Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

Statistical analysis of gene expression data related to breast cancer diagnosis

Working document

Note no. SAMBA/19/14

Authors Marit Holden and Lars Holden

Date 1. juni 2014

Norsk Regnesentral

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
office address
Gaustadalleen 23a
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Bankkonto · bank account
8200.01.48888
Org.nr. · enterprise no.
NO 952125001 VAT

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Title	**Statistical analysis of gene expression data related to breast cancer diagnosis**
Authors	**Marit Holden and Lars Holden**
Date	1. jun. 2014
Year	2014
Publication number	SAMBA/19/14

Abstract

This note is a working document and is based on work in progress.

The note describes methods for and results of analyzing gene expression data related to breast cancer diagnosis. The hypothesis is that the genes related to the stages of cancer development could be differentially expressed over time, perhaps in a small but consistent manner. We have started developing methods for testing whether there is such a development in time, and for identifying groups of genes with similar behavior, or functional form, the last years before diagnosis. Hence, we are looking for weak signals from a large number of genes in contrast to stronger signals from a few genes. We have also proposed a method for using information from such groups of genes for predicting whether a case has breast cancer with or without spread.

From the preliminary results described in this note we conclude that it is important to normalize the data before further analysis. However, normalizing the data may also remove trends we are looking for, and we have observed that the results presented are sensitive to choice of normalization method. Therefore different normalization methods should be tested and evaluated to decide which method is best suited for our dataset. This is outside the scope of this note and should be further examined in later work.

The dataset consists of $\log_2$-transformed gene expression values in blood cells related to breast cancer. The developed methods have been tested on several version of this dataset as the data are continuously updated when new information becomes available (for example when new individuals are diagnosed with cancer or the quality of the data is improved) and because different subsets of the dataset have been selected dependent of what information we wanted to include in the analyses. This, and slightly different choices in the preprocessing steps, resulted in different subsets of genes selected for the different versions of the dataset. We have observed that the results are sensitive to the subset of genes selected. Later this will be further examined to find the procedure for selecting genes to be included in the statistical analyses that is best suited for our dataset.

For some of the preliminary analyses we conclude that there is a significantly high number of genes that increase or decrease monotonically in gene expression the years before diagnosis in the stratum where we a priori expect it is most likely to observe a signal. We expect a more homogeneous dataset for persons participating in a screening program and expect a stronger signal from patients with spread. However, the signal is still weak. Using information from the identified groups of genes when predicting spread or not spread, we were able to identify about 1/3 of the cases without spread and no or few false negatives. The preliminary methods will be further developed later, and they will also be tested on a dataset with improved quality where more optimal preprocessing procedures and normalization methods are used.

Keywords	Gene expression data; Breast cancer;Development in time; Curve groups; Hypothesis testing using randomization; Prediction of diagnosis
Availability	Open
Project number	220 633
Research field	Bioinformatics
Number of pages	74

Table of Content

1 Introduction

This note is a working document and is based on work in progress. The work is performed in close cooperation with the University of Tromsø and professor Eiliv Lund. His group has provided all the data.

The note describes methods for and results of analyzing gene expression data related to breast cancer diagnosis. The hypothesis is that the genes related to the stages of cancer development could be differentially expressed over time, perhaps in a small but consistent manner. We have started developing methods for testing whether there is such a development in time, and for identifying groups of genes with similar behavior, or functional form, the last years before diagnosis. Hence, we are looking for weak signals from a large number of genes in contrast to stronger signals from a few genes. We also propose a method for using information from such groups of genes for predicting whether a case has breast cancer with or without spread.

The dataset consists of log2-transformed gene expression values in blood cells related to breast cancer. The developed methods have been tested on several version of this dataset as the data are continuously updated when new information becomes available (for example when new individuals are diagnosed with cancer or the quality of the data is improved) and because different subsets of the dataset have been selected dependent of what information we wanted to include in the analyses. This, and slightly different choices in the preprocessing steps, resulted in different subsets of genes selected for the different versions of the dataset. Each section in this note describes the analysis of one version of the dataset. See Table 1.

Table 1 The datasets used in the study. For a dataset with four strata, there is one stratum for each combination of screening / not screening and spread / not spread. For a dataset with two strata, there is one stratum for screening and not spread, and one stratum for not screening and not spread. (Not) screening means that the case in a case control pair did (not) participate in the screening program for breast cancer.

Section	Description of dataset	#case-control pairs	#genes	#years(periods)	#strata
2	Before diagnosis, invasive	251	9060	4 (3)	4
3	Before diagnosis, invasive	251	8552	4 (3)	4
	At diagnosis Hcc1, invasive	65		1 (1)	
4	Before diagnosis, invasive	249	6952	5 (3)	4
	At diagnosis Hcc1, invasive	64		1 (1)	
	At diagnosis Hcc2, invasive	42		1 (1)	
	At diagnosis Hcc3, invasive	53		1 (1)	
5	Before diagnosis, invasive[1]	249	8130	5 (3)	4
	At diagnosis Hcc1, invasive	64		1 (1)	
	At diagnosis Hcc2, invasive	42		1 (1)	
	At diagnosis Hcc3, invasive	53		1 (1)	
	Before diagnosis, insitu[1]	49		5 (3)	2
	At diagnosis Hcc1, insitu	2		1 (1)	
	At diagnosis Hcc2, insitu	6		1 (1)	
	At diagnosis Hcc3, insitu	3		1 (1)	
6	Before diagnosis, invasive[1]	467	8952	8 (4)	4
	Before diagnosis, insitu[1]	79		6 (4)	2

[1] The data have been produced in three different runs (run1, run2 and run3).

2 Data for four years before diagnosis

2.1 Data

The data are $\log_2$-transformed gene expression values, $D^*_{g,p,c}$, where g=1,…,9060 (gene), p=1,…,251 (case-control pair), and c=1 (case), 2 (control). Let s(p) be the stratum and t(p) the time (days to diagnosis) for case-control pair p. There are four different strata, one for each combination of screening / not screening and spread / not spread. For each stratum s and pair p, we assume that the true gene expression depends on time, and that this time dependency is described by the smooth function $d^*_{s,g,p,c}$(t). We further assume that the function $d^*_{s,g,p,c}$ is the sum of two other functions, $e^*_{s,g,c}$ that is independent of properties of the pair p, and $f^*_{s,g,p}$that is independent of whether the sample is a case or a control. This means that we assume the following model for the $\log_2$-transformed gene expression data:

$$D^*_{g,p,c} = e^*_{s(p),g,c}(t(p)) + f^*_{s(p),g,p}(t(p)) + \varepsilon^*_{g,p,c}\,,$$

where $\varepsilon^*_{g,p,c}$ is model and measurement error. We will study the difference in gene expression for the cases compared to their matched controls. We will therefore analyze the data $D_{g,p} = D^*_{g,p,1} - D^*_{g,p,2}$. From the formula above it follows that we assume the following model for $D_{g,p}$:

$$\begin{aligned} D_{g,p} = D^*_{g,p,1} - D^*_{g,p,2} &= e^*_{s(p),g,1}\big(t(p)\big) - e^*_{s(p),g,2}\big(t(p)\big) + \varepsilon^*_{g,p,1} - \varepsilon^*_{g,p,2} \\ &= e_{s(p),g}\big(t(p)\big) + \varepsilon_{g,p}, \end{aligned}$$

where $\varepsilon_{g,p}$ is model and measurement error.

We use the following names for datasets for the four strata: NScrNSpr, NScrYSpr, YScrYSpr and YScrNSpr, where NScr means “not screening”, YScr means “screening”, NSpr means “not spread” and YSpr means “spread”. See Table 2 for a summary of the number of observations in each stratum. We use the following abbreviations for NScrNSpr, NScrYSpr, YScrYSpr and YScrNSpr, respectively: NN, NY, YY and YN.

Table 2 Number of observations in each year before diagnosis.

