DOI 10.1007/s13748-012-0021-y

REGULAR PAPER

Conformal predictors in early diagnostics of ovarian and breast cancers

Dmitry Devetyarov · Ilia Nouretdinov · Brian Burford · Stephane Camuzeaux · Aleksandra Gentry-Maharaj · Ali Tiss · Celia Smith · Zhiyuan Luo · Alexey Chervonenkis · Rachel Hallett · Volodya Vovk · Mike Waterfield · Rainer Cramer · John F. Timms · John Sinclair · Usha Menon · Ian Jacobs · Alex Gammerman

Received: 7 November 2011 / Accepted: 10 June 2012 / Published online: 8 July 2012

Abstract The work describes an application of a recently developed machine-learning technique called Mondrian predictors to risk assessment of ovarian and breast cancers. The analysis is based on mass spectrometry profiling of human serum samples that were collected in the United Kingdom Collaborative Trial of Ovarian Cancer Screening. The work describes the technique and presents the results of classification (diagnosis) and the corresponding measures of confidence of the diagnostics. The main advantage of this approach is a proven validity of prediction. The work also describes an approach to improve early diagnosis of ovarian and breast cancers since the data in the United Kingdom Collaborative Trial of Ovarian Cancer Screening were collected over a period of 7 years and do allow to make observations of changes in human serum over that period of time. Significance of improvement is confirmed statistically (for up to 11 months for ovarian cancer and 9 months for breast cancer). In addition, the methodology allowed us to pinpoint the same mass spectrometry peaks as previously detected as carrying statistically significant information for discrimination between healthy and diseased patients. The results are discussed.

Keywords Proteomics · Mass spectrometry · Artificial intelligence · Prediction with confidence · Early diagnosis · Conformal predictors · Ovarian cancer · Breast cancer · Biological markers

1 Introduction

Recent advances in the analysis of the human serum proteome aim to establish novel disease biomarkers that would allow early detection of diseases. The current techniques include analysis of serum using mass spectrometry (MS). The output of MS is a large volume of high-dimensional data (Fig. 1), and it requires modern methods of data analysis.

Most known techniques are usually good in accuracy of classification (diagnosis) but suffer from a lack of a measure of confidence in the diagnosis; therefore, it is difficult to estimate risk of incorrect diagnosis of a patient. This work describes a novel machine-learning technique called Mondrian predictors [1], also known as category-based confidence machines, that addresses this problem by introducing measures of confidence that would allow us to estimate a risk of misclassification. The Mondrian predictors were applied to a subset of the United Kingdom Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) biobank which contains serum samples and data on cancers in a cohort of 202,638 women participating in this trial. Women were recruited between 2001 and 2005, and those in the cancer antigen 125 (CA125) screening (multimodal) group underwent annual screening with repeat samples collected if an abnormality

D. Devetyarov · I. Nouretdinov · B. Burford · Z. Luo ·
A. Chervonenkis · V. Vovk · A. Gammerman
Computer Learning Research Centre, Royal Holloway,
University of London, Egham, UK

S. Camuzeaux · A. Gentry-Maharaj · R. Hallett · M. Waterfield ·
J. F. Timms · J. Sinclair · U. Menon · I. Jacobs
EGA Institute for Women's Health, University College London,
London, UK

A. Tiss · C. Smith · R. Cramer
BioCentre, University of Reading, Reading, UK

A. Tiss · C. Smith · R. Cramer
Department of Chemistry, University of Reading, Reading, UK

I. Nouretdinov (✉)
Department of Computer Science, Royal Holloway,
University of London, Egham, Surrey, TW20 0EX, UK
e-mail: ilia@cs.rhul.ac.uk

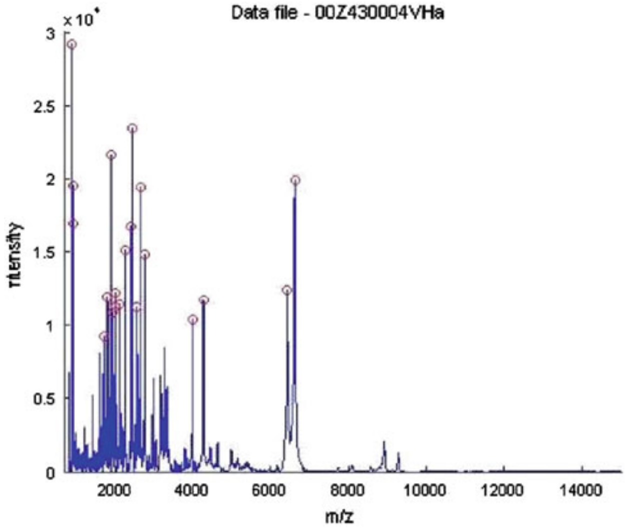

Fig. 1 Example of a spectrum with identified peaks

was detected [2]. Women were followed up through cancer registry and postal questionnaires. The unique feature of this trial was that the women were screened annually for up to 5 years. Two case–control sets of samples from women diagnosed to have ovarian and breast cancer, respectively, and healthy (no cancer at follow up) controls were undertaken. Control samples were matched for trial centres and date when the cancer sample was taken to minimise differences in sample processing. The serum samples underwent pre-fractionation using a reversed-phase batch extraction protocol prior to MALDI-TOF MS data acquisition [3,4]. In this work, we analysed ovarian cancer and breast cancer data sets.

This work is based on the theory of hedged (confident) algorithmic learning [1]. One of the major advantages of hedged algorithms is that they can be used for solving high-dimensional problems without requiring any parametric statistical assumptions about the source of data (unlike traditional statistical techniques); the only assumption made is independent identically distributed (i.i.d.): the examples are generated from the same probability distribution independently of each other. Another advantage of conformal predictor is that it also allows to make estimation of confidence in the classification of individual examples.

The algorithm itself is based on testing each classification hypothesis about a new example whether it is conforming to the i.i.d. assumption. This requires application of a test for randomness based on a non-conformity measure (NCM) which is a way of ranking objects within a set by their relative strangeness. The defining property of NCM is its independence of the order of examples, so any computable functions with this property can be used. Conformal predictor is valid under any choice of NCM, however, it can be more efficient if NCM is appropriate. Concrete meaning of efficiency (performance measure) depends on the problem type and interpretation of output.

NCM for classification is also usually based on an underlying learning algorithm. For example, it can be k-Nearest-Neighbours (kNN) algorithm. Although is usually applied to clean data, where all attributes are informative (such as USPS handwritten digits [5]), with an additional step of feature selection it may be used in less-clean cases, for example, in the work on machine-learning in functional clustering [6], where only few attributes (gene expressions) are useful for separation between diseases, we embedded a step of feature selection (based on a T test) into a version of kNN NCM.

Another useful underlying algorithm is Support Vector Machine (SVM) [7,8]. In the work on diagnostic using microarrays [9], an SVM-based NCM with a feature-selection step was used together with another NCM based on Nearest Centroid that is in some sense a 'limit' version of Nearest Neighbours. In various medical applications, good performance was also shown by NCM based on genetic algorithms [10], and neural networks [11,12].

However, NCM for some special data are not directly based on standard underlying algorithms: in the work [13] they apply algorithm with 'conformity' between a set and a new example represented by a number of attributes (voxels) showing good separability, e.g., by two-sample T test.

We are now working with mass spectrometry that has its own specific characteristics as well. Although it is high-dimensional (many peaks are identified), we know from practice that normally only few of them (called biomarkers) react to the disease. Thus, we developed a version of conformal predictor with a special NCM that is based on a search within high-dimensional data for a simple decision rule that involves only few biomarkers.

