Bayesian and deep-learning models applied to the early detection of ovarian cancer using multiple longitudinal biomarkers

Luis Abrego¹,² | Alexey Zaikin¹,² | Ines P. Marino³ | Mikhail I. Krivonosov⁴,⁵ | Ian Jacobs¹ | Usha Menon⁶ | Aleksandra Gentry-Maharaj¹,⁶ | Oleg Blyuss¹,⁷,⁸

¹Department of Women’s Cancer, EGA Institute for Women’s Health, University College London, London, UK
²Department of Mathematics, University College London, London, UK
³Department of Biology and Geology, Physics and Inorganic Chemistry, Universidad Rey Juan Carlos, Madrid, Spain
⁴Research Center for Trusted Artificial Intelligence, Ivanov Institute for System Programming of the Russian Academy of Sciences, Moscow, Russia
⁵Institute of Biogerontology, Lobachevsky State University, Nizhny Novgorod, Russia
⁶MRC Clinical Trials Unit, University College London, London, UK
⁷Wolfson Institute of Population Health, Queen Mary University of London, London, UK
⁸Department of Pediatrics and Pediatric Infectious Diseases, Institute of Child’s Health, Sechenov First Moscow State Medical University (Sechenov University), Moscow, Russia

Abstract

Background: Ovarian cancer is the most lethal of all gynecological cancers. Cancer Antigen 125 (CA125) is the best-performing ovarian cancer biomarker which however is still not effective as a screening test in the general population. Recent literature reports additional biomarkers with the potential to improve on CA125 for early detection when using longitudinal multimarker models.

Methods: Our data comprised 180 controls and 44 cases with serum samples sourced from the multimodal arm of UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). Our models were based on Bayesian change-point detection and recurrent neural networks.

Results: We obtained a significantly higher performance for CA125–HE4 model using both methodologies (AUC 0.971, sensitivity 96.7% and AUC 0.987, sensitivity 96.7%) with respect to CA125 (AUC 0.949, sensitivity 90.8% and AUC 0.953, sensitivity 92.1%) for Bayesian change-point model (BCP) and recurrent neural networks (RNN) approaches, respectively. One year before diagnosis, the CA125–HE4 model also ranked as the best, whereas at 2 years before diagnosis no multimarker model outperformed CA125.

Conclusions: Our study identified and tested different combination of biomarkers using longitudinal multivariable models that outperformed CA125 alone. We
Ovarian cancer is the most lethal of all gynecological cancers. When detected at early stage, the survival is much more encouraging (5-year survival of >93% for Stage I disease) than when diagnosed at an advanced stage (5-year survival of 13% for Stage IV).\(^1\) Despite the extensive efforts to improve treatment over the last 20 years, although there have been modest improvements in survival, these have not had a significant impact. The major efforts in detecting ovarian and tubal cancer have spanned decades. Screening trials using Cancer Antigen 125 (CA125) interpreted using a cutoff have not shown any mortality benefit.\(^2\)–\(^4\) Algorithm-based approaches to screening in the UK trial have demonstrated that longitudinal CA125 can lead to earlier detection (stage shift with multimodal screening) with no impact on mortality from the disease.\(^5\) More recent data however suggest that there may be longer survival in women diagnosed with the most lethal subtype of ovarian cancer, the high-grade serous cancers (HGSC) in the screened (multimodal group) compared to the control group. Since its identification in 1981, CA125 has been used in clinical practice and investigated in screening trials. Clinical decisions have been made based on the patients’ risk of having a change point in the serial CA125 with respect to their baseline. A statistical method to determine in a probabilistic way such a risk was developed (Risk of Ovarian Cancer Algorithm, ROCA).\(^6\)–\(^7\) It has not been until more recently (despite many international efforts) that HE4 had been identified as the second-best marker in screening after CA125 (73% vs. 86% for CA125).\(^8\)–\(^12\) The efforts have since focused on exploring the value of a combination of CA125, HE4, and other promising markers in combination. Furthermore, p53 autoantibodies have been shown to detect ovarian/tubal cancers in women with ovarian/tubal cancers which do not express CA125 (16% and many months prior to diagnosis, lead time of 22 months).\(^13\) The interpretation of multiple markers in longitudinal samples is challenging unless sophisticated mathematical modeling is applied. We have previously shown that a method of mean trends (MMT) algorithm has a comparable performance to the ROCA which was used in the UK trial (UK Collaborative Trial of Ovarian Cancer, UKCTOCS).\(^14\) Here, we extend the observation made in\(^14\)–\(^16\) and describe a novel approach to interpretation of multiple markers in samples preceding diagnosis to assess if any of these can improve on sensitivity of the ROCA or offer potential advantage on lead time to detection of ovarian/tubal cancer.

