www.oncotarget.com Oncotarget, 2018, Vol. 9, (No. 32), pp: 22717-22726

Research Paper

Parenclitic networks for predicting ovarian cancer

Harry J. Whitwell[1], Oleg Blyuss[2], Usha Menon[3], John F. Timms[3] and Alexey Zaikin[3,4]

[1]Chemical Engineering, Imperial College London, London, United Kingdom

[2]Wolfson Institute of Preventive Medicine, Queen Mary University of London, London, United Kingdom

[3]Institute for Women's Health, University College London, London, United Kingdom

[4]Department of Mathematics, University College London, London, United Kingdom

Correspondence to: *Harry J. Whitwell,* ***email:*** *h.whitwell@imperial.ac.uk*
John F. Timms, ***email:*** *john.timms@ucl.ac.uk*
Alexey Zaikin, ***email:*** *alexey.zaikin@ucl.ac.uk*

Keywords: *parenclitic; network; ovarian cancer; biomarker; serum*

Received: *March 13, 2018* ***Accepted:*** *April 07, 2018* ***Published:*** *April 27, 2018*

Copyright: Whitwell et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License 3.0 (CC BY 3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

ABSTRACT

Prediction and diagnosis of complex disease may not always be possible with a small number of biomarkers. Modern 'omics' technologies make it possible to cheaply and quantitatively assay hundreds of molecules generating large amounts of data from individual samples. In this study, we describe a parenclitic network-based approach to disease classification using a synthetic data set modelled on data from the United Kingdom Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) and serological assay data from a nested set of samples from the same study. This approach allows us to integrate quantitative proteomic and categorical metadata into a single network, and then use network topologies to construct logistic regression models for disease classification. In this study of ovarian cancer, comprising of 30 controls and cases with samples taken <14 months to diagnosis (*n* = 30) and/or >34 months to diagnosis (*n* = 29), we were able to classify cases with a sensitivity of 80.3% within 14 months of diagnosis and 18.9% in samples exceeding 34 months to diagnosis at a specificity of 98%. Furthermore, we use the networks to make observations about proteins within the cohort and identify GZMH and FGFBP1 as changing in cases (in relation to controls) at time points most distal to diagnosis. We conclude that network-based approaches may offer a solution to the problem of complex disease classification that can be used in personalised medicine and to describe the underlying biology of cancer progression at a system level.

INTRODUCTION

Personalised medicine is hailed as the next significant step in the treatment and monitoring of diseases. Some cancer treatments, for example Trastuzumab in HER-2 positive breast cancer, are already tailored to individuals based on personal expression data [1]. Given the rate at which technology is advancing for high-throughput molecular analysis, it is inevitable that samples will be routinely taken from a patient and analysed by a range of 'omic technologies [2]. Whilst this will allow truly personalised medicine, there will be huge challenges in analysing large amounts of multi-dimensional and longitudinal data. Given the technological advances of high-throughput, multi-omic technologies, a conceivable ideal would be to take all available data and identify changes that indicate an early stage malignancy or accurately predict the formation of such. The cancer genome atlas has shown that there are many possible combinations of changes responsible for the onset of particular cancer types [3]. Hence there is a need for a procedure that considers the changes in a system as a whole, i.e. a network biomarker. Herein, we present an algorithm for generating parenclitic networks that are optimised towards biomarker identification. The technique we present can also inform on the underlying biology [4],

as well as provide network models for prediction. Thus, we believe this could be an important progression in the advancement of personalised medicine.

Parenclitic networks, first described by Zanin *et al.* in 2014 [5, 6], are established by determining differences between pairs of analyte measurements (protein, mRNA etc.) in a control data set and individual case samples. If the difference is above a threshold, then a linkage between the analytes is created. This is repeated for every pair of analytes to generate a network. In [5], the differences were calculated by plotting a linear regression through a control population for a specific analyte pair and then calculating the perpendicular distance from the line of regression to the sample point for the same pair. We first show that a better approach, at least from the perspective of constructing a network-biomarker, is to use 2-dimensional kernel density estimation (2DKDE) as an underlying model for the control distribution. The difference is determined by the area under the density distribution in the controls for the sample marker-pair, thus if the sample marker pair lies in a region of low density, the area under the density distribution will be larger, and the inferred distance is greater. We have also developed the network approach to allow the inclusion of categorical as well as continuous data, allowing testing of cancer risk variables on the networks. In our approach, we take all available information and reduce it to network topological features that are then built into logistic regression models.

We apply the approach to the prediction of ovarian cancer (OC) using multiple protein measurements made in pre-diagnosis serum samples from a cohort of type II OC cases and matched controls. Ovarian cancer is the 6th most frequently occurring cancer in women and 10 year survival is only 35% for both Type I and Type II cancers combined [7]. Type II cancers are much more aggressive than Type I, and are responsible for the majority of deaths. Cancer antigen 125 (CA125) is the current best biomarker for OC, however it is only predictive in late stages of the disease when survival is between 5–19% [8]. A recent report from the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS), which used serial serum CA125 measurements, showed that whilst there was no significant benefit to mortality, there was a stage shift at diagnosis [9]. This supports that screening for OC may be a sensible approach in the future, but better models are required for earlier detection that will translate into improved mortality.

Herein, we employ parenclitic networks for disease prediction in a small cohort of ovarian cancer cases and controls at two time points with respect to clinical diagnosis. We also generate average (mode) networks for each time group that highlight differences in the protein-networks between cases and controls at these two time points.

RESULTS

Linkages between analytes in an individual are based on how the analyte pair is predicted to deviate from a control population. Zanin *et al.* [5] demonstrated that this can be performed by plotting a linear regression through a sample set of data and, for each individual, calculating the perpendicular distance (normalised by standard deviation (z-score)) from the regression for each analyte pair. To implement this in a general algorithm, one must assume that all pairs of analytes will be both correlative and follow a linear model. However, in biological samples, this is often not the case. For example, in Figure 1, we have plotted MUC16/CA125, an OC biomarker, against androgen receptor or folate receptor gamma, which form a non-correlative or a bimodal distribution respectively (Figure 1A, 1B). In neither case was the average distance able to differentiate between cases and controls. When repeating for all combinations of markers with MUC16, the mean *P*-value was 0.42 (SD = 0.31). We have overcome this by using 2-dimensional kernel density estimation to predict the density of two markers in a non-cancer control population. Sample deviation is determined based on the estimated density for the analyte pair (see methods and Supplementary Information (SI)). For the same combinations of markers, we were able to differentiate between cases and controls (Figure 1C and 1D) with a mean *P*-value of 3.17×10^{-7} (SD = 2.01×10^{-6}). Therefore, we employed 2DKDE in our algorithm for linkage assignment.

Preliminary investigations with a longitudinal, synthetic data set modelled on CA125 (see Supplementary Information 2), showed that topological features of the networks can be used to detect changes within the data set (see Supplementary Figures 1, 2) at a given threshold. Not all topological features are maximally discriminative between cases and controls at the same threshold so we cycle through a number of thresholds to iteratively determine the optimum for each topological feature (Figure 2A and 2C, Supplementary Figures 3, 4). These descriptors can be combined into a multi-parameter logistic regression for disease prediction. We tested this procedure in an OC data set comprising type II OC cases and controls ($n = 30$), where each individual has two serum samples taken <14.5 months (late, $n = 30$) or >34.5 months (early, $n = 29$) prior to diagnosis. Protein quantification for each sample was performed by proximity extension assay for a panel of 92 cancer-related proteins (Olink Oncology II panel). A second data set, comprising 120 controls was also assayed with the same panel and used to generate the kernel density estimates (for further description of both data sets, see methods and Supplementary Tables 1-2). The best model for each time group generated using the parenclitic methodology was then compared with logistic models generated using the raw data (raw data logistic regression, RDLG) after Monte Carlo cross-validation. At a specificity of >98%, the best parenclitic model had a higher sensitivity in both early and late groups (Table 1).

In the late group (<14 months to diagnosis), the best parenclitic model used "number of connections to MUC16" (Figure 2B), with the RDLG model using the raw values for MUC16. Whilst the sensitivity of the parenclitic

model (80.3%) was higher than the RDLG model (76.7%), suggesting that the parenclitic approach favours discrimination, there was no improvement on case/control discrimination within this data-set. (Figure 3A).

In the early group (>34.5 months to diagnosis), the parenclitic model comprised of alpha-centrality (Figure 2D) and oral contraceptive pill (OCP) use (as a categorical variable), whereas in the RDLG model, glypican 1 (GPC1) was used. In this group, the higher sensitivity achieved for the parenclitic model (18.9%) compared to the RDLG model (16.3%) translated into 8 predicted cases in the parenclitic model, compared to 6 in the RDLG model (Figure 3B). Furthermore, of these, only 1 was diagnosed by both tests, suggesting that the parenclitic and RDLG models are identifying orthogonal features and combining them in a decision tree may provide improved discrimination at early time points.

