 months using peak 3. The main and conditional *P*-values given in Table V have been adjusted for multiple hypothesis testing (see (33) and Lemma 2 in Supplementary Information 1). Figure 2 illustrates the performance of the best classification rules log C − 2 log $I(2)$ and log C − log $I(3)$, respectively, in comparison with the performance of log C. The horizontal axis shows time to diagnosis, the vertical axis 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 and that 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 prior to diagnosis. Figure 3 shows the median dynamics of these rules 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 the values from the rules combining CA125 with peak intensity are elevated earlier than when using CA125 alone and that this is followed by the exponential growth of CA125 closer to diagnosis. Notably, combining both peaks with CA125 did not improve the accuracy compared to the single peak models.

Peak identification: Identification of peak 3 (m/z 9297) was achieved using a thoroughly executed fractionation and MS-based analysis of UKCTOCS samples, ensuring that the peak of interest was retained after each preparation step (27). Briefly, MALDI MS spectra were acquired before fractionation and for each fraction. These were compared with ESI MS spectra for all fractions, making sure that the major peaks also appear as major peaks within the respective ESI MS spectra. The fractions in which peak 3 was obtained were also digested and further analysed by MS/MS. From this analysis, three peptides were obtained that identified connective tissue-activating peptide III (CTAPIII) as the major component of the relevant fractions. CTAPIII is a bioactive cleavage product of platelet basic protein/C-X-C motif chemokine 7. As also both MALDI and ESI MS of the undigested fractions revealed that the peak(s) close in mass to CTAPIII account for the main ion signal intensities in these spectra, it can be concluded that within these fractions, the MALDI MS peak 3 is in fact CTAPIII. Figure S1 in the Supplementary Information shows example spectra acquired throughout the analysis. In addition, using GPMAW software for matching sequences to *m/z* values obtained from the ESI MS spectra, we found only CTAPIII (full sequence) to fit the obtained MS peak. For further confirmation, we also generated the theoretical isotopic distribution of CTAPIII which matched closely to that found experimentally. Despite intensive effort, peak 2 (m/z 7772) eluded identification using these methods. Instead, we have relied on literature searches and the fact that the peak is relatively isolated in spectra and frequently occurring (Table II). The identity of peak 2 as platelet factor 4 (PF4/CXCL4; average molecular weight 7769.18 Da) was thus inferred from two studies where it was identified by off-line LC fractionation and ESI MS/MS following SELDI-TOF MS serum profiling (34) and by immuno-capture after MALDI-TOF MS serum profiling (35). Although inferred, it is highly likely that peak 2 in our samples represents the chemokine PF4.

Verification: In an attempt to verify the PF4 data, an ELISA was used to measure its levels in 173 of the case samples and 173 matched controls, and associations with case control status,

time to diagnosis, clotting time, tumour stage, age, hormone replacement therapy (HRT) use and CA125 level were examined. There was no significant difference when all cases and controls were compared with median values of 6,958 IU/ml (range=1,642-15,096) for cases and 6,847 IU/ml (range=1,610-15,223) for controls (*P*=0.69). However, PF4 levels were significantly lower in cases in the 6-12 and 12-24 months time to diagnosis groups *versus* those in the 0-6 months time group, with *P*-values of 0.037 and 0.012, respectively (Figure 4A). There was also a trend for lower PF4 levels in cases *versus* controls in the distant time groups, but these were not significant at the 95% confidence level. There was no difference in PF4 level with HRT use, age or clotting time, whilst samples from cases with stage IV tumours at diagnosis had lower levels of PF4 *versus* those with other stages (*e.g. P*=0.027 *versus* stage I) and controls (*e.g. P*=0.013 *versus* all controls). Finally, there was a clear rise in CA125 in cases in the lead up to diagnosis (Figure 4B), whereas PF4 showed no consistent change (Figure 4C). There was no correlation between PF4 and CA125 levels for either the case or control groups (Figure 4D), although examination of only samples with low CA125 (<30 U/ml) did reveal lower PF4 levels in cases *versus* controls that approached significance (*P*=0.065).

Discussion

The purpose of this study was to assess if low mass serum polypeptides carry information to aid in the early diagnosis of ovarian cancer. We showed that two MALDI peaks (identified as CTAPIII and inferred as PF4, respectively), when combined with serum CA125, provided significantly earlier detection of cancer. CA125 alone gave significant prediction up to 9 months prior to diagnosis, similar to a previous report (36), with MS peak information not adding significantly to this. At greater than 9 months prior to diagnosis, CA125 performance was significantly complemented by adding MS peak information, with confident detection up to 15 months using all peaks and 15 and 11 months using peaks 2 and 3, respectively.

We wanted to confirm the altered expression of these peaks using orthogonal assay methods. Peak 2, speculatively identified as PF4, was assayed using a commercial ELISA. There was no significant difference between cases and controls when all samples were considered, but levels of PF4 were lower in distant *versus* proximal cases, as well as there being a trend for lower PF4 in cases *versus* controls at the distant time points, and in women later diagnosed with stage IV cancer. Whilst not confirmatory from a statistical standpoint, these data give some support for reduced serum PF4 levels early in cancer development. For CTAPIII, there were no suitable immunoassays available which would be specific for CTAPIII without also recognising the nine other processed products of platelet basic protein with their overlapping sequences *i.e.* TC-2, CTAPIII(1-81), beta-thromboglobulin, neutrophil-activating peptide 2 (NAP2)(74), NAP2(73), NAP2, TC-1, NAP2(1-66) and NAP2(1-63).

CTAPIII is a chemokine released into the circulation from platelet alpha granules. It is known to stimulate DNA synthesis, mitosis, glycolysis, intracellular cAMP accumulation, prostaglandin E2 secretion and synthesis of hyaluronic acid and sulfated glycosaminoglycan in target tissues. PF4 is also released from platelet alpha granules and possibly leukocytes and has a major role in neutralising the anticoagulant effect of heparin through binding. It is also chemotactic for neutrophils and monocytes and inhibits endothelial and activated T-cell proliferation. From our data, it would appear that the secretion of both chemokines is suppressed in the early stages of ovarian cancer possibly through a host response to tumour development. This suppression may support tumour growth through the modulation of inflammatory and immunogenic processes, coagulation and angiogenesis, and as such may not be specific to ovarian cancer (see below).

Perhaps a weakness of this study is the fact that CA125 was used as part of the diagnostic procedure making interpretation of results for a combined biomarker panel difficult. However, limiting the analysis to only those cases which were screen negative and who developed symptomatic disease would greatly diminish the value of this unique preclinical sample set and the impact of the study. Although the median age was significantly different between the sample

groups (*P*=0.01), it has no clinical significance in this post-menopausal population. It is known that HRT use is a risk factor for ovarian cancer (37) and also that HRT use can have a profound effect on the serum proteome (38). However, there was no significant difference in HRT use between the case and control groups used here (Table I) and indeed no correlation between HRT use (or age) and PF4 level (measured by ELISA). We thus conclude that HRT use is not a confounding factor in this study. The other significant difference between the groups was in the use of the oral contraceptive pill (*P*=0.005), which was higher in the control group. This may be expected as the pill is known to confer long-term protection against ovarian cancer (39).

