Predicting breast cancer metastasis from blood gene expression using time-to-diagnosis data to derive signature estimates

Einar Holsbø,[1, 2] Vittorio Perduca,[2] Lars Ailo Bongo,[1] Eiliv Lund,[3] Etienne Birmelé[2]

1 Department of Computer Science, UiT Arctic University of Norway
2 MAP5, Université Paris Descartes
3 Department of Community Medicine, UiT Arctic University of Norway

Our data at a glance

```
dim(gene_expression)
## [1]    88 12404

summary(days_to_diagnosis)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     6.0   117.8   189.5   186.8   269.2   358.0

summary(metastasis)
## FALSE  TRUE
##    66    22

table(metastasis, stratum)
##           stratum
## metastasis screening interval clinical
##      FALSE         43       10       13
##      TRUE           6        6       10
```

The data belong to the prospective Norwegian Women and Cancer study. The women provided blood samples and questionnaire info on enrollment. The Cancer Registry of Norway provides information about disease progression.

We have data on 88 cases with age-matched controls, the values in the expression matrix are $\log_2$(case/control), ie log fold change.

We're trying to predict metastasis T/F from these gene expression values. The cancers belong to one of three strata: detected at a screening, detected between two screenings, or detected at a clinic in women who did not attend a screening in at least two years.

Predictive models, performance metrics

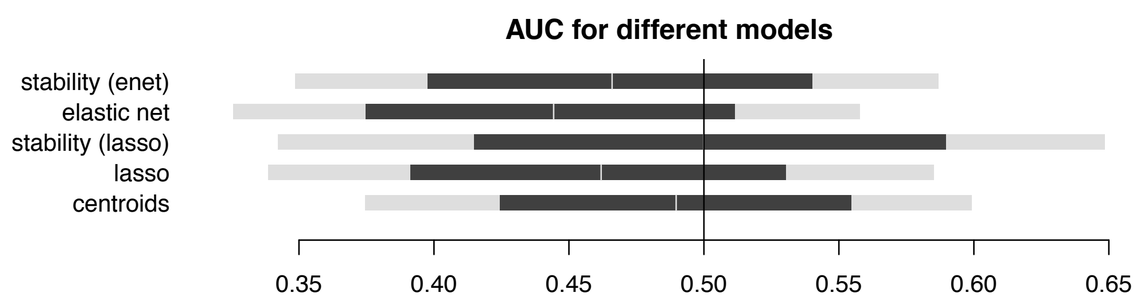
We evaluate a handful of predictive models: different flavors of penalized logistic regression and nearest centroids. We fit models with and without time-to-diagnosis preselection of genes and compare.

We measure predictive power in terms of area under the ROC curve (same as Mann-Whitney U statistic) and signature stability in terms of the Jaccard index between signatures derived from partially overlapping data.

We calculate uncertainy estimates by repeated cross-validation (1500 resamplings). Tuning parameters and preselection come from a separate cross-validation procedure nested in the repeated cross-validation. In other words, every step of the modelling is isolated from the test data fold.

If there is a signal, it's weak and noisy

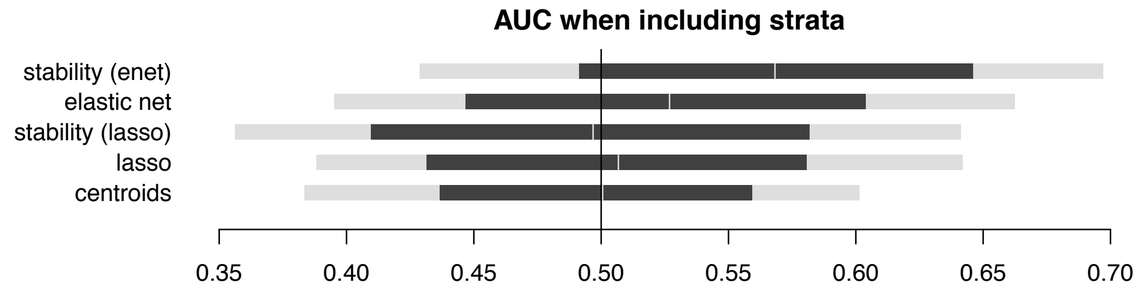
The reason we use repeated cross-validation is that there is enough variance in the model fitting procedure that two cross-validations will provide very different point estimates and confidence intervals. There is also a strong conflicting signal in some genes as evidenced in some models giving AUC estimates below random guess. See the figure below.