Parenclitic networks can also be used to inform on changes within biological systems without the arbitrary cut off of a *P*-value or fold change. Thus, we generated a modal network to investigate differences between cases and controls and between the early and late OC groups. The modal network was produced by generating networks for each individual and taking the

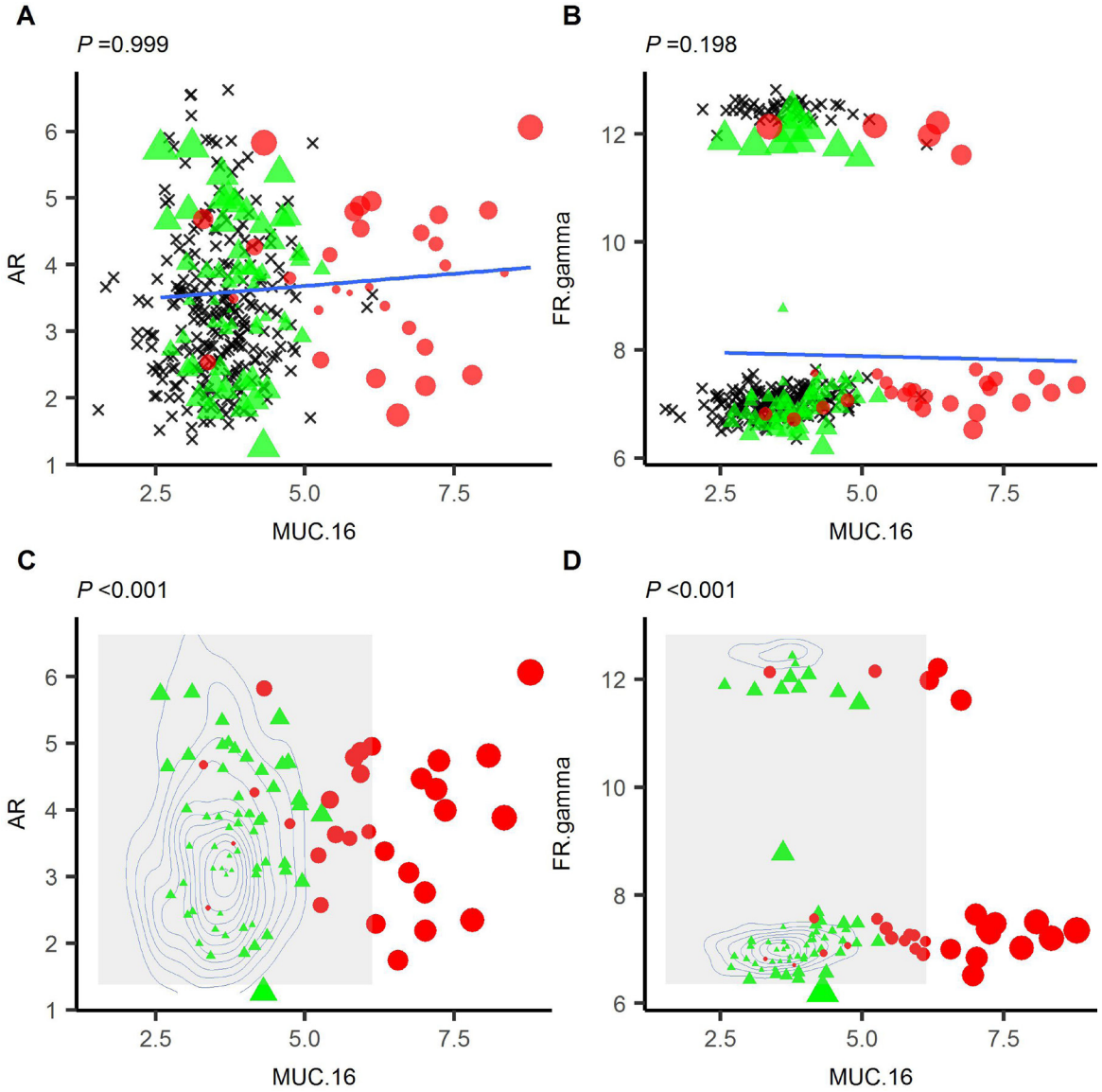


Figure 1: 2D-Kernal density estimation-based distance can differentiate cases (red circle) and controls (green triangle) where linear-regression based distance cannot. Linear regression was plotted using a control set (black crosses) for MUC16 vs AR (**A**) or MUC16 vs folate receptor gamma (**B**). Cases and controls are then overlaid and the perpendicular distance from each point determined. It was not possible to differentiate cases from controls using the linear regression ($P = 0.999$ or 0.198 for AR or folate receptor gamma respectively). 2DKDE estimation of the same distributions was performed (**C** and **D**) and distance calculated for each case and control based on the density of the underlying distribution. In both cases, it was possible to differentiate cases and controls in this manner.

modal state of a linkage – i.e. if the modal state between any two markers was connected, they were connected in the modal-network. Networks were generated across a range of thresholds with community membership identified based on edge-betweenness clustering. Well connected markers are found in the centre of clusters and these "hub-centres" represent proteins that are the most different between cases and controls (Figure 4). In samples from the late group, i.e. closest to diagnosis, WDCF2 (HE4) and MUC16 (CA125) were clearly at the centre of the largest hub. Midkine (MK) was also well-connected in a subsidiary cluster, although it is not a clear hub-centre (Figure 4A). In samples taken furthest from diagnosis (early, >34.5 months), granzyme-H (GZMH) and fibroblast growth factor binding protein 1 (FBF BP1) are well-connected centres and thus may indicate involvement in the onset of OC or response to tumorigenesis (Figure 4B).

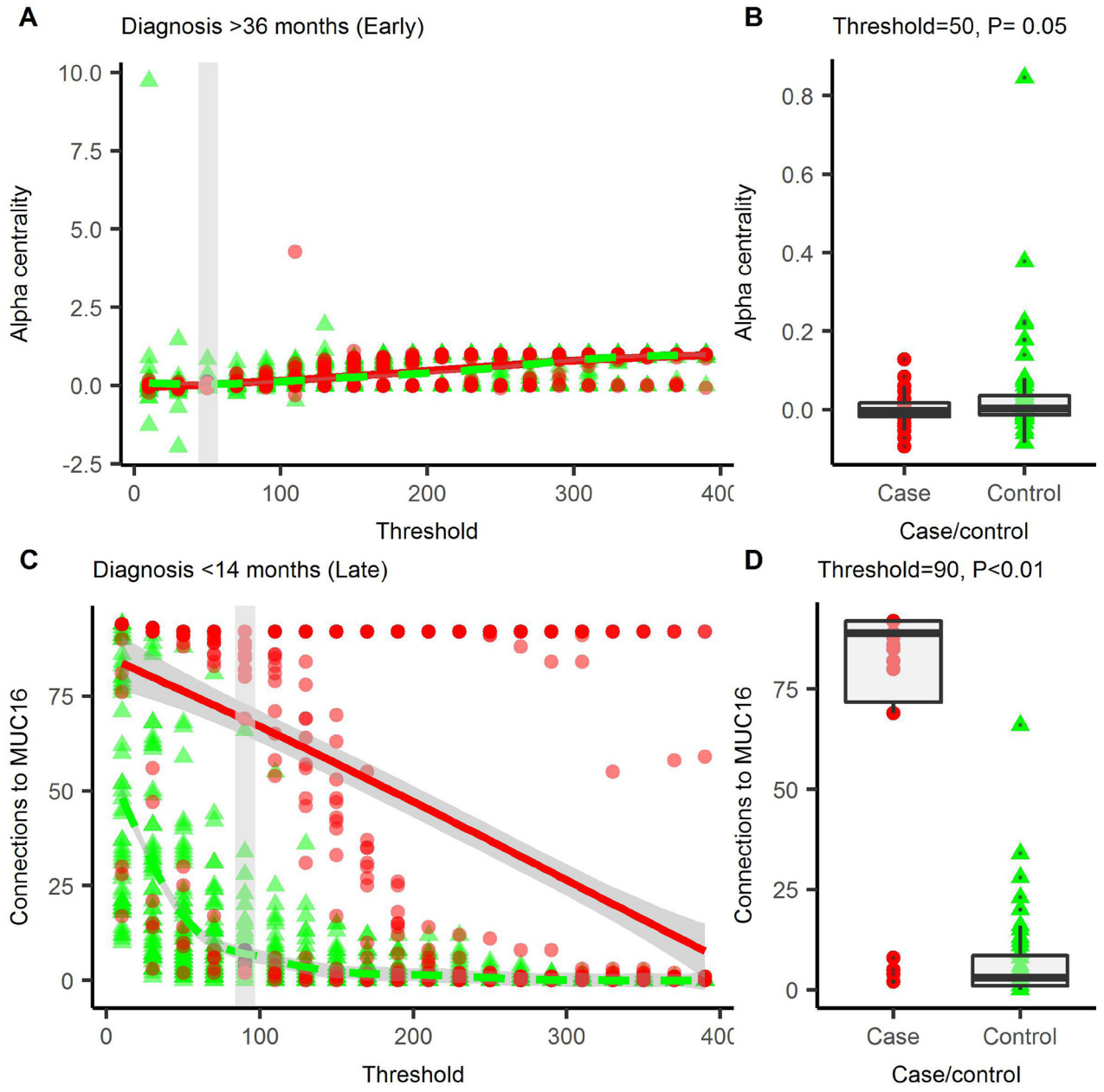


Figure 2: Parenclitic networks are generated across a range of thresholds. At each threshold, the network is described using a number of topological indices, the index value at each threshold is presented for connections to MUC16 (**A**) and alpha-centrality (**C**), which featured in prediction models for OC in the early, and late groups respectively. For each index, the smallest threshold that gives the greatest degree of differentiation between cases and controls was used to build the logistic regression models. For connections to MUC16, this was at a threshold of 90 (**B**) and for alpha-centrality, 50 was optimal (**D**). Red circles = cases, green triangles = controls, solid line = trend line. In A and C, the grey box is the optimal threshold; this data is shown in B and D.

Table 1: Area under the ROC curve (AUC), specificity and sensitivity of the best performing parenclitic and RDLG models for early (>34.5 months to diagnosis) and late (<14.5 months to diagnosis) groups

Group	Type	Model	AUC	Specificity	Sensitivity
Early	Parenclitic	-7.44 (Alpha-centrality) -0.16 (Pill use)	0.624	1	0.189 ± 0.124
	RDLG	-2.09 (GPC1)	0.663	1	0.163 ± 0.087
Late	Parenclitic	0.07 (Connections to MUC16)	0.904	1	0.803 ± 0.077
	RDLG	2.14 (MUC16)	0.904	1	0.767 ± 0.082

Models were generated using all combinations of 1, 2 or 3 markers and following cross-validation the best models were selected based initially on sensitivity and secondly on AUC. Specificity was set to be >98% and the maximum is reported for the given sensitivity. Models were generated by logistic regression of either the parenclitic topologies or raw data and model coefficients are provided.

Whilst GO-term enrichment analysis shows that there is some enrichment for serine proteases within the group (data not shown), the protein pool is small and biased towards "cancer-related" proteins and consequently we do not hold much weight by this observation and have not investigated this further.