It has been proposed that tumour-specific exopeptidases may generate surrogate peptide markers of cancer *ex vivo* during coagulation (28). Despite this, we recently showed that such peptides do not make useful biomarkers for ovarian cancer diagnosis (32). This result seems at odds with the findings of the present study. However, it is important to note that both CTAPIII and PF4 are released into the circulation in their processed forms and appear not to arise or be subject to proteolysis during serum preparation: we have found no evidence that smaller fragments of these proteins are generated (27) that may explain their reduced levels in pre-diagnosis sera. We therefore hypothesise that these proteins are altered in the earlier stages of tumour development but do not maintain differential expression close to diagnosis. This may explain why we failed to identify changes in these proteins in serum samples from clinically diagnosed cases and controls using the same profiling strategy (32) (40). Here no models incorporating peak intensities were accurately able to discriminate cases from benign or healthy controls, and peak information failed to add to the performance of CA125.

Although a diagnostic test for ovarian cancer based on SELDI-TOF MS assays, includes CTAPIII as an up-regulated protein (8) (41), we have failed to reproduce this using the same strategy (unpublished data). Notably, both CTAPIII and PF4 have been identified as putative biomarkers down-regulated in serum samples from acute lymphoblastic leukaemia cases where they were similarly identified by off-line LC fractionation and ESI-MS/MS following SELDI-TOF

MS profiling (34). PF4 was also identified from MALDI-TOF MS serum profiling experiments as a putative biomarker of pancreatic cancer and was shown also to be down-regulated in the samples from cases (35). Although this data implies the poor specificity of these proteins in detecting ovarian cancer, it does support their possible negative roles in cancer progression.

In conclusion, our discovery study shows that the period of significant discrimination in advance of diagnosis can be extended from 9 to 15 months if CA125 is combined with certain MALDI MS peaks which we identify as the chemokine CTAPIII and putatively as the chemokine PF4. This data supports a link between platelet function and tumour development in the early stages of ovarian cancer. Further work will be required to validate these findings to assess the potential of these markers for early ovarian cancer detection.

Conflict of interest statement

IJ has a consultancy arrangement with Becton Dickinson in the field of tumour markers and ovarian cancer but not involving work directly related to this study. UM has a financial interest through UCL Business and Abcodia Ltd in the third party exploitation of clinical trials biobanks developed through research at UCL. This does not involve any interests directly related to this work. The other Authors have no conflicts of interest to declare.

Acknowledgements

Research was undertaken within the Women's Health Theme of the UCLH/UCL Comprehensive Biomedical Research Centre, the Computer Learning Research Centre, Royal Holloway, University of London and the University of Reading. UCLH/UCL Comprehensive Biomedical Research Centre received a proportion of funding from the Department of Health's NIHR Biomedical Research Centres funding scheme. The work was supported by MRC grant G0301107 and EPSRC grants EP/E00053/1 and EP/F002998/1.

References

1 IARC. GLOBOCAN 2008. Cancer Incidence and Mortality Worldwide in 2008. http://globocan.iarc.fr/. Accessed in 2011.

2 CRUK. Cancer statistics: Ovarian cancer survival statistics. http://info.cancerresearchuk.org/cancerstats/types/ovary/survival/. Accessed in 2011.

3 Menon U, Gentry-Maharaj A, Hallett R, Ryan A, Burnell M, Sharma A, Lewis S, Davies S, Philpott S, Lopes A, Godfrey K, Oram D, Herod J, Williamson K, Seif MW, Scott I, Mould T, Woolas R, Murdoch J, Dobbs S, Amso NN, Leeson S, Cruickshank D, McGuire A, Campbell S, Fallowfield L, Singh N, Dawnay A, Skates SJ, Parmar M and Jacobs I: Sensitivity and specificity of multimodal and ultrasound screening for ovarian cancer, and stage distribution of detected cancers: results of the prevalence screen of the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). Lancet Oncol *10:* 327-340, 2009.

4 Partridge E, Kreimer AR, Greenlee RT, Williams C, Xu JL, Church TR, Kessel B, Johnson CC, Weissfeld JL, Isaacs C, Andriole GL, Ogden S, Ragard LR and Buys SS: Results from four rounds of ovarian cancer screening in a randomized trial. Obstet Gynecol *113:* 775-782, 2009.

5 Brioschi PA, Irion O, Bischof P, Bader M, Forni M and Krauer F: Serum CA 125 in epithelial ovarian cancer. A longitudinal study. Br J Obstet Gynaecol *94:* 196-201, 1987.

6 Bast RC, Jr., Badgwell D, Lu Z, Marquez R, Rosen D, Liu J, Baggerly KA, Atkinson EN, Skates S, Zhang Z, Lokshin A, Menon U, Jacobs I and Lu K: New tumor markers: CA125 and beyond. Int J Gynecol Cancer *15 Suppl 3:* 274-281, 2005.

7 Timmerman D, Testa AC, Bourne T, Ameye L, Jurkovic D, Van Holsbeke C, Paladini D, Van Calster B, Vergote I, Van Huffel S and Valentin L: Simple ultrasound-based rules for the diagnosis of ovarian cancer. Ultrasound Obstet Gynecol *31:* 681-690, 2008.

8 Fung ET: A recipe for proteomics diagnostic test development: the OVA1 test, from biomarker discovery to FDA clearance. Clin Chem *56:* 327-329, 2010.

9 Wulfkuhle JD, Paweletz CP, Steeg PS, Petricoin EF, 3rd and Liotta L: Proteomic approaches to the diagnosis, treatment, and monitoring of cancer. Adv Exp Med Biol *532:* 59-68, 2003.

10 Ludwig JA and Weinstein JN: Biomarkers in cancer staging, prognosis and treatment selection. Nat Rev Cancer *5:* 845-856, 2005.

11 Chatterjee SK and Zetter BR: Cancer biomarkers: knowing the present and predicting the future. Future Oncol *1:* 37-50, 2005.

12 Kozak KR, Su F, Whitelegge JP, Faull K, Reddy S and Farias-Eisner R: Characterization of serum biomarkers for detection of early stage ovarian cancer. Proteomics *5:* 4589-4596, 2005.

13 Su F, Lang J, Kumar A, Ng C, Hsieh B, Suchard MA, Reddy ST and Farias-Eisner R: Validation of candidate serum ovarian cancer biomarkers for early detection. Biomark Insights *2:* 369-375, 2007.

14 Diamandis EP: Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. Mol Cell Proteomics *3:* 367-378, 2004.

15 Coombes KR, Morris JS, Hu J, Edmonson SR and Baggerly KA: Serum proteomics profiling–a young technology begins to mature. Nat Biotechnol *23:* 291-292, 2005.

16 Karsan A, Eigl BJ, Flibotte S, Gelmon K, Switzer P, Hassell P, Harrison D, Law J, Hayes M, Stillwell M, Xiao Z, Conrads TP and Veenstra T: Analytical and preanalytical biases in serum proteomic pattern analysis for breast cancer diagnosis. Clin Chem *51:* 1525-1528, 2005.

