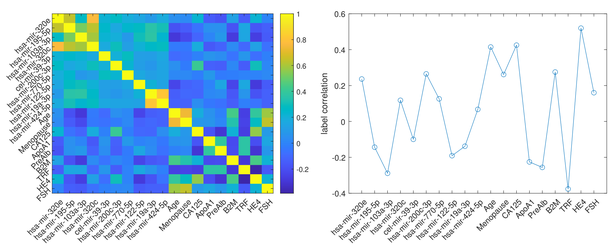
Improving the diagnostic accuracy of an ovarian cancer triage test using a joint miRNA-protein model

James W. Webber [1] Laura Wollborn [1] Sudhanshu Mishra [1] Allison Vitonis [3] Daniel W. Cramer [2,3] Ryan Phan [4] Todd Pappas [4] Dipanjan Chowdhury [2,5] Kevin M. Elias [1,2]

[1]Division of Gynecologic Oncology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA USA 02115, [2]Harvard Medical School, Boston, MA USA, [3]Obstetrics and Gynecology Epidemiology Center, Department of Obstetrics and Gynecology, Brigham Women's Hospital, Boston, Massachusetts, USA, [4]Aspira Women's Health, Trumbull, CT, [5]Department of Radiation Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

Motivation

Protein, miRNA, and metadata are largely linearly independent, and also have high correlation to ovarian cancer indivdually. See figure 1.



(a) correlation matrix (b) cancer correlations

Figure 1. Feature correlation matrix (A), and feature-cancer label correlation plot (B).

1. The feature correlation matrix has an approximate block diagonal structure. The miRNA are self-correlated, and so are the proteins and metadata, but the miRNA and proteins do not exhibit a high degree of inter-correlation.
2. Many of the miRNA (e.g., miR-200c) and proteins (e.g., HE4) have high correlation to the cancer labels. Thus, the miRNA and proteins have good synergy, and are well-suited to train a linear classification model.
3. The selection of the miRNA features was done using forward regression. More details are provided in the methods section.

Goals

The goals of this work are summarized as follows:

- **Can we improve a currently deployed ovarian cancer triage test?** - the current test is based on proteins and metadata. We aim to investigate whether miRNA can improve the current test.
- **Testing and comparison** - to do this, we compare model performance on internal and external validation sets provided by Aspira Women's Health (AWH) and Brigham and Women's Hospital (BWH).
- **Indentifying patients for surgery** - the patients are presenting with an ovarian mass, and the goal is to determine who should be recommended for surgery. We aim to investigate whether miRNA can improve model sensitivity, particularly among early-stage and serous cancers.

Data sets considered

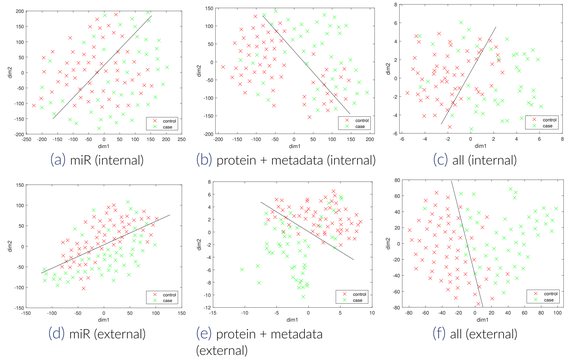
- **Data** - We test our hypothesis on data provided by AWH, and BWH.
- **Properties** - The data is comprised of a training cohort, and two validation cohorts, one internal and one external. See table 1. The patients considered are women, mostly presenting with ovarian masses.

Cohort	n (total sample number)	controls	cases
Training (AWH)	468	277	191
Internal validation (AWH)	100	56	44
External validation (BWH)	110	59	51

Table 1. Data properties per cohort, and the providers of each cohort.

TSNE plots of validation sets

The TSNE plots of the validation sets show the combined features offer the best linear seperation between cases and controls.



(a) miR (internal) (b) protein + metadata (internal) (c) all (internal)

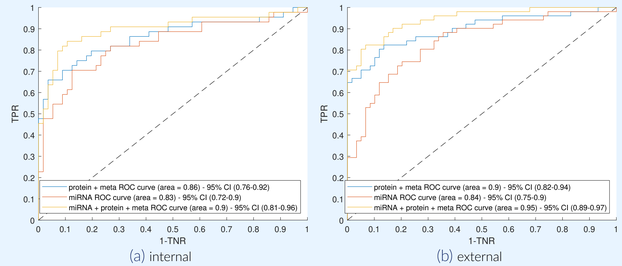
(d) miR (external) (e) protein + metadata (external) (f) all (external)

Figure 2. TSNE plots of internal (top row) and external (bottom row) validation sets, using different feature sets.