2 | METHODS

2.1 | Change-point detection algorithm

2.1.1 | Joint multivariable fully Bayesian model

Biomarker levels \(Y_{ijk}\) are modeled using a hierarchical Bayesian model (Figure S1). Here, subject-specific variables are indexed by \(i = 1, 2, \ldots, n_0, n_0 + 1, \ldots, N\), where \(n_0\) is the number of controls and the remaining subjects account as cases. A specific biomarker is indexed by \(k = 1, 2, \ldots, K\). Each patient \(i\) has a set of screening visits \(t_{ij}\) from zero up to time of last measurement \(d_i\) (in years), where \(j = 1, 2, \ldots, T_i\).

Following the assumptions introduced in\(^7\) to model the cancer progression based on fully Bayesian screening, the longitudinal observations of biomarkers vary based on the nature of the patient. For control patients, the biomarker levels are expected to randomly fluctuate around a constant mean \(\theta_{ik}\). That is expressed as \(Y_{ijk} = \theta_{ik} + \epsilon_{ijk}\), with \(\epsilon_{ijk} \sim N(0, \sigma_{ik}^2)\). For cases, we define a binary indicator \(I_{ik}\) to distinguish between two different model assumptions in the evolution of biomarkers. If \(I_{ik} = 0\), then we assume that the marker level does not increase after the onset of cancer and follows the same behavior modeled for controls. If \(I_{ik} = 1\), the marker levels vary around a mean \(\theta_{ik}\) until an unobserved change-point time, defined by \(\tau_{ik}\). From this change point, we expect a positive slope \(\gamma_{ik}\) of the biomarker levels up to diagnosis. That is \(Y_{ijk} = \theta_{ik} + \gamma_{ik}(t_{ij} - \tau_{ik})^+ + \epsilon_{ijk}\) where \((\cdot)^+\) is the positive part of the expression.
Let \( S = \{\theta_{ik}, I_i, \log(\gamma_{ik}), \tau_{ik}\} \) be the set of subject-specific parameters. Thus, the probability density function of the observations conditional on the set of parameters \( S \) is

\[
P(Y | S ; t) = \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{j=1}^{T_i} \phi \left( \frac{Y_{ijk} - \theta_{ik}}{\sigma_k} \right) \times \prod_{i=1}^{N_{0}} \prod_{k=1}^{K} \prod_{j=1}^{T_i} \phi \left( \frac{Y_{ijk} - \theta_{ik}}{\sigma_k} \right) \]  

where \( Y \) denotes the set of values \( Y_{ijk}, t \) the set of screening times \( t_0 \) for each patient, and \( \phi \) is the standard normal probability density function.

A key feature of the Bayesian hierarchical model that we adopt in this paper is the statistical dependence among the levels of different biomarkers, following the methodology proposed by.\(^{17}\) This element is now compared to earlier work in.\(^{7,15,16}\) Dependence is explicitly introduced for the binary indicators \( I_i = \{I_{ik}\}_{k=1,\ldots,K} \), which are assumed to form a Markov random field (MRF). Their joint probability mass function (pmf) has the form

\[
P(I_i) \propto \exp \left\{ \mu_I \sum_{k=1}^{K} I_{ik} + \eta_I (I_i^T R I_i) \right\}\]  

where \( R \) is an upper triangular matrix weighted by a coupling coefficient \( \eta_I \). The parameter \( \mu_I \) controls the sparsity of the model, given that not all biomarkers may increase during the onset of disease. From Expression (2), it follows that the probability of a change point in the level of the biomarker \( k \) for patient \( i \) given all the other markers is

\[
P\{I_{ik} \mid I_{ik}^k \neq k'\} = \frac{\exp \{I_{ik} F(I_{ik})\}}{1 + \exp \{F(I_{ik})\} \} \]  

where \( F(I_{ik}) = \mu_I + \eta_I \sum_{k \neq k} I_{ik} \). According to Equation (3), a change point in one biomarker, for example \( I_{ik} = 1 \), implies an increase of the probability of having a change point in the remaining biomarkers whenever \( \eta_I > 0 \). If we set \( \eta_I = 0 \), then the biomarkers are independent (decoupled), and the probability of a change point for each single biomarker is reduced to a Bernoulli distribution with a mean parameter of \( 1/ (1 + \exp(−\mu_I)) \).

The mean level for the \( i \)-th patient and \( k \)-th biomarker is assumed Gaussian, \( \theta_{ik} \sim N\left(\mu_{0k}, \sigma_{0k}^2\right) \) where \( \mu_{0k} \) is itself Gaussian,

\[
\mu_{0k} \sim N\left(\mu_{0k}, \sigma_{0k}^2\right) \]  

The variance \( \sigma_{0k}^2 \) an inverse gamma distribution,

\[
\sigma_{0k}^2 \sim IG\left(a_{0k}, b_{0k}\right) \]  

The hyperparameters \( (\mu_{0k}, \sigma_{0k}^2, a_{0k}, b_{0k}) \) in (4) and (5) are assumed deterministic. The binary indicators \( I_i \) follow a Markov random field distribution, Equation (2),

\[
\exp(\mu_I) / (1 + \exp(\mu_I)) \sim Beta(p_1, p_2) \]  

\[
\eta_I \sim Beta(p_3, p_4) \]  