Above: Initial AUC estimates for five different models. The darker region is the .75 uncertainty interval, while the lighter region shows the .95 interval. All but the lasso stability selection have a median AUC below random guess (at .5) over 1500 resamplings.

We suspect that this comes from having cancers with different characteristics in our data set. The between-screenings cancers are likely to be pretty aggressive as they have appeared in a two-year period between screenings. The at-screening cancers are likely to be caught early but to be less aggressive than those caught between screenings. In both these cases we are dealing with small, young cancers. Finally, the clinical cancers are likely to be pretty old, as they were discovered by the women themselves.

If we include the cancer strata in our models, this anomaly goes away. In other words, this is probably a kind of Simpson's paradox:



We have performed simulations to describe this anomaly, please see
`github.com/3inar/degenerate_auc`

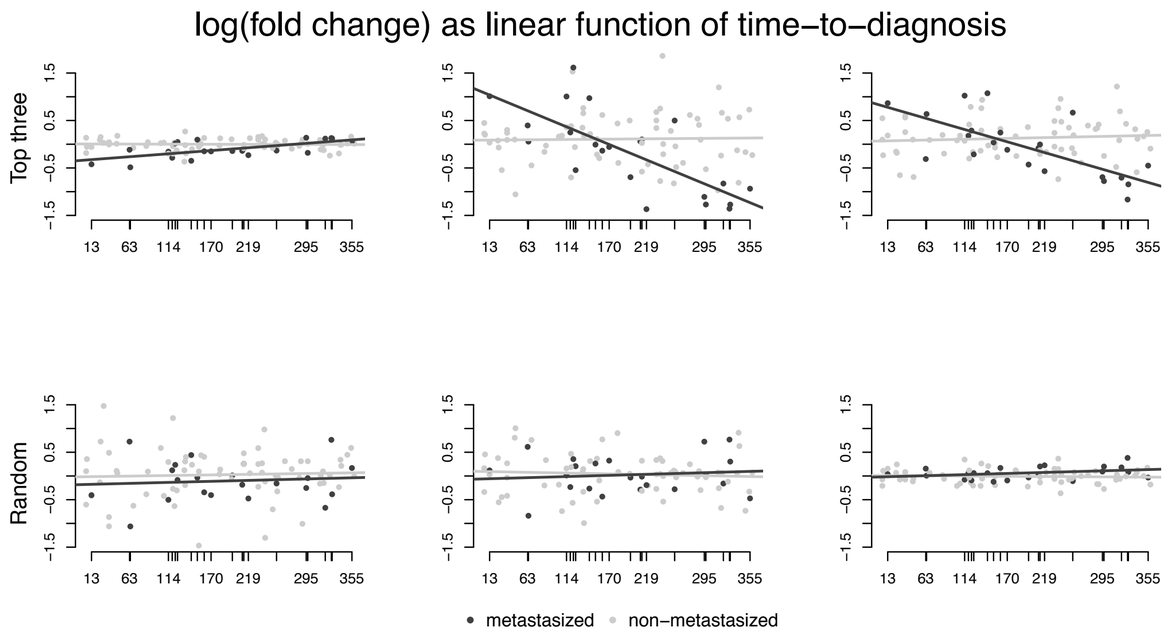
However, the signal is either weak or non-existent and it would be more useful if we could avoid using the stratum information in a "blood sample before diagnosis" type situation.

