Heliyon

Received:
13 August 2015
Revised:
16 December 2015
Accepted:
23 December 2015

Heliyon (2016) e00059

Cytokines and cell adhesion molecules exhibit distinct profiles in health, ovarian cancer, and breast cancer

Matthew P.A. Henderson [b], **Holger Hirte** [c], **Sebastien J. Hotte** [c], **Peter A. Kavsak** [a,*]

[a] *Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Canada*

[b] *Department of Pathology and Laboratory Medicine, University of Ottawa, Ottawa, Canada*

[c] *Juravinski Cancer Centre, Medical Oncology, Hamilton Ontario, Canada*

* Corresponding author at: Hamilton Regional Laboratory Medicine Program, Juravinski Hospital and Cancer Centre (Core Lab Section), 711 Concession Street, Hamilton, ON, Canada.
E-mail address: kavsakp@mcmaster.ca (P.A. Kavsak).

Abstract

Objective: We examined a panel of cytokines and cell adhesion molecules in an attempt to identify cancer specific profiles.
Design and methods: Cytokines and cell adhesion arrays (Randox Ltd.) were measured in samples from women with a histological diagnosis of ovarian cancer ($n = 42$) or breast cancer ($n = 60$) or cancer free ($n = 32$). Random forest analysis was used for classification.
Results: Ovarian cancer subjects were classified with a sensitivity of 85.7% (95% CI 50–100) and a specificity of 84.2% (95% CI 69.4–93.4). Breast cancer subjects were classified with a sensitivity of 70.8% (95% CI 47.1–86.4) and a specificity of 96.4% (95% CI 82.1–100).
Discussion: Cytokine and cell adhesion molecule profiles provide additional information that may be useful for cancer characterization of female cancers.

Keywords: Statistics, Laboratory medicine, Medical informatics, Cancer diagnostics

http://dx.doi.org/10.1016/j.heliyon.2015.e00059

1. Introduction

Clinical proteomics and bioinformatics have spurred increased research in the field of cancer biomarkers. The ideal biomarker would indicate both the presence of malignancy as well as the identity of the tissue of origin. We examined a panel of cytokines and cell adhesion molecules (CAM) using previously characterised biochip arrays in an attempt to identify breast and ovarian cancer specific profiles [1].

Cytokines are a diverse group of proteins comprised of hematopoietic growth factors, interferons, lymphokines, and chemokines [2]. Cytokines act as mediators of cell-to-cell communication. Uncontrolled cytokine expression contributes to: tumour growth and metastasis, immunosuppression and angiogenesis. Cytokine expression is not cancer specific and can be up-regulated during inflammation and wound repair [2]. CAMs are cell surface proteins involved in cell-to-cell and cell-to-extracellular matrix interactions. Altered CAM expression contributes to: tumour cell motility, tumour cell invasion and angiogenesis. CAM expression is not cancer specific and can be induced by cytokines [3].

Among women, breast cancer is the most commonly diagnosed cancer after non-melanoma skin cancer, and it is the second leading cause of cancer deaths after lung cancer. In 2015, an estimated 231,840 new cases will be diagnosed, and 40,290 deaths from breast cancer will occur in the United States [4]. In American women, ovarian cancer is the ninth most common cancer, with an estimated 21,290 new cases in 2015, but is the fifth most deadly, with an estimated 14,180 deaths in 2015 [4]. The availability of good biomarkers could assist in early detection, which in turn could contribute to improved prognosis.

Although expression of cytokines and CAMs is not cancer specific we set out to determine if there are breast and ovarian cancer specific cytokine and CAM plasma profiles. These profiles could then be leveraged to determine a patient's cancerous tissue of origin. The random forest algorithm was used to address this question because it is resistant to over-fitting, provides estimates of variable importance and generates a classifier that can be applied to future data sets [5].

2. Methods

2.1. Samples and biochemical analysis

After ethics approval by the Hamilton Health Sciences research ethics board, EDTA plasma samples were obtained from cancer free women ($n = 32$; after informed consent), and the Ontario Tumour Bank (OTB), [1, 6]. All cases are

Table 1. Tumour staging (TNM system) of participants in this study.