Stratum\ year	4	3	2	1
YScrYSpr	3	5	11	7
NScrYSpr	4	5	17	16
YScrNSpr	12	22	39	44
NScrNSpr	4	10	25	27

2.2 Normalization of the data

We will try to reduce the variation in the data represented by the error term $\varepsilon_{g,p}$ by comparing the data for the different case-control pairs. We have studied $D_{g,p}$ statistically in order to find out whether it is possible to remove some of the variation included in $\varepsilon_{g,p}$.

First, we compared the data by comparing the boxplots of the data for the different case-control pairs. Boxplots for some case-controls pairs are shown in Figure 1, where each mean value is plotted as a red point. We observe that for each case-control pair the mean and median values are almost equal. We also observe that the distribution of $D_{g,p}$ for a given case-control pair p varies from pair to pair. If these observed differences vary with time either for all data or for some strata, we want to keep these differences in the data. However, if the observed differences are assumed to be noise, we should remove this noise.

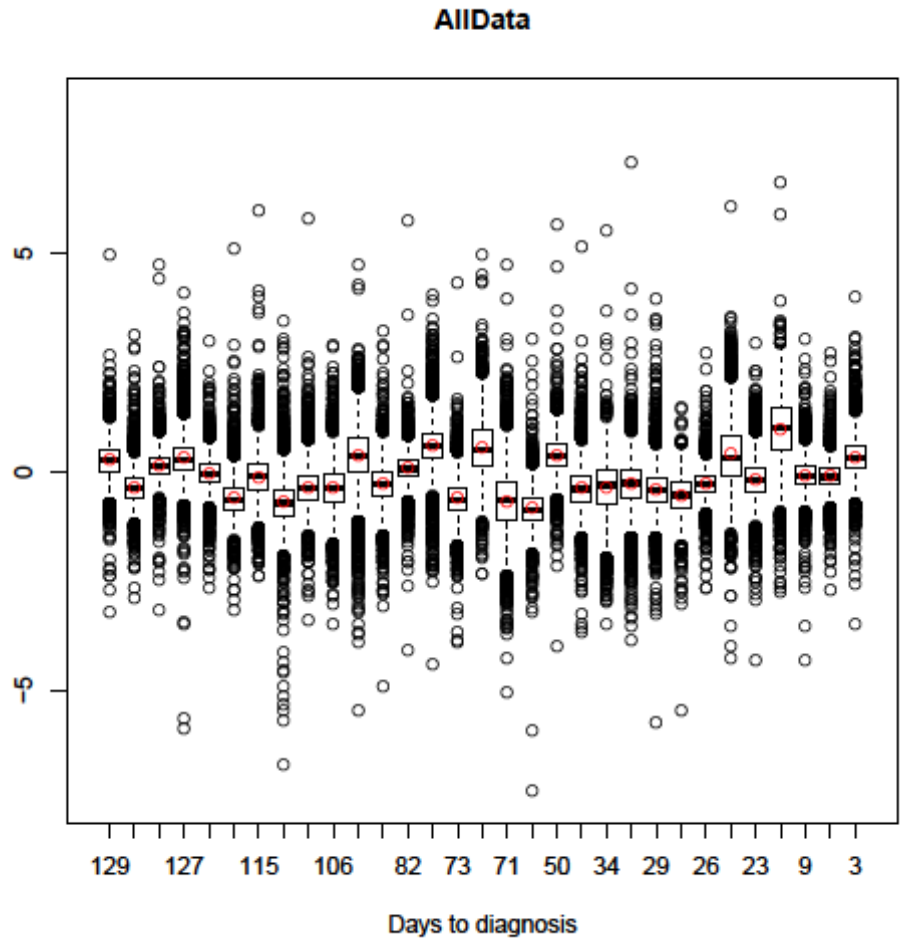

Figure 1 Boxplots of some case-control pairs. Mean values are indicated as red points.

To check whether the observed differences in the boxplots vary with time or stratum we have plotted different summary statistics for the different case-control pairs. Figure 2 (Figure 3) shows the mean (median) value. We observe that the mean (median) value has the same distribution for each of the four strata. The distribution is close to normal, but with a slightly heavier tail to the right. We see no time development in any of the four strata. This makes it natural to believe that differences in the mean (median) value are mainly white noise.

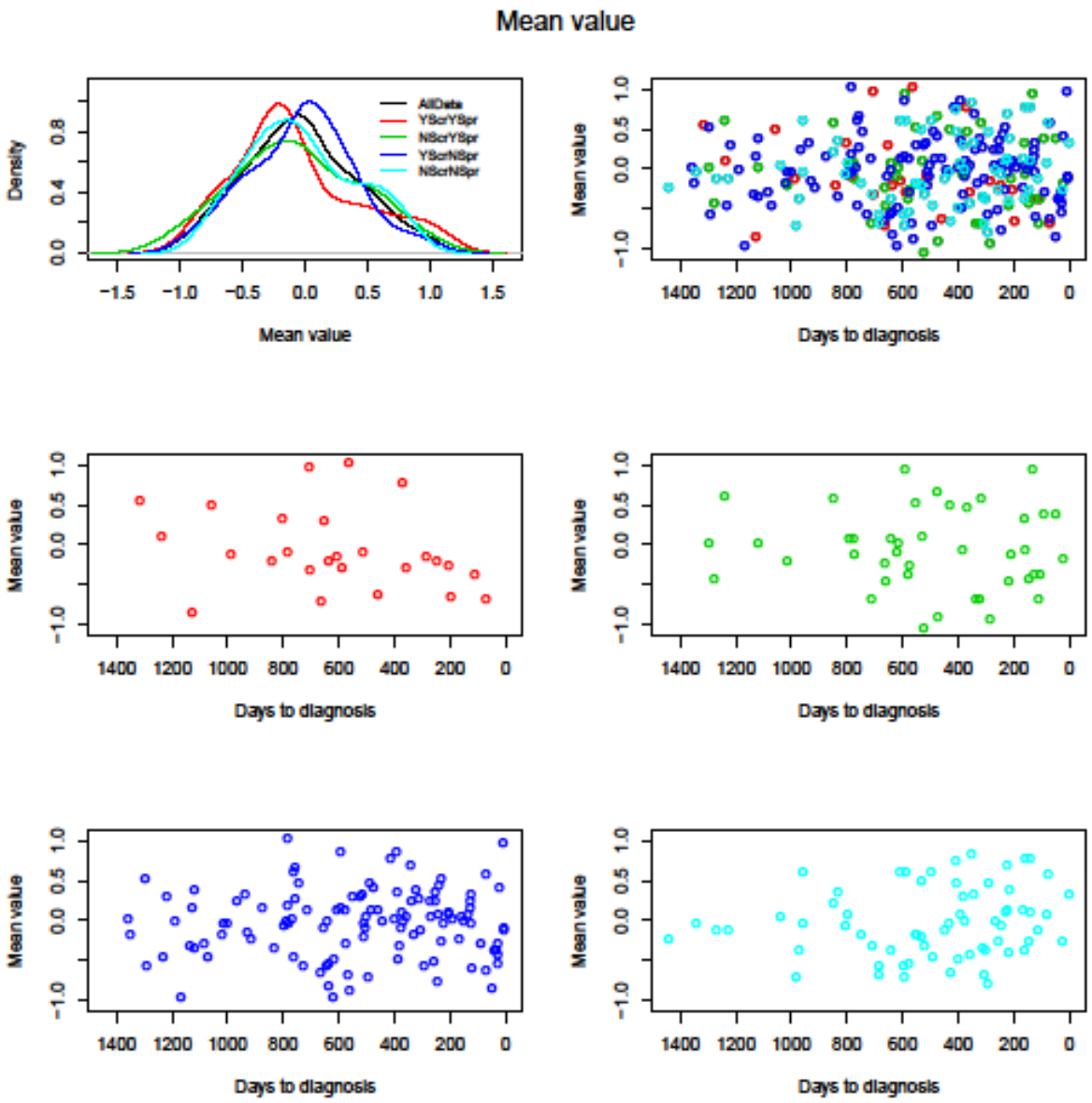

Figure 2 Difference in the mean value for the case-control pairs for all the data and each stratum separately. It is close to a normal distribution with slightly heavier right tail and no identifiable trend.