This work first outlines the background and introduces the main ideas of conformal predictors and its extension to Mondrian predictors. We then describe the data and classification rules and present the results.

2 Methods

The framework we are going to deploy in the analysis of MS data is the one of conformal predictors [1,14]. It represents a new generation of algorithms with reliability measures.

2.1 Conformal predictors

Let us assume that we are given a training set of patients with diagnoses

$$(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$$

where $x_i \in X$ is a vector of features which describe a patient and $y_i \in Y$ is a diagnosis out of a finite set of possible diagnoses (classes). Our goal is to predict the diagnosis y_n for a new patient x_n. We will denote a combination of a patient and a diagnosis as $z_i = (x_i, y_i) \in Z = X \times Y$.

The general idea of conformal predictors is the following: when we have a new patient, x_n, we try every possible diagnosis y as a candidate for patient's diagnosis and see how well the resulting pair (x_n, y) conforms with $(x_1, y_1), \dots, (x_{n-1}, y_{n-1})$. The ideal case is when exactly one diagnosis conforms with the rest of the sequence and all others do not. We can then be confident in predicting this diagnosis.

Firstly, we need to define the notion of a nonconformity measure, which is the core of conformal predictors. A specific nonconformity measure depends on a particular algorithm and can be based on many well-known machine-learning algorithms. This nonconformity measure will assign some value α_i (nonconformity score) to every patient in the sequence $z_1, \dots, z_n$ including a new patient with diagnosis and will evaluate 'nonconformity' between a set and its element:

$$\alpha_i := A_n(\wr z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n \int, z_i), i = 1, \dots, n, \quad (1)$$

where $\wr \dots \int$ denotes a multiset.

When we consider a diagnosis hypothesis $y_n = y$ and after we calculated the corresponding nonconformity scores $\alpha_1, \dots, \alpha_n$ for a full sequence with diagnosis y for the last patient, a natural way to compare α_n with the other α_is is to look at the ratio of patients which conform with the other patients at most as much as the new one, that is, to calculate

$$p_n(y) = \frac{|\{i = 1, \dots, n-1 : \alpha_i \geq \alpha_n\}| + 1}{n}. \quad (2)$$

This ratio is called the p value associated with the possible diagnosis y for x_n. Thus, we can complement each candidate diagnosis with a p value, which shows how well a new patient with this possible diagnosis conforms with the rest of the sequence in comparison with other patients. The last thing which needs to be set is a significance level $0 < \epsilon < 1$, which is an error rate we are willing to tolerate.

Finally, the p values calculated above can produce a region predictor: the conformal predictor determined by the nonconformity measure $A_n, n = 1, 2, \dots$, and a significance level ϵ is defined as the function $\Gamma : Z^* \times X \times (0, 1) \to 2^Y$ (2^Y is the set of all subsets of Y) such that the prediction set $\Gamma^{(\epsilon)}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$ is defined as the set of all candidate diagnoses $y \in Y$ such that $p_n(y) > \epsilon$. Thus, for any finite sequence of diagnosed patients, $(x_1, y_1, \dots, x_{n-1}, y_{n-1})$, a new undiagnosed patient x_n and a significance level ϵ, the conformal predictor outputs a region prediction $\Gamma^{(\epsilon)}$—a set of possible diagnoses for the new patient.

The main advantage of conformal predictors is their validity: in the long run the frequency of errors made by a conformal predictor (i.e., cases when prediction set Γ^ϵ does not contain the real diagnosis) does not exceed ϵ (this is subject to the assumption that all examples are drawn independently from the same distribution, which is called the i.i.d. assumption). This point is different from the methods (such as logistic regression) which produce probabilistic estimates that rely on assumptions that are stronger than i.i.d. (see Appendix for the discussion of logistic regression).

While validity is guaranteed, we have to optimize efficiency—the ability of conformal predictors to produce as small region predictions as possible.

2.1.1 Alternative way of presenting the results

However, prediction sets are dependent on selected significance level ϵ. If there are several such levels, prediction sets form a nested sequence: prediction set for a smaller ϵ always covers a prediction set for a larger ϵ. If Y is finite, we can summarise all these outputs in one. It is enough to order all possible labels by their p values and to set thresholds for ϵ at which the cardinality of prediction set changes. If Y is binary, this is the same as to output a prediction for each example by choosing the highest p value and to complement each such prediction with two indicators: confidence and credibility. Confidence is equal to 1 less the second maximum p value, it is the complement to 1 of the smallest ϵ at which the prediction set is certain (contains at most one element). Credibility is the maximum value of all possible p values, or the smallest ϵ at which the prediction set is empty.

High confidence means that the alternative diagnoses are excluded by having a low p value, high credibility checks whether the prediction itself does not have a very small p value. Thus, a prediction is considered to be reliable if its confidence is close to 1 and its credibility is not close to 0. If its credibility is low, this means that the new patient is not typical for any class presented in the training set.

2.2 Mondrian predictors

Conformal predictors allow us to obtain a guaranteed error rate which does not exceed the significance level ϵ. However, we may encounter problems in medical diagnosis, when we know that certain patients are easier to correctly classify than others (for example, men are more easily diagnosed than women, or it is more likely to misclassify a healthy patient than a diseased one). In this case, conformal predictors will guarantee the overall error rate; they may result in higher actual error rate on harder groups of patients and lower on easier groups of patients. However, it would be good to guarantee the error rate within these groups.

In the current work, we have two classes: healthy and diseased patients. There are two types of errors in this case. It is not always clear in advance what type of error is more important: to misclassify a healthy patient or to misclassify a diseased one. If we keep both error rates on a guaranteed level, then the same guarantee will be true for any weighted mixture of them.

Mondrian predictors [1,14], which are the development of conformal predictors, allow us to tackle this problem. They split all possible patients into categories and set significance levels ϵ_k, one for each category k. Mondrian predictors can guarantee that in the long-run patients of each category k are misclassified with frequency at most ϵ_k.

One of the simplest examples could be a taxonomy conditioned on diagnoses, when each category corresponds to a certain diagnosis and comprises only patients with this diagnosis. Another possibility is division in categories based on features and their combinations, e.g., patients can be grouped by age. Finally, taxonomies can get even more complex: they can be based on combinations of features, diagnoses and even ordinal numbers of patients in the sequence.

In comparison with conformal predictors, the difference in constructing Mondrian predictors is that we compare the nonconformity score of (x_n, y) not with all patients in the sequence but only with patients of the same category:

$$p_n(y) = \frac{|\{i = 1, \ldots, n-1 : \kappa_i = \kappa_n \ \& \ \alpha_i \geq \alpha_n\}| + 1}{|\{i = 1, \ldots, n-1 : \kappa_i = \kappa_n\}| + 1}, \quad (3)$$

where $\kappa_i, i = 1, \ldots, n-1$ is the category of (x_i, y_i); κ_n is a category of (x_n, y).

Finally, any Mondrian predictor is conditionally valid: in the long run, the frequency of errors made by the machine (i.e., cases when prediction set does not contain a real diagnosis) on patients in category k does not exceed ϵ_k for each k.

Thus, Mondrian predictors allow us to solve two main problems.