Individual random effects for the rate log(\( \gamma_{ik} \)) are assumed to follow an independent normal distribution log(\( \gamma_{ik} \)) \sim N\left(\mu_{rk}, \sigma_{rk}^2\right) where

\[
\mu_{rk} \sim N\left(\mu_{1k}, \sigma_{1k}^2\right) \]  

\[
\sigma_{rk}^2 \sim IG(\alpha_{rk}, \beta_{rk}) \]  

The individual change point \( \tau_{ik} \) is modeled as a truncated normal distribution as in,\(^{7}\)

\[
\tau_{ik} \sim TN[\nu_{ik}, \sigma_{ik}^2](\nu_1 - \tau^*, \nu_2) \]  

where the mean is centered at \( \mu_{rk} = 2 \) years and the variance is \( \sigma_{rk}^2 = 0.75 \). The distribution is truncated at \([\nu_1 - \tau^*, \nu_2]\) reflecting the preclinical duration, which is assumed to be of \( \tau^* = 5 \) years.

Finally, the conditional variance of the \( k \)-th biomarker level, \( \sigma_{rk}^2 \) is assumed to follow an inverse gamma prior distribution IG(\( a, b \)) as in,\(^{15,16}\)

From the description above, we define by \( K = \{\mu_{0k}, \mu_{rk}, \mu_I, \eta_I, \sigma_{0k}^2, \sigma_{rk}^2, \sigma_{ik}^2\} \) as the set of parameters specific to each biomarker \( k \). For a given set of observations \( Y = \{Y_{ijk}\} \), the likelihood of the parameters \( K \) and \( S \) can be readily obtained from Equation (1). Thus, the likelihood function \( L(K, S) \) is Equation (1) for a given set of observations \( Y \), that is
If we let $P_0(K, S)$ denote the a priori probability of the model parameters, as described through Equations (2–9), the a posteriori probability distribution of $K$ and $S$ given the data $Y$ has the form

$$P_Y(K, S) \propto L(K, S)P_0(K, S)$$

(12)

This posterior distribution contains all the statistical information relevant for the model. Below, we discuss methods for the numerical approximation of $P_Y(K, S)$. The hyperparameter values are chosen following clinical considerations discussed in (Table S1).

2.1.2 Procedure

The posterior distribution of all the unknown parameters can be approximated using a Markov chain Monte Carlo (MCMC) algorithm, following the procedure proposed and described in full detail in. Here, we highlight the main considerations for the iteration process. First, the choice between the types of sampling is based on whether the full conditional distributions can be easily calculated or not. For the biomarker-specific parameters, the posterior distribution of the subset of parameters $\{\theta_{ik}, \sigma_{ik}, \mu_{ik}, \sigma^2_{ik}, \theta_{ijk}, \mu_{ijk}, \sigma^2_{ijk}\}$ can be obtained using Gibbs sampling at each step of the iteration. In addition, $\{\mu_j, \eta_j\}$ are determined from Metropolis–Hasting sampling.

For the subject-specific parameters, $\theta_{ik}$, we use a Gibbs sampler. To get draws from the full conditionals of $I_{ik}, Y_{ik}$, and $\tau_{ik}$, we use a reversible-jump step. This comes from the construction of the change-point parameter: $I_{ik}$ provides either $\theta_{ik}$ for $I_{ik} = 0$ (for $i = 1, 2, \ldots, N$) or $\{\theta_{ik}, Y_{ik}, \tau_{ik}\}$ for $I_{ik} = 1$ (for $i = n_0 + 1, \ldots, N$).

**Initialization**

For each parameter in $K$, we draw an initial sample from its a priori distribution (similarly, for each patient and each parameter in $S$). We initialize the MCMC iteration by sampling from the priors of the model parameters (Table S1). Then, we estimate the posterior accordingly using MCMC as described above in this subsection.

**Iteration**

In this paper, we generate two independent chains, each with different initial values, to assess the convergence to the same stationary distribution for each unknown parameter using trace plots. We also ensured convergence using the Gelman–Rubin statistic. For each chain and unknown parameter, we simulate 40,000 samples with a burn-in period fixed at 5000 samples. The remaining samples from the two chains are combined to generate 70,000 samples in total. These were used for the calculation of the average of the joint probabilities to get an estimate of $P(\{Y_{i1k}, \ldots, Y_{ijnk}\} | \alpha_i)$ for each patient at each screening time and biomarker $k$. Here, $\alpha_i = \{0, 1\}$ indicates whether the patient $i$ is a control or has ovarian cancer.