Disease	Progression						
	ND	T0	T1	T2	T3	T4	TX
Breast	0	0	17	34	4	5	0
Healthy	32	0	0	0	0	0	0
Ovarian	0	3	10	8	17	0	4

pathologically verified and have clinical information, such as histology, patient history, and systemic therapy. The EDTA plasma samples obtained from the OTB were from women with a histological diagnosis of ovarian cancer ($n = 42$) or breast cancer ($n = 60$). Tumour staging (*TNM system*) of participants in this study is summarized in Table 1. Samples were measured for cytokines: IL-1 α, IL-1 β, IL-2, IL-4, IL-6, IL-8, IL-10, VEGF, MCP-1, EGF, TNF-α, and IFN-γ using the "Cytokine Array 1" and adhesion molecules: VCAM-1, ICAM-1, P-selectin, E-selectin, and L-selectin using the "Cell Adhesion Molecule Array" on the Evidence Investigator™ platform (Randox Ltd.). The analytical performance of these assays has been described previously [1].

2.2. Training the random forest algorithm

The patients were randomly assigned to the test (40%) or training (60%) data set. Random forest analysis of the training set was used to identify important variables for classification of the test set (Figure 3). The random forest was trained to classify patient samples into one of four classes: breast cancer, ovarian cancer, cancer free or unknown based on the concentration of cytokines and cell adhesion molecules in the training set. The random forest procedure generated 1000 bootstrapped classification trees based on the training data set.

2.3. Classification of the test data set

The random forest algorithm defaults to majority rules classification. To improve diagnostic specificity a two part classification rule was created for this study. Classification thresholds were established based on the calculated probability of belonging to a given class and the highest threshold that did not reduce test efficiency was selected. This approach was taken to preserve test sensitivity where possible while optimizing specificity. Test efficiency was calculated as indicated in equation (1): where test efficiency (E), true positive (TP), true negative (TN), false positive (FP), false negative (FN).

$$E = \frac{TP + TN}{TP + TN + FP + FN} \cdot 100 \tag{1}$$

During the optimization process test efficiency was calculated iteratively as the classification threshold was incremented from 0–1 in steps of 0.001 (Figure 4). To maximize specificity the point with optimal efficiency and the highest probability threshold was selected as the classification threshold (Table 3).

An "unknown" class was added to the algorithm for subjects that did not meet the classification threshold of any other class. This permitted the following two part classification rule for classification of the test data set:

- Only the class with the highest probability was considered, this probability was then compared to the classification threshold for that class.
- Samples with probability below the threshold were classified as "unknown".

2.4. Software

All data analysis, and graphing was done using the R programming language [7] and the random forest [8], boot [9], and ggplot [10] packages. R scripts are available for download at the following url: https://github.com/hendersonmpa/chemokines.