Time-to-diagnosis preselection

To select genes that are likely to be predictive of metastasis, we fit gene-wise linear models where we regress gene expression on time and metastasis according to the following:

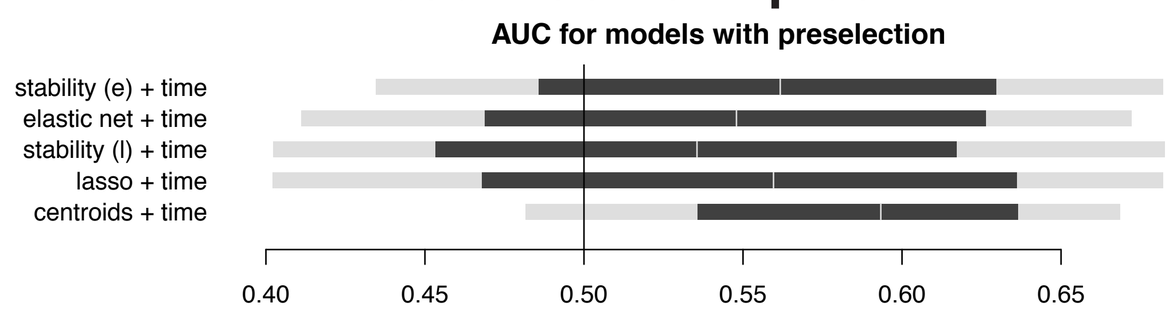
$$\text{expression} = \beta_0 + \beta_1 \text{time} + \beta_2 \text{metastasis} + \beta_3 \text{time} \times \text{metastasis} + \text{error}$$

We use the largest t-statistic from any of these coefficients as that particular gene's score and pick the *n* genes with the highest scores as a first filtering. For the penalized regression methods we set *n=200*, while for centroids we use *n=50*. These models pick out the following interesting patterns of gene expression over time: cases and controls diverging over time and constant difference between cases and controls. The interaction with metastasis allows metastatic and non-metastatic cancers to have different patterns within the same gene.



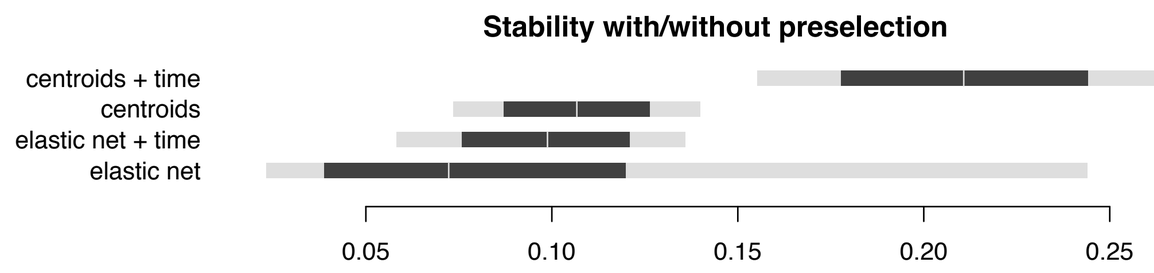
Above: The three top-raked genes compared to three randomly selected genes. Dark grey is metastasized pairs, light grey non-metastasized. The lines describe the linear model above. The top three models all show metastasized cases diverging from their controls while the less aggressive non-metastasized look more or less the same.

There is some evidence of predictive power in these blood samples



Above: AUC for models with time-to-diagnosis preselection. The darker region is the .75 uncertainty interval, while the lighter region shows the .95 interval. There appears to be some predictive power. The simplest model, centroids, does better and does so with less variance. The vertical line at .5 denotes the threshold for random guess. AUC=1 is perfect prediction.

Time-preselection improves stability



Above: Stability of the final signatures selected by two different models with and without preselection. This is the set overlap between models fit on 75% overlapping data. The darker region is the .75 uncertainty interval, while the lighter region shows the .95 interval. The clear winner is centroids with time-to-diagnosis preselection. This uses the 50 highest-ranked genes. The non-time-preselection centroids ranks genes by simple gene-wise t-test and uses the top 50. In general the signatures aren't very stable.You can expect to have 10 genes (one fifth) common to two centroids+time models fit on partially overlapping data.

In summary

Gene expression in blood is noisy and might even be paradoxical. We observe some suggestion that there is information predictive of metastasis in blood gene expression before diagnosis if you choose your genes carefully. Using the time-to-diagnosis information from our cohort we were able to build models that were more accurate and more stable than out-of-the-box models.

email: `einar@cs.uit.no`
twitter: `@0xeinar`
github: `github.com/3inar`

This poster is available online at `3inar.github.io/talks/`