DISCUSSION

Parenclitic networks have been utilised for a number of approaches including both biological discovery [5] and cancer detection [4, 6]. In these cases, the methodology underlying edge/connection creation within the network has relied on multiple linear regression models. In biological data sets, the assumptions for performing a linear regression are not always met. In Figure 1**,** we provide examples of non-correlating and bimodal distributions in which distance from a linear regression between a disease-predictor (MUC16) and a non-predictor (AR, folate receptor gamma) is not able to differentiate between cases and controls. The method we have developed describes the data with a greater degree of fidelity by using 2DKDE and thus can achieve a better estimation of distance between an individual and a population. Indeed, we distinguished between cases and

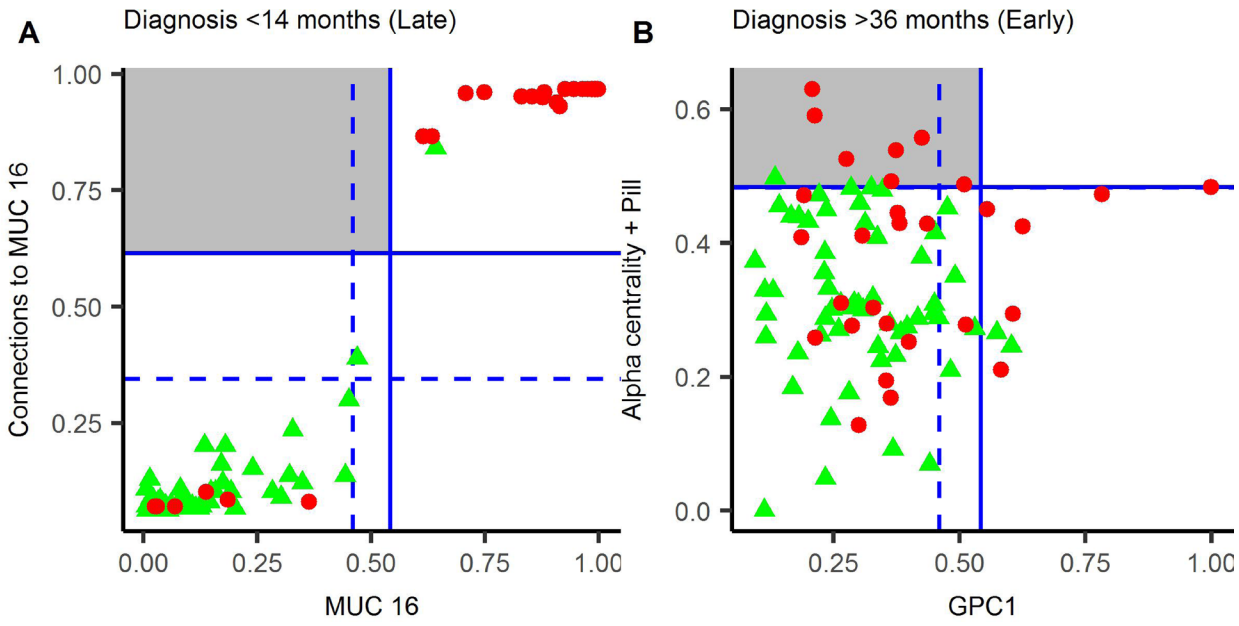


Figure 3: Parenclitic-based models versus RDLG-models for late (**A**) and early (**B**) pre-diagnosis groups. The predicted value for each model is plotted with horizontal and vertical lines representing a diagnostic threshold giving 95% (dashed line) or 98% (solid line) specificity. Cases are represented by red circles and controls by green triangles, hence cases in the top left quadrant (grey square) were detected by the parenclitic-model only and cases in the bottom right quadrant by RDLG-models only. Points in the top right quadrant are detected by both models, OCP = oral contraceptive pill.

controls for all pairs of distributions that included MUC16. The 2D density matrices require a large number of samples to describe the distribution accurately [10, 11], hence in this study, we used 249 independent control samples for deriving the density matrices and we would envisage that with fewer samples, this methodology would not work as effectively. This is being further investigated with synthetic data.

MUC16 (CA125) is currently the best diagnostic marker for OC, however, the specificity and sensitivity of the marker is hampered by elevated expression in a number of benign diseases and other cancer types [12]. Whilst the marker may be elevated at early stages of the disease, recent large studies of annual screening showed there was no improvement in mortality for those diagnosed from elevated CA125 or otherwise [13, 14]. In using a network approach, we hoped to be able to pick up on subtle changes in protein expression prior to diagnostic levels of CA125 [3]. As the data set was small, we aimed to avoid overfitting by limiting the number of possible markers in a logistic regression to 3 and cross-validated all results by the Monte Carlo method. At both time points, the parenclitic-based model had higher sensitivity than the best model generated from standard linear regression (Table 1). In the late group (close to diagnosis), this increase was marginal and the network-topological descriptor that features in the model was dependent on the expression of MUC16. Therefore it is unlikely that at time points closest to diagnosis we would be able to improve greatly upon current models that use this marker. In the networks, MUC16 was very highly connected close to diagnosis in cases, and this was also evident in the modal network (Figure 4A).

Mortality from OC is improved by diagnosis at earlier stages, therefore a test that can detect OC earlier is potentially of significant impact. In the raw data, we found that glypican-1 (GPC1) was the most predictive marker in the early (>34.5 months to diagnosis) group (sensitivity = 16.3%). GPC1 was recently identified as elevated in exosomes prepared from the serum of patients with pancreatic cancer [15] and may be involved in disease progression as a mediator of angiogenesis [16], although interestingly, a reduction of GPC1 was found to be predictive of disease in this case. Whilst the AUC was lower for the parenclitic model than the RDLG model, we did show improved sensitivity (sensitivity = 18.9%), a more clinically relevant measure, within this group and were able to detect a greater number of cases than the RDLG model. The best model found through the network topologies was reduced alpha-centrality (Figure 2D) combined with use of contraceptive pill as a categorical variable. Alpha-centrality is a universal measure of eigenvector centrality and is related to number of degrees (connections). Use of the contraceptive pill is known to reduce the risk of OC, particularly in non-BRCA1/2 mutation carriers [17, 18].

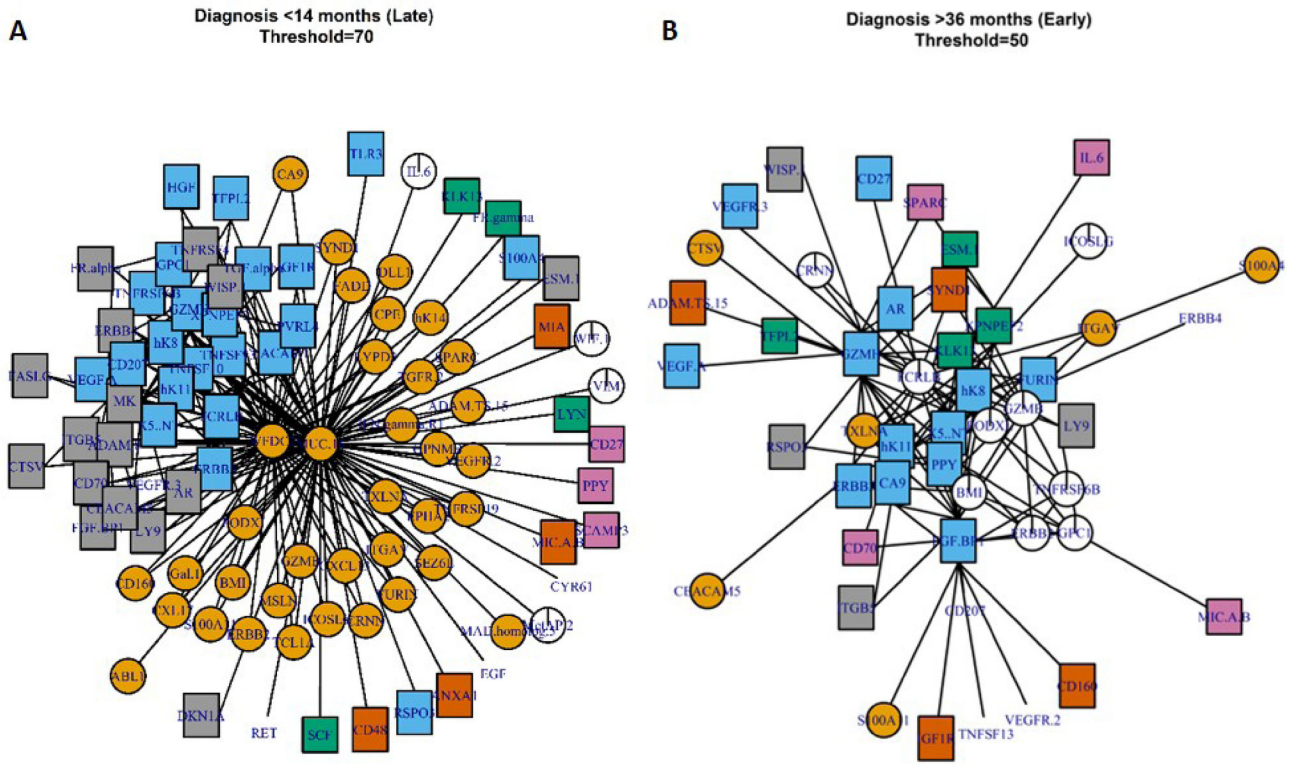


Figure 4: Modal-networks represent the average networks for individuals at <14.5 months to diagnosis (**A**) or >34.5 months to diagnosis (**B**). Colours and shapes represent community membership as determined from clustering by edge-betweenness topology.

In the early group (>34.5 months to diagnosis), we showed that the best parenclitic model and best RDLG model detected different cases, with only a single patient overlapping between them (Figure 3B). This raises the possibility of employing a decision-tree based test, and indeed, doing so within this data set would give a sensitivity of 48.3%. Whilst we do not have sufficient statistical power within this data set to validate this finding, it warrants further investigation.

Early stage OC is often asymptomatic, or symptoms are easily confused with other ailments and therefore it is usually diagnosed at a later stage when symptoms are more pronounced. In order to diagnose OC at earlier stages, a screening program coupled with surgery and chemotherapy may be successful [19]. With screening comes the opportunity to increase the power of statistical models for disease prediction by exploiting longitudinal models that consider an individual's baseline. Combining parenclitic methodologies with longitudinal data analysis may provide further sensitivity. In our synthetic data-set, modelled on CA125 in UKCTOCS, we observed immediate changes in a number of topological indices at the initiation of "cancer" (SF 2). In future, we hope to investigate longitudinal changes in the networks using real data-sets from UKCTOCS.