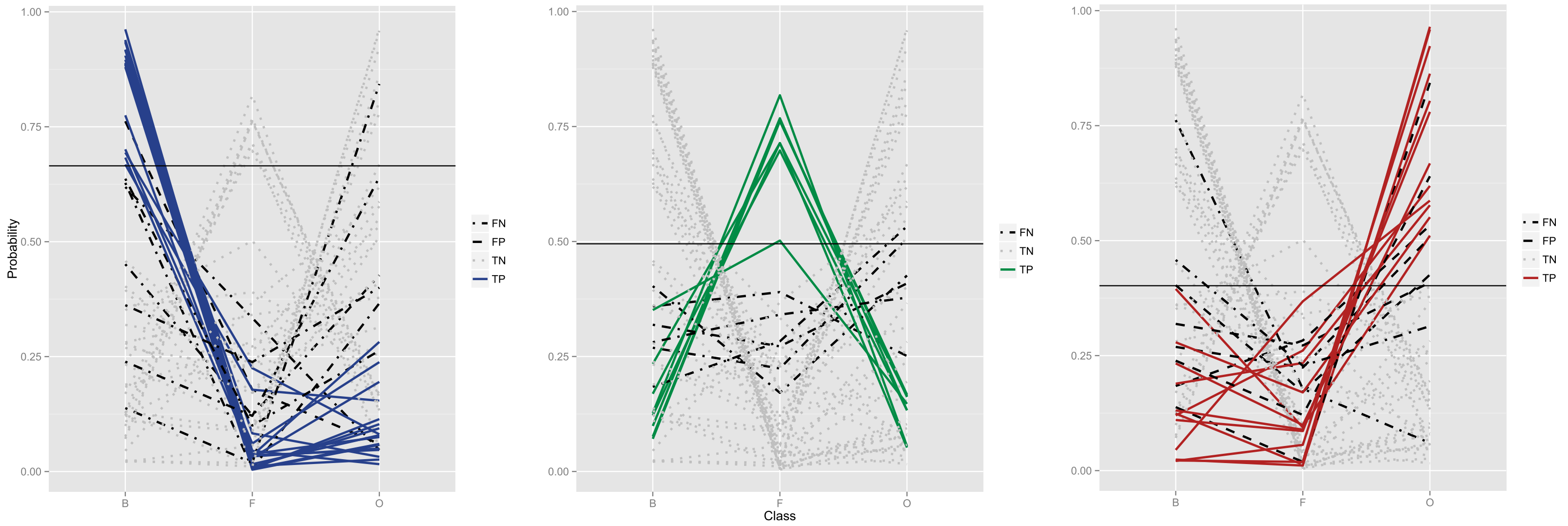
3. Results

3.1. Analyte selection

The normalized distribution of cytokines and cell adhesion molecules in the study subjects is shown in Figure 2. The random forest method was applied to the training set and each variable was left out in succession to determine what impact that variable has on classification accuracy. Subsequent analysis was performed using only analytes that contributed to classification accuracy: TNF-α, L-selectin, IL-1α, P-selectin, IL-2, ICAM-1, IL-4, and VCAM-1 (Figure 3).

3.2. Classification of the test data set

The optimal threshold probability was used to classify subjects from the test data set. The resulting predicted classes are presented in Figure 1 as parallel co-ordinates plots. In the parallel co-ordinates plots each line traces the probability of that individual belonging to the respective class: breast cancer, cancer free or ovarian cancer. The parallel co-ordinates plot for the breast cancer classification shows that the random forest performs well, with most true positives far exceding the high threshold probability for classification (Figure 1a). The errors in classification occurred between ovarian cancer and cancer free classes, importantly



(a) Breast cancer

(b) Cancer free controls

(c) Ovarian cancer

Figure 1. Parallel co-ordinates plots for the three classes: breast cancer (blue), cancer free controls (green) and ovarian cancer (red). The threshold probablity for classification in each group is represented by a horizontal black line. FN: false negative, FP: false postive, TN: true negative, TP: true postive.

http://dx.doi.org/10.1016/j.heliyon.2015.e00059

Table 2. Concordance table for predicted classification of the test data set. The values in the "Adjusted" column are the classification error with samples classified as "unknown" removed.

Observed	Predicted				Error	Adjusted
	Breast	Cancer free	Ovarian	Unknown		
Breast	17	0	2	5	0.29	0.08
Cancer free	0	8	4	2	0.43	0.29
Ovarian	1	0	12	1	0.14	0.07

Table 3. Summary of the Random Forest algorithm classification accuracy using the optimal classification threshold. Bootstrapped confidence intervals are provided. The asterisk indicates that intervals could not be calculated as there was no variation between the bootstrapped data sets.

	Train (n)	Test (n)	Threshold	Efficiency	Sensitivity	95% CI	Specificity	95% CI
Breast cancer	36	24	0.665	86.6	70.8	47.1–86.4	96.4	82.1–100
Cancer free	18	14	0.495	87.8	57.1	29.6–81.8	100.0	*–*
Ovarian cancer	28	14	0.402	91.5	85.7	50–100	84.2	69.4–93.4

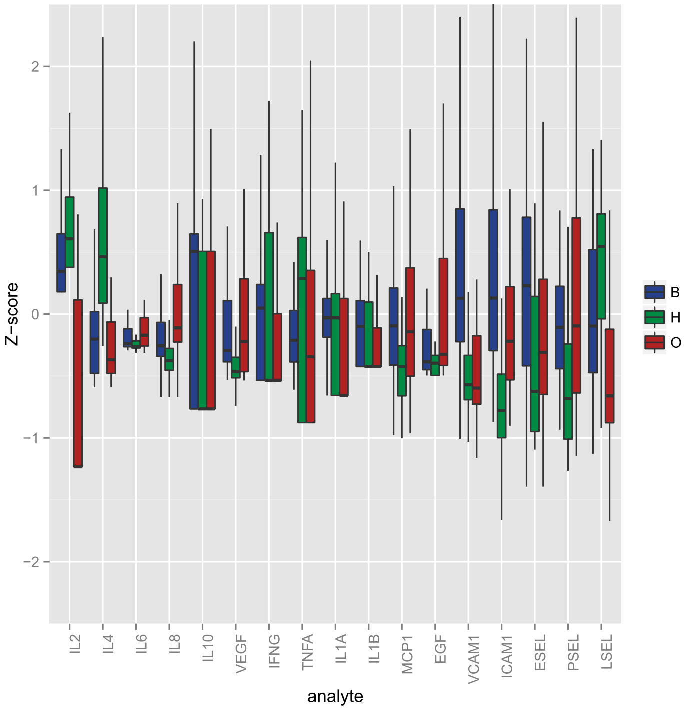


Figure 2. Boxplot summary of analyte concentration z-scores grouped by diagnosis: breast cancer (blue), healthy (green), and ovarian cancer (red).

no subjects with known cancer were assigned to the cancer free class. The cancer free class had no false positives, however, 4 cancer free subjects were classified as ovarian cancer (Figure 1b). Only two subjects were classified as unknown, both of which were cancer free. It is apparent from Figure 1 and Table 2 that the classification was poorest for the cancer free subjects.