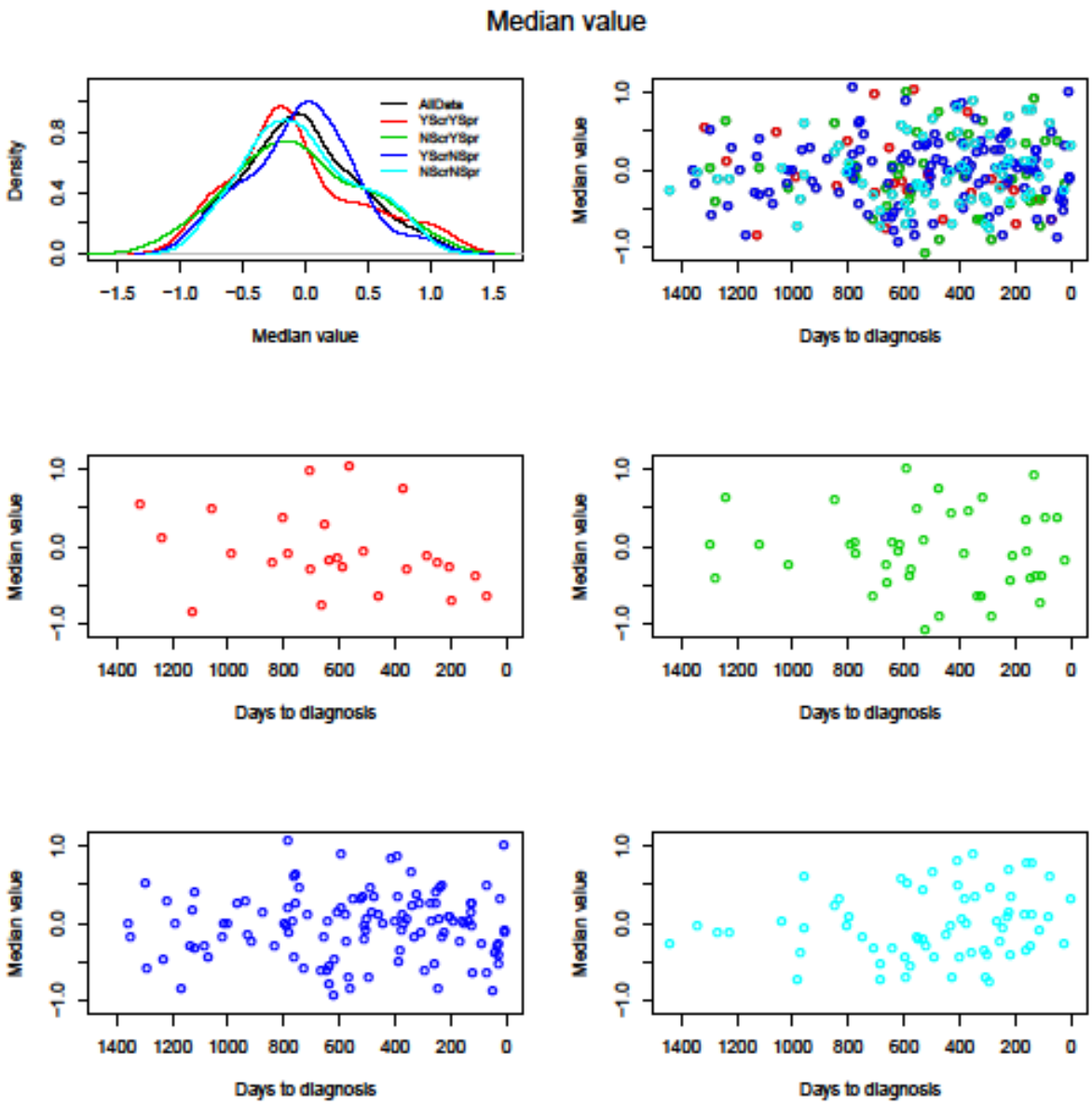

Figure 3 Difference in the median value for the case-control pairs for all the data and each stratum separately. It is close to a normal distribution with slightly heavier right tail and no identifiable trend.

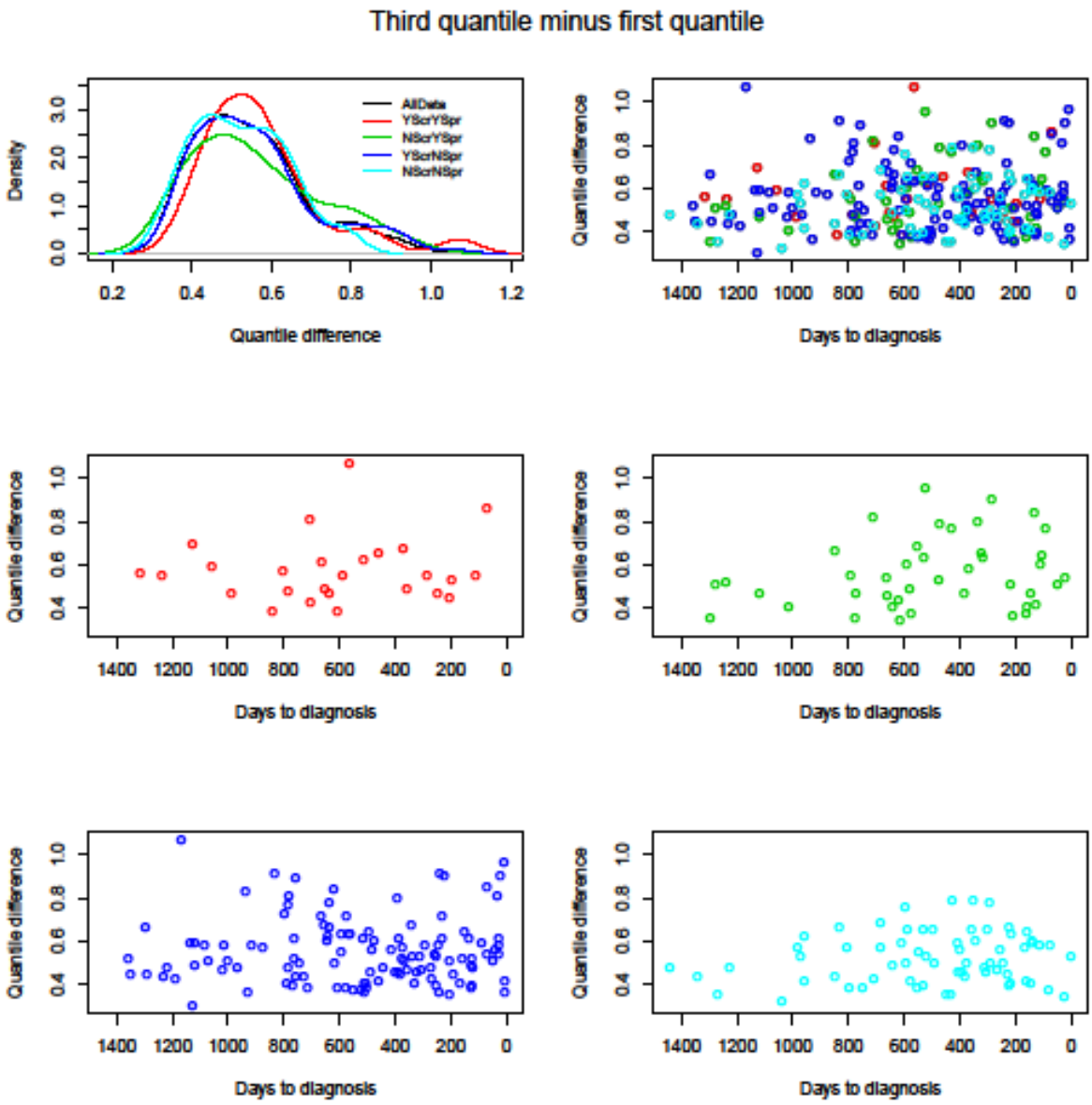

Figure 4 Difference between third and first quantile for the case-control pairs for all the data and each stratum separately. It is close to a normal distribution with slightly heavier right tail and no identifiable trend.

