Supplementary Methods

FLT4 assay

A sandwich ELISA kit was purchased for the quantification of soluble FLT4 (DuoSet, R&D Systems, Minneapolis, MN). This assay failed to validate when tested as serum FLT4 did not serially dilute in parallel with the assay standard. Four samples tested for parallelism gave results ranging 36.8-38.5% (acceptable limit is ≤15%). Further investigation showed that this may be due to the tagged FLT4 protein used as a standard initially and when the original DuoSet standard was replaced with a polyhistidine-tagged recombinant protein (Sino-Biological, Beijing), parallelism to within acceptable limits was seen. Therefore, the FLT4 ELISA assay was validated using the monoclonal antibody pair from the DuoSet and the protein from Sino Biological. The antibodies were initially tested in a checkerboard analysis to determine the ideal working concentrations, testing the capture antibody at 2, 4 and 6 µg/mL diluted in PBS and bound to a Nunc Maxisorp plate (Nalge Nunc International; Rochester, NY). The detection antibody was tested at 0.25, 0.5 and 1 µg/mL diluted in PBS-T (PBS, 0.1% v/v Tween 20). Several dilutions of recombinant protein (0, 0.313, 5.0 and 20 ng/mL) diluted in PBS-T + 1% w/v BSA were used to analyse the antibody performance. Streptavidin HRP (R&D systems) at a 1 in 200 dilution in PBS-T, tetramethylbenzadine (TMB; Sigma-Aldrich, St. Louis, MO) and 1N sulphuric acid (Sigma-Aldrich) were all used to complete the assay.

Antibody concentrations were selected as 2 µg/mL in PBS and 0.25 µg/mL in PBS-T for the capture and detection respectively. The standard curve range was determined by running serial dilutions of recombinant protein from 30 ng/mL to 0.078 ng/mL and refined to a range of 12.5 ng/mL to 0.625 ng/mL. In parallel, multiple dilutions of serum samples with expected high and low FLT-4 concentrations were tested to determine a suitable sample dilution factor of 1 in 10.

In the final ELISA procedure, 100 µL of capture antibody was added to each well of a 96 well plate and incubated overnight at 2-8°C, before being washed (four times with 400 µL PBS-T; wells were left to soak for 5 seconds between buffer dispense and aspiration). 100 µL sample (diluted 1 in 10), standards and quality control (QC) samples were added and incubated at room temperature on an orbital plate shaker at 400 rpm for two hours. Plates were washed and 100 µL of detection antibody added and incubated with shaking for one hour. Plates were washed and 100 µL of streptavidin-HRP diluted 1 in 200 was added and incubated with shaking for 30 minutes. Plates were washed for a final time, 100 µL of TMB was added and incubated protected from light on the bench for 20 minutes. The reaction was stopped by adding 50 µL 1 N sulphuric acid, shaking the plate for 2 minutes and measuring the OD of the wells at 450 nm and corrected to 570 nm. All sample and reagent dilutions were performed with PBS-T + 1% w/v BSA.

The validation of the FLT4 ELISA was performed using serum samples from patients with diagnosed ovarian cancer. The validation included: parallelism assessment; recovery analysis; intra- and inter- assay precision; lower limit of quantification determination; interference analysis; the assessment of hook effect and dilution linearity.

Parallelism assessment was conducted using three serum samples serially diluted from the initial dilution factor (1 in 10) three times to a final dilution of 1 in 80. Parallelism was deemed acceptable when the back calculated concentrations of each dilution gave a between dilution coefficient of variance of ≤15%. All samples diluted parallel with results between 2.6 - 12.6%.

Recovery testing was performed using three serum samples spiked with low and high concentrations (final assay concentrations of 1.4 ng/mL and 6.7 ng/mL respectively) of recombinant FLT4 diluted in PBS-T + 0.1% w/v HSA. Spiked samples were compared to samples spiked to 10% of the total volume with recombinant protein diluent (PBS-T + 0.1%
w/v HSA) and recombinant protein spiked into sample diluent to obtain the ‘true’ spike concentration. Recovery was found to be acceptable when the spiked sample adjusted for the endogenous FLT4 concentration was within 20% of the ‘true’ spike concentration. All samples recovered to between 81.5 – 104.9%.

Intra- and inter-assay precision were assessed using QC samples made from both ovarian cancer and healthy patient serum samples. Intra- and inter-assay precision, assessed on both QC samples run in duplicate using five independent dilutions on six plates, were found to be 4.2% and 3.6% (intra-) and 11.2% and 10.0% (inter-) for low and high QC samples respectively.

Lower limit of quantification (LLoQ) was assessed by comparing independent duplicates of a dilution of the second low standard, low standard and two dilutions below the lowest standard to the standard curve. This was performed over six days and each concentration was assessed for imprecision and inaccuracy. The LLoQ of the assay was given as the concentration where the CV between the six measurements was ≤20% and the mean concentration over the five measurements was 80-120% of the expected concentration, which in this case was 0.625 ng/mL.