**Screening**

The screening methodology in a $N'$ patient from a testing cohort is based on the computation of the posterior probability of ovarian cancer $P(o_{N'Y'})$. The variable $Y_{N'}$ denotes the longitudinal time series of different biomarkers for the patient up to time $t_j$, that is, the sequence $Y_{N'} = \{Y_{N'(j,k)}; j = 1, 2, \ldots; k = 1, 2, \ldots, K\}$.

$$P(o_{N'} = 1 | Y_{N'}) = \frac{P(Y_{N'} | o_{N'} = 1)P(o_{N'} = 1)}{\sum_{i\in\{0,1\}} P(Y_{N'} | o_{N'} = i)P(o_{N'} = i)}$$

(13)

where $P(o_{N'})$ is the prior prevalence estimated from population data (Annual Incidence of Ovarian Cancer in the United Kingdom by 5-year age group, 2016–2018), and $P(Y_{N'} | o_{N'})$ is estimated from the posterior predictive distribution for the $N'$ patient at each screening time using the training data of $N$ patients. Algorithm implementation was developed in RStudio, using the supporting codes provided by.
problem, which is common in traditional RNNs, making it generally more stable during training. It may suffer from computational complexity and potential overfitting, especially when dealing with small datasets. The latter, however, was mitigated by employing cross-validation and dropout, a regularization technique.

More precisely, the LSTM equations can be written as follows,

\[ c_{ijk} = f_{ijk} \odot c_{ijk-1} + i_{ijk} \odot \bar{c}_{ijk} \]  
\[ h_{ijk} = o_{ijk} \odot \tanh(c_{ijk}) \]  

where \( c_{ijk} \) and \( h_{ijk} \) denote the cell state and hidden state for patient \( i \), time step \( j \), and biomarker \( k \), respectively; \( \odot \) denotes point-wise multiplication. The cell state and hidden state at a fixed time step are given by vectors of size \( 1 \times H \), where \( H \) is the number of hidden neurons.

The cell state of the LSTM at each time step is controlled by input and forgetting mechanisms using the equations shown below. That is, the forget gate unit \( f_{ijk} \) modulates the effect of the cell state of the previous step \( c_{ijk-1} \). Similarly, the external input gate \( i_{ijk} \) weights the contribution of the candidate cell via \( \bar{c}_{ijk} \). Finally, the memory information in the hidden state is controlled by the output gate \( o_{ijk} \).

The so-added gates \( \bar{c}_{ijk}, i_{ijk}, f_{ijk}, \) and \( o_{ijk} \) are defined by the following equations.

\[ \bar{c}_{ijk} = \tanh(h_{ijk-1} U_c + Y_{ijk} W_c + b_c) \]  
\[ i_{ijk} = \sigma(h_{ijk-1} U_i + Y_{ijk} W_i + b_i) \]  
\[ f_{ijk} = \sigma(h_{ijk-1} U_f + Y_{ijk} W_f + b_f) \]  
\[ o_{ijk} = \sigma(h_{ijk-1} U_o + Y_{ijk} W_o + b_o) \]  

where \( \sigma(x) = 1/(1 + \exp(-x)) \) denotes the sigmoid function. The terms \( W_c, W_f, W_i, \) and \( W_o \) denote the weight matrices, with dimension \( 1 \times H \); the terms \( U_c, U_f, U_i, \) and \( U_o \) denote the kernel matrices of size \( H \times H \); vectors \( b_c, b_f, b_i, \) and \( b_o \) denote the bias of size \( 1 \times H \). The terms \( W, U, \) and \( b \) are learned during training.

Thus, for each time step \( j \), we obtain a temporal sequence of hidden states \( (h_{i1k}, h_{i2k}, ..., h_{iTk}) \) corresponding to the \( i \)-th subject and biomarker \( k \). Here, the state of the network at the last step is denoted \( h_{iTk} \), which corresponds to the last sample of the patient under consideration.

Next, we can concatenate the last hidden state associated to the output of each LSTM cell from the \( K \) different biomarkers. We also include the last hidden state associated to the LSTM processing the screening age \( t_{ij} \) as a longitudinal feature. That is,

\[ h_i = [h_{iT,1}, h_{iT,2}, ..., h_{iT,K}, h_{iT,0}] \]  

where \( h_i \) is the last hidden state for \( i \)-th patient, resulting from the concatenation of the last hidden state of each one of the \( K \) biomarkers under consideration, and an additional one associated with the age of the patient \( h_{iT,0} \).

To define \( (20) \), we can consider multiple combinations of biomarkers to analyze how the joint interaction of them impacts the model classification performance. In this paper, whichever is the case, we always include the age of the patient in our model as a feature. The resulting vector is of size \( 1 \times (H_1 + H_2 + \cdots + H_K + H_0) \), where each \( H_k \) denotes the number of hidden neurons used for each LSTM associated with each one of the features, that is, biomarkers \( k = 1, 2, \ldots, K \) and age, respectively.

The final output of the proposed model is

\[ \hat{\delta}_i = \sigma(W_i \tilde{h}_i + b_i) \]  

where \( \tilde{h}_i \) denotes the last hidden state after dropout, \( W_i \) is a weight vector of size \( (H_1 + H_2 + \cdots + H_K + H_0) \times 1 \) and \( b_i \) is a scalar bias. We optimize the weight matrices and biases using the cross-entropy loss.

\[ L(o_i, \hat{\delta}_i) = \frac{1}{N} \sum_{i=1}^{N} - (o_i \log(\hat{\delta}_i) + (1 - o_i) \log(1 - \hat{\delta}_i)) \]  

where \( N \) is the number of patients, \( o_i \) is the true label of subject \( i \) (0 for controls and 1 for cases), and \( \hat{\delta}_i \) is the estimated probability of risk of ovarian cancer.