– We can guarantee not only an overall accuracy but also a certain level of accuracy within each category of patients. In particular, we can preset the level of accuracy within groups of healthy and diseased samples, which is similar to specificity and sensitivity. This will allow avoiding classifications when small number of errors on healthy samples is compensated by high number of errors on diseased ones or the other way around. Therefore we use Mondrian predictors.
– If we preset different significance levels for categories, we can treat them in a different way, e.g., put analogue of sensitivity first and consider a misclassification of a diseased sample more serious than misclassification of a healthy sample.

3 Data

The methodology based on a Mondrian predictor was applied to the data sets from the UKCTOCS study, which was designed to provide data on the effect of ovarian-cancer screening on mortality. It is the world's largest ovarian-cancer screening research programme and involves sample collection of 200,000 women aged 50–74 years. In this research, we have analysed the available ovarian-cancer and breast-cancer data sets.

The data pertain to serum samples collected from patients diagnosed with the disease (we will call them cases) and healthy patients (they will be referred to as controls). Originally, each case was accompanied by two controls matched on patient age, sample collection location and sample collection date/time, among other factors. For this reason, in each data set, the number of controls is twice as large as the number of cases:

– 104 cases and 208 controls in the ovarian-cancer data set (312 samples in total);
– 54 cases and 108 controls in the breast-cancer data set (162 samples in total).

The samples were analysed by MS and its output by the use of a Mondrian predictor. The MS data of ovarian and breast cancers were provided by the University of Reading and University College London, respectively.

MS is an attractive analytical tool, because it enables researchers to simultaneously analyse hundreds of biomolecules. Matrix assisted laser desorption/ionisation-time of flight (MALDI-TOF) MS, one of several possible techniques, has revealed the complexity of the low-molecular weight proteomes of serum and plasma.

The MS data we have submitted to our methodology are represented as intensities at *m/z* (mass to charge ratio) values. Preprocessing steps, including peak identification, applied in this work can be found in a separate work [15]. The identified peaks are sorted by their frequency: the greater the number of mass spectra containing a peak, the higher is the rank of that peak. We consider a certain number of the most frequent peaks only. Throughout the article, peak numbers are used; the lower the peak number, the more common the peak is. Please note that sets of peaks vary for different data sets, therefore, peaks with the same number from various data sets have different *m/z* values.

Several biomarkers for ovarian cancer have been identified, but none so far have been adopted for screening. The most extensively assessed biomarker is CA125 that is typically elevated in the blood of some ovarian-cancer patients. However, the potential role of this protein for the early detection of ovarian cancer is unproven and still subject to clinical trials. One of the main problems related to the use of CA125 is its low predictive ability at early-stages of the disease. Another problem is that CA125 can be produced by other mesothelium-derived tissues [15], and therefore, may also be elevated in women with benign gynaecological conditions and other types of cancer (such as breast, bladder, pancreatic, liver, lung) [16]. Therefore, CA125 deployment lacks sensitivity: if the level of CA125 is elevated, an operation is needed to confirm the disease. Thus, it is thought that CA125 alone may not be accurate enough for detection of early-stage ovarian cancer.

In this study, we aim to verify whether it is possible to improve the ability of CA125 to discriminate between ovarian cancer and healthy patients in early stages of the disease. In addition, we attempt to identify certain mass spectral peaks which could, in combination with CA125, result in accurate ovarian-cancer diagnosis well in advance of the moment of clinical diagnosis.

Thus, each mass spectrum of the ovarian-cancer data set is also assigned a level of CA125, and we will make predictions of the diagnosis based not only on MALDI-TOF MS data but also on CA125 levels.

Finally, each sample is assigned a non-negative value $T(\tau)$—time to diagnosis confirmed by histology/cytology. Controls are assigned the same value $T(\tau)$ as the case they match. We will refer to this value as time to diagnosis and the moment of diagnosis confirmed by histology/cytology as the moment of diagnosis. Since we have the information regarding when each sample was taken, we can consider sets of samples taken in different time slots before the moment of diagnosis.

4 Algorithms

Practically, any known machine-learning algorithm can be plugged into a Mondrian predictor, and thus, result in a new algorithm of prediction with confidence. In our research, we used a set of linear discriminant functions. This section describes the application of linear rules within the framework of Mondrian predictors used for discrimination between MS samples taken from healthy and diseased patients.

Every patient description, x_i, comprises M features $x_i(q), q = 1, \ldots, M$. In the case of the breast-cancer data, these features are intensities of the M most frequent peaks $x_i(q) = I(q), q = 1, \ldots, M$. For the ovarian-cancer data, the features are the $(M-1)$ most frequent peaks and biomarker CA125 $x_i(M) = C_i$. Diagnoses, y_i, are equal to 0 for controls and 1 for cases.

When designing a new Mondrian predictor we will use simple linear rules of the following type (see Algorithm 1):

$$\sum_{k=1}^{m} v_k \log I(q_k) > \theta, \tag{4}$$

where m is a fixed (usually small) number of peaks in a linear combination, $I(q_k)$ is the intensity of peak q_k; $v_k \in \mathbb{R}$; $k = 1, \ldots, m$ are weights; $\theta \in \mathbb{R}$ is a threshold. A rule classifies a patient as diseased if it returns the value true, healthy otherwise.

Algorithm 1 Mondrian predictor based on linear rules

```
Require:
  $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$ — sequence of patients with diagnoses, $x_i = \{x_i(1), \ldots, x_i(M)\}$, $y_i \in \{0, 1\}$
  $x_n$ — patient without a diagnosis
  $x_i(j)$ — intensity of the peak $j$ for the patient $i$
  $m$ — number of peaks in a linear rule
  $V_k \subseteq \mathbb{R}, k = 1, \ldots, m$ — set of possible weights in linear rules
  $Q_k \subseteq \{1, \ldots, M\}, k = 1, \ldots, m$ — set of possible peak numbers in linear rules
  for all $y \in \{0, 1\}$ do
    $y_n := y$
    $z_n := (x_n, y)$
    for $i := 1, \ldots, n$ such that $y_i = y$ do
      for $v_1 \in V_1$ do
        for $q_1 \in Q_1$ do
          …
          for $v_m \in V_m$ do
            for $q_m \in Q_m$ do
              $\Theta = \{-\infty\} \cup \{\sum_{k=1}^{m} v_k \log x_j(q_k),\ j = 1, \ldots, m\}$
              for $\theta \in \Theta$ do
                Compute predictions $\hat{y}_j, j = 1, \ldots, n$, provided by a linear rule with parameters $(v_1, \ldots, v_m, \theta, q_1, \ldots, q_m)$:
                for j := 1, …, n do
                  if $\sum_{k=1}^{m} v_k \log x_j(q_k) > \theta$ then
                    $\hat{y}_j := 1$
                  else
                    $\hat{y}_j := 0$
                  end if
                end for
                $\mathrm{TPR}(v_1, \ldots, v_m, \theta, q_1, \ldots, q_m) := \frac{|j=1,\ldots,n\colon y_j=1 \ \&\ \hat{y}_j=y_j|}{|j=1,\ldots,n\colon y_j=1|}$
                $\mathrm{TNR}(v_1, \ldots, v_m, \theta, q_1, \ldots, q_m) := \frac{|j=1,\ldots,n\colon y_j=0 \ \&\ \hat{y}_j=y_j|}{|j=1,\ldots,n\colon y_j=0|}$
              end for
            end for
          end for
        end for
        …
      end for
      $\{\tilde{v}_1, \ldots, \tilde{v}_m, \tilde{\theta}, \tilde{q}_1, \ldots, \tilde{q}_m\} := \arg\max_{\substack{v_1 \in V_1, \ldots, v_m \in V_m, \\ q_1 \in Q_1, \ldots, q_m \in Q_m \\ \theta \in \mathbb{R}}} (\min(\mathrm{TPR}(v_1, \ldots, v_m, \theta, q_1, \ldots, q_m), \mathrm{TNR}(v_1, \ldots, v_m, \theta, q_1, \ldots, q_m)))$
      if $y = 0$ then
        $\alpha_i := \sum_{k=1}^{m} \tilde{v}_k \log x_n(\tilde{q}_k)$
      else
        $\alpha_i := -\sum_{k=1}^{m} \tilde{v}_k \log x_n(\tilde{q}_k)$
      end if
    end for
    $p_n(y) := \frac{|\{i=1,\ldots,n-1\colon y_i=y \ \&\ \alpha_i \geq \alpha_n\}|+1}{|\{i=1,\ldots,n-1\colon y_i=y\}|+1}$
  end for
  Compute a diagnosis for $x_n$: $y_{\mathrm{pred}} := \arg\max_{y \in \{0,1\}} p_n(y)$
  Compute its confidence as $1 - \min_{y \in \{0,1\}} p_n(y)$
  Compute its credibility as $\max_{y \in \{0,1\}} p_n(y)$
```