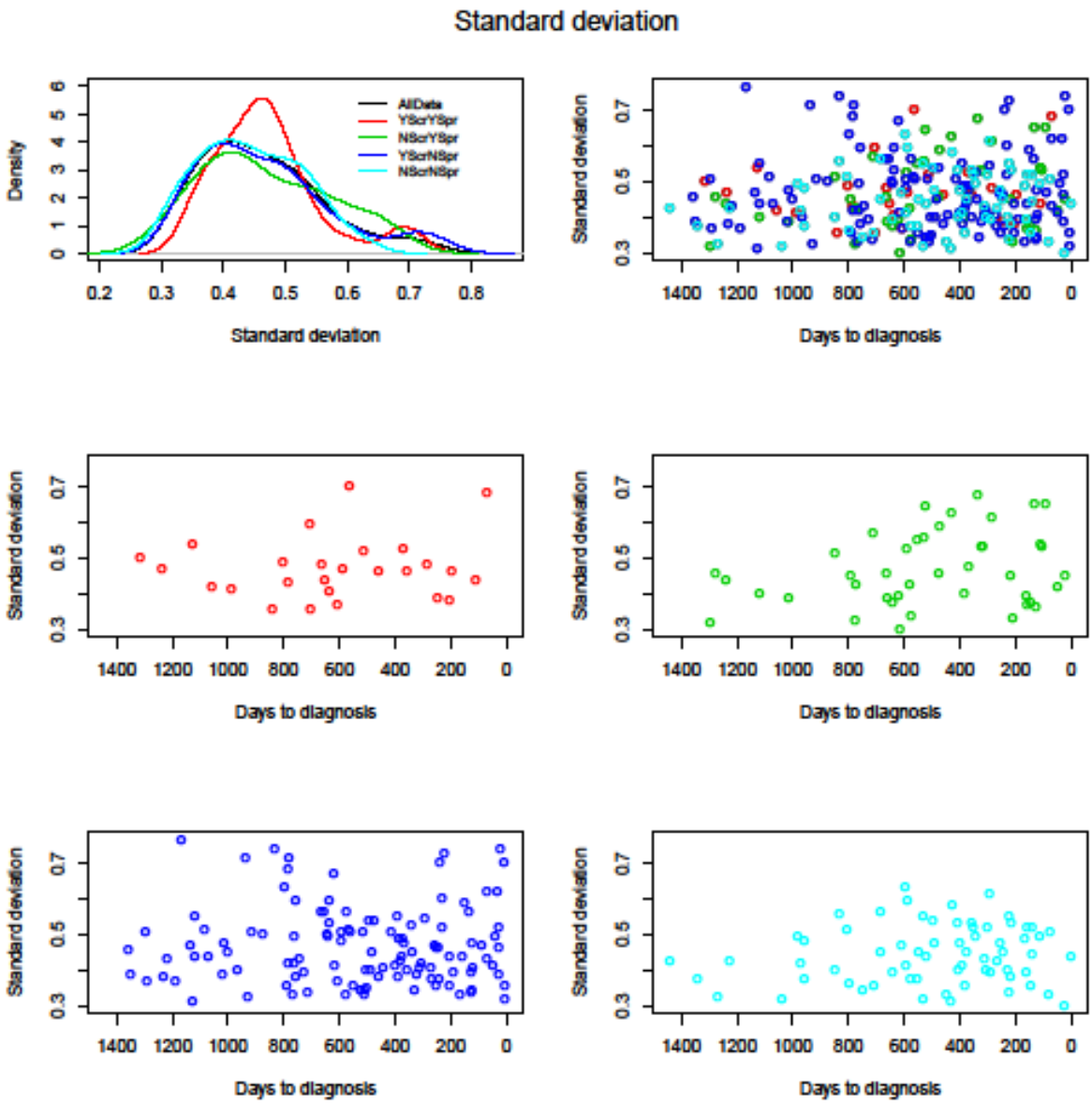

Figure 5 Standard deviation for the gene expressions for each case-control pair for all the data and each stratum separately. It is close to a normal distribution with slightly heavier right tail and no identifiable trend.

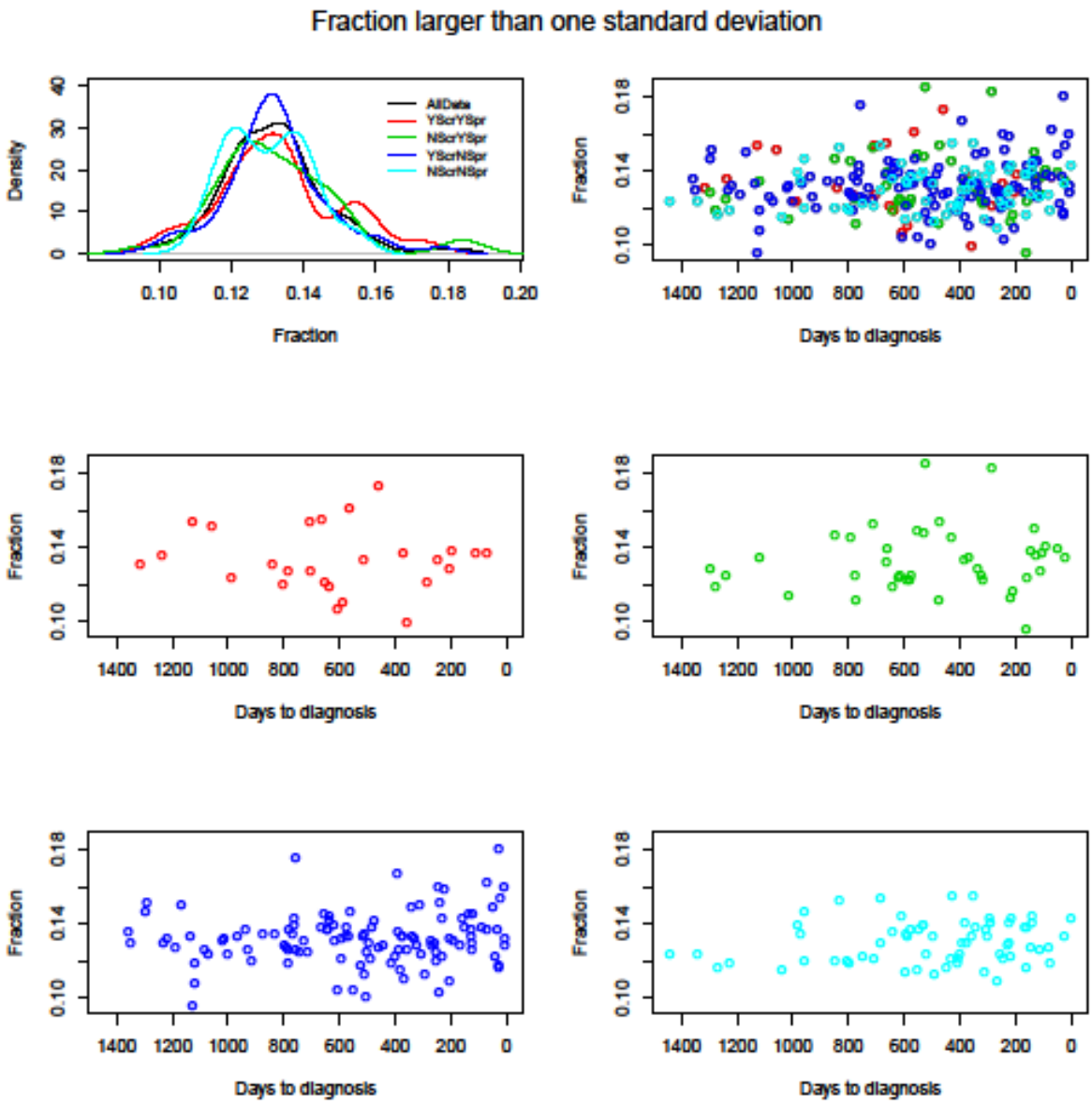

Figure 6 Fraction larger than one standard deviation for all the case-control pairs for all the data and each stratum separately. It is close to a normal distribution with slightly heavier right tail and no identifiable trend.