The RNNs must be adequately trained before they can be used for the classification of unknown subjects. For the training phase, we used batch gradient descent with dynamic learning rates updated through the Adam optimizer. The hyperparameter tuning number is defined by the number of hidden neurons and dropout rate. Meanwhile, the learning rate and number of epochs are set fixed as meta-parameters (Table S2). The weight matrix for the recurrent state is initialized by a random orthogonal matrix, while for the inputs we use a weight matrix initialized using the Glorot’s scheme. The bias of each transformation is initialized at zero.

As part of the preprocessing step, the features are standardized during training by obtaining the mean and variance for each one of them. The same scaling transformation used during training is then applied for validation. Finally, the input data for the LSTM network were masked and padded to handle variable sequence lengths. Deep-learning algorithms were implemented in Python 3.8.8 using TensorFlow version 2.11.0 and Keras version 2.11.0.
2.3 Simulation, model selection, and evaluation

The detection scheme for each patient at each screening time is based on the soft classification of ovarian cancer using multiple correlated longitudinal biomarkers (CA125, HE4, and glycodelin). Our methodology is based on fully Bayesian screening based on change-point models (BCP) and LSTM-based models (RNNs).

As a preprocessing step for both methodologies, the biomarkers are transformed into the form $Y = \log(Z + 4)$, where $Z$ is a particular biomarker.\(^{7,15,37}\)

To evaluate the screening performance in both approaches, we use stratified 5-fold cross-validation with two repetitions (outer loop). This approach is particularly helpful in reducing the bias in performance evaluation. In this way, each fold divides the data into one set of training data and testing data preserving the proportion between cases and controls. Our main evaluation metrics included area under the ROC curve (AUC) and sensitivity (at 90% specificity). Significance was determined using the permutation test for mean of paired differences between two models.

For the Bayesian method, we estimate the posterior distribution of the parameters at each fold using $N$ patients in the training data. In turn, $P(Y_{N^+} | o_{N^+} = 0)$ and $P(Y_{N^+} | o_{N^+} = 1)$ are calculated from the posterior predictive distribution through these biomarker levels for the patient $N^+ > N$ in the testing data. The probability of having ovarian cancer is then calculated using (13). As mentioned before, convergence of the MCMC chains was determined using the Gelman–Rubin statistic for each of the posterior parameters (extracted from training data).

When considering deep-learning-based models, for each of the 10 iterations obtained from the outer loop we perform hyperparameter tuning on hidden neurons and dropout rate for model selection. The inner loop consists of a 10-fold cross-validation (3 repetitions). Once the optimal set of hyperparameters is selected, we estimate the probability of having ovarian cancer on the data held out from training for each patient (and each longitudinal observation). We repeat this step for every outer fold. Flow chart describing the design of the study is presented in Figure S3.

2.4 Lead-time analysis

Patients developing cancer show detectable preclinical elevations of biomarkers. The literature reports that patients with ovarian cancer show abnormal rise in these biomarkers approximately 3 years before diagnosis, with detectable elevations becoming apparent in the last year before diagnosis. We assess the potential value of using models based on multiple biomarkers in detecting cancer at earlier stages than it would be if diagnosed clinically.\(^{12,14,38,39}\) In the results section, we determine the model-based lead time of each one of the proposed screening tests. This is defined as the interval from being correctly classified as case by the diagnostic test to the actual clinical diagnosis. To discriminate patients, we detect the earliest observation per patient considered as abnormal using a threshold at 90% specificity.\(^{40}\) We then calculate the interval from this screening time point up to the time of diagnosis. This procedure is repeated over multiple outer folds in the cross-validation procedure, from which we get different statistics for each of the models under consideration.

3 RESULTS

3.1 Data

The data consist of 224 patients with serum samples sourced from the multimodal arm of UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS, number ISRCTN22488978; NCT00058032\(^{41}\)). This includes 180 controls (healthy subjects) and 44 cases (diagnosed patients). The eligible patients attended for screening and had an annual serum CA125 level measured as a baseline and transvaginal ultrasound in women as a second-line test. Additional biomarker assays human epididymis 4 (HE4) and glycodelin (PAEP) were performed within a subset of the serial samples from the general population of the UKCTOCS trial.

The screening age range of cases is [52.0, 77.4] years with an average of 65.3 years. The dataset includes the biomarkers history per patient for up to 5 years. Out of the 44 cases, 12 cases are screened for 1 year prior to clinical diagnosis, 12 cases for 2 years, and 20 cases for up to 3–5 years. In addition, from this set of cases 10 have 2 samples, another additional 10 have 3 samples, and 24 cases have 5 samples. For controls, the screening age is within the range of [50.3, 78.8] years with an average age of 63.6 years. Each patient has four to five observations (for 2 and 178 controls, respectively).