To design a nonconformity measure, we first need to define the taxonomy of a Mondrian predictor. We will consider the taxonomy $\kappa(n, (x_n, y_n)) = y_n$, i.e., the taxonomy which consists of two categories that correspond to two different diagnoses: the category of healthy patients and the category of diseased patients. Such taxonomy will allow us to guarantee the error rate within classes of healthy patients and diseased patients, which is analogous to controlling sensitivity and specificity. Hence, p values are calculated as in equation 3 with $k_i = y_i$, i.e., the p-value is calculated as the ratio of healthy (diseased) patients which conform with the the other patients at most as much as the new one to the total number of healthy (diseased) patients.

It also appears to be more natural to deploy a Mondrian predictor rather than conformal predictor with this taxonomy. This will be easily seen from the nonconformity measure.

The nonconformity measure is calculated as follows. We fix the number m of peaks used in a rule, so a rule can include any m of M most frequent peaks. We then consider a set of possible linear rules of type (4) where parameters of the rules can possess the following values: $\theta \in \mathbb{R}$, $v_k \in V_k \subseteq \mathbb{R}$, $q_k \in Q_k \subseteq \{1, \ldots, M\}$, $k = 1, \ldots, m$.

We compare quality of rules by maximum of sensitivity and specificity, in order not to improve one of them at the expense of another. In terms of a ROC curve, we approach a point where the sensitivity is equal to the specificity.

Out of these rules, we select the following one:

$$\{\tilde{v}_1, \ldots, \tilde{v}_m, \tilde{\theta}, \tilde{q}_1, \ldots, \tilde{q}_m\} = \arg \max_{\substack{v_1 \in V_1, .., v_m \in V_m, \\ q_1 \in Q_1, .., q_m \in Q_m \\ \theta \in \mathbb{R}}} (\min(\text{TPR}(v_1, \ldots, v_m, \theta, q_1, \ldots, q_m), \text{TNR}(v_1, \ldots, v_m, \theta, q_1, \ldots, q_m))),$$

where

$$TPR(v_1, \ldots, v_m, \theta, q_1, \ldots, q_m)$$

and

$$TNR(v_1, \ldots, v_m, \theta, q_1, \ldots, q_m)$$

are sensitivity (true positive rate) and specificity (true negative rate) of rule (4) with parameters

$$(v_1, \ldots, v_m, \theta, q_1, \ldots, q_m),$$

respectively, on the set of patients including a new patient with a new hypothetical diagnosis. If there is more than one set of parameters which provide maximum of the arg max expression, we choose the one with the smallest absolute values of parameters giving priorities in the following order: $v_1, \ldots, v_m, q_1, \ldots, q_m, \theta$.

We can then define the nonconformity score of a new patient with a candidate diagnosis on the basis of the chosen rule. The value of the chosen linear combination $\sum_{k=1}^{m} \tilde{v}_k \log I(\tilde{q}_k)$ is used as a nonconformity score for healthy patients or as a value negative to a nonconformity score for diseased patients. Thus, when calculating a p value, we compare the value of the chosen linear combination for the new patient with the value of the same combination for patients with the same diagnosis. If a new patient was healthy, the larger the value of the linear combination, the more nonconformal the patient is, and the other way around if a patient is diseased.

Note that a rule itself reflects a hypothesis about biomarkers relevant for diagnosis and their relative weight. Therefore, the algorithm includes a kind of embedded feature (biomarker) selection.

In our experiments, the significance level is the same for the classes of healthy and diseased patients. Leave-one-out cross-validation is performed: each patient (x_i, y_i) is considered as if it was a new test sample, and all the remaining patients in the data are treated as the training set.

5 Results

5.1 Early detection

It is shown [17] that for the analysed diseases there are certain time slots when MS profile peaks carry statistically significant information for discrimination between controls and cases, i.e., we can reject the null hypothesis that the diagnosis is independent of the information contained in peak intensities at significance level of 5 % well in advance of the moment of diagnosis.

To investigate how long in advance of the moment of diagnosis accurate predictions can be provided, we consider different time slots of fixed length (6 months for ovarian cancer and 12 months for breast cancer) shifting away from the moment of diagnosis. These time slots finish 1, 2, 3, …months in advance of the moment of diagnosis.

After fixing the time slot, we pick all the patients whose measurements were taken in this time slot together with matched controls. For the ovarian-cancer data, if several measurements of the same patients fall in this time slot, we consider only the one closest to the moment of the diagnosis, eliminating the others together with corresponding controls.

We then apply designed Mondrian predictors to patient measurements in time-slots moving away from the moment of the diagnosis. We expect prediction accuracy to deteriorate as the time slot is moving away since we assume that, further from the moment of diagnosis, mass spectra contain less information useful for discrimination between cases and controls.

5.2 Ovarian-cancer results

For the ovarian-cancer data, we consider the simplest possible combinations (4) of CA125 and one peak ($m = 2$);

q_1 corresponds to CA125 level ($Q_1 = \{M\}$), $Q_2 = \{1, \ldots, M\}$, $v_1 \in V_1 = \{0, 0.5, 1, 2\}$ is a CA125 weight, $v_2 \in V_2 = \{-1, 0, 1\}$ is a peak weight.

V_1 and V_2 were selected in the same way as analogous parameters in our experiments related to our previous works [17,15]. Our experience had shown that because of the small number of samples, any additional terms in the rules (4) are either useless or would bring overfitting.

At first, we will demonstrate how prediction with confidence works. For each patient, Mondrian predictor provides two p values, corresponding to 'healthy' and 'diseased' hypotheses. On the basis of these p values, we calculate confidence and credibility for each patient as described in Sect. 2.1. After assigning every patient with two p values, we predict the diagnosis with the highest p value.

Table 1 represents several examples of p values, confidence and credibility for ovarian-cancer measurements taken not earlier than 6 months in advance of the moment of diagnosis. If confidence is close enough to 1 and credibility is not close to 0, the prediction is considered to be reliable.