The power of parenclitic analysis is that as well as being compatible for use in predictive models, it may also be possible to investigate the underlying biology of cancer progression [5]. The advantages with using our parenclitic-distance measure over conventional techniques is that it avoids the use of arbitrary fold-change or *P*-value cut-offs [20]. On the other hand, we do have to make a cut-off for distance thresholds and, although this is somewhat subjective, it is done with the advantage of seeing all the data represented in a single figure, allowing the researcher a greater level of insight when selecting the degree of network complexity. Model network analysis of OC in late stages highlighted MUC16 (CA125) and WFDC2 (HE4) as being responsible for most of the differences within the network (Figure 4A) and this is in line with these being the best reported serological markers of OC [21–24]. In the late-group network, there are two other clusters (blue or grey squares) in which MK is well-connected, although not an obvious hub-centre. Midkine is a growth factor with a diverse role in cell growth and development [25] as well as being a marker and potential mediator of cisplatin resistance [26, 27]. That all three proteins are well-connected within 14 months to diagnosis, but not at 36 months, suggests their involvement in the progression of OC in response to early oncogenic events. Inhibition of MK in a number of tumor cell lines leads to reduced growth [28] and there is growing interest in this protein in a number of applications. The function of MK in ovarian cancer specifically is not well-understood, although its epigenetic modification in ovarian cancer cell lines has been linked to acquired cisplatin resistance [29]. Perhaps further investigation into the function of MK in ovarian cancer would be prudent.

In the early-group modal network, there was no single dominating hub. Whilst there was some enrichment of serine proteases in the connected proteins, the protein pool is small (93 proteins) and manually selected as "cancer-related", therefore, GO-term enrichment analysis provided no useful information. However, GZMH and FGFBP1 are both well connected within the network, indicating their differential expression between cases and controls. GZMH has been reported as upregulated in some OC-tumor infiltrating leukocytes [30] as well as serologically in breast cancer [31] and this may be indicative of enhanced tumor-specific leukocyte activity in the early stages of disease. FBFBP-1, to the knowledge of the authors, has not been reported as elevated during ovarian cancer, indeed staining could not be seen in histological samples of ovarian tissue held by the human protein atlas (https://www.proteinatlas.org/ENSG00000137440-FGFBP1/tissue/ovary) [32]. It is, however, involved in cell proliferation and angiogenesis and is associated with a number of other cancers [33, 34]. Both GZMH and FGFBP1 would be interesting to investigate further in the context of early OC development.

CONCLUSIONS

We have developed a novel method to generate parenclitic networks that uses high-fidelity 2D density distributions to measure the differences between co-variables in a data set. Using topological indices achieved better discrimination of pre-diagnosis OC cases and controls than raw data logistic-regression models. Modal-networks showed differences GZMH and FGFBP1 between cases and controls at >34.5 months prior to diagnosis and they may be relevant in the development of OC. Parenclitic networks can be used in biomarkers models and improve upon the sensitivity of simple, linear regression-based models. In the context of ovarian cancer, we were able to detect, with a sensitivity of 18.3% and a specificity >98%, OC in individuals >34.5 months to diagnosis. Interestingly, the parenclitic-based and RDLG models diagnosed non-overlapping cases at early time points and this raises the option of validating a decision-tree based method in a larger cohort. We aim to validate our current models in larger data sets for early detection of OC and postulate that this approach could be exploited in future data-driven personalised medicine.

MATERIALS AND METHODS

Sample set

A nested set of OC cases and controls were taken from the multi-modal screening arm of UKCTOCS as part of the Predicting Risk of Ovarian Malignancies, Improved Screening and Early detection (PROMISE)

study. 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 was provided. Trial participants at enrolment were post-menopausal women aged 50-74 who had no family history of ovarian cancer. Women subsequently diagnosed with ovarian cancer were identified by cross-referencing with the Health and Social Care Information Centre cancer registry and death codes, with diagnosis confirmed by review of histopathology reports. Paired serum samples from 29 early (>34.5 months to clinical diagnosis) and 30 late (<14.5 months to clinical diagnosis) type II OC cases (mostly high-grade serous) randomised to the multi-modal arm of UKCTOCS. Samples from 30 controls with no history of cancer were matched to cases by age (±5 years), collection date and collection centre. Epidemiological data (OCP use (ever), hormone replacement therapy use (at randomisation), body mass index and age) was available for these women. For more information, see Supplementary Tables 1–2. Single samples from an independent set of 249 non-cancer control women from UKCTOCS were also selected as base controls to build density distributions.

Serum protein measurements

Ninety two cancer-related analytes were measured in individual serum samples using Olink's multiplex immunoassay Oncology II panel. Based on the proximity extension assay [35], this validated platform uses matched antibody pairs linked to DNA reporters. When binding to their target, the pair gives rise to amplicons which are quantified by RT-PCR, providing high sensitivity and accuracy. Known cancer antigens including MUC16, growth factors, receptors, angiogenic factors and adhesion regulators are measured [36]. Data was returned as log2-transformed normalised expression values.

Parenclitic network and model testing

For a detailed description of methods, see supplementary information. In brief, each subject/sample is represented by a graph (network) that includes both continuous and categorical data. Edges are set based on how pairs of covariates differ from a cloud of controls (base controls) by determining the distance based on two dimensional kernel density estimation (2DKDE). The greater the deviation of a pair of covariates from the region of greatest density in the base controls, the greater the distance. The inclusion of categorical-continuous data pairs is achieved by calculating the deviation from a set of base controls of the same categorical value. Once all distances are calculated, a threshold is applied so that only edges between deviating covariates remain, such that each subject/sample is represented by a graph of unique topology. Classification is then performed by combining indices of topology into logistic regression models with performance being assessed by sensitivity at >98% specificity and the area under the ROC curve (AUC). The top 5 parenclitic and RDLG models (selected based on sensitivity and then AUC) were further validated by Monte Carlo cross-validation by splitting the data 50/50 100 times. All analysis was performed in R Studio (1.0.143) running R version 3.4.0. Networks were generated using the igraph package (version 1.0.1).

Modal-network construction

Modal networks were derived by generating parenclitic networks for each sample at a range of thresholds. For each threshold, the networks were combined into a single network where nodes are connected, if they were connected in more than half of the individual networks (i.e. the modal stage of connectedness). The analytes were then clustered using the cluster_edge_betweenness function in the igraph package (version 1.0.1) and community membership determined based on maximum modularity score in the same function.

Abbreviations

2DKDE: 2-dimensional kernel density estimation (2DKDE); OC: ovarian cancer; CA125/MUC 16: cancer antigen 125; UKCTOCS: United Kingdom Collaborative Trial of Ovarian Cancer Screening; RDLG: raw data logistic regression; OCP: oral contraceptive pill; ROC: receiver operating characteristic; AUC: area under the ROC curve.

Author contributions

AZ and JFT conceived and designed the study. UM, JFT and AZ secured grant funding. UM selected and provided samples. HW processed samples and developed parenclitic methods. HW and OB carried out data analysis and model testing. HW, AZ and JFT drafted and edited the manuscript. All authors commented on and approved the final version.

ACKNOWLEDGMENTS

The authors would like to thank Simeon Draganov for proof reading.

CONFLICTS OF INTEREST

UM has a financial interest through UCL Business and Abcodia Ltd in the third party exploitation of clinical trials biobanks which have been developed through the research at UCL. None of the other authors have conflicts of interest.

FUNDING

The study was funded through Cancer Research UK PRC Programme Grant A12677 and by the Eve Appeal Gynaecological Cancer Research Trust through the Predicting Risk of Ovarian Malignancies, Improved Screening and Early detection (PROMISE) study. It was supported by the National Institute for Health Research (NIHR) University College London Hospitals (UCLH) Biomedical Research Centre. UKCTOCS was core-funded by the Medical Research Council, Cancer Research UK, and the Department of Health with additional support from the Eve Appeal, Special Trustees of Bart's and the London, and Special Trustees of UCLH. We thank all the trial participants and all staff involved in the UKCTOCS trial.

REFERENCES

1. Jackson SE, Chester JD. Personalised cancer medicine. Int J Cancer. 2015; 137:262–6. https://doi.org/10.1002/ijc.28940.
2. Da Fontoura Costa L, Oliveira ON, Travieso G, Rodrigues FA, Ribeiro P, Boas V, Antiqueira L, Palhares Viana M, Correa Da Rocha LE. Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications. Adv Phys. 2008; 60:329–412. Available from https://arxiv.org/pdf/0711.3199.pdf.
3. Collins FS, Barker AD. Mapping the Cancer Genome. Sci Am. 2007; 296:50–7. https://doi.org/10.1038/scientificamerican0307-50.
4. Karsakov A, Bartlett T, Ryblov A, Meyerov I, Ivanchenko M, Zaikin A. Parenclitic Network Analysis of Methylation Data for Cancer Identification. Ruan J, editor. PLoS One. Public Library of Science. 2017; 12:e0169661. https://doi.org/10.1371/journal.pone.0169661.
5. Zanin M, Alcazar JM, Carbajosa JV, Paez MG, Papo D, Sousa P, Menasalvas E, Boccaletti S. Parenclitic networks: uncovering new functions in biological data. Sci Rep. Nature Publishing Group. 2014; 4:5112. https://doi.org/10.1038/srep05112.
6. Zanin M. Using complex networks for refining survival prognosis in prostate cancer patient. F1000Research. 2016; 5:2675. https://doi.org/10.12688/f1000research.8282.1.
7. Quaresma M, Coleman MP, Rachet B. 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971–2011: a population-based study. Lancet. 2015; 385:1206–18. https://doi.org/10.1016/S0140-6736(14)61396-9.
8. Cramer DW, Bast RC, Berg CD, Diamandis EP, Godwin AK, Hartge P, Lokshin AE, Lu KH, McIntosh MW, Mor G, Patriotis C, Pinsky PF, Thornquist MD, et al. Ovarian cancer biomarker performance in prostate, lung, colorectal, and ovarian cancer screening trial specimens. Cancer Prev Res (Phila). NIH Public Access. 2011; 4:365–74. https://doi.org/10.1158/1940-6207.CAPR-10-0195.
9. Jacobs IJ, Menon U, Ryan A, Gentry-Maharaj A, Burnell M, Kalsi JK, Amso NN, Apostolidou S, Benjamin E, Cruickshank D, Crump DN, Davies SK, Dawnay A, et al. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. Lancet. Elsevier. 2015; 387:945–56. https://doi.org/10.1016/S0140-6736(15)01224-6.
10. Seaman E, Millspaugh JJ, Kernohan BJ, Brundige GC, Raekede KJ, Gitzen RA. Effects of Sample Size on Kernel Home Range Estimates. J Wildl Manage. 1999; 63:739–47. Available from https://fresc.usgs.gov/products/papers/595_Seaman.pdf.
11. Silverman BW. Density estimation for statistics and data analysis. Chapman and Hall; 1986. 175 p. Available from https://www.crcpress.com/Density-Estimation-for-Statistics-and-Data-Analysis/Silverman/p/book/9780412246203.
12. Jacobs I, Bast RC. The CA 125 tumor-associated antigen: a review of the literature. Hum Reprod. 1989; 4:1–12. Available from http://www.ncbi.nlm.nih.gov/pubmed/2651469.
13. Jacobs IJ, Menon U, Ryan A, Gentry-Maharaj A, Burnell M, Kalsi JK, Amso NN, Apostolidou S, Benjamin E, Cruickshank D, Crump DN, Davies SK, Dawnay A, et al. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. Lancet. 2015; 387:945–56. https://doi.org/10.1016/S0140-6736(15)01224-6.
14. Johnson N. Two large randomised trials show ovarian cancer screening has minimal impact on survival. BJOG. 2016. https://doi.org/10.1111/1471-0528.14052.
15. Melo SA, Luecke LB, Kahlert C, Fernandez AF, Gammon ST, Kaye J, LeBleu VS, Mittendorf EA, Weitz J, Rahbari N, Reissfelder C, Pilarsky C, Fraga MF, et al. Glypican-1 identifies cancer exosomes and detects early pancreatic cancer. Nature. 2015; 523:177–82. https://doi.org/10.1038/nature14581.
16. Aikawa T, Whipple CA, Lopez ME, Gunn J, Young A, Lander AD, Korc M. Glypican-1 modulates the angiogenic and metastatic potential of human and mouse cancer cells. J Clin Invest. 2008; 118:89–99. https://doi.org/10.1172/JCI32412.
17. Ferris JS, Daly MB, Buys SS, Genkinger JM, Liao Y, Terry MB. Oral contraceptive and reproductive risk factors for ovarian cancer within sisters in the breast cancer family registry. Br J Cancer. 2014; 110:1074–80. https://doi.org/10.1038/bjc.2013.803.
18. Roberts HE. Gynaecological cancer and the contraceptive pill. Obstet Gynaecol. 2004; 6:75–9. https://doi.org/10.1576/toag.6.2.75.26981.
19. Narod SA, Sopik V, Giannakeas V. Should we screen for ovarian cancer? A commentary on the UK Collaborative