Within the cohort of cases, 35 patients present primary invasive epithelial ovarian cancer (average age of 64.4 years within a range of [52.0, 77.4] years), 6 cases of malignant neoplasm of peritoneum (average 68.5 years within [62.5, 76.3] years), and 3 cases of primary invasive fallopian tube cancer (average 68.2 years within [64.4, 70.8] years).

The available information indicates that of the 44 cases, 16 are in the early stage (I–II) and 28 late stages...
The morphology of cancers was predominantly serous ($n=27$ cases). Other types include papillary ($n=2$), endometrioid ($n=3$), clear cell ($n=2$), carcinosarcoma ($n=3$), carcinoma ($n=2$), neoplasm malignant ($n=1$), and adenocarcinoma ($n=4$).

In addition to CA125, human epididymis 4 (HE4) and glycodelin (PAEP) were selected based on previous studies. All serum samples were assayed by ELISA (enzyme-linked immunosorbent assay). Previous reports discarded possible confounding effects of sample processing using correlation between the concentration of any of the samples and time between sample collection and spin.

Longitudinal observations of biomarkers in case patients prior to clinical diagnosis are displayed in Figure 1. Loess curves were fit to reflect the mean levels over time. It is observed that for case patients the biomarker levels start to slowly rise between 1 and 2 years prior to diagnosis. This is particularly noticeable for CA125 and HE4, while for glycodelin the rise is to a limited extent. This rise becomes more pronounced within 1 year before diagnosis, in which all the biomarkers show recognizable elevations. In addition, from a qualitative perspective, the rate of increase within 1 year appears to be led by CA125 followed by glycodelin and then HE4.

### 3.2 Multivariable longitudinal models

Next, we evaluate the performance of different models in the detection of cases. The simulation studies consider multiple scenarios, including three joint multivariable screening tests that combine CA125 levels with other biomarkers (HE4 and glycodelin), as well as three single biomarker tests. More specifically, we have considered the following scenarios:

- m(1,2,3): CA125-HE4-glycodelin
- m(1,2): CA125-HE4

![Figure 1](A–C) Box plots of protein biomarker levels (Y1: CA125, Y2: HE4, and Y3: glycodelin) in controls and cases. For each panel, the cases are grouped in time ranges: year, between 1 and 2 or more than 2 years before diagnosis. (D–E) Biomarker levels in cases for CA125, HE4, and glycodelin for cases and controls, respectively. Loess curves fit has been added to depict the trend prior to clinical diagnosis. As observed, the biomarkers for case patients exhibit a significant increase within the final year. The biomarker levels have been log-transformed in all the plots.
• m(1,3): CA125-glycodelin
• u(1): CA125
• u(2): HE4
• u(3): glycodelin

Notation m(i,j,k) (or m(i,j)) indicates a joint multivariable test that relies on the biomarkers i, j, and k; while u(i) refers to a test that uses the i-th biomarker alone. The biomarkers are numbered as CA125 (1), HE4 (2), and glycodelin (3). We use two classification methodologies based on Bayesian change-point and recurrent neural network models.

The risk of ovarian cancer is estimated for each patient’s longitudinal observation starting from two visits up to all the screening period. That is, for each screening time point $t_i^*$ starting from the second visit, the model makes

![Figure 2](https://onlinelibrary.wiley.com/doi/10.1002/cam4.7163)
The combination of CA125 with HE4 provides improvement of the AUC score and sensitivity over the univariate tests using the proposed methodologies. To test the significance of such improvement, we used the permutation test for mean of paired differences with respect to CA125 for these two metrics using BCP and RNNs. For the AUC, we get the one-sided p-values of 0.043 and 0.002, respectively. Meanwhile, for the sensitivity, we obtain 0.031 and 0.062.

The second best-performing scheme is m(1,2,3) (CA125, HE4, glycodelin). In a similar way, the permutation test for the AUC provides p = 0.234 and 0.002, while for the sensitivity p = 0.016 and 0.125, using BCP and RNNs, respectively.

In Table 2a, using case patients only, we build a contingency table, to compare the best two diagnostic tests (joint multivariable and univariate, respectively) based on data from all the outer folds. The tests used a threshold set at 90% specificity. As observed, the joint multivariable model m(1,2) (CA125 and HE4) attains higher sensitivity than the reference standard, CA125. This holds for both RNNs and BCP methodologies. Applying a McNemar test is unfeasible in this case as the power is dependent on discordant pair sample size, which in our case is rather small. However, the detection rate in addition to the results shown above suggests that the combination of CA125 and HE4 has potential to improve current tests based on CA125 alone.

Next, we study the ability of our algorithms to detect the ovarian cancer at earlier stages. In Figure 3, we show the summary statistics of the AUC and sensitivity (90% specificity) estimated by screening the last time point available per patient using the complete longitudinal history, 1 and 2 years prior to clinical diagnosis. We observe that both metrics decrease as we screen patients with biomarker levels taken at earlier times.