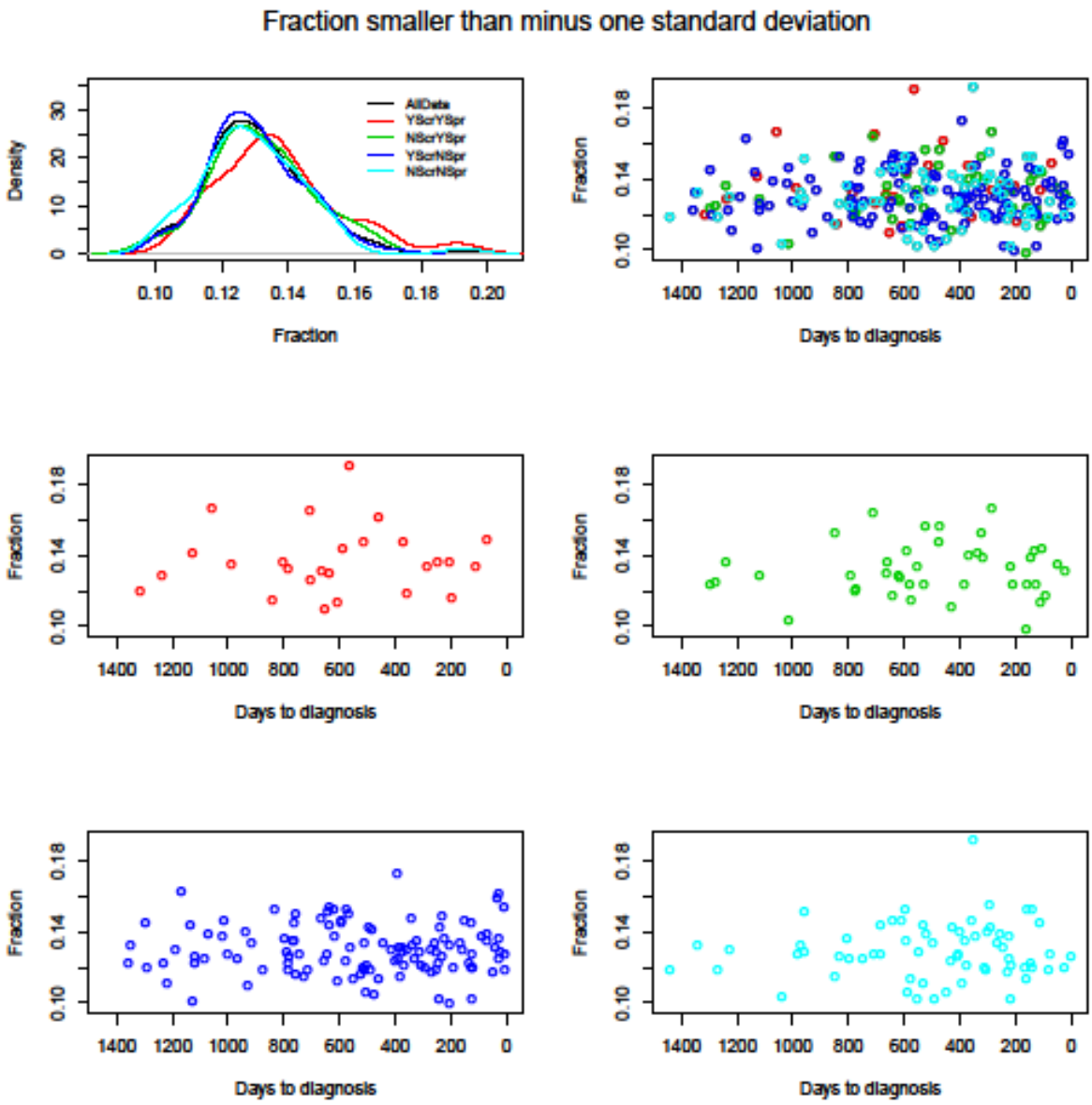

Figure 7 Fraction smaller than minus one standard deviation for the case-control pairs for all the data and each stratum separately. It is close to a normal distribution with slightly heavier right tail and no identifiable trend.

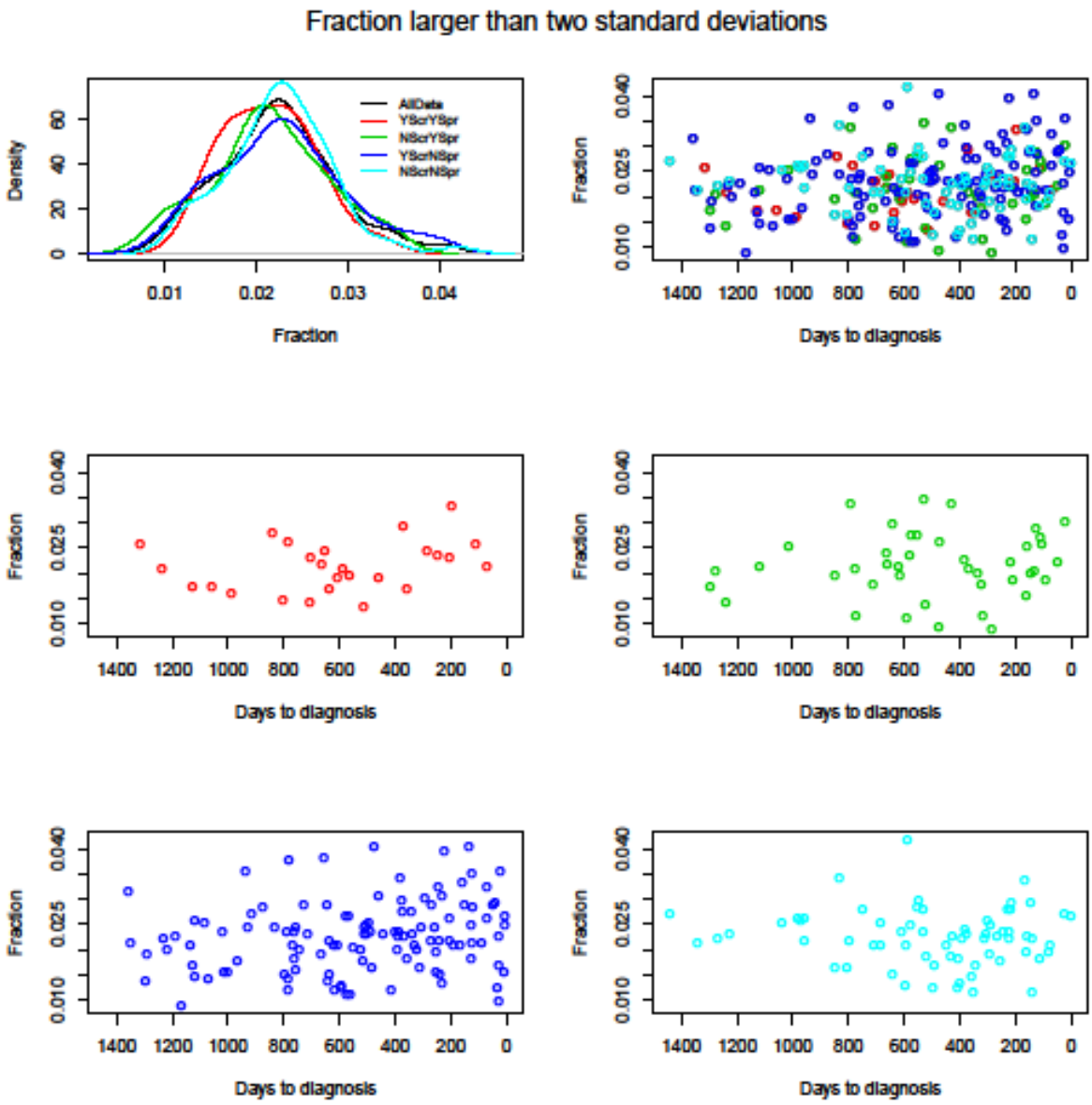

Figure 8 Fraction larger than two standard deviations for the case-control pairs for all the data and each stratum separately. It is close to a normal distribution with slightly heavier right tail and no identifiable trend.