The first part of interference was assessed by looking at common endogenous serum interferents. These included complement, rheumatoid factor, bilirubin, triglycerides and haemoglobin. Haemolysis interference was tested by spiking pooled haemolysate (produced by washing EDTA plasma derived erythrocytes in PBS then lysing in MilliQ water) into a pool of serum samples to a haemoglobin concentration of 5 mg/mL. The sample was compared to the same pool of serum spiked with MilliQ water to the same ratio as the haemolysate spike. Interference was detected if the two samples differed by >20%. All other interferents were analysed by using serum samples with high concentrations of the relevant interferent spiked at 10% of the total sample volume to a final concentration of 5.4 ng/mL of recombinant FLT4. Samples were analysed as per recovery testing and interference was detected if the spiked sample adjusted for the endogenous FLT4 concentration was >20% of the true spike concentration. None of the interferents tested interfered with the assay performance.

Interference was also assessed by spiking in differing concentrations of VEGF-C and VEGF-D (10, 50, 100, 500 and 1000 ng/mL), in PBS-T/0.1% w/v HSA, into serum samples and comparing results with samples spiked with PBS-T/0.1% w/v HSA. Interference was considered present if spiked VEGF affected FLT4 concentrations by >20%. It was found that VEGF-C concentrations >100 ng/mL resulted in a reduction of measured FLT-4. VEGF-D had no effect.

Hook analysis was performed by spiking recombinant protein into serum samples to a final concentration of 5 µg/mL. Samples were then diluted 1 in 10 as per the assay protocol, making the spike concentration 500 ng/mL (40 times the top standard). Dilution linearity was assessed when samples were then diluted 1 in 100 to bring the sample back into the assay range (5 ng/mL). This test was mimicked in buffer as a comparison. It was confirmed that there was no hook effect, as spiked samples were not detected within the assay range. Samples diluted back into the range to the expected concentration (determined by the buffer test), it was therefore accepted that samples dilute linearly.

Upper limit of normal based on a 95th percentile was 72.2 ng/mL with no significant differences being seen between males and females or with age (n=72, 36 females and 36 males, age range 32-75 and 30-75 and median age 59 and 53 respectively).
**Assay of other protein biomarkers**

Soluble mesothelin-related peptides (SMRPs) were quantified using the FDA-approved MESOMARK® assay (Fujirebio Diagnostics) following the manufacturer’s instructions. The manufacturer’s data indicated that 99% of healthy controls have concentrations ≤ 1.5 nM. Serum AGP and CA-125 concentrations were measured using routine clinical assays by the Department of Clinical Biochemistry and Immunology (Leeds General Infirmary, UK). Serum AGP was measured on a Behring Nephelometric Analyser (BNA II, Siemens) and CA-125 levels were measured using the Siemens ADVIA Centaur CA-125 II assay. The normal ranges for these assays are: AGP age <50 female 0.4 – 1.0 g/L, male 0.6 – 1.2 g/L and age > 50 both sexes 0.8 – 2.0 g/L; CA-125 <35 kU/L.

Samples for validation cohorts I and II were analysed in January 2012 and June 2013 respectively with appropriate QC.

**Calculation of biomarker index**

The data for each marker was split into intervals containing 1/6th of the data resulting in 5 potential cut-points per marker. Each cut-point was considered in turn and the marker was dichotomized above and below the cut-point. This ‘low’ and ‘high” concentration data was used to produce separate Cox PH regression models: one for data above the cut-point and one for data below the cut-point. A $Z$-statistic was then calculated using the coefficients of the two models and a pooled estimate of their standard errors (se) according to the equation:

$$Z = (\beta_{+ve} - \beta_{-ve}) / \text{pooled se (} \beta_{+ve} - \beta_{-ve})$$

where $\beta_{+ve}$ is the coefficient from the ‘high’ model and $\beta_{-ve}$ is the coefficient from the ‘low’ model. This was repeated for the all cut-points and p-values calculated similarly. The cut-point yielding the lowest p-value was considered optimal and taken forward to be used in the predictive index. This was repeated for each marker, resulting in an optimal cut-point being determined for each.

The index was formed by considering the marker concentrations for each patient. A score=1 was given if the patient’s marker concentration was on the side of the cut-point which favoured bevacizumab in terms of PFS in the model, or a score=0 otherwise. This was repeated for each marker and every patient to form an index on the scale 0-4 for each patient. This index was then dichotomized at each potential cut-point and Cox PH models produced for each with dichotomized index, treatment arm and an interaction term between index and treatment arm included as covariates. Models with a significant interaction term were subsequently investigated further.

**Internal validation of biomarker index**

In order to perform an internal validation of the predictive ability of the biomarker index bootstrap resampling was used to estimate the optimism in Harrell’s concordance index (C-index). This process involved building the final Cox proportional hazards model with $R=1000$ bootstrap resamples of the marker and survival data. For each bootstrap sample a new boot model was constructed and the C-index estimated on the bootstrap resample ($C_{boot}$) used to derive the model and also on the full original dataset ($C_{full}$). The optimism (O) in the original model’s predictive ability was then estimated as $C_{full} - C_{boot}$. This was repeated for each resample and then the optimism in the model defined as the mean of O over all bootstrap resamples. This value was then subtracted from the C-index originally estimated for the model to give a corrected estimate of predictive potential.
This is analogous to using a training set and test set, but in this situation the bootstrap resample is the training set and the full data set is being used as the test set. This approach is described in detail elsewhere for prognostic models and more generally (Harrell 1996; Efron 1983).