### Table 1

<table>
<thead>
<tr>
<th>Model</th>
<th>Sensitivity (95% CI)</th>
<th>p-value</th>
<th>AUC (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m(1,2,3)</td>
<td>0.978 0.949, 1.0</td>
<td>0.016</td>
<td>0.966 0.941, 0.991</td>
<td>0.234</td>
</tr>
<tr>
<td>m(1,2)</td>
<td>0.967 0.934, 1.0</td>
<td>0.031</td>
<td>0.971 0.946, 0.996</td>
<td>0.043</td>
</tr>
<tr>
<td>m(1,3)</td>
<td>0.954 0.905, 1.0</td>
<td>0.062</td>
<td>0.957 0.924, 0.990</td>
<td>0.187</td>
</tr>
<tr>
<td>u(1)</td>
<td>0.908 0.853, 0.963</td>
<td>0.455</td>
<td>0.949 0.908, 0.990</td>
<td></td>
</tr>
<tr>
<td>u(2)</td>
<td>0.851 0.761, 0.941</td>
<td>0.527</td>
<td>0.949 0.920, 0.978</td>
<td></td>
</tr>
<tr>
<td>u(3)</td>
<td>0.821 0.741, 0.901</td>
<td>0.883</td>
<td>0.930 0.905, 0.955</td>
<td></td>
</tr>
<tr>
<td>RNNs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m(1,2,3)</td>
<td>0.954 0.917, 0.991</td>
<td>0.125</td>
<td>0.978 0.962, 0.994</td>
<td>0.002</td>
</tr>
<tr>
<td>m(1,2)</td>
<td>0.967 0.934, 1.0</td>
<td>0.062</td>
<td>0.987 0.979, 0.995</td>
<td>0.002</td>
</tr>
<tr>
<td>m(1,3)</td>
<td>0.942 0.903, 0.981</td>
<td>0.313</td>
<td>0.973 0.951, 0.995</td>
<td>0.006</td>
</tr>
<tr>
<td>u(1)</td>
<td>0.921 0.864, 0.978</td>
<td>0.953</td>
<td>0.953 0.928, 0.978</td>
<td></td>
</tr>
<tr>
<td>u(2)</td>
<td>0.853 0.773, 0.933</td>
<td>0.938</td>
<td>0.955 0.935, 0.975</td>
<td>0.455</td>
</tr>
<tr>
<td>u(3)</td>
<td>0.843 0.769, 0.917</td>
<td>0.984</td>
<td>0.941 0.917, 0.965</td>
<td>0.872</td>
</tr>
</tbody>
</table>

*Note: One-sided p-values are determined from permutation test on the difference of mean AUC (or sensitivity) between diagnostic tests with respect to our baseline (CA125). Bold values indicate p < 0.05 considered as statistically significant.*
Before 1 year of diagnosis, CA125–HE4 ranks as the best joint multivariable model. Using BCP, AUC scores 0.782 (with 47.0% sensitivity), whereas for RNNs, AUC scores 0.8 (47.3% sensitivity). In the univariate case, CA125 has AUC equal to 0.769 (54.9% sensitivity), using HE4 has AUC equal to 0.714 (52.7% sensitivity) using BCP. Using RNNs, CA125 alone provided higher sensitivity and AUC equal to 0.782 (with 47.0% sensitivity), whereas for RNNs, AUC scores 0.782 (47.0% sensitivity). In the univariate case, CA125 has AUC equal to 0.769 (54.9% sensitivity), using HE4 has AUC equal to 0.714 (52.7% sensitivity). In the univariate case, CA125 has AUC equal to 0.769 (54.9% sensitivity), using HE4 has AUC equal to 0.714 (52.7% sensitivity).

The estimated mean lead time based on joint multivariable tests spanned from 1.6 to 1.9 years, while median lead times were 1.6 and 1.7 years before diagnosis, CA125 alone provided higher sensitivity and AUC equal to 0.782 (with 47.0% sensitivity), whereas for RNNs, AUC scores 0.8 (47.3% sensitivity). In the univariate case, CA125 has AUC equal to 0.769 (54.9% sensitivity), using HE4 has AUC equal to 0.714 (52.7% sensitivity). In the univariate case, CA125 has AUC equal to 0.769 (54.9% sensitivity), using HE4 has AUC equal to 0.714 (52.7% sensitivity).

The present study addresses two distinct issues—methodological, related to the integration of multiple longitudinal biomarkers into a single model, and practical, concerning the enhancement of performance rates in ovarian cancer detection through longitudinal data analysis.

We compared two longitudinal algorithms allowing the integration of more than one biomarker—a Bayesian change-point model and a LSTM architecture of the recurrent neural networks approach. Our findings show that the combination of longitudinal CA125 and HE4 levels outperforms the CA125-only model with both the change-point model and the LSTM method, highlighting the complementary nature of HE4. The multimarker model did not improve the lead time but provided higher area under the ROC curve (AUC) and sensitivity at a fixed specificity, potentially improving early cancer detection. Importantly, this work illustrates the advantage of the methodology for the simultaneous analysis of multiple longitudinal biomarkers that may prove useful in any scenario where such biomarkers emerge, particularly in early cancer detection.