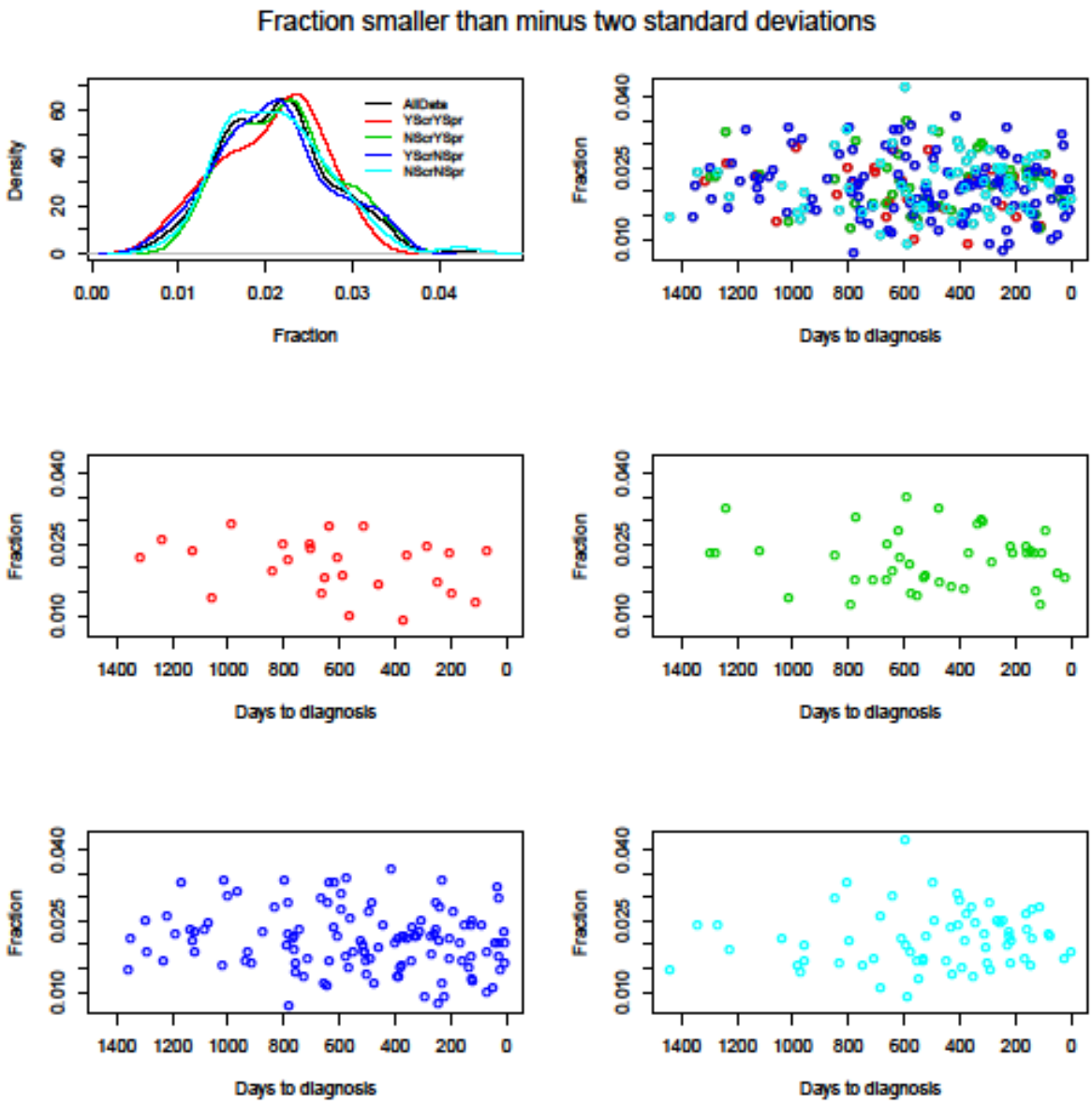

Figure 9 Fraction smaller than two standard deviations for the case-control pairs for all the data and each stratum separately. It is close to a normal distribution with slightly heavier right tail and no identifiable trend.

Figure 4-Figure 9 analyze the variation and tail behavior of the data for the different case-control pairs. We use the difference between the 3rd and 1st quantile and the standard deviation (σ) for measuring variation, while the tail behavior is measured by counting the number of data points larger than $mean + \sigma$ and $mean + 2\sigma$ and smaller than $mean - \sigma$ and $mean - 2\,\sigma$. Figure 4-Figure 9 show the same normal distribution, with a heavier tail to the right, for all four strata and no time development. This indicates that differences in the variance and tail behavior of $D_{g,p}$ between the case-control pairs are mainly white noise.

We may reduce the observed white noise using quantile normalization. This keeps the heavy tails in the data and reduces the $\varepsilon_{g,p}$ term. Boxplots after quantile normalization are shown in

Figure 10. We observe that the mean value is equal to the median value. Below, results will be given both for normalized and not normalized data.

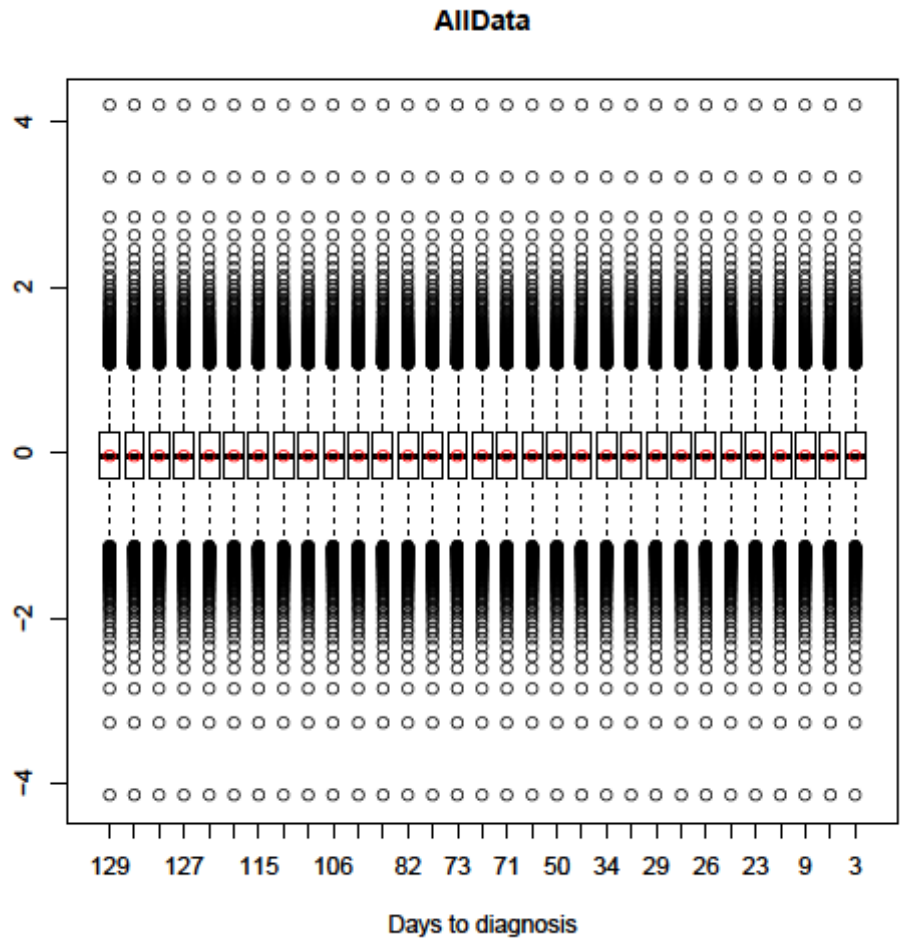

Figure 10 Boxplots of the case-control pairs in Figure 1 after normalization. Mean values are indicated as red points.