Trial of Ovarian Cancer Screening (UKCTOCS) randomized trial. Gynecol Oncol. 2016; 141:191–4. https://doi.org/10.1016/j.ygyno.2016.02.020.

20. Amrhein V, Korner-Nievergelt F, Roth T. The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. PeerJ. 2017; 5:e3544. https://doi.org/10.7717/peerj.3544.
21. Li J, Dowdy S, Tipton T, Podratz K, Lu WG, Xie X, Jiang SW. HE4 as a biomarker for ovarian and endometrial cancer management. Expert Rev Mol Diagn. 2009; 9:555–66. https://doi.org/10.1586/erm.09.39.
22. Van Gorp T, Cadron I, Despierre E, Daemen A, Leunen K, Amant F, Timmerman D, De Moor B, Vergote I. HE4 and CA125 as a diagnostic test in ovarian cancer: prospective validation of the Risk of Ovarian Malignancy Algorithm. Br J Cancer. Cancer Research UK. 2011; 104:863–70. https://doi.org/10.1038/sj.bjc.6606092.
23. Hellström I, Raycraft J, Hayden-Ledbetter M, Ledbetter JA, Schummer M, McIntosh M, Drescher C, Urban N, Hellström KE. The HE4 (WFDC2) protein is a biomarker for ovarian carcinoma. Cancer Res. 2003; 63:3695–700. Available from http://www.ncbi.nlm.nih.gov/pubmed/12839961.
24. Moore RG, McMeekin DS, Brown AK, DiSilvestro P, Miller MC, Allard WJ, Gajewski W, Kurman R, Bast RC, Skates SJ. A novel multiple marker bioassay utilizing HE4 and CA125 for the prediction of ovarian cancer in patients with a pelvic mass. Gynecol Oncol. 2009; 112:40–6. https://doi.org/10.1016/j.ygyno.2008.08.031.
25. Muramatsu T. Midkine, a heparin-binding cytokine with multiple roles in development, repair and diseases. Proc Japan Acad Ser B. 2010; 86:410–25. https://doi.org/10.2183/pjab.86.410.
26. Zhang D, Ding L, Li Y, Ren J, Shi G, Wang Y, Zhao S, Ni Y, Hou Y. Midkine derived from cancer-associated fibroblasts promotes cisplatin-resistance via up-regulation of the expression of lncRNA ANRIL in tumor cells. Sci Rep. 2017; 7. https://doi.org/10.1038/s41598-017-13431-y.
27. Wu X, Zhi X, Ji M, Wang Q, Li Y, Xie J, Zhao S. Midkine as a potential diagnostic marker in epithelial ovarian cancer for cisplatin/paclitaxel combination clinical therapy. Am J Cancer Res. 2015; 5:629–38. Available from http://www.ncbi.nlm.nih.gov/pubmed/25973302.
28. Masui M, Okui T, Shimo T, Takabatake K, Fukazawa T, Matsumoto K, Kurio N, Ibaragi S, Naomoto Y, Nagatsuka H, Sasaki A. Novel Midkine Inhibitor iMDK Inhibits Tumor Growth and Angiogenesis in Oral Squamous Cell Carcinoma. Anticancer Res. 2016; 36:2775–81. Available from http://www.ncbi.nlm.nih.gov/pubmed/27272788.
29. Zeller C, Dai W, Steele NL, Siddiq A, Walley AJ, Wilhelm-Benartzi CSM, Rizzo S, van der Zee A, Plumb JA, Brown R. Candidate DNA methylation drivers of acquired cisplatin resistance in ovarian cancer identified by methylome and expression profiling. Oncogene. 2012; 31:4567–76. https://doi.org/10.1038/onc.2011.611.
30. Chifman J, Pullikuth A, Chou JW, Bedognetti D, Miller LD. Conservation of immune gene signatures in solid tumors and prognostic implications. BMC Cancer. 2016; 16:911. https://doi.org/10.1186/s12885-016-2948-z.
31. Tahbaz-Lahafi B, Akbari ME, Amir-Rassouli H, Rahimipour A, Azargashb E, Nafissi N, Jahani F. Granzyme h serum levels variations with both reproductive hormone receptors, and related hormone receptors in breast cancer patients. Iran J cancer Prev. 2014; 7:48–52. Available from http://www.ncbi.nlm.nih.gov/pubmed/25250148.
32. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015; 347:1260419. https://doi.org/10.1126/science.1260419.
33. Tassi E, Henke RT, Bowden ET, Swift MR, Kodack DP, Kuo AH, Maitra A, Wellstein A. Expression of a Fibroblast Growth Factor–Binding Protein during the Development of Adenocarcinoma of the Pancreas and Colon. Cancer Res. 2006; 66:1191–8. https://doi.org/10.1158/0008-5472.CAN-05-2926.
34. Ray PE, Al-Attar A, Liu XH, Das JR, Tassi E, Wellstein A. Expression of a Secreted Fibroblast Growth Factor Binding Protein-1 (FGFBP1) in Angioproliferative Kaposi Sarcoma. J AIDS Clin Res. 2014; 5. https://doi.org/10.4172/2155-6113.1000309.
35. Assarsson E, Lundberg M, Holmquist G, Björkesten J, Bucht Thorsen S, Ekman D, Eriksson A, Rennel Dickens E, Ohlsson S, Edfeldt G, Andersson AC, Lindstedt P, Stenvang J, et al. Homogenous 96-Plex PEA Immunoassay Exhibiting High Sensitivity, Specificity, and Excellent Scalability. Hoheisel JD, editor. PLoS One. 2014; 9:e95192. https://doi.org/10.1371/journal.pone.0095192.
36. Mahboob S, Ahn SB, Cheruku HR, Cantor D, Rennel E, Fredriksson S, Edfeldt G, Breen EJ, Khan A, Mohamedali A, Muktadir MG, Ranganathan S, Tan SH, et al. A novel multiplexed immunoassay identifies CEA, IL-8 and prolactin as prospective markers for Dukes' stages A-D colorectal cancers. Clin Proteomics. 2015; 12:10. https://doi.org/10.1186/s12014-015-9081-x.

Parenclitic networks for predicting ovarian cancer

SUPPLEMENTARY MATERIALS

Supplementary information 1: Algorithm

Building and testing of models in R includes the following steps

1. We subset data including $N_{individuals}$ with N variables/parameters into two sets:

a. set I, a set of controls to construct 2DKDE for all pairs of analytes;

b. set II, a set of controls and cases to build the network and topological models.

2. Training set I is used to construct the 2DKDE matrix filled for every pair of analytes by calling the kde2d function from the Massx package. The size of the matrix is chosen to be $N_{bins} \times N_{bins,}$ $N_{bins} = 16$ but 2DKDE values are known to be independent of this parameter [1]. We generate $N_{bins} \times N_{bins}$ contour matrices.

3. The networks are generated for each individual in a data frame of set II. This network is a graph formed by N nodes/analytes and binary edges between them. Edges can be between continuous-continuous, continuous-categorical, or categorical-categorical parameters (not used in this analysis). In each, graph nodes are formed by analytes, continuous or categorical, and the binary edge is calculated as following:

a. For each individual, the value of the edge between two continuous analytes (i, j), $i, j = 1, ..., N$ forming nodes is calculated from the contourmatrix for this pair of parameters. Here we have two cases:

i. If the sample, i.e., a pair of parameters (x_i, y_j) from data set II corresponding to an individual, is within the matrix of the density estimation, the edge is equal to the volume under the densities that are higher than the density where the sample is within the distribution.

ii. We estimate distance as the normalised ratio between two distances, $\frac{D_{max}}{D_{edge}}$ where $D_{max} = \sqrt{(x_i - x_{max})^2 + (y_j - y_{max})^2}$ and is the distance between the nearest point on the contour matrix $(x_{nearest,} y_{nearest})$ which lies on the straight line between the point with the maximal density $(x_{max,} y_{max})$ and investigated sample point (x_i, j) and is the closest point to (x_i, y_j).