17 Villanueva J, Philip J, Chaparro CA, Li Y, Toledo-Crow R, DeNoyer L, Fleisher M, Robbins RJ and Tempst P: Correcting common errors in identifying cancer-specific serum peptide signatures. J Proteome Res *4:* 1060-1072, 2005.
18 Banks RE, Stanley AJ, Cairns DA, Barrett JH, Clarke P, Thompson D and Selby PJ: Influences of blood sample processing on low-molecular-weight proteome identified by surface-enhanced laser desorption/ionization mass spectrometry. Clin Chem *51:* 1637-1649, 2005.
19 Baumann S, Ceglarek U, Fiedler GM, Lembcke J, Leichtle A and Thiery J: Standardized approach to proteome profiling of human serum based on magnetic bead separation and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Clin Chem *51:* 973-980, 2005.
20 Findeisen P, Sismanidis D, Riedl M, Costina V and Neumaier M: Preanalytical impact of sample handling on proteome profiling experiments with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Clin Chem *51:* 2409-2411, 2005.
21 Timms JF, Arslan-Low E, Gentry-Maharaj A, Luo Z, T'Jampens D, Podust VN, Ford J, Fung ET, Gammerman A, Jacobs I and Menon U: Preanalytic influence of sample handling on SELDI-TOF serum protein profiles. Clin Chem *53:* 645-656, 2007.
22 Thorpe JD, Duan X, Forrest R, Lowe K, Brown L, Segal E, Nelson B, Anderson GL, McIntosh M and Urban N: Effects of blood collection conditions on ovarian cancer serum markers. PLoS One *2:* e1281, 2007.
23 Baggerly KA, Morris JS, Edmonson SR and Coombes KR: Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. J Natl Cancer Inst *97:* 307-309, 2005.
24 Diamandis EP: Serum proteomic profiling by matrix-assisted laser desorption-ionization time-of-flight mass spectrometry for cancer diagnosis: next steps. Cancer Res *66:* 5540-5541, 2006.
25 Hortin GL: The MALDI-TOF mass spectrometric view of the plasma proteome and peptidome. Clin Chem *52:* 1223-1237, 2006.
26 Koomen JM, Li D, Xiao LC, Liu TC, Coombes KR, Abbruzzese J and Kobayashi R: Direct tandem mass spectrometry reveals limitations in protein profiling experiments for plasma biomarker discovery. J Proteome Res *4:* 972-981, 2005.
27 Tiss A, Smith C, Menon U, Jacobs I, Timms JF and Cramer R: A well-characterised peak identification list of MALDI MS profile peaks for human blood serum. Proteomics *10:* 3388-3392, 2010.
28 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 and Tempst P: Differential exoprotease activities confer tumor-specific serum peptidome patterns. J Clin Invest *116:* 271-284, 2006.
29 Gammerman A, Vovk V, Burford B, Nouretdinov I, Luo Z, Chervonenkis A, Waterfield M, Cramer R, Tempst P, Villanueva J, Kabir M, Camuzeaux S, Timms J, Menon U and Jacobs I: Serum proteomic abnormality predating screen detection of ovarian cancer. Computer J *52:* 326-333, 2009.
30 Menon U, Skates SJ, Lewis S, Rosenthal AN, Rufford B, Sibley K, Macdonald N, Dawnay A, Jeyarajah A, Bast RC, Jr., Oram D and Jacobs IJ: Prospective study using the risk of ovarian cancer algorithm to screen for ovarian cancer. J Clin Oncol *23:* 7919-7926, 2005.
31 Tiss A, Smith C, Camuzeaux S, Kabir M, Gayther S, Menon U, Waterfield M, Timms J, Jacobs I and Cramer R: Serum peptide profiling using MALDI mass spectrometry: avoiding the pitfalls of coated magnetic beads using well-established ZipTip technology. Proteomics *7 Suppl 1:* 77-89, 2007.
32 Timms JF, Cramer R, Camuzeaux S, Tiss A, Smith C, Burford B, Nouretdinov I, Devetyarov D, Gentry-Maharaj A, Ford J, Luo Z, Gammerman A, Menon U and Jacobs I: Peptides

generated *ex vivo* from serum proteins by tumor-specific exopeptidases are not useful biomarkers in ovarian cancer. Clin Chem *56:* 262-271, 2010.
33 Holm S: A simple sequentially rejective multiple test procedure. Scand J Stat *6:* 65-70, 1979.
34 Shi L, Zhang J, Wu P, Feng K, Li J, Xie Z, Xue P, Cai T, Cui Z, Chen X, Hou J, Zhang J and Yang F: Discovery and identification of potential biomarkers of pediatric acute lymphoblastic leukemia. Proteome Sci *7:* 7, 2009.
35 Fiedler GM, Leichtle AB, Kase J, Baumann S, Ceglarek U, Felix K, Conrad T, Witzigmann H, Weimann A, Schutte C, Hauss J, Buchler M and Thiery J: Serum peptidome profiling revealed platelet factor 4 as a potential discriminating Peptide associated with pancreatic cancer. Clin Cancer Res *15:* 3812-3819, 2009.
36 Anderson GL, McIntosh M, Wu L, Barnett M, Goodman G, Thorpe JD, Bergan L, Thornquist MD, Scholler N, Kim N, O'Briant K, Drescher C and Urban N: Assessing lead time of selected ovarian cancer biomarkers: a nested case–control study. J Natl Cancer Inst *102:* 26-38, 2010.
37 Zhou B, Sun Q, Cong R, Gu H, Tang N, Yang L and Wang B: Hormone replacement therapy and ovarian cancer risk: a meta-analysis. Gynecol Oncol *108:* 641-651, 2008.
38 Pitteri SJ, Hanash SM, Aragaki A, Amon LM, Chen L, Busald Buson T, Paczesny S, Katayama H, Wang H, Johnson MM, Zhang Q, McIntosh M, Wang P, Kooperberg C, Rossouw JE, Jackson RD, Manson JE, Hsia J, Liu S, Martin L and Prentice RL: Postmenopausal estrogen and progestin effects on the serum proteome. Genome Med *1:* 121, 2009.
39 Beral V, Doll R, Hermon C, Peto R and Reeves G: Ovarian cancer and oral contraceptives: collaborative reanalysis of data from 45 epidemiological studies including 23,257 women with ovarian cancer and 87,303 controls. Lancet *371:* 303-314, 2008.
40 Tiss A, Timms JF, Smith C, Devetyarov D, Gentry-Maharaj A, Camuzeaux S, Burford B, Nouretdinov I, Ford J, Luo Z, Jacobs I, Menon U, Gammerman A and Cramer R: Highly accurate detection of ovarian cancer using CA125 but limited improvement with serum matrix-assisted laser desorption/ionization time-of-flight mass spectrometry profiling. Int J Gynecol Cancer *20:* 1518-1524, 2011.
41 Podust V, Zhang Z, Fung ET, Bast R, Chan DW and Song J: Biomarker for Ovarian and Endometrial Cancer: Hepcidin. United States: Vermilllion, Inc. (Fremont, CA, US), The John Hopkins University (Baltimore, MD, US), Board of Regents, The University of Texas System (Austin, TX, US), 2009.