2.3 Null model

We make the following assumptions

1. $D_{g,p}$ is stationary in time, i.e. $D_{g,p}$ is independent of the time t(p).
2. There is no difference between the strata.

We use the randomization algorithm $D_{g,r(p)}$ where we use the same permutation for all the genes. This is a robust randomization algorithm that maintains the correlation between the genes for the same person and we make no assumption on this.

When we analyze the data for only one stratum, we generate data from the null model by randomizing data only for the case-control pairs that are in this stratum.

2.4 Hypothesis testing of functional form

We want to find out whether there is a significant development in time. We divide all genes into *K*=6 curve groups based on the order of the average of the data in year 3 and 4, year 2 and

year 1 before diagnosis. These averages are denoted $\overline{D}_{g,3}$, $\overline{D}_{g,2}$ and $\overline{D}_{g,1}$, respectively. If e.g. $\overline{D}_{g,1} > \overline{D}_{g,2} > \overline{D}_{g,3}$, gene g belongs to curve group 123 indicating an increasing gene expression in time when approaching time of diagnosis (year 3 and 4 before diagnosis has order 1, year 2 order 2, and year 1 order 3). Year 3 and 4 are merged since there are so few observations in these years. Let G_i denote the number of genes in curve group i. The probability of each curve group depends on the number of data each year and is computed by simulating from the null model as described above. Since the genes for the same person are correlated, also the curve groups are correlated between the genes.

2.4.1 Test number of genes in curve groups

We want to find out whether there are more genes in the same curve group than expected. This would indicate a significant development in time for some of these genes, e.g. $D_{g,p}$ increases with time for a higher number of genes than expected. We have simulated some datasets for each stratum using the randomization algorithm described above to find the distribution of the number of genes in each curve group. We have tested whether the number of genes in the different curve groups in the data $D_{g,p}$ are different from those in the simulated datasets, $D_{g,r(p)}$. This test for each curve group gave some significant p-values. But we believe it is better to focus on the number of genes with a stronger signal.

The challenge then is to find a stronger test that will find a specific functional form for groups of genes. It is possible to use curve groups combined with a Mann-Whitney test. We expect that most members of a curve group are in the curve group by coincidence. We hope to identify a smaller number of genes that have a specific curved form. Let $M_{a,i}$ be the number of genes in curve group i where the Mann-Whitney test gives a p-value smaller than a. Define $M_a = \sum_i M_{a,i}$. The test is performed as follows:

> For a gene g, find the time periods with the smallest and highest average in a time period. Find the ranking of the data $D_{g,p}$ in these two time periods and use the Mann-Whitney test to find the probability for this ranking of the data for these two time periods assuming the same distribution for all the measurements.

The rank test is designed such that if for example most measurements for one of the time period are higher than most measurements for the other time period, this gives a small p-value. Notice that this p-value is the correct p-value when comparing the measurements in two arbitrary time periods. When we select the time periods with the smallest and the largest average, the “correct” p-value is larger.

The variables M_a and $M_{a,i}$ represent the number of genes with a quite strong functional form. We may find these variables in the data and in the simulated datasets obtained by the randomization $D_{g,r(p)}$. For each simulated dataset, we find the curve group for each gene and the number of genes with a strong functional form in this curve group. Then we compare the number of genes with a strong functional form in the original dataset and the simulated datasets. If the number of genes with a strong functional form is larger in the original dataset than the number of genes with a strong functional form in 95% of the simulated datasets, we consider this curve group as significant. The p-value is set to K/N+1/(2N) where K is the

number of simulations with a larger number of genes with a strong functional form than was observed for the data and N is the number of simulation.

Table 3 P-values obtained when testing whether there is a significant development in time or not. The gene expression data are not normalized. P-values below 0.01 are highlighted in yellow, while p-values between 0.05 and 0.01 are highlighted in blue. A p-value close to 1 means that there are fewer genes than expected in the curve group. The test for each curve group is based on $\boldsymbol{M_{a,i}}$ and the global test is based on $\boldsymbol{M_a}$.

	Not normalized gene expression data							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global test	0.13	5755	0.57	2323	0.71	2224	0.52	2808
123	0.71	67	0.73	57	0.58	121	0.29	423
132	0.41	168	0.69	76	0.91	27	0.87	41
312	0.87	21	0.42	195	0.24	652	0.31	336
321	0.044	3175	0.092	1773	0.25	667	0.70	75
231	0.070	2309	0.37	194	0.51	169	0.93	19
213	0.92	15	0.85	28	0.23	588	0.064	1914

Table 4 P-values obtained when testing whether there is a significant development in time or not. The gene expression data are normalized. P-values below 0.01 are highlighted in yellow, while p-values between 0.05 and 0.01 are highlighted in blue. A p-value close to 1 means that there are fewer genes than expected in the curve group. The test for each curve group is based on $\boldsymbol{M_{a,i}}$ and the global test is based on $\boldsymbol{M_a}$.

	Normalized gene expression data							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global	0.18	3940	0.84	2599	0.54	3400	0.71	2966
123	0.040	1128	0.25	708	0.32	671	0.80	384
132	0.48	441	0.88	360	0.83	376	0.48	573
312	0.67	344	0.67	474	0.77	400	0.50	561
321	0.022	1259	0.46	559	0.27	717	0.80	384
231	0.67	362	0.91	245	0.38	596	0.46	476
213	0.58	406	0.89	253	0.32	640	0.27	588

See Table 3 for results for data before normalization and Table 4 for results for data after normalization. We have used a=0.1 as this seems to give most significant values. Our interpretation is that there are few/no genes with a very strong signal, but there is a larger group of genes with quite strong signal since smaller a-values result in less significant results. The signal is so weak that if we analyze the genes one by one, we will not be able to separate signal from noise. We get significant results since we analyze many genes with a low signal in the same direction simultaneously. For data before normalization, we get significance for only one curve group and stratum. The data after normalization shown in Table 4 gives slightly stronger results with two significant values, where one is the same curve group as before normalization. Two significant values in 6x4=24 tests are not impressive. But since these two curve groups are where we a priori expected the strongest signal, we consider this as a

valuable result. YScrYSpr is the most homogenous stratum and the significant curve groups are for monotonically increasing or decreasing gene expressions.

The test indicates that we have around 1260 (1130) genes where the gene expressions decrease (increase) slightly in time. For the data before normalization we get less significant results. For these data the test indicates that we have around 3175 (67) genes where the gene expressions decrease (increase) slightly in time. These are the number of genes with strong functional form. The tendency for all members of the curve group is the same.

2.5 Models for spread/not spread

Our focus here is to identify whether it is possible to predict spread/not spread at time of diagnosis based on the data. In these tests we have used the same $\log_2$ case-control differences as in the previous tests.

2.5.1 Regression models

We have made covariates based on genes where we expect a difference between spread and not spread. Define the set of genes $G'_{C,S_1,S_2} = G_{C,S_1} \backslash G_{C,S_2}$. This set of genes includes for curve group C, the genes G_{C,S_1} with a strong functional form for stratum S_1, and it does not include the genes G_{C,S_2} with strong functional form for stratum S_2. For each set of genes G, we let $\#G$ denote the number of genes in the set. We then define the 24 covariates $\bar{X}_{1,C,p} = \frac{1}{\#G'_{C,S_1,S_2}} \sum_{g \in G'_{C,S_1,S_2}} D_{g,p}$ and the corresponding 24 covariates (4 time periods gives 24 combinations) $\bar{X}_{2,C,p} = \frac{1}{\#G'_{C,S_2,S_1}} \sum_{g \in G'_{C,S_2,S_1}} D_{g,p}$. In $\bar{X}_{1,C,p}$ we summarize over genes that have a strong functional form in S_1, but not in S_2, and in $\bar{X}_{2,C,p}$ we summarize over genes that have a strong functional form in S_2, but not in $S_1$for curve group C. Note that by including both $\bar{X}_{1,C,p}$ and $\bar{X}_{2,C,p}$, the two strata S_1 and S_2 are treated symmetrically. Note also that the number of genes in some of these curve groups is small.