We will demonstrate this in detail in several examples from Table 1. The columns represent a measurement ID, true diagnosis, predicted diagnosis, p values for 'healthy' and 'diseased' diagnoses, confidence and credibility. For instance, patient with measurement ID 141100 in Table 1 has two p values, one of which is close to 1 (0.99) and another close to 0 (0.01). This results in high confidence of 0.99 and high credibility of 0.99 and identifies the prediction as reliable: only one diagnosis conforms well the rest of the set. If this patient was classified as a case (diagnosis value of 1), this would mean that an event of probability ≤ 1 % occurred. For this reason, we expect the patient to be healthy, which is correct. In contrast, patient with measurement ID 146384 has low p values close to each other (0.12 and 0.13), which means neither of the diagnoses is likely to be correct and, hence, there is not enough information to confidently classify the patient. Thus, these p values do not produce confidence close enough to 1 (0.88) or high credibility (0.13). As a result, the output prediction for the patient with measurement ID 146384 is indeed incorrect.

Table 2 shows the accuracy of Mondrian predictors in different time slots. The table demonstrates that Mondrian predictors are reasonably accurate well in advance of the moment of diagnosis. For example, the accuracy in the time slot of 10–16 months (the latest time slot when CA125 on its own does not carry statistically significant information for disease discrimination [15]) is 70.2 %. This is quite good given that diagnosis is made not later than 10 months in advance before the diagnosis is confirmed by histology/cytology. For comparison, when we make predictions with the same method of measurements just before the moment of diagnosis (in a 0–6 time slot), the accuracy is equal to 92.2 %.

When we combine results for different time slots, we can estimate how Mondrian predictors perform in early ovarian-cancer diagnosis. In general, Mondrian predictors produce predictions with accuracy higher than 66 % up to 11 months in advance of the moment of diagnosis. As we move away from the moment of diagnosis, accuracy of predictions decreases. Low accuracy 6, 7 and 8 months in advance may be explained by a small number of samples in this period (below 70 samples for any time slot).

To estimate statistical significance of achieved accuracy, we calculated p values that reject the null hypothesis that the assignment of labels is independent of MS peak intensities and CA125 levels. The p values we calculated by the use of the Monte-Carlo method: we estimate how possible it is to make the prediction of same quality by chance. Suppose there is no real dependence between true diagnosis and peak intensities. Such a situation can be simulated by reshuffle of the labels without changing the feature information. Monte-Carlo method answers the questions: what will be the accuracy in this case? In what percentage of cases it will be as good as the current one or even better?

For each time slot, we consider Mondrian predictor's accuracy. For a large number $N = 500$ of times, we calculate the statistics, the accuracy of the Mondrian predictor applied to the data in the same time slot but with randomly permuted labels. Accuracy here is just amount of true diagnoses. We count a number of times n when the statistics is at least as high as the accuracy calculated on true labels. The p value is then defined as $(n + 1)/(N + 1)$.

These p values are presented in the last column of Table 2, which shows that the accuracy achieved in the time slots finishing 0–6 and 9–11 months in advance is significant at the level of 5 %. Analogous p values were calculated for linear combinations of CA125 and MS peaks without the framework of Mondrian predictors. We obtained values similar to the ones calculated for Mondrian predictors. In particular, p

Table 1 Examples of the output of Mondrian predictors applied to the ovarian-cancer data in a 0–6 month time slot: true and predicted diagnoses, p values for both diagnoses, confidence and credibility for several patients

Measurement ID	True diagnosis	Predicted diagnosis	p value for 0	p value for 1	Confidence	Credibility
141100	0	0	0.99	0.01	0.99	0.99
146384	0	1	0.12	0.13	0.88	0.13
232604	1	0	0.51	0.28	0.72	0.51
245401	1	1	0.01	0.97	0.99	0.97

Table 2 Accuracy of Mondrian predictors applied to the ovarian-cancer data set (CA125 and 5 most frequent peaks) in the leave-one-out mode in different time slots

Timeslot	Samples	Accuracy (%)	Sensitivity (%)	Specificity (%)	p value
0–6	204	92.2	91.2	92.7	0.002
1–7	168	89.9	89.3	90.2	0.002
2–8	141	83.7	83.0	84.0	0.002
3–9	108	78.7	80.6	77.8	0.002
4–10	81	79.0	74.1	81.5	0.002
5–11	69	73.9	73.9	73.9	0.002
6–12	60	66.7	65.0	67.5	0.050
7–13	51	68.6	64.7	70.6	0.060
8–14	51	66.7	70.6	64.7	0.102
9–15	60	73.3	75.0	72.5	0.020
10–16	84	70.2	71.4	69.6	0.004
11–17	84	66.7	67.9	66.1	0.050

values were below 5 % in the same time slots, which demonstrates that CA125 and MS peaks carry information which allows statistically significant discrimination between ovarian-cancer patients and controls.

As mentioned before, the feature of the ovarian-cancer data set is that ovarian-cancer cases can have several measurements taken at different moments. For this reason, we can observe the change in the output of Mondrian predictors for this data set. As an illustration, we will consider several ovarian-cancer cases that have measurements taken over a long period of time and will show how confidence and credibility are changing when the patient is approaching the moment of diagnosis.

We select patients with at least three measurements. For each measurement, we train the Mondrian predictor on the samples in the earliest 6-month time slot containing the measurement leaving out the measurement itself. For example, if a measurement was taken 6.5 months in advance, we consider the time slot from month 12 to month 6. We then apply the Mondrian predictor to the left-out measurement and output a prediction, its confidence and credibility. Dynamics of confidence and credibility for measurements of several patients is shown in Table 3.

Table 3 Dynamics of confidence and credibility for measurements taken for two ovarian-cancer cases

Case ID	Months	Prediction	Confidence	Credibility
ID	in advance		(%)	(%)
39	10	1	89.5	67.9
	4	1	90.9	44.4
	2	1	99.0	66.0
	1	1	99.1	76.8
42	24	1	69.0	71.4
	15	0	45.0	78.1
	3	1	98.6	100.0

We can trust the prediction if its confidence is close to 1 (i.e., all p values for alternative diagnoses are close to 0) and its credibility is not close to 0 (i.e., the maximum p value is not close to 0). This implies that if a Mondrian predictor makes correct predictions about the case, we expect confidence to be approaching 100 % when measurements are getting closer to the moment of diagnosis. Meanwhile, credibility is expected not to be getting close to 0 %. Table 3 demonstrates that patient 39 confirms our expectations.

Patient 42 represents a more interesting example: we make an erroneous prediction 15 months in advance. However, its confidence is not close to 100 %, which reflects that we cannot be sure in this prediction. When we make a final prediction for this patient 3 months in advance, both confidence and credibility are close to 100 %.

Overall statistic of conformal predictor output in terms of prediction sets is presented in Table 7. It shows that the number of certain non-empty predictions (for which prediction set consists of exactly one label) increases for ovarian cancer as time becomes closer to diagnosis.

5.3 Breast-cancer results

The same approach was applied to the set of breast-cancer patients and matched controls, which was taken from the UKCTOCS trial. We consider cut-off rules (4) with one peak involved ($m = 1$) with $Q_1 = \{1, \ldots, M\}$ and $v_1 \in V_1 = \{-1, 1\}$, a weight that determines whether the peak has higher or lower intensities for cases.

Firstly, this approach allows us to complement each diagnosis of prediction with measures of confidence and credibility. This is demonstrated in Table 4, which contains p values, confidence and credibility for some breast cancer and healthy patients whose measurements were taken no earlier than 12 months in advance of the moment of diagnosis.