Previous research in ovarian cancer has mainly concentrated on multiple biomarkers at a single time point or on longitudinal CA125, the best-performing individual biomarker. However, the UKCTOCS trial demonstrated that monitoring CA125 alone does not provide significant mortality benefits. Consequently, there is an urgent need to explore additional potential biomarkers that could improve detection rates. Our previous work focused on the analysis of longitudinal HE4, CA72-4, and anti-TP53. The findings suggested that these biomarkers offer limited additional value to longitudinal CA125. However, the study was limited to using only the MMT approach thus emphasizing the significance of the present work.

The primary limitation of our study is the small sample size, which we attempted to address through the nested cross-validation approach. Both cancer cases and controls were randomly selected from the complete dataset. Since the UKCTOCS was a randomized trial, this selection process should mitigate potential confounding factors and biases. Another limitation is that only two approaches have been tested in this study, and the biomarker panel comprised only three proteins. We focused on currently available approaches for analyzing time-series biomarker data in the cancer setting, utilizing the available dataset. While we incorporated some of the most prominent biomarkers, it is important to note that future studies may achieve improved performance with a larger panel of biomarkers and potentially new statistical and AI methodologies. The primary strength of our work is the use of the unique dataset obtained from the UKCTOCS trial.
FIGURE 3 Each box plot displays estimates of (A, C) AUC and (B, D) sensitivity (90% specificity) comparing changes at different models (joint multivariable m(1,2,3), m(1,2), and m(1,3), and univariate u(1), u(2), and u(3)) and time periods (left to right: all screening period, 1 and 2 years before diagnosis). Each plot results from the cross-validation procedure. Figures (A, B) correspond to Bayesian change point and (C, D) to recurrent neural networks.
In conclusion, our investigation provides evidence of enhanced performance using a combination of longitudinal biomarkers compared with the best individual longitudinal CA125. To the best of our knowledge, this is the first investigation in which statistical and artificial intelligence (AI) approaches have been employed and compared for the analysis of multiple longitudinal biomarkers, not only in the context of ovarian cancer but also in other healthcare settings. The findings from the current work could be applied to facilitate early detection, risk stratification, and the prevention and treatment of various diseases. The enhanced early detection capabilities of the CA125–HE4 multimarker model hold the potential to significantly improve patient outcomes by enabling timely diagnosis, assisting healthcare providers in more accurate clinical decision-making, and informing policymakers in shaping effective strategies for ovarian cancer screening and management.

These findings now warrant blinded validation in a larger longitudinal sample set to assess the potential for early detection in ovarian cancer screening.

**AUTHOR CONTRIBUTIONS**

**Luis Abrego:** Formal analysis (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal).

**Alexey Zaikin:** Formal analysis (equal); funding acquisition (supporting); writing – original draft (equal); writing – review and editing (equal).

**Ines P. Marino:** Formal analysis (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal).

**Ian Jacobs:** Conceptualization (equal); methodology (equal); writing – review and editing (equal).

**Usha Menon:** Conceptualization (equal); data curation (equal); supervision (equal); writing – review and editing (equal). 

**Aleskandra Gentry-Maharaj:** Conceptualization (equal); data curation (equal); project administration (equal); writing – original draft (equal); writing – review and editing (equal).

**Oleg Blyuss:** Conceptualization (equal); formal analysis (equal); funding acquisition (lead); methodology (equal); supervision (equal); writing – original draft (equal); writing – review and editing (equal).

**Mikhail I. Krivonosov:** Writing – original draft (equal); writing – review and editing (equal).

**ACKNOWLEDGMENTS**

This study was funded by Cancer Research UK and EPSRC joint award EDDCPJT/100022. We thank all trial participants and all staff involved in the UKCTOCS trial. IPM acknowledges the support of grant PID2021-125159NB-I00 (TYCHE) funded by MCIN/AEI/10.13039/501100011033 and by ‘ERDF A way of making Europe’. MIK acknowledges support by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730321PSQ0002) and the agreement with the Ivanovsk Institute for System Programming of the Russian Academy of Sciences dated November 2, 2021, No. 70-2021-00142. OB acknowledges support from Barts Charity (G-001522).

**CONFLICT OF INTEREST STATEMENT**

U.M. and I.J. declare financial interest through UCL Business and Abcodia Ltd in the third-party exploitation of clinical trial biobanks, which have been developed
through research at UCL. The remaining authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT
Raw assay data, excepting CA125, are available upon request.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE
This nested case–control study within UKCTOCS was approved by the Joint UCL/UCLH Research Ethics Committee A (Ref. 05/Q0505/57). Written informed consent was obtained from donors, and no data allowing identification of patients were provided. The study was performed in accordance with the Declaration of Helsinki.

ORCID
Oleg Blyuss https://orcid.org/0000-0002-0194-6389

REFERENCES

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Abrego L, Zaikin A, Marino IP, et al. Bayesian and deep-learning models applied to the early detection of ovarian cancer using multiple longitudinal biomarkers. *Cancer Med*. 2024;13:e7163. doi:10.1002/cam4.7163