Tables

Table I. Baseline characteristics of cases and controls with significance testing.

Baseline characteristics	Median (25^{th} - 75^{th} centiles)		*P*-value
	Controls (n=585)	**Cases (n=159)**	**Mann-Whitney test**
Age (years) at randomisation	59.8 (55.6-65.2)	61.7 (57.3-66.4)	***0.01***
Age at last period (years)	50.1 (46.8-52.7)	50.7 (46.2-52.6)	0.873
Duration of HRT use in those on HRT at randomisation (years)	7.8 (4.5-10.7)	8.9 (4.8-13.8)	0.263
Duration of OCP use (years) in those who had used it	5 (2-10)	5 (2-10)	0.639
Miscarriages (pregnancies <6 months)	0 (0-1)	0 (0-1)	0.656
No. of children (pregnancies >6 months)	2 (2-3)	2 (1-3)	0.167
Height (cm)	162.6 (157.5-165.1)	162.6 (157.5-167.6)	0.809
Weight (kg)	67.1 (60.8-76.2)	68.7 (60.3-76.2)	0.601
	Number (%)		**Fisher's exact test**
Ethnicity:			0.59
White	571 (97.6%)	155 (97.5%)	
Black	4 (0.7%)	0 (0.0%)	
Asian	5 (0.9%)	2 (1.3%)	
Other	2 (0.3%)	1 (0.6%)	
Missing	3 (0.5%)	1 (0.6%)	
Hysterectomy	101 (17.3%)	27 (17.0%)	1.000
Ever use of OCP	327 (55.9%)	69 (43.4%)	***0.005***
Use of HRT at recruitment	126 (21.5%)	41 (25.8%)	0.284
Maternal history of ovarian cancer	30 (5.1%)	5 (3.1%)	0.693
Maternal history of breast cancer	102 (17.4%)	31 (19.5%)	0.450

HRT, Hormone replacement therapy; OCP, oral contraceptive pill.

Table II. List of the 67 most frequently occurring MALDI MS peaks in the dataset showing their peak number assignment, processed m/z values and frequency of occurrence.

Peak number	*m/z* Value (Da)	Frequency (out of 11048)	Peak number	*m/z* Value (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	11036	38	9724.2	6658
5	5909.2	11029	39	5583.3	6570
6	6636	10994	40	4285.7	6517
7	4647.8	10904	41	2993.5	6460
8	4057.2	10878	42	2606	6422
9	3243.9	10739	43	2274.3	6358
10	3959	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	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	4910
21	3510.1	9870	55	6810	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 III. Statistical analysis of peaks for discrimination of cases and controls. *P*-values are presented for each of 67 peaks distributed by time-to-diagnosis time slots (t + 6 months prior to diagnosis). *P*-values of <0.05 are indicated in bold italics. For adjusted *P*-values a threshold of 0.05/67= 0.00075 should be considered as significant at the 95% confidence level.