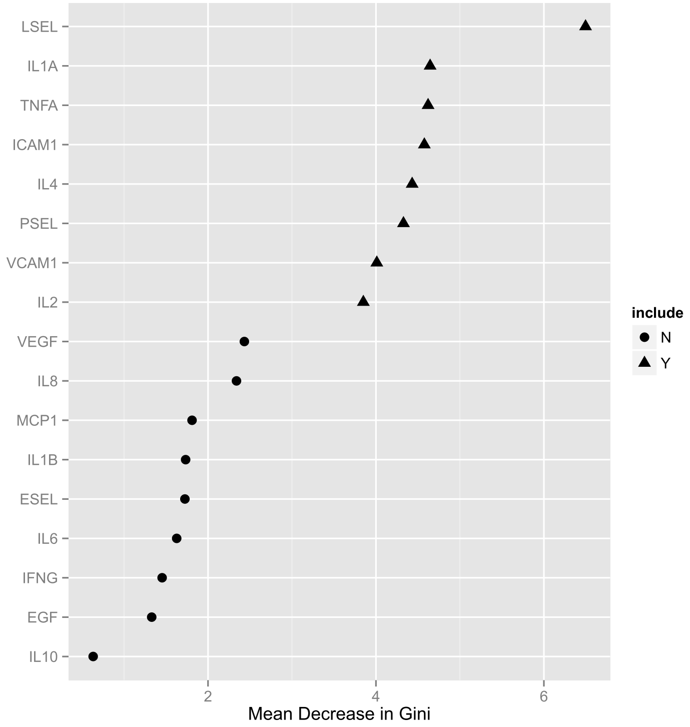


Figure 3. Variable importance plot for random forest analysis on the training data set. A mean decrease in accuracy of 0.02 was used as a cut-off for inclusion (triangles) of the variable in subsequent analysis.

3.3. Classifier performance

The performance of the classifier on the test data set is presented in Table 3. The sensitivity and specificity of the classifiers with accompanying boot strapped confidence intervals were calculated at the optimal probability thresholds. As expected given the classification results (Table 2) the sensitivity for cancer free controls is poor while the other two classes have adequate sensitivity and specificity.

4. Discussion

In this study, we investigated whether cytokines and cell adhesion molecules exhibit distinct profiles in ovarian cancer, breast cancer and cancer free controls. Our results show that using the random forest classification algorithm a panel of cytokines and cells adhesion molecules can distinguish between cancer free control subjects and those with ovarian and breast cancer with promising sensitivity and specificity.

The pathophysiological link between cytokines and cell adhesion molecules and cancer has been studied extensively [2, 3]. With the advent of multiplexed assays

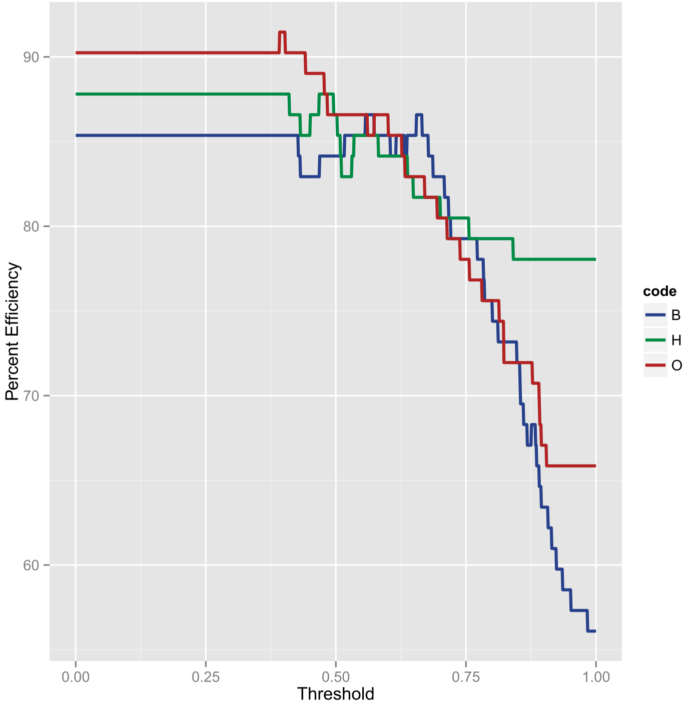


Figure 4. Test efficiency for classification of breast cancer (blue), healthy (green), and ovarian cancer (red). Test efficiency was calulated for the training data at each vote threshold from 0 to 1 in increments of 0.001.

for these analytes, groups have begun investigating the diagnostic potential of these markers [1, 11, 12]. To date however no group has demonstrated more than bi-variate (health vs disease state) classification. In practice, patients may present with uncertain tissue of origin therefore the three way classification presented here represents a more challenging and realistic diagnostic situation.