The combined feature set (all) is most optimal for training a linear classification model.

ROC curve plots

The combined model offers the best AUC score on the internal and external validation sets. In particular, on the external set, the ROC curve offered by the combined model is always above the ROC curves compared against.



(a) internal (b) external

Figure 3. ROC plots on internal and external validation sets.

Sensitivity and specificity

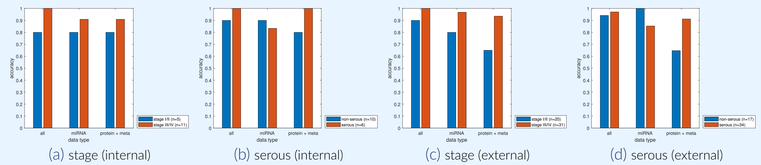
The sensitivity and specificity scores are tabulated in table 2. The combined features offer a highly sensitive model, whereas the protein + metadata model is more specific.

Feature set	spec (internal)	sens (internal)	spec (external)	sens (external)
protein + meta	.80	.77	**.78**	.82
miRNA	.75	.80	.49	.90
protein + meta + miRNA	**.84**	**.86**	.66	**.96**

Table 2. Sensitivity and specificity scores with cancer probability threshold corresponding to the point on the training ROC curve closest to $(0, 1)$ (see figure 3).

Breakdown by stage and serous vs non-serous cancers

The combined model offers the best accuracy when identifying serous and early-stage cancers.



(a) stage (internal) (b) serous (internal) (c) stage (external) (d) serous (external)

Figure 4. Accuracy scores for serous/non-serous and early/late-stage cancers.

Feature set	early OC	late OC	serous OC	non-serous OC	serous & early OC
protein + meta	.65	.82	.85	1	.67
miRNA	.80	.97	.91	.65	.56
protein + meta + miRNA	**.90**	**1**	**.97**	**.94**	**.89**

Table 3. External set - accuracy scores broken down by cancer stage and serous/non-serous cancers.

The goal of the model is to triage patients for surgery, and thus a highly sensitive model is desired. The combined data model is most optimal in this regard.

Methods

The original triage test is based on 7 proteins (e.g., CA-125) and age and menopausal status. We introduce a panel of 180 miRNA. To reduce the miRNA dimension, we use forward regression [3, 1]. Let $\hat{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m] \in \mathbb{R}^{n \times m}$ be the miRNA data, and let $Y \in \{0,1\}^n$ be the cancer labels. Then, we implement forward regression [3, 1]:

1. Initialize $S_0 = \emptyset$ and the number of features, k.
2. **for** every integer $j \in [0, k-1]$ **do**
3. Let $\mathbf{x}_{i_j}$ be a variable maximizing $R^2_{Y,S_j \cup \{i_j\}}$, and set $S_{j+1} = S_j \cup \{i_j\}$.
4. Output S_k,

where

$$R^2_{Y,S} = \frac{\text{Var}(Y) - E\left[(Y - Y')^2\right]}{\text{Var}(Y)}, \tag{1}$$

and $Y' = \sum_{j=0}^{k-1} v_{i_j} \mathbf{x}_{i_j}$, where $\mathbf{v} = (v_{i_0}, \ldots, v_{i_{k-1}})^T$ is defined $\mathbf{v} = \arg\min_{\mathbf{v}} \left\| [\mathbf{x}_{i_0}, \ldots, \mathbf{x}_{i_{k-1}}]\mathbf{v} - Y \right\|_2^2$. The output, S_k, identifies the optimal miRNA features, which are combined with the proteins and metadata. $k = 10$ was identified as the optimal k using cross-validation. Then, we train a softmax function [2] to map X to Y

$$P(y = j \mid \mathbf{x}) = \frac{e^{\mathbf{x}^T \mathbf{w}_j + j}}{\sum_{i=0}^{n_c - 1} e^{\mathbf{x}^T \mathbf{w}_i + i}}, \tag{2}$$

where $j \in \{0, 1\}$ is the class label, $\mathbf{x} \in \mathbb{R}^{k+9}$ is a patient sample (i.e., one row of X), and the $(\mathbf{w}_{j,j})$ are weights and biases to be trained. Here y denotes the class label assigned to $\mathbf{x}$. The class with the highest probability P is then chosen for membership.

References

[1] Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *arXiv preprint arXiv:1102.3975*, 2011.

[2] Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

[3] Alan Miller. *Subset selection in regression*. CRC Press, 2002.