t	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No. cases	68	56	47	36	27	23	20	17	17	20	28	28	28	30	25	20
Peak 1	0.507	0.299	0.188	0.427	0.567	0.515	0.330	0.722	0.080	0.192	0.467	0.462	0.471	0.279	0.624	0.520
2	0.714	0.626	0.185	0.560	0.269	0.352	0.091	0.335	0.174	0.193	***0.009***	***0.021***	0.103	***0.045***	***0.042***	***0.042***
3	0.506	0.407	0.195	0.290	0.723	0.689	0.185	0.333	0.529	0.341	0.051	0.107	0.308	0.093	***0.045***	0.094
4	0.162	0.086	0.074	0.190	0.575	0.678	0.515	0.525	0.175	0.195	0.304	0.309	0.455	0.409	0.627	0.702
5	0.112	***0.008***	0.071	0.194	0.411	0.823	0.700	0.868	0.331	0.093	0.309	0.191	0.183	0.089	0.171	0.519
6	0.067	0.086	0.189	0.185	0.268	0.216	0.198	0.327	0.173	0.527	0.188	0.102	0.103	0.282	0.628	0.520
7	0.865	0.934	0.907	0.998	0.993	0.974	0.942	0.871	0.533	0.524	0.108	0.185	0.310	0.171	0.183	0.097
8	0.160	0.297	0.512	0.704	0.930	0.696	0.520	0.326	0.330	0.094	***0.023***	***0.021***	***0.024***	***0.043***	***0.040***	0.088
9	0.312	0.395	0.287	0.297	0.409	0.822	0.941	0.959	0.872	0.521	0.191	***0.048***	***0.048***	***0.044***	0.088	0.197
10	0.788	0.886	0.843	0.571	0.734	0.525	0.516	0.717	0.528	0.339	0.625	0.944	0.946	0.827	0.887	0.937
11	0.157	0.207	0.287	0.419	0.583	0.690	0.711	0.722	0.718	0.339	0.104	0.053	0.099	0.285	0.312	0.519
12	0.316	0.516	0.523	0.420	0.578	0.688	0.715	0.874	0.871	0.849	0.619	0.466	0.624	0.713	0.775	0.701
13	0.513	0.405	0.285	0.560	0.570	0.826	0.847	0.724	0.864	0.849	0.468	0.188	0.189	0.165	0.182	0.187
14	0.970	0.963	0.837	0.699	0.417	0.109	0.344	0.713	0.712	0.523	0.625	0.758	0.625	0.571	0.631	0.851
15	0.618	0.729	0.191	0.110	***0.039***	0.103	0.199	0.524	0.526	0.852	0.873	0.764	0.623	0.837	0.786	0.701
16	0.853	0.985	0.909	0.421	0.569	0.359	0.528	0.529	0.719	0.852	0.761	0.940	0.942	0.839	0.776	0.702
17	0.985	0.993	0.988	0.895	0.929	0.927	0.941	0.532	0.327	0.334	0.463	0.308	0.194	0.278	0.459	0.347
18	0.945	0.879	0.840	0.704	0.724	0.524	0.524	0.710	0.720	0.855	0.769	0.763	0.628	0.282	0.304	0.091
19	0.612	0.634	0.639	0.423	0.850	0.697	0.342	0.332	0.720	0.516	0.471	0.622	0.865	0.830	0.785	0.846
20	0.408	0.812	0.640	0.693	0.851	0.923	0.846	0.719	0.872	0.847	0.768	0.464	0.467	0.712	0.779	0.704
21	0.312	0.933	0.947	0.979	0.575	0.832	0.941	0.530	0.713	0.519	0.186	0.190	0.193	0.412	0.301	0.340
22	0.612	0.624	0.633	0.694	0.406	0.205	***0.042***	0.079	0.172	0.193	0.320	0.104	0.194	0.287	0.463	0.524
23	0.408	0.821	0.747	0.977	0.927	0.993	1.000	0.962	0.875	0.849	0.874	0.765	0.762	0.912	0.955	0.946
24	0.795	0.809	0.637	0.810	0.925	0.972	0.855	0.875	0.528	0.701	0.630	0.460	0.623	0.572	0.890	0.848
25	0.611	0.517	0.836	0.293	0.268	0.518	0.524	0.327	0.169	0.340	0.053	0.101	0.186	0.277	0.184	0.096
26	0.795	0.934	0.946	0.814	0.847	0.836	0.707	0.528	0.524	0.333	0.471	0.626	0.871	0.835	0.893	0.942
27	0.704	0.727	0.951	0.899	0.931	0.688	0.700	0.718	0.532	0.703	0.763	0.870	0.765	0.408	0.454	0.338
28	0.607	0.299	0.630	0.696	0.567	0.832	0.845	0.954	0.724	0.853	0.881	0.874	0.767	0.567	0.771	0.527
29	0.791	0.879	0.946	0.816	0.924	0.921	0.849	0.863	0.720	0.700	0.761	0.771	0.876	0.563	0.463	0.527
30	0.910	0.813	0.384	0.899	0.724	0.519	0.850	0.952	0.875	0.943	0.768	0.624	0.469	0.175	0.177	***0.038***
31	0.107	***0.028***	***0.039***	0.111	0.417	0.204	0.340	0.524	0.870	0.940	0.944	0.942	0.876	0.917	0.885	0.712
32	0.611	0.517	0.836	0.293	0.268	0.518	0.524	0.327	0.169	0.340	0.053	0.101	0.186	0.277	0.184	0.096
33	0.418	0.510	0.636	0.555	0.570	***0.048***	0.200	0.519	0.314	0.524	0.313	0.762	0.772	0.415	0.459	0.336
34	0.320	0.135	0.502	0.820	0.727	0.973	0.938	0.989	0.873	0.523	0.192	0.101	0.195	0.089	0.091	0.196
35	0.857	0.887	0.903	0.296	0.576	0.525	0.333	0.328	0.323	0.519	0.626	0.625	0.876	0.710	0.774	0.704
36	0.857	0.811	0.639	0.420	0.847	0.521	0.709	0.871	0.989	0.997	0.878	0.622	0.459	0.420	0.301	0.095
37	0.908	0.883	0.953	0.950	0.723	0.693	0.702	0.723	0.529	0.707	0.624	0.768	0.871	0.712	0.770	0.704
38	0.621	0.981	0.642	0.188	0.257	0.209	0.201	0.181	0.335	0.705	0.877	0.879	0.870	0.829	0.885	0.844
39	0.107	0.624	0.639	0.559	0.851	0.993	0.983	0.725	0.714	0.528	0.314	0.098	0.185	0.420	0.457	0.515
40	0.413	0.299	0.392	0.426	0.725	0.519	0.702	0.870	0.872	0.708	0.767	0.940	0.977	0.911	0.885	0.938
41	0.408	0.727	0.952	0.421	0.928	0.970	0.935	0.726	0.332	0.186	0.317	0.050	***0.050***	0.166	0.461	0.519
42	0.411	0.815	0.636	0.893	0.845	0.835	0.519	0.329	0.526	0.337	0.320	0.468	0.620	0.710	0.774	0.853
43	0.858	0.929	0.949	0.812	0.921	0.927	0.940	0.989	0.954	0.984	0.979	0.994	0.999	0.963	0.953	0.848
44	0.415	0.508	0.828	0.699	0.844	0.993	0.984	0.990	0.961	0.848	0.768	0.315	0.463	0.420	0.457	0.527
45	0.790	0.814	0.632	0.948	0.973	0.999	0.938	0.719	0.328	0.336	0.471	0.198	0.463	0.576	0.890	0.849
46	0.233	0.141	0.196	0.058	0.574	0.687	0.530	0.524	0.717	0.705	0.760	0.616	0.763	0.709	0.630	0.695
47	0.997	0.981	0.974	0.893	0.728	0.683	0.522	0.719	0.173	0.336	0.760	0.873	0.978	0.912	0.985	0.985
48	0.616	0.728	0.830	0.558	0.414	0.352	0.527	0.521	0.516	0.519	0.464	0.472	0.470	0.282	0.452	0.339
49	0.712	0.720	0.833	0.814	0.736	0.830	0.699	0.716	0.335	0.344	0.620	0.763	0.872	0.559	0.773	0.714
50	0.973	0.934	0.394	0.896	0.732	0.693	0.845	0.872	0.958	0.941	0.626	0.461	0.457	0.273	0.181	0.095
51	0.162	0.519	0.835	0.809	0.970	0.973	0.983	0.954	0.872	0.699	0.310	0.315	0.310	0.272	0.184	0.339
52	0.110	0.054	0.504	0.561	0.564	0.927	0.852	0.714	0.321	0.344	0.465	0.318	0.467	0.164	0.176	0.192
53	0.318	0.212	0.278	0.295	0.571	0.921	0.982	0.990	0.959	0.849	0.766	0.318	0.317	0.571	0.623	0.708
54	0.945	0.931	0.992	0.810	0.972	0.829	0.856	0.957	0.870	0.944	0.941	0.945	0.871	0.712	0.774	0.698
55	0.862	0.963	0.907	0.810	0.722	0.921	0.980	0.957	0.991	0.941	0.946	0.942	0.942	0.915	0.893	0.946
56	0.234	0.517	0.286	0.111	0.156	0.112	0.097	0.081	0.172	0.193	0.194	0.189	0.189	0.169	0.300	0.334
57	0.114	0.209	0.069	0.429	0.269	0.350	0.942	0.864	0.528	0.523	0.620	0.767	0.468	0.278	0.458	0.340
58	0.859	0.734	0.639	0.291	0.155	***0.018***	***0.014***	***0.029***	0.173	0.517	0.766	0.621	0.620	0.572	0.626	0.335
59	0.230	0.517	0.745	0.812	0.841	0.923	0.850	0.724	0.331	0.336	0.197	0.315	0.463	0.275	0.464	0.336
60	0.788	0.625	0.745	0.698	0.929	0.685	0.847	0.875	0.720	0.710	0.625	0.628	0.630	0.279	0.300	0.191
61	0.317	0.139	0.268	0.189	0.260	0.829	0.844	0.958	0.874	0.853	0.874	0.466	0.626	0.563	0.631	0.519
62	0.992	0.992	0.905	0.894	0.855	0.699	0.852	0.873	0.719	0.524	0.623	0.766	0.872	0.715	0.889	0.940
63	0.111	0.093	0.641	0.810	0.410	0.928	0.849	0.956	0.957	0.705	0.872	0.771	0.768	0.719	0.619	0.852
64	0.613	0.886	0.948	0.813	0.931	0.684	0.849	0.871	0.720	0.938	0.873	0.871	0.765	0.705	0.766	0.510
65	0.237	0.135	0.185	0.417	0.845	0.975	0.935	0.868	0.719	0.522	0.471	0.618	0.767	0.568	0.779	0.846
66	0.067	0.211	0.073	0.058	0.725	0.927	0.943	0.723	0.526	0.530	0.321	0.203	0.311	0.570	0.794	0.704
67	0.861	0.622	0.734	0.977	0.991	0.973	0.940	0.869	0.718	0.853	0.943	0.621	0.316	0.093	0.180	***0.039***