Note that alternatively this distance can be estimated from the linear regression (original version of the parenclitic analysis [2, 3]) or from the Mahalonobic distance [4]. Using the linear regression for each pair of analytes, m_i and m_j, we build a linear regression based on the control group set I:

$$m_j = \alpha_{i,j} + \beta_{i,j} x m_i \quad (4)$$

where $\alpha_{i,j}$ and $\beta_{i,j}$ are regression coefficients. Next we build complete weighted graph for each cancer-positive and negative sample from set II, excluding the control group of set I, such that each vertex corresponds to a particular analyte, and edges are weighted according to

$$w_{i,j} = \frac{|x_j - (\alpha_{i,j} + \beta_{i,j} x_i)|}{\sigma_{i,j}} \quad (5)$$

where x_i and x_j are respective analyte levels for the sample (i, j), and $\sigma_{i,j}$ is the standard deviation of errors in the linear regression model for control objects. In the second case, we implement the Mahalanobis distance [4], which, essentially, measures separation between data sets. In particular, the edge weight is as:

$$w_{i,j} = \sqrt{(x_{i,j} - \mu_{i,j})^T S_{i,j}^{-1} (x_{i,j} - \mu_{i,j})} \quad (6)$$

where, as before, x_i and x_j are analyte levels in an investigated sample, μ_i and μ_j are the analyte levels in the control data set I, $(x_{i,j} = x_i, x_j)$, $\mu_{i,j} = (E(m_i), E(m_j))$ and $S_{i,j} = cov(m_i, m_j)$, where $E(.)$ is the expectation and the $cov(.)$ covariance.

b. Distances between continuous-categorical variables are calculated by generating a density estimation of the continuous data using points from training data set I with the value of the categorical variable equal to the value of the investigated individual. It means that we estimate the deviation only from the corresponding group characterised by the same value of the categorical. Then the distance is calculated as above.

c. Edges between categorical data points are not used in this version but can be potentially calculated from a density estimation of the continuous data by selecting a significant pair of variables (e.g. most predictive pair by AUC) and using the points from training data set I with the value of the categorical equal to these values of the investigated individual. Again in this way we estimate the distance from the group with the same categorical only evaluated on the plane of the most predictive continuous analytes.

d. A threshold is applied to the distances and they are coded 1 (higher than threshold) or 0. A copy of the

network is saved in a new data frame (distance) that contains only those that are greater than the threshold value. This is used for calculating some of the topological indexes later.

4. Each individual is represented by calculating the topological indices for the original and binary graph, both representing the same patient. The number of edges in constructed graphs makes a straightforward machine learning classification task intractable in practice, also manifesting a huge imbalance between the number of features and available samples. Therefore, we utilize a number of topology metrics, widely used for characterising complex networks [5–7] a network of chemicals linked by chemical reactions, and the Internet, a network of routers and computers connected by physical links. While traditionally these systems have been modeled as random graphs, it is increasingly recognized that the topology and evolution of real networks are governed by robust organizing principles. This article reviews the recent advances in the field of complex networks, focusing on the statistical mechanics of network topology and dynamics. After reviewing the empirical data that motivated the recent interest in networks, the authors discuss the main models and analytical tools, covering random graphs, small-world and scale-free networks, the emerging theory of evolving networks, and the interplay between topology and the network's robustness against failures and attacks. A. Quantities of interest in percolation theory 60 B. General results 60 1. The subcritical phase (p\u03fdp c, appropriately generalised for weighted graphs $G := (V, E)$ with |E| vertices and edges:

a. Mean, standard deviation and maximal values of edge weights.
b. Mean, standard deviation and maximal values of vertex degree.
c. Mean, variance and maximal values of shortest path lengths.
d. Alpha-centrality
e. Mean, variance, and maximal values of the page rank.
f. The distance $d(v_i,v_j)$ between the nodes $v_{i,j} \in V$, defined as the sum of the edge weights in the shortest path.
g. The diameter of the graph G as the maximal distance between a pair of vertices.
h. The degree centrality $C_D(G)$ of the graph defined as the normalised graph (G) degree centrality $H(G)$

$$C_D(G) = \frac{H(G)}{H_{max}} \quad (7)$$

which is

$$H(G) = \sum_{i=1}^{|v|} | C_D(v^*) - C_D(v_i) | \quad (8)$$

based on the node degree centrality $C_D(v) = \frac{\deg(v)}{|V|}$, and where v^* is the node with the maximal degree centrality, and $H_{max} = (|V|-1)(|V|-2)$ is the maximal graph degree centrality, obtained for the star topology.

i. Graph or network efficiency [8] $E_c(G)$ defined as

$$E_C(G) = \frac{C_C(G)}{|V|(|V|-1)} \quad (9)$$

based on the graph centrality measure $C_c(G) = \sum_{i \neq j} \frac{1}{d(v_i, v_j)}$

j. Betweenness centrality $C_B(vk)$ of a node as the number of the shortest paths the particular node belongs to:

$$C_B(vk) = \sum_{k \neq i \neq j} \frac{\sigma_{v_i v_j}(vk)}{\sigma_{v_i, v_j}} \quad (10)$$

where $\sigma_{vi,vj}$ is the number of the shortest paths between the nodes v_i and v_j, among which $\sigma_{vi,vj}(vk)$ passing through vk.

k. Google maximal page rank index.

l. Robustness. This is the number of nodes that need to be removed for the network to be unconnected.

These quantities in one way or another should reflect the expected differences between the sample classes. For instance, increasing separation of data sets produces greater edge weights and may result in substantial decrease of the graph diameter. Likewise, nodes with large centrality scores signify their key role in class distinction and give us information on biological importance.

5. Topological indexes are calculated for networks at a range of thresholds (see 3. d.). The threshold at which each topological index provides the best classification of cases and controls (by AUC) in set II are collated and used to generate logistic regression models.

6. The best combination for logistic regression is chosen by comparing all possible combinations. To avoid overfitting the number of indices included in the model is found by forward-backward elimination or limited by a threshold. Categorical variables can be added as extra features for generating the logistic regression.

Supplementary information 2: generation of synthetic data

Data was generated to mimic the behaviour of ovarian cancer biomarkers obtained in the UKCTOCS screening trial [9]. The best biomarker currently used for ovarian cancer detection is CA125, and has been shown to follow a Bayesian Hierarchical Changepoint model [10]. The model is based on the principle that each woman without cancer has her own baseline level of CA125 and all her CA125 values fluctuate around this level. Assume that we have a set of N controls indexed as $i = 1,\ldots,N$.

The number of samples collected for each control is not necessarily the same so we use k_i to denote the number of data points for the *i*-th patient. The *j*-th sample for the *i*-th patient is denoted $Y_{i,j}$ and is taken at $Age_{i,j}$. The model for controls is specified as $Y_{i,j} \sim N(\theta_i, \theta^2)$. with constant mean $E(Y_{i,j} \mid Age_{i,j}) = \theta_i$ and variance σ^2. Previous studies indicate that 85% of cases of ovarian cancer generate a rise in CA125. Therefore the parameter I_i is included in the model for cases, which indicates whether a particular case produces an elevated CA125 ($I_i = 1$) or not ($I_i = 0$). The model for cases with $I_i = 0$ coincides with the model for controls, while in cases with increased CA125, marker levels increase sharply over time after the change-point τ_i at rate γ_i. Thus the mean for cases is modeled by a piecewise linear function:

$$E\left(Y_{i,j|Age_{i,j},I_i=1}\right) = \theta_i + \gamma_i (Age_{i,j} - \tau_i)^+$$

where $(...)^+$ is the positive part of the expression.

Three artificial biomarkers were generated: one followed a hierarchical distribution with the same parameters and prior distributions as for CA125 and two others followed the same distributions with slightly different priors. All parameters are fully described in [10].

Controls were generated from $N(\theta_i, \sigma^2)$, where $\theta_i \sim N(\mu_\theta, \sigma_\theta^2)$, $\mu_\theta \sim N(m_\theta, v_\theta^2)$, $\sigma_\theta^2 \sim IG(a_\theta, b_\theta)$, $\sigma^2 \sim IG(a, b)$, Here $m_\theta = \{2.75, 3.05, 3.72\}$, $v_\theta = \{0.1, 0.082, 0.04\}$, $a_\theta = \{2.04, 2.04, 2.04\}$, $b_\theta = \{0.065, 0.065, 0.065\}$, $a = \{13.95, 13.95, 13.95\}$, $b = \{0.97, 0.97, 0.98\}$, for artificial CA125 and two other artificial markers correspondingly. Biomarker behaviour in cases is the same as in controls apart from some extra parameters: $\log(\gamma_i) \sim N(\mu_\gamma, \sigma_\gamma^2)$, $\mu_\gamma \sim N(m_\gamma, v_\gamma^2)$, , and probability of a changepoint π_i. Here, $\mu_\gamma = \{0.4, 0.248, 0.218\}$, $\theta_\gamma^2 = 0.165, 0.172, 0.141$, $a_\gamma = \{2.8, 2.8, 2.8\}$, $b_\gamma = \{1.58, 1.58, 1.58\}$, $\pi_i = \{0.85, 0.75, 0.65\}$ for the three markers respectively.

Based on these distributions, we have generated serial data for 3 artificial biomarkers: one mimics CA125 and two others mimic possible ovarian cancer biomarkers. 27 other "random" analytes were also generated. A further 9 categorical variables were also included: 1 was predictive in 80% of cases, 2 were predictive in 60% of cases with the rest having no disease-association. In total, 300 patients were generated, including 150 cases and 150 controls. Every patient had 9 annual serial measurements (within a 9 year window) for each of the artificial biomarkers.