The random forest algorithm allowed refinement of the initial set of 17 potential markers down to 8 based on the markers contribution to classification accuracy. The 8 selected markers: TNF-α, L-selectin, IL-1α, P-selectin, IL-2, ICAM-1, IL-4, and VCAM-1 contain both cytokines and cell adhesion molecules in equal frequency. Statistical analysis showed that although correlation of each of the above markers with each class was modest when combined via the random forest algorithm the panel showed very promising classification efficiency.

Machine learning algorithms are easily applied to biochemical data sets with current statistical analysis software [13]. This may lead to undesirable classification results. One example is the default majority rules classification rule used by the random forest algorithm [5]. Applied naively the random forest algorithm will force each subject into one of the available classes. In this manuscript we modified the classification algorithm to select highest threshold probability that preserved test efficiency. This approach resulted in greater specificity and the opportunity to classify borderline subjects into an “unknown” class.

The classification algorithm was most successful in subjects with ovarian cancer. This is encouraging as ovarian cancer is a challenging diagnosis relying on physical examination, imaging and ultimately a tissue diagnosis with tumour tissue obtained at the time of staging surgery. While this study is a promising proof of concept further studies with a larger and more diverse set of training data would allow classification based on histologic subtypes of breast and ovarian cancer. The more homogeneous training classes will improve classification. In addition, inclusion of recognized tumour markers such as carcinoembryonic antigen, CA-125, CA15-3, p53, HE4 and soluble HER2 will improve the utility of this classification model.

Declarations

Author contribution statement

Pete Kavsak, Matthew Henderson: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Sebastien Hotte, Holger Hirte: Conceived and designed the experiments; Analyzed and interpreted the data

Funding statement

This work was supported by the Ontario Institute for Cancer Research.

Competing interest statement

The authors declare no conflict of interest.

Additional information

Data associated with this study has been deposited at: https://github.com/hendersonmpa/chemokines

References

[1] P. Kavsak, M. Henderson, P. Moretto, H. Hirte, K. Evans, D. Wong, W. Korz, S.J. Hotte, Biochip arrays for the discovery of a biomarker surrogate in a

phase I/II study assessing a novel anti-metastasis agent, Clin. Biochem. 42 (10–11) (2009) 1162–1165.

[2] B.E. Lippitz, Cytokine patterns in patients with cancer: a systematic review, Lancet Oncol. 14 (6) (2013) e218–e228.

[3] N. Makrilia, A. Kollias, L. Manolopoulos, K. Syrigos, Cell adhesion molecules: role and clinical significance in cancer, Cancer Investig. 27 (10) (2009) 1023–1037.

[4] American Cancer Society, Cancer Facts and Figures 2015, American Cancer Society.

[5] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[6] P. Kavsak, M. Henderson, A. Lee, H. Hirte, E. Young, J. Gauldie, Cytokine elevations in acute coronary syndrome and ovarian cancer: a mechanism for the up-regulation of the acute phase proteins in these different disease etiologies, Clin. Biochem. 41 (7–8) (2008) 607–610.

[7] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, 2011.

[8] A. Liaw, M. Wiener, Classification and regression by randomforest, R News 2 (3) (2002) 18–22.

[9] A. Canty, B. Ripley, Bootstrap R (S-Plus) Functions, rpackage version 1.3-3, 2011.

[10] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer, New York, 2009.

[11] S. Lawicki, G. Bedkowska, M. Szmitkowski, VEGF, M-CSF and CA 15-3 as a new tumor marker panel in breast malignancies: a multivariate analysis with ROC curve, Growth Factors 31 (3) (2013) 98–105.

[12] E. Gorelik, D. Landsittel, A. Marrangoni, F. Modugno, L. Velikokhatnaya, M. Winans, W. Bigbee, R. Herberman, A. Lokshin, Multiplexed immunobead-based cytokine profiling for early detection of ovarian cancer, cancer epidemiology, Biomark. Prev. 14 (4) (2005) 981–987.

[13] M.P.A. Henderson, G.R. Pond, P.A. Kavsak, Statistical and analytical approaches for assessing biomarkers: new approaches, new technologies, with the same-old rigor for evaluation, Clin. Biochem. 45 (3) (2012) 187–188.