Table IV. Initial statistical analysis of CA125 and MS peak information.

t[a]	No. cases	CA125 only		CA125 and all peaks				
		No. errors	CA125 *P*-value	No. errors	Peak used	*Weight v/w*	Main *P*-value	Conditional *P*-value
0	68	2	0.0001	1	1	1	0.0001	0.3164
1	56	4	0.0001	2	7	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	0.001	4	15	-1	0.0028	0.6973
7	17	6	0.0071	4	1	-1	0.0141	0.6034
8	17	5	0.0021	3	1	-1	0.0019	0.1523
9	20	7	0.0042	5	2	-1	0.0076	0.4497
10	28	14	0.0503	6	3	-1	0.0003	0.0013
11	28	15	0.1028	8	3	-1	0.0042	0.0078
12	28	17	0.3164	10	2	-2	0.0585	0.0658
13	30	16	0.0895	10	2	-2	0.0168	0.0428
14	25	16	0.4661	8	2	-2	0.0304	0.0206
15	20	13	0.5211	6	2	-2	0.0464	0.0609
16	10	6	0.4406	2	67	+1/0	0.4101	0.5066

[a]Columns in order are: time *t* to diagnosis in months; the number of cases (triplets) with measurement taken between *t* and *t* + 6 months prior to diagnosis; the number of errors when classifying triplets with CA125 alone and associated *P*-value; the minimal number of errors when classifying with $w \log C + v \log I(p)$; most often used peak (see Table II); the ratio of weights *v*/*w* used in the rule; the main *P*-value for overall significance of this result; the conditional *P*-value for significance of non-CA125 component. Shaded areas show significance at *P*-value <0.05 for CA125 and main *P*-values beyond 9 months.

Table V. Experimental results for triplet classification with a fixed peak (peak 2 or 3) and CA125. Results for prediction by CA125 alone (columns 3-4), CA125 and peak 2 (columns 5–8) or CA125 and peak 3 (columns 9–12) are shown for each time to diagnosis group.

t	No. cases	CA125 only		CA125 and peak 2				CA125 and peak 3			
		Error[b]	CA125 *P*-value	Rule[a]	Error[b]	Main *P*-value	Conditional *P*-value[c]	Rule[a]	Error[b]	Main *P*-value	Conditional *P*-value[c]
0	68	2	0.0001	2CA-p02	2	0.0007	1	2CA-p03	2	0.0015	1
1	56	4	0.0001	2CA-p02	4	0.0007	1	CA+p03	4	0.0015	1
2	47	6	0.0001	2CA-p02	5	0.0007	1	2CA-p03	5	0.0015	1
3	36	8	0.0001	2CA-p02	7	0.0007	1	2CA-p03	7	0.0015	1
4	27	7	0.0001	2CA-p02	6	0.0007	1	CA+p03	6	0.0015	1
5	23	7	0.0008	CA-p02	6	0.0026	1	2CA-p03	6	0.0104	1
6	20	6	0.001	CA-p02	5	0.0066	1	CA-p03	6	0.0681	1
7	17	6	0.0071	CA-p02	4	0.0112	1	CA-p03	5	0.1451	1
8	17	5	0.0021	CA-p02	4	0.0132	1	CA-p03	4	0.0296	1
9	20	7	0.0042	CA-p02	5	0.0066	0.5461	CA-p03	5	0.0133	1
10	28	14	0.0503	CA/2-p02	7	0.0007	0.0046	CA-p03	6	0.0015	0.003
11	28	15	0.1028	CA/2-p02	9	0.0053	0.0132	CA-p03	8	0.0059	0.0074
12	28	17	0.3164	CA/2-p02	10	0.0217	0.0296	CA-p03	10	0.0725	0.0725
13	30	16	0.0895	CA/2-p02	10	0.0046	0.0092	CA-p03	10	0.0222	0.0237
14	25	16	0.4661	CA/2-p02	8	0.0099	0.0072	CA-p03	10	0.4456	0.268
15	20	13	0.5211	CA/2-p02	6	0.0145	0.0072	CA-p03	8	0.8542	1
16	10	6	0.4406	CA/2+p02	5	1	1	CA/2+p03	5	1	1

[a]The 'rule' columns show the best rules selected: CA in these columns means log *C*, p02 means log *I*(2), p03 means log *I*(3). [b]The number of errors is shown for the best rule with the associated main *P*-value and conditional *P*-value. [c]Conditional *P*-value shows the improvement achieved by adding information from peak 2 or 3 beyond 9 months. *P*-values were multiplied by adjustment coefficients $4\zeta(2)=6.579$ for peak 2 and $9\zeta(2)=14.804$ for peak 3. The same table with non-adjusted *P*-values is shown in Supplementary Information, Table S1.

Figure Legends

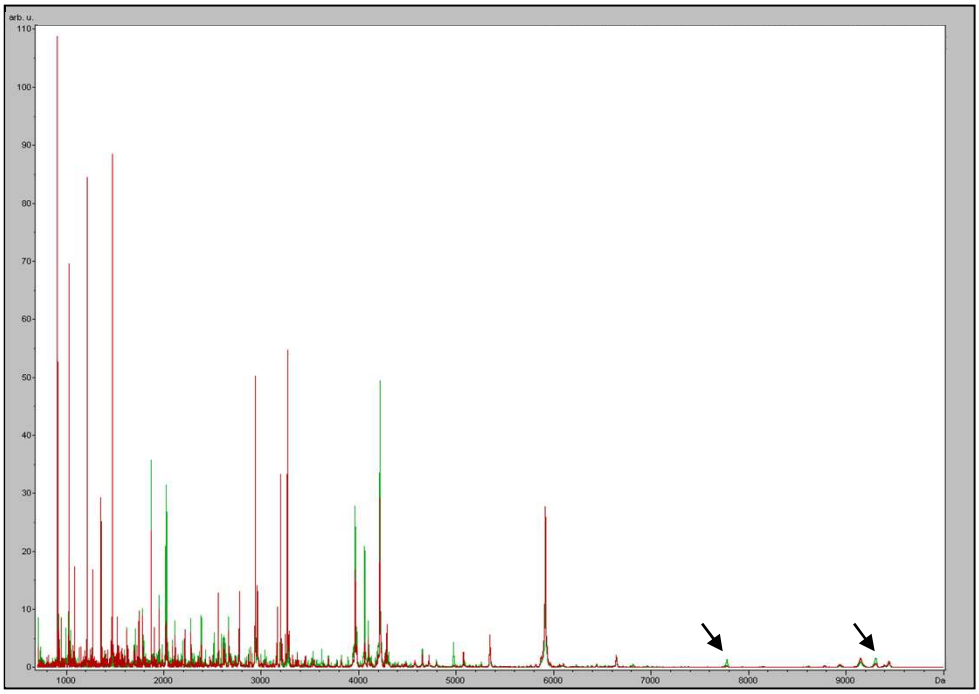
Figure 1. A: Processed full-mass range MS spectra for example case (red) and healthy control (green) samples. The spectra are averaged from 10 replicate acquisitions per sample. Peak 2 (m/z 7772) and peak 3 (m/z 9297) are indicated by arrows. **B:** Processed MS data for peak 2 (m/z 7772) and peak 3 (m/z 9297) plotted for all samples. Red is controls, blue is cases.