Table 4 The output of Mondrian predictors applied to the breast-cancer data in the time slot of 0–12 months in advance: true and predicted diagnoses, p values for both diagnoses, confidence and credibility form some patients

Measurement ID	True diagnosis	Predicted diagnosis	p value for 0	p value for 1	Confidence	Credibility
1832	0	1	0.29	0.30	0.71	0.30
77217	0	0	1.00	0.05	0.95	1.00
195604	1	1	0.08	0.95	0.92	0.95

Secondly, Mondrian predictors result in accurate predictions well in advance of the moment of diagnosis (detailed results are presented in Table 5). Mondrian predictors achieve an accuracy higher than 70 % up to 9 months in advance of the moment of breast-cancer diagnosis. However, there is no apparent decreasing trend in accuracy; it fluctuates in the range of 70.4–77.8 %. It falls to 71.9 % in the latest time slot (0–12 months), because the number of examples available for the experiments falls down to 57. This is also the reason why the certainty rate Table 7 decreases as the time of diagnosis is approaches. At slot (10–22 months) it falls to 48.2%, this is because selection if best rule becomes unstable: according to Table 6, top peak 19 is selected only in 87 %, so selection is not robust even to change of one example. Monte-Carlo p values shown in the last column of Table 5 demonstrate statistical significance of achieved accuracy up to 8 months in advance of the moment of diagnosis.

However, we see that BC data is, in general, less-informative than OC possibly, because OC has a strong biomarker CA125 in addition to MS data. This reflects both in lower accuracy (Table 5 compared to Table 2) and lower certainty rate (Table 7).

5.4 Informative peaks and comparison to related research

In parallel, another approach was applied to the UKCTOCS data sets in another work [15], which is written from the medical point of view and utilises the already developed methods of early diagnosis and peak identification. The research is devoted to statistical analysis, whereas this work describes machine-learning approach that complements each prediction with its confidence. In addition, statistical analysis was carried out in a different experimental setting: the data were normalized against such factors as age, sample collection time and location, storage and transportation conditions. All measurements were grouped in triplets comprising one diseases patient and two healthy controls matched by these factors. Thus, when making predictions, we had additional information about diagnosis distribution: we knew that exactly one patient was diseased in a triplet. We will refer to this research and the corresponding work as triplet analysis. Triplet analysis pinpointed MS profile peaks that allowed statistically significant (containing essential information in addition to CA125) discrimination at the 5 % level between cases and controls long in advance of the moment of diagnosis of ovarian cancer. We demonstrated that mass spectra from the low molecular weight serum proteome carry information useful for early detection.

The triplet analysis [15] of the ovarian and breast cancers allowed us to determine statistically significant peaks which could be potential candidates for biomarkers. We identified certain MS profile peaks that carry statistically significant information for the diagnosis of the diseases. In the current research, we do not analyse statistical significance of

Table 5 Accuracy of Mondrian predictors applied to the breast-cancer data set (20 most frequent peaks) in the leave-one-out mode in different time slots

Time slot	Number of samples	Accuracy (%)	Sensitivity (%)	Specificity (%)	p value
0–12	57	71.9	73.7	71.1	0.028
1–13	72	77.8	79.2	77.1	0.002
2–14	78	76.9	76.9	76.9	0.002
3–15	78	76.9	76.9	76.9	0.002
4–16	72	77.8	79.2	77.1	0.002
5–17	72	75.0	75.0	75.0	0.002
6–18	60	73.3	75.0	72.5	0.012
7–19	57	71.9	73.7	71.1	0.040
8–20	51	70.6	70.6	70.6	0.026
9–21	54	70.4	72.2	69.4	0.066
10–22	54	48.2	55.6	44.4	0.535
11–23	54	70.4	72.2	69.4	0.058

Table 6 Top peaks pinpointed by Mondrian predictors in different time slots for the ovarian and breast-cancer data sets

Month	Ovarian cancer		Breast cancer	
	Top peak	Peak frequency(%)	Top peak	Peak frequency(%)
0	1	96.1	19	100.0
1	1	83.0	19	100.0
2	1	72.7	19	100.0
3	2	56.0	19	100.0
4	2	98.2	19	100.0
5	1	95.7	19	100.0
6	1	69.2	19	100.0
7	4	94.1	19	100.0
8	3	73.5	19	100.0
9	3	100.0	19	100.0
10	3	100.0	19	87.0
11	3	100.0	19	100.0
12	2	85.7	19	100.0
13	3	95.0	6	78.4
14	3	85.3	15	100.0
15	2	89.2	14	67.5
16	5	63.3	14	67.5

Table 7 The percentage of non-empty certain predictions output by category-based confidence machines at significance levels $\epsilon = 5, 10, 20$ % in different time slots for the ovarian-cancer and breast-cancer data sets

Time slot	Ovarian cancer			Breast cancer		
	$\epsilon = 5$ (%)	$\epsilon = 10$ (%)	$\epsilon = 20$ (%)	$\epsilon = 5$ (%)	$\epsilon = 10$ (%)	$\epsilon = 20$ (%)
0–6	91.7	94.1	81.4	5.3	15.8	50.9
1–7	78.6	94.6	84.5	8.3	23.6	62.5
2–8	58.2	80.1	90.1	7.7	24.4	76.9
3–9	25.9	48.2	88.9	7.7	24.4	76.9
4–10	27.2	56.8	91.4	8.3	19.4	93.1
5–11	21.7	44.9	71.0	8.3	29.2	88.9
6–12	18.3	41.7	58.3	10.0	36.7	90.0
7–13	9.8	21.6	62.8	5.3	14.0	59.7
8–14	2.0	29.4	60.8	3.9	17.7	62.8
9–15	18.3	33.3	73.3	1.9	14.8	64.8
10–16	11.9	29.8	67.9	1.9	14.8	35.2
11–17	9.5	22.6	58.3	1.9	16.7	50.0

particular peaks. However, Mondrian predictors indirectly pinpointed informative peaks. Despite the different nature of these methods, observed mostly the same peaks as the ones that carry statistically significant information for discrimination between controls and cases according to the triplet analysis.

We will consider the time slots when Mondrian predictors produced high accuracy on the data sets. In addition, for ovarian cancer, we are especially interested in time slots starting from month 10, because this is the first time-slot when CA125, on its own, does not provide statistically significant discrimination between cases and controls.

Mondrian predictors help us identify informative peaks in the following way. When we run leave-one-out procedure, for each possible diagnosis we choose the best rule $w \log(C) + v \log I(p) > \theta$ (for ovarian cancer) or $v \log I(p) > \theta$ (for breast cancer), which contains a peak. The selected peak may not be the same for every possible diagnosis and every possible left out patient, but in the time slots we are examining, the same peak was selected as a part of the best rule, that is, we choose the same weights and peak number when leaving out a patient: these are peak 19 for breast cancer in time slots finishing with months 0–9, 11, 12 and peak 3 for ovarian cancer in time slots finishing with months 9–11. The detailed results for ovarian and breast-cancer measurements taken in different time slots are represented in Table 6. The table shows the peak which was selected most often ('top peak') and how often it was selected ('peak frequency').

Table 8 summarises all peaks selected by two different approaches: triplet analysis and Mondrian predictors. Those peaks are shown that were selected in time slots of high interest: slots finishing with months 0–9 for breast cancer and 10–11 for ovarian cancer. Table 8 demonstrates that Mondrian predictors pinpoint the same peaks as identified as carrying statistically significant information in the triplet setting.