REFERENCES

1. Venables WN, Ripley BD. Modern Applied Statistics with S. Springer; 2002.
2. Zanin M, Alcazar JM, Carbajosa JV, Paez MG, Papo D, Sousa P, Menasalvas E, Boccaletti S. Parenclitic networks: uncovering new functions in biological data. Sci Rep. Nature Publishing Group. 2014; 4:5112. https://doi.org/10.1038/srep05112.
3. Karsakov A, Bartlett T, Ryblov A, Meyerov I, Ivanchenko M, Zaikin A. Parenclitic Network Analysis of Methylation Data for Cancer Identification. Ruan J, editor. PLoS One. Public Library of Science. 2017; 12:e0169661. https://doi.org/10.1371/journal.pone.0169661.
4. Huberty CJ, Huberty, J. C. Mahalanobis Distance. Wiley StatsRef: Statistics Reference Online. Chichester, UK: John Wiley & Sons, Ltd. 2014. https://doi.org/10.1002/9781118445112.stat06485.
5. ka Albert R, szló Barabá si AL. Statistical mechanics of complex networks. 2002. Available from http://barabasi.com/f/103.pdf.
6. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D. Complex networks: Structure and dynamics. Phys Rep. 2006; 424:175–308. https://doi.org/10.1016/j.physrep.2005.10.009.
7. Freeman LC. Centrality in social networks conceptual clarification. Soc Networks. North-Holland. 1978; 1:215–39. https://doi.org/10.1016/0378-8733(78)90021-7.
8. Latora V, Marchiori M. Efficient Behavior of Small-World Networks. Phys Rev Lett. 2001; 87:198701. https://doi.org/10.1103/PhysRevLett.87.198701.
9. Jacobs IJ, Menon U, Ryan A, Gentry-Maharaj A, Burnell M, Kalsi JK, Amso NN, Apostolidou S, Benjamin E, Cruickshank D, Crump DN, Davies SK, Dawnay A, et al. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. Lancet. Elsevier. 2015; 387:945–56. https://doi.org/10.1016/S0140-6736(15)01224-6.
10. Skates SJ, Pauler DK, Jacobs IJ. Screening Based on the Risk of Cancer Calculation from Bayesian Hierarchical Changepoint and Mixture Models of Longitudinal Markers. Journal of the American Statistical Association. Taylor & Francis, Ltd.American Statistical Association. 429–39. https://doi.org/10.2307/2670281.

Supplementary Table 1: Demographics and clinical pathology of the ovarian cancer cohort

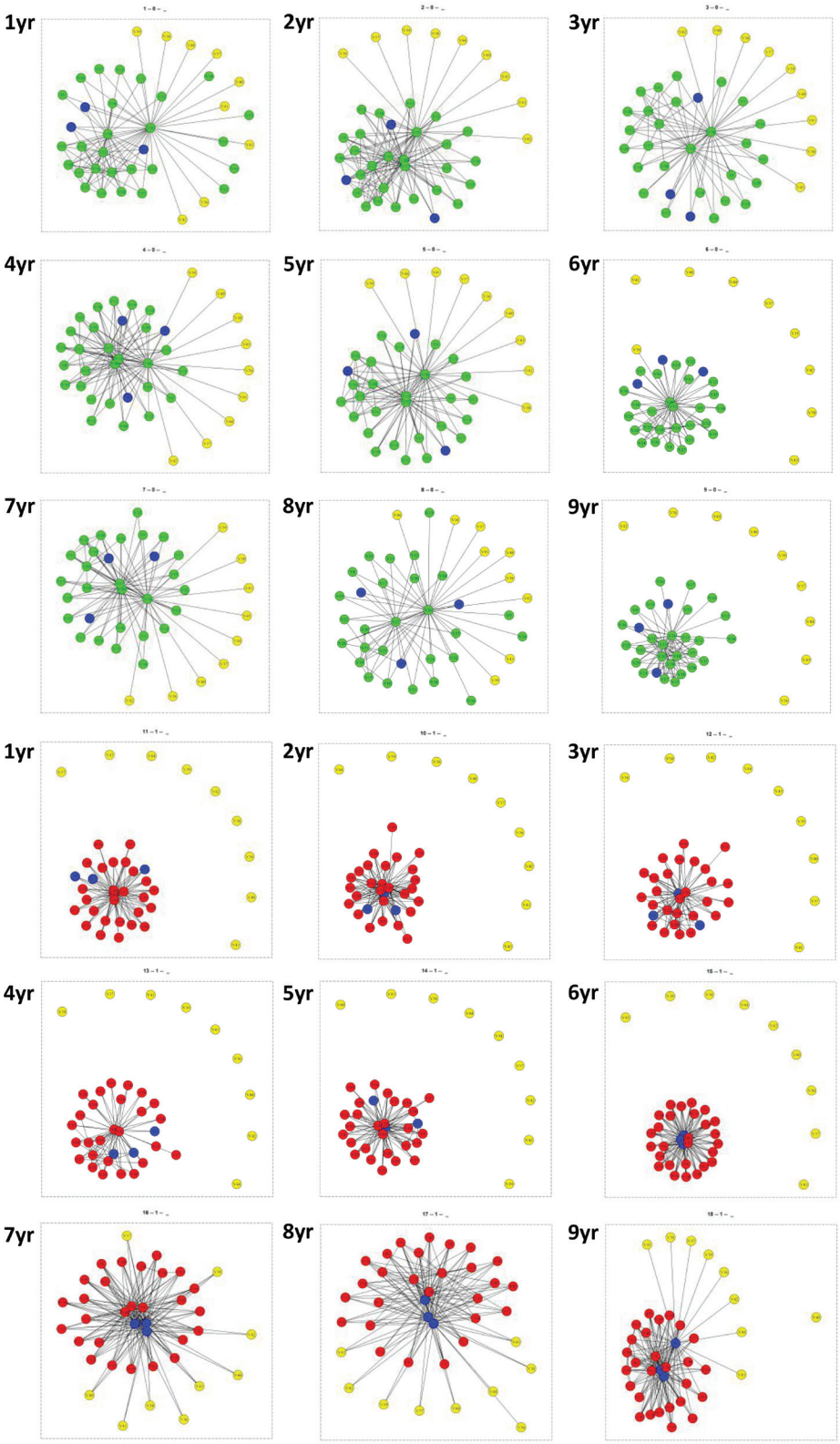
	Cases	Controls	*P*
No. individuals	30	30	
No. samples	59	59	
Median age at sample draw–yrs (range)	65.2 (50.6–79.7)	64.6 (51.4–78.6)	n.s.
Median time to spin–hrs (range)	22.1 (2.6–46.8)	22.6 (1.7–47.9)	n.s.
Median time to diagnosis–months (range)	14.3 (2.9–97.5)		
Median BMI at recruitment–kg/m^2 (range)	25.0 (19.2–43.2)	24.8 (20.2–40.1)	n.s.
HRT use at recruitment (no.)	6	8	n.s.
OCP use at recruitment (no.)	12	18	n.s.
Grade and morphology			
High grade serous	23		
High grade endometrioid	3		
High grade not specified	3		
High grade carcinosarcoma	1		
FIGO Stage			
I	7		
II	9		
III	14		

Not significant, n.s..

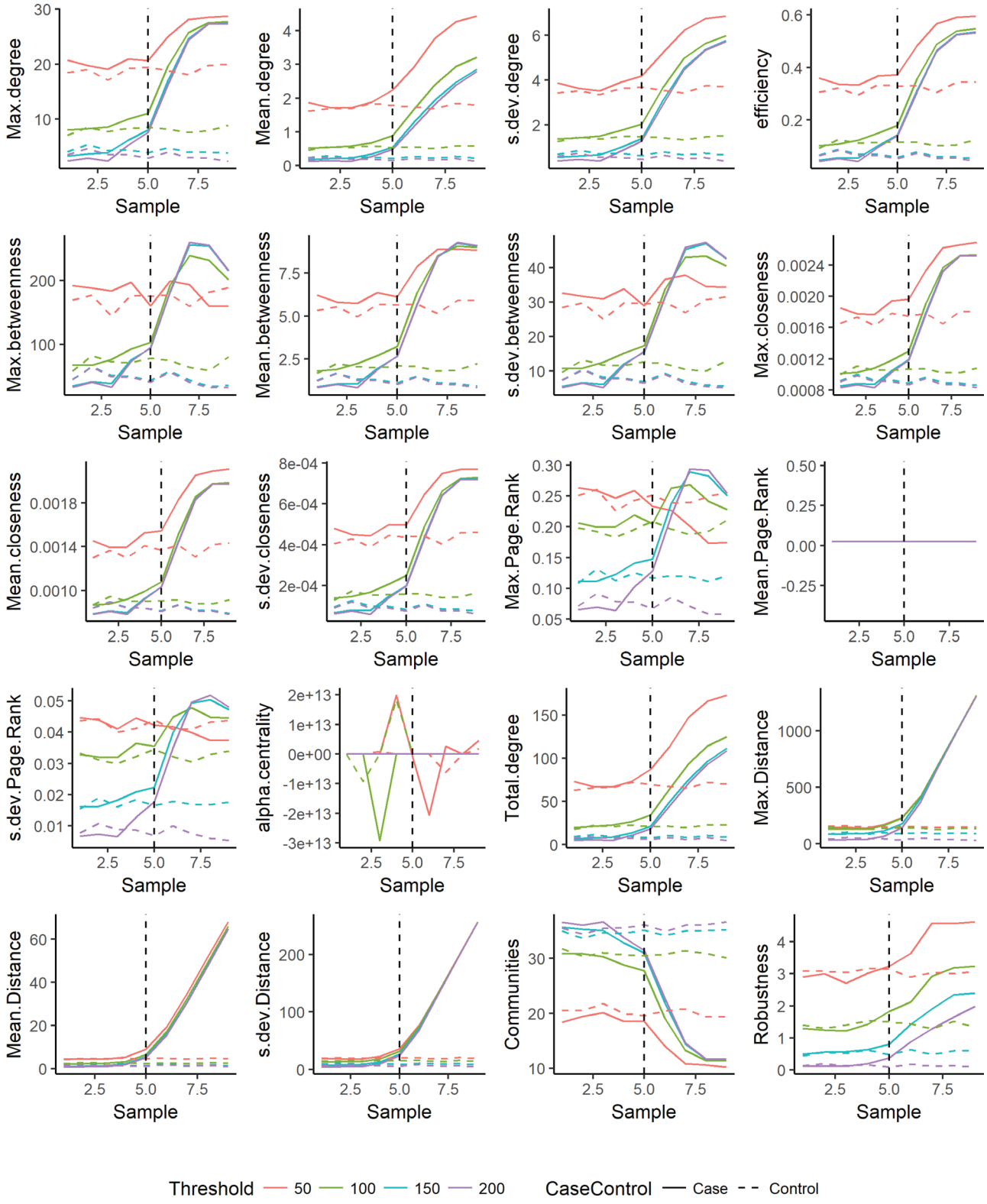
Supplementary Table 2: Demographics and clinical pathologies of ovarian cancer cases, separated into late (<14.5 months to diagnosis) and early (>34.5 months to diagnosis) groups