Figure 2. A: Comparison of rules log C and log C − 2 log $I(2)$ for peak 2 on time/case scale. **B:** Comparison of rules log C and log C − log $I(3)$ for peak 3 on time/case scale. A circle means that a triplet was correctly classified by both rules. A cross means misclassification in both cases. A triangle upwards shows improvement and downwards shows deterioration after addition of the − log $I(p)$ component. The figures demonstrate that most samples where addition of a peak to CA125 is beneficial (marked as upward triangles) are in the interval of 13–16 months before the diagnosis (dashed vertical lines).

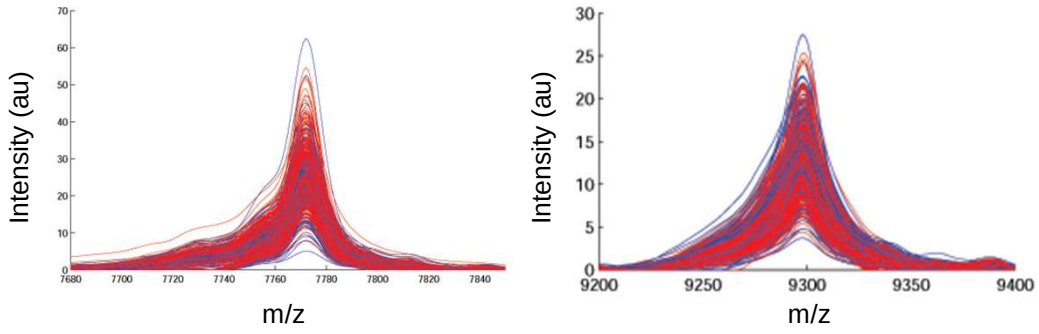
Figure 3. A: Median dynamics of rules log C and log C − 2 log $I(2)$ for cases only. **B:** Median dynamics of rules log C and log C − log $I(3)$ for cases only.

Figure 4. A: Scatter dot plots for serum levels of platelet factor 4 (PF4) measured by ELISA for cases and matched controls in groups with different times to diagnosis. The horizontal bars indicate mean values. Significant changes between case groups are indicated. **B:** Continuous time to diagnosis data plotted for CA125. LOWESS curve fitting was applied to the CA125 data (solid lines) **C:** Continuous time to diagnosis data plotted for PF4. **D:** PF4 and CA125 data were plotted against one another (for cases only). Linear regression curve fitting was applied to the PF4/CA125 plot (dashed line).

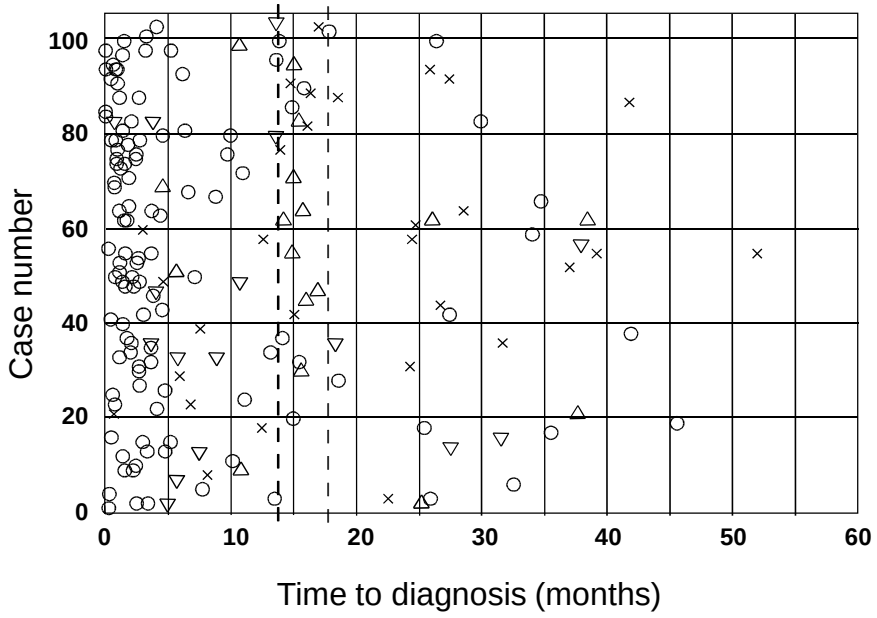
A


arb. u.
110
100
90
80
70
60
50
40
30
20
10
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
Da

B


Intensity (au)
70
60
50
40
30
20
10
0
7680
7700
7720
7740
7760
7780
7800
7820
7840
m/z
Intensity (au)
30
25
20
15
10
5
0
9200
9250
9300
9350
9400
m/z

A


Case number
0
20
40
60
80
100
0
10
20
30
40
50
60
Time to diagnosis (months)

B

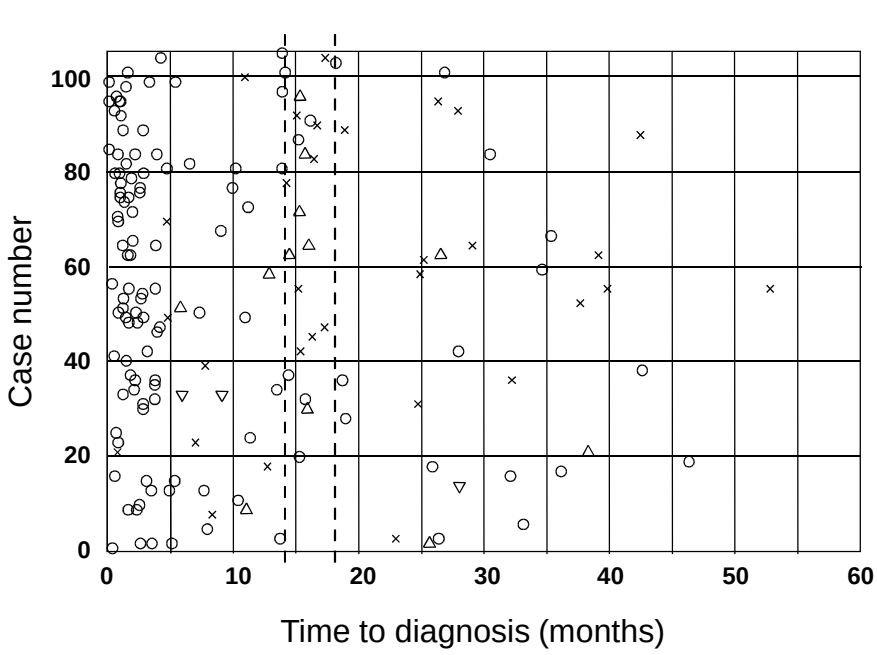
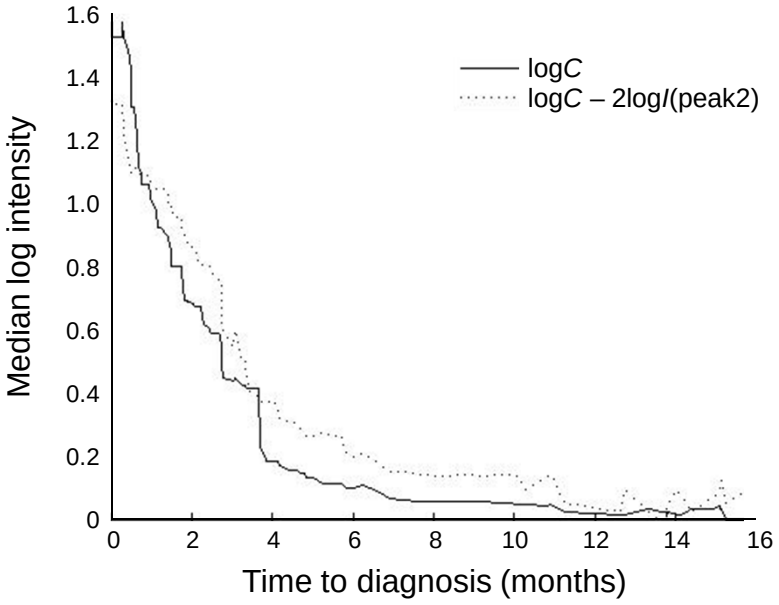