For the ovarian-cancer data, both methods select peak 3 in time slots finishing with month 10 or 11. These are the time slots when CA125 on its own does not carry statistically significant information as shown in the triplet analysis [15]. Ovarian-cancer peak 3 was also observed in research on other data sets [18]. In addition, peak 3 coincides with peak 7 previously found in the analysis of similar serial ovarian-cancer samples and controls in the pilot [17,19] trial which preceded UKCTOCS. Peak 3 is identified as CTAP III [20] and peak 2 is is potentially platelet factor 4 (PF4).

The predictive ability of CA125 on its own and in combination with peak 3 is demonstrated in Fig. 2. The figure illustrates that the combination of CA125 with peak 3 starts growing earlier than $\log C$; CA125 growth at the moments close to diagnosis is quicker due to the exponential growth of CA125. Graphs with similar behaviour for a combination of CA125 with another peak were presented in the pilot [17] trial.

For the breast-cancer data, we observe the dynamics of selected peak 19, whose intensities are supposed to be lower for cases rather than for controls according to our research. In Fig. 3, the solid line represents the median dynamics of peak 19 for breast-cancer cases, the dashed line shows the peak 19 median calculated for all breast-cancer controls. The values in the figure were calculated for measurements within a 9-month window ending with the month shown on the horizontal axis. One can see from Fig. 3 that peak 19 median intensity drops about 15 months in advance of the moment of the diagnosis, which confirms our hypothesis about predictive ability of peak 19 and explains the results we obtained using this peak when discriminating between breast-cancer cases and controls. Peak 19 is preliminarily identified as either ApoCI or ApoCII (or their combination).

Table 8 Numbers of the most important peaks selected with different methods for the ovarian and breast-cancer data sets

Method	Ovarian cancer	Breast cancer
Triplet analysis	2, 3	19
Mondrian predictors	3	19

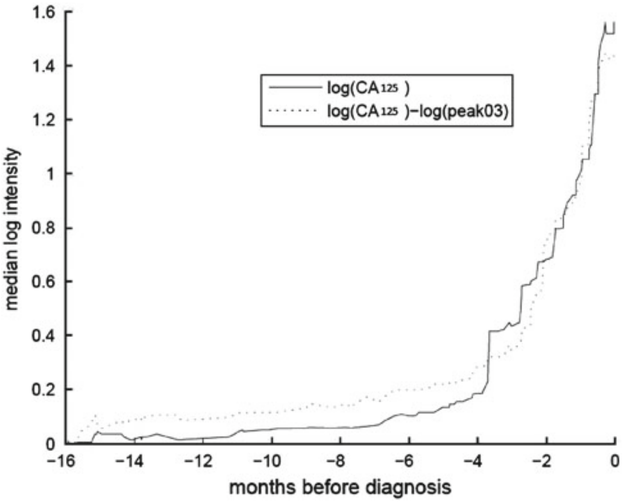

Fig. 2 Median dynamics of intensity for rules $\log \text{CA125}$ and $\log \text{CA125} - \log(\text{Peak } 3)$ (for ovarian-cancer cases only)

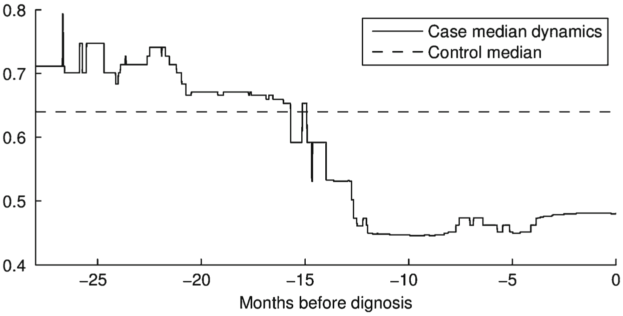

Fig. 3 Median dynamics of intensity for peak 19 in the breast-cancer data for cases and the median of peak 19 for controls

6 Discussion

This work introduced the methodology of providing predictions with confidence for MS-based proteomics. First, the framework of Mondrian predictors allowed us to complement each prediction with certain information reflecting our confidence in each prediction. Second, application of Mondrian predictors to the ovarian and breast-cancer experimental data demonstrated that Mondrian predictors result in high accuracy well in advance of the moment of the disease diagnosis. The accuracy of the proposed methods on the ovarian-cancer data rises from 66.7 % at 11 months in advance of the moment of diagnosis to up to 92.2 % just before the moment of diagnosis. When applied to the breast-cancer data, the methods allowed us to achieve accuracy of 70.4–77.8 % for up to 9 months in advance of diagnosis.

We constructed a special NCM to take into account data specific nature (mass spectrometry) and additional aim (biomarker identification). However, it might be very straightforward in the part of search: we used overall scanning within the set of possible rules. It was done to make general idea more clear, but it may be one of future tasks to replace it with a more practical way of search. In addition, we assumed that only few biomarkers might be informative for the prediction. Alternative to this is a possible influence of the disease on many peaks in larger or smaller degree; search methods as SVM are more relevant here as there are less restrictions on the rule set. But in this case, a clear interpretation of results (list of found biomarkers) becomes a more complex task.

Acknowledgments We are very grateful to the many volunteers throughout the UK who participated in the trial. We would also like to thank Dr. Andy Ryan for his help with identifying the samples for this nested case control study and Mr. Jeremy Ford for his help with retrieving and aliquoting samples. The current study was funded by the MRC. UKCTOCS was core funded by the Medical Research Council, Cancer Research UK and the Department of Health with additional support from the Special Trustees of UCLH, Eve Appeal and Special Trustees of Bart's. The work at UCL was done at UCLH/UCL within the "women's health theme" of the NIHR UCLH/UCL Comprehensive Biomedical Research Centre supported by the Department of Health. This work was also supported by EraSysBio+ grant funds from the European Union, BBSRC and BMBF "Living with uninvited guests: comparing plant and animal responses to endocytic invasions" to the Salmonella Host Interactions Project European Consortium, SHIPREC; MRC Grant G0802594 (Application of conformal predictors to functional magnetic resonance imaging research); MRC Grant G0301107 (Proteomic analysis of the human serum proteome); Veterinary Laboratories Agency (VLA) of Department for Environment, Food and Rural Affairs (Defra) on Machine learning algorithms for analysis of large veterinary datasets; EU FP7 grant O-PTM-Biomarkers (2008–2011); and by Grant 'Development of new Venn prediction methods for osteoporosis risk assessment' from the Cyprus Research Promotion Foundation.

Appendix: logistic regression

Here, we illustrate why methods outputting probabilities (logistic regression) are less-applicable than conformal predictors. We applied a well-known method of logistic regression in leave-one-out to some parts of OC and BC data sets used in this work. The data sample are sorted by probabilites predicted by the logistic regression method and then cumulative prediction (dashed lines) are compared to cumulative true values (solid lines) are presented in the Fig. 4. Gaps between dashed and solid lines shows that probabilities are not close to real one in average. Hence, the logistic regression model is less-suitable for the data than the i.i.d. assumption made for the conformal predictor.