	Cases (Late)	Cases (Early)	*P*
No. individuals	30	29	
No. samples	30	29	
Median age at sample draw–yrs (range)	66.9 (55.8–79.7)	62.3 (50.6–55.8)	n.s.
Median time to diagnosis–months (range)	3.8 (2.86–14.3)	58.7 (34.9–97.5)	n.s.
Median BMI at recruitment–kg/m^2 (range)	25.1 (19.2–43.2)	25.0 (19.2–43.2)	n.s.
HRT use at recruitment (no.)	6	6	n.s.
OCP use at recruitment (no.)	12	12	n.s.
Grade and morphology			
High grade serous	23	22	
High grade endometrioid	3	3	
High grade not specified	3	3	
High grade carcinosarcoma	1	1	
FIGO Stage			
I	7	7	
II	9	9	
III	14	14	

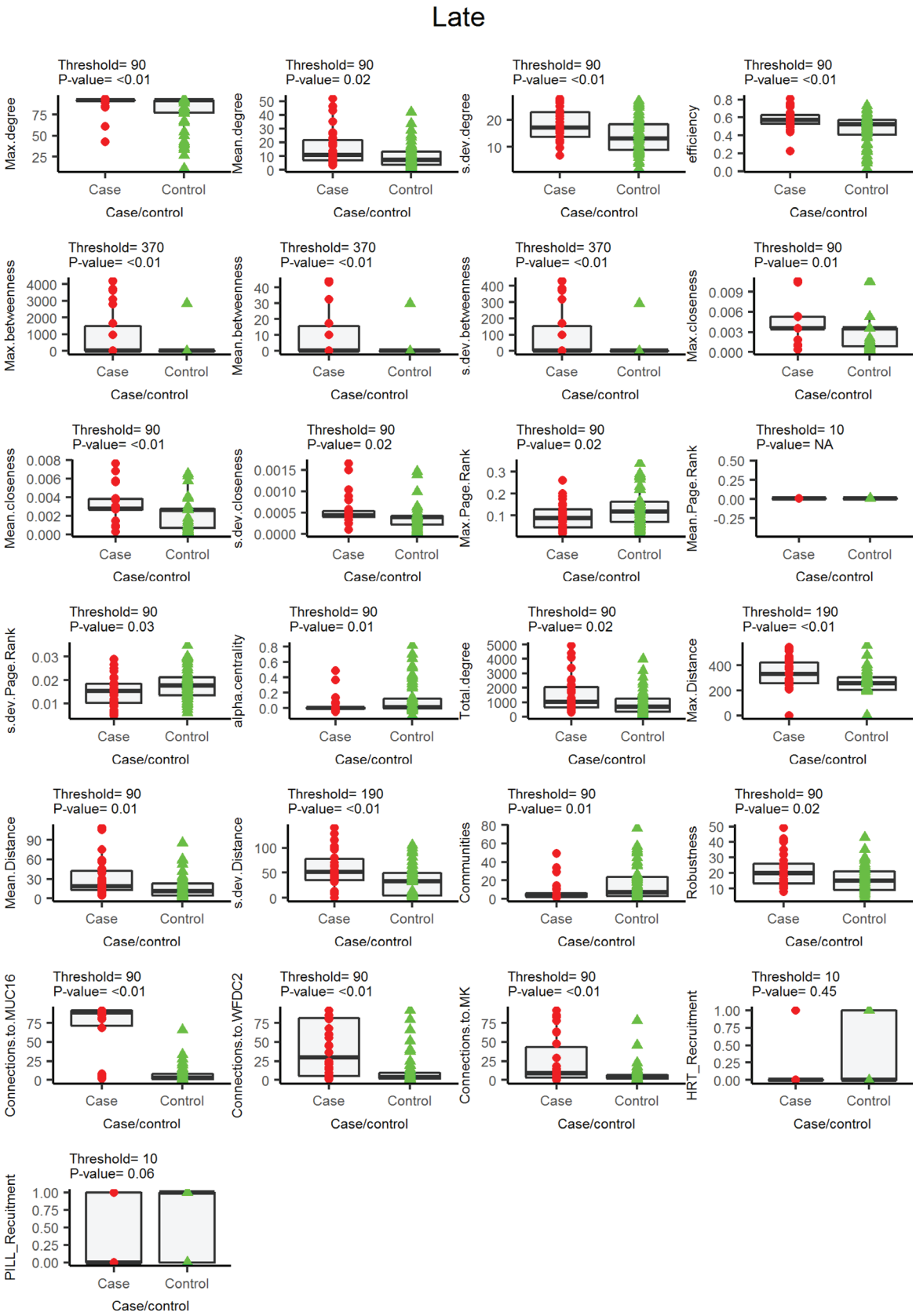
Not significant, n.s..



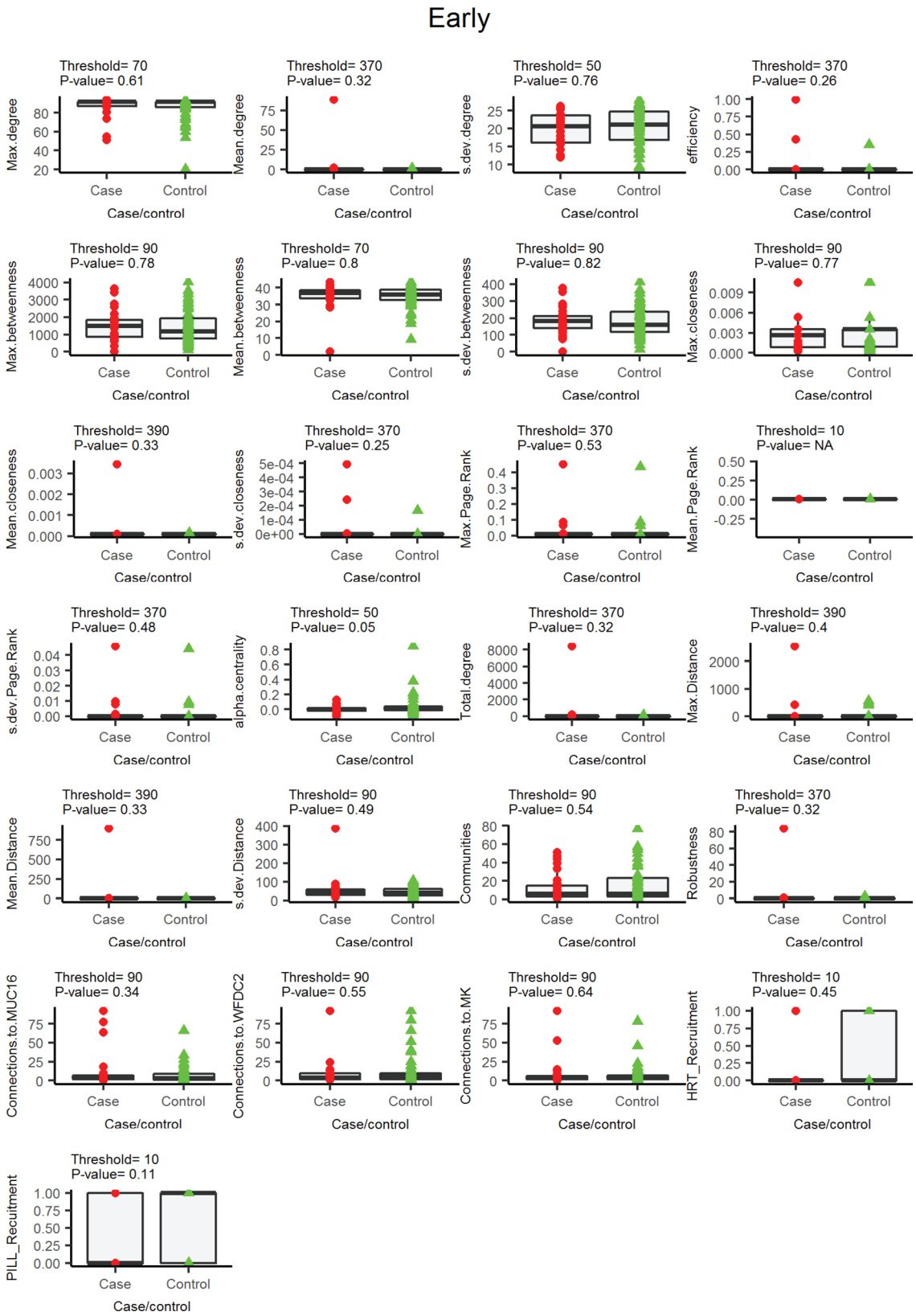
Supplementary Figure 1: Parenclitic networks were generated for "controls" and "cases" from a synthetic data set. From years 1–4, all markers behaved randomly. At year 5, cases developed cancer and markers were modelled for a further 4 years. One marker in the data set mimicked CA125 and 3 others were partially predictive (blue nodes) (for full description, see Supplementary Information). Yellow nodes represent categorical data, green nodes are random markers in controls and red nodes are random markers in cases. Presented is a representative case and control with networks generated from years 1–9. In cases, but not controls, there is an increase in the number of connected nodes after the onset of cancer, furthermore, predictive nodes (blue) become more heavily connected and move towards the centre of the networks.



Supplementary Figure 2: Parenclitic networks generated from synthetic data were described by 20 topological indexes at 4 thresholds. Each individual had 9 samples with cancer being initiated in cases at year 5 (vertical dashed line). Presented are the mean topological value for each individual over their 9 samples. Cases are shown as solid lines and controls as dashed lines, colours indicate different thresholds. For further information as to how the thresholds are applied and indices generated, see Supplementary Information. At the initiation of "cancer" in cases, topological features change. For some thresholds, the magnitude of the change is different and thus there is no single threshold that fits all topological features.



Supplementary Figure 3: In late samples (<14.5 months to diagnosis), parenclitic analysis was performed across a range of thresholds and each network described by a number of topological features. The best threshold for each topology was determined by comparing AUC values. This figure shows the topology values at its optimal threshold for cases and controls from the OC data set. *P*-values were determined by Wilcoxon test.



Supplementary Figure 4: In early samples (>36 months to diagnosis), parenclitic analysis was performed across a range of thresholds and each network described by a number of topological features. The best threshold for each topology was determined by comparing AUC values. This figure shows the topology values at its optimal threshold for cases and controls from the OC data set. *P*-values were determined by Wilcoxon test.