Case number
0
20
40
60
80
100
0
10
20
30
40
50
60
Time to diagnosis (months)

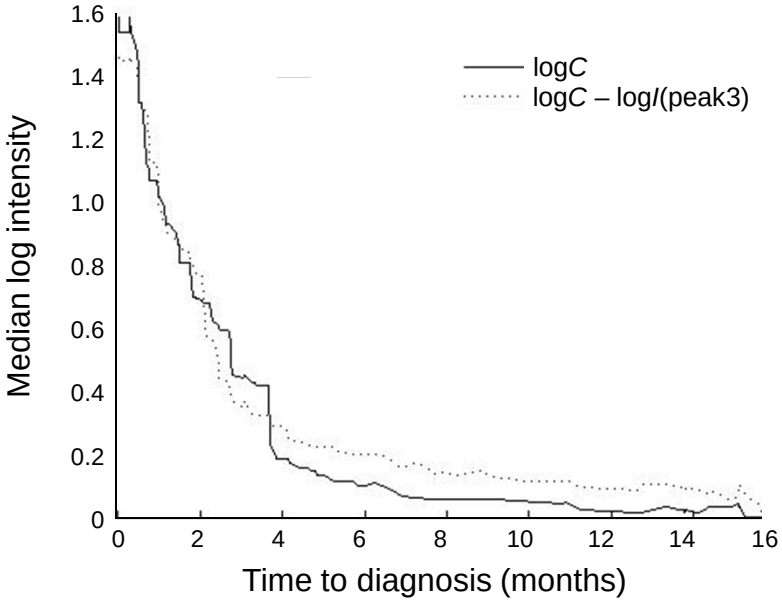
Figure 2

A



Peak 2 (*m/z* 7772)

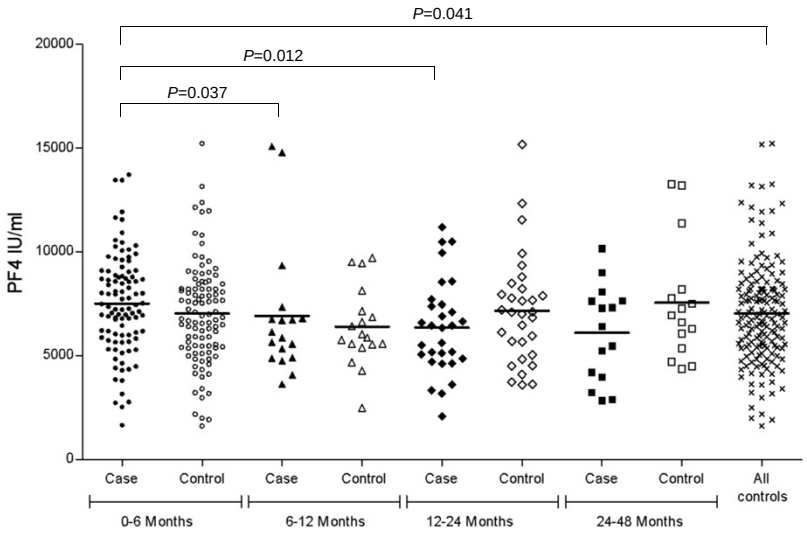
B



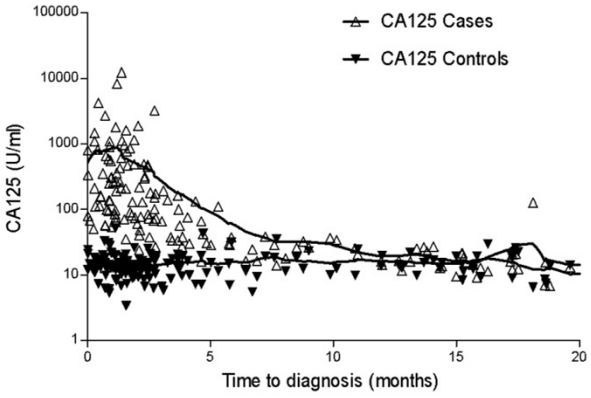
Peak 3 (*m/z* 9297)

Figure 3

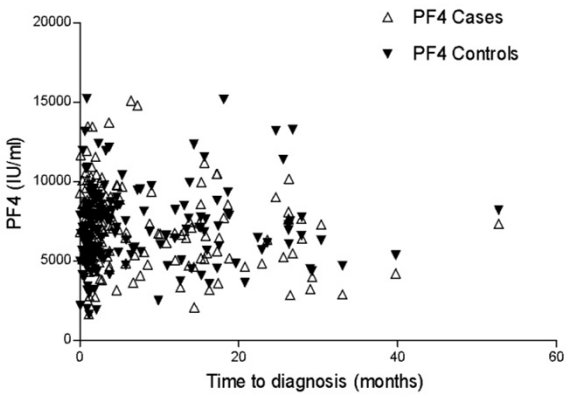
A



B



C



D

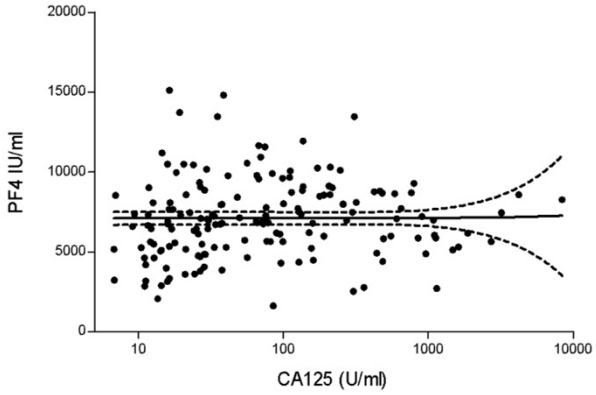


Figure 4