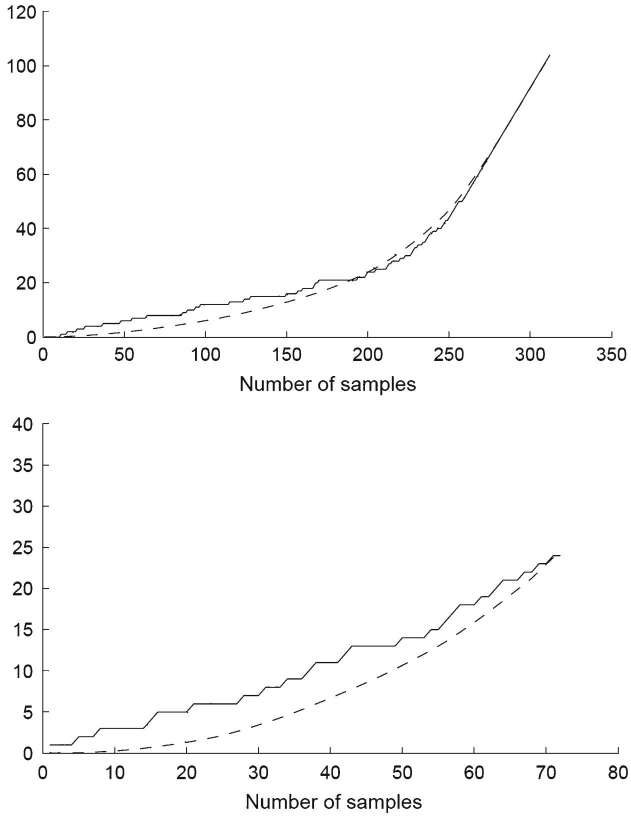

Fig. 4 Cumulative logistic regression predictions for the OC data (all samples); for the BC data (samples in the 5–17 month time slot)

Conflict of interest IJ has consultancy arrangements with Becton Dickinson, who have an interest in tumour markers and ovarian cancer. They have provided consulting fees, funds for research, and staff but not directly related to this study. No other financial disclosures.

Ethical approval UKCTOCS is registered as an International Standard Randomized Controlled Trial, number ISRCTN22488978. The study was approved by the Joint UCL/UCLH Committees on the Ethics of Human Research (MREC reference 05/Q0505/57). All trial participants gave informed written consent for secondary studies.

References

1. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005)
2. 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, M.W., Scott, I., Mould, T., Woolas, R., Murdoch, J., Dobbs, S., Amso, N.N., Leeson, S., Cruickshank, D., McGuire, A., Campbell, S., Fallowfield, L., Singh, N., Dawnay, A., Skates, S.J., Parmar, M., 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)
3. Timms, J.F., 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., Jacobs, I.: Peptides generated ex vivo from abundant serum proteins by tumour-specific

txopeptidases are not useful biomarkers in ovarian cancer. Clin. Chem. **56**, 262–271 (2010)

4. Tiss, A., Timms, J.F., Smith, C., Devetyarov, D., Gentry-Maharaj, A., Camuzeaux, S., Burford, B., Nouretdinov, I., Ford, J., Luo, Z., Jacobs, I., Menon, U., Gammerman, A., Cramer, R.: Highly accurate detection of ovarian cancer using CA125 but limited improvement with serum MALDI-TOF MS profiling. Int. J. Gynecol. Cancer **20**, 1518–1524 (2010)
5. Nouretdinov, I., Vovk, V., Vyugin, M., Gammerman, A.: Pattern recognition and density estimation under the general i.i.d. assumption. Lect. Notes Artif. Intell. **2111**, 337–353 (2001)
6. Nouretdinov, I., Burford, B., Gammerman, A.: Application of inductive confidence machine to ICMLA competition data. In: Proceedings of The Eighth International Conference on Machine Learning and Applications, pp. 435–438 (2009)
7. Nouretdinov, I., Li, G., Gammerman, A., Luo, Z.: Application of conformal predictors to tea classification based on electronic nose. In: Proceedings of Artificial Intelligence Applications and Innovations, pp. 303–310 (2010)
8. Gammerman, A., Nouretdinov, I., Burford, B., Chervonenkis, A., Vovk, V., Luo, Z.: Clinical mass spectrometry proteomic diagnosis by conformal predictors. Stat. Appl. Genetics Mol. Biol. **7**(2-13)(2008). Available at: http://www.bepress.com/sagmb/vol7/iss2/art13
9. Bellotti, A., Luo, Z., Gammerman, A., Van Delft, F.W., Saha, V.: Qualified predictions for microarray and proteomics pattern diagnostics with confidence machines. Int. J. Neural Syst. **15**(4), 247–258 (2005)
10. Lambrou, A., Papadopoulos, H., Gammerman, A.: Reliable confidence Measures for medical diagnosis with evolutionary algorithms. IEEE Trans. Inf. Technol. Biomed. **15**(1), 93–99 (2011)
11. Papadopoulos, H., Gammerman, A., Vovk, V.: Reliable diagnosis of acute abdominal pain with conformal prediction. Eng. Intell. Syst. **17**(2–3), 127–137 (2009)
12. Lambrou, A., Papadopoulos, H., Kyriacou, E., Pattichis, C.S., Pattichis, M.S., Gammerman, A., Nicolaides, A.: Assessment of stroke risk based on morphological ultrasound image analysis with conformal prediction. In: Proceedings of the 6th IFIP International Conference on Artificial Intelligence Applications and Innovations. IFIP AICT **339**, 146–153 (2010)
13. Nouretdinov, I., Costafreda, S.G., Gammerman, A., Chervonenkis, A., Vovk, V., Vapnik, V., Fu, C.H.Y.: Machine learning classification with confidence: Application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression. Neuroimage **56**(2), 809–813 (2011)
14. Vovk, V., Lindsay, D., Nouretdinov, I., Gammerman, A.: Mondrian confidence Machine (On-line Compression Modelling Project, working paper 4): Technical Report. Computer Learning Research Centre, Royal Holloway, University of London, UK (2003) http://www.vovk.net/cp/04.jpg
15. Timms, J.F., Menon, U., Devetyarov, D., Tiss, A., Camuzeaux, S., McCurry, K., Nouretdinov, I., Burford, B., Smith, C., Gentry-Maharaj, A., Hallett, R., Ford, J., Luo, Z., Vovk, V., Gammerman, A., Cramer, R., Jacobs, I.: Early detection of ovarian cancer in prediagnosis samples using CA125 and MALDI MS peaks. Cancer Genomics Proteomics **8**(6), 289–305 (2011)
16. Brioschi, P.A., Irion, O., Bischof, P., Bader, M., Forni, M., Krauer, F.: Serum CA 125 in epithelial ovarian: A longitudinal study cancer. Br. J. Obstet. Gynaecol. **94**, 196–201 (1987)
17. 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., Jacobs, I.: Serum proteomic abnormality predating screen detection of ovarian cancer. Comput. J. **52**, 326–333 (2008)
18. Nouretdinov, I., Burford, B., Luo, Z., Gammerman, A.: Data Analysis of 7 Biomarkers: Technical Report. Computer Learning Research Centre, Royal Holloway, University of London, UK (2008) http://www.clrc.rhul.ac.uk/projects/proteomics_reports.htm
19. Menon, U., Skates, S.J., Lewis, S., Rosenthal, A.N., Rufford, B., Sibley, K., Macdonald, N., Dawnay, A., Jeyarajah, A., Bast, R.C. Jr., Oram, D., Jacobs, I.J.: Prospective study using the risk of ovarian cancer algorithm to screen for ovarian cancer. J. Clin. Oncol. **23**, 7919–7926 (2005)
20. Tiss, A., Smith, C., Menon, U., Jacobs, I., Timms, J.F., Cramer, R.: A well-characterised peak identification list of MALDI MS profile peaks for human blood serum. Proteomics **10**, 3388–3392 (2010)