We then use logit regression with covariates $\bar{X}_{1,C,p}$ and $\bar{X}_{2,C,p}$ for the significant curve groups where the response variable is equal to 1 if $p \in S_1$ and 0 if $p \in S_2$. Only pairs from one time period are included when estimating the model.

We have tested prediction of spread using these 48 covariates with different regressions methods on the dataset described in Section 3. We have selected a subset of covariates for the models either by using variable selection methods like Lasso or by selecting the covariates for the curve groups that have a much larger number of genes than expected. We have also tried selecting the covariates that have the most significant differences in gene expression values between the two strata. Neither of the resulting regression models gave sufficiently good predictions. Since the results were not promising, this was not tested on the dataset described in Section 2, but a description of the test has been included here to motivate the test described below.

2.5.2 Nearest neighbor models

In this section we focus on the two covariates with the most significant difference between the stratum with spread, YScrYSpr, and the stratum with not spread, YScrNSpr. For each curve group C we have identified the gene sets $G'_{C,YY,YN}$ and $G'_{C,YN,YY}$. For all case-control pairs we then compute the covariates $\bar{X}_{YY\backslash YN,C,p}$ and $\bar{X}_{YN\backslash YY,C,p}$ as the average gene expression value

for the genes in $G'_{C,YY,YN}$ and $G'_{C,YN,YY}$, respectively. For each time period, curve group and corresponding gene set, we then perform a t-test in order to find out whether these average values are significantly different between the two strata. The subscript YY\YN of $\bar{X}_{YY\backslash YN,C,p}$, shows that this covariate is based on genes that have a strong functional form in the YScrYSpr stratum, and not a strong functional form in the YScrNSpr stratum. Similarly for the YN\YY subscript of $\bar{X}_{YN\backslash YY,C,p}$.

We have identified the two curve groups with the most significant difference in values between spread and not spread for each time period. For time period 1 the two most significant curve groups are 213-YY\YN and 231-YY\YN. The numbers correspond to the numbers in Table 4. In both cases these are genes belonging to a curve group for YScrYSpr and not the same curve group for YScrNSpr. Figure 11 shows the difference between spread and not spread for the two most significant covariates for each of the three time periods separately. Notice that for time period 1 and 3 the points for spread are in the upper left corner while we do not have the same separation in time period 2.

We have used these covariates in a regression model without success. We have therefore tested an alternative model, a nearest neighbor model, that includes non-linear properties.

Based on two most significant covariates for each time period, we have developed a method for predicting spread or not spread. This method is tested using a leave one out approach. The test is as follows: We leave out one observation, find the genes in each curve group and identify the two covariates with most significant difference in a t-test for spread and not spread. Then we compute the covariates for the observation that is left out and compare these covariates with the corresponding covariates for all the other observations in the same time period. If all the other observations within a circle with radius r have the same diagnosis (spread/not spread), then we predict that the left out observation has the same diagnosis. The results are shown in Table 5. Note that we are able to identify about 1/3 of the negative (not spread) with few or no false negatives in period 1 and 3. We do not know why this is more difficult in period 2. Notice that we with radius 0.5 are able to identify in the first time period 16 of 44 negatives and for the third time period 11 of 34 negatives with no false negatives.

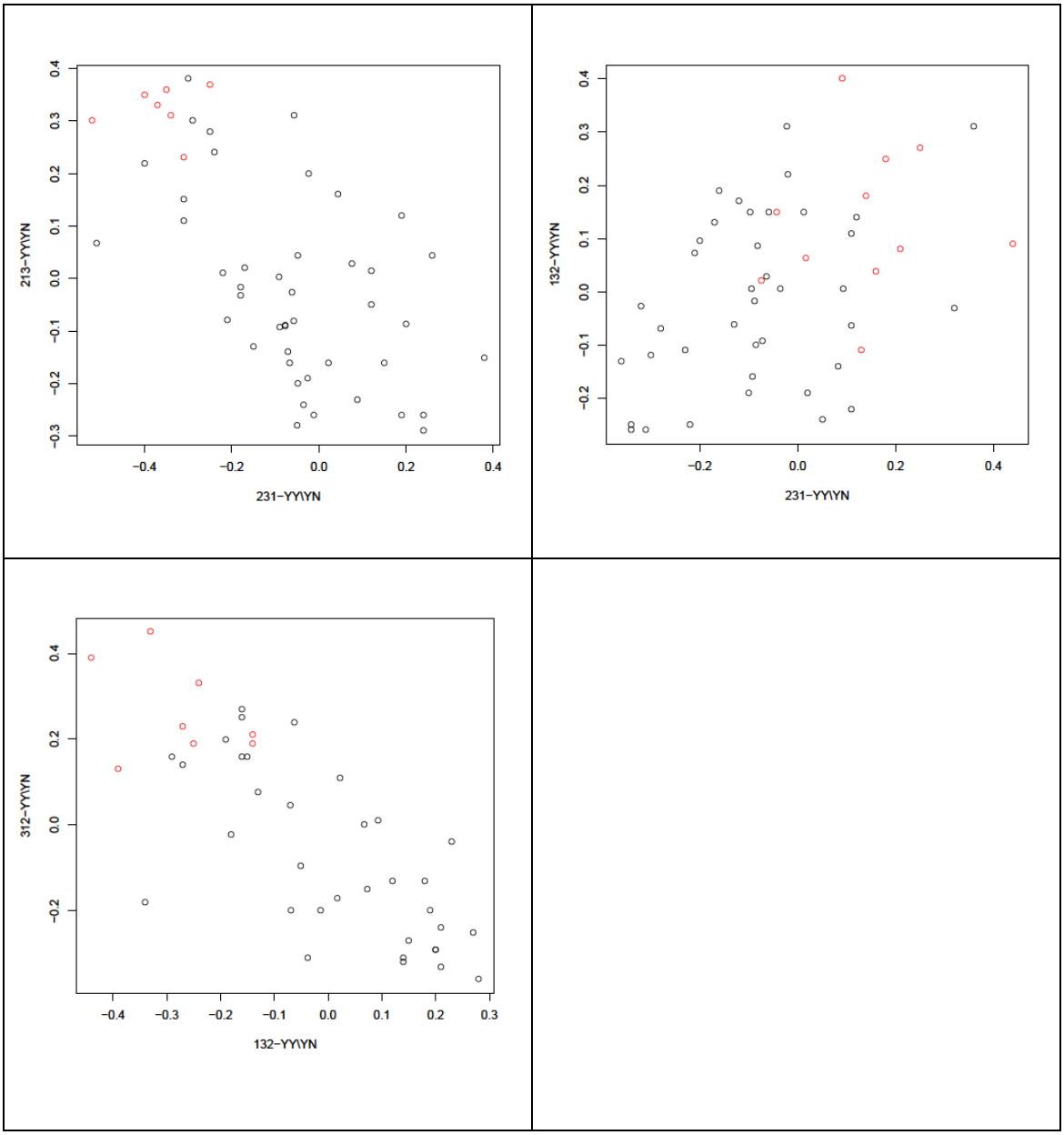

Figure 11 Case control pairs with spread (red) and not spread (black) for the two most important covariates for time period 1 (upper left), 2 (upper right) and 3 (lower left), respectively.

Table 5 Prediction of spread. Classification based on similar values in all points in a circle with the specified radius. The predicted values are divided into the categories: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) and No predictions. No predictions are for observations with no other observations within the circle or there are both observations with and without spread within the circle.

First time period (P 7, N 44)							
Most significant covariates		Radius	TP	TN	FP	FN	No pred.
Covariate	Frequency	0.10	0	32	1	5	12
231-YY\YN	47	0.15	0	34	0	6	11
213-YY\YN	51	0.20	0	35	0	6	10
123-YY\YN	4	0.25	0	32	0	6	13
		0.30	0	30	0	4	17
		0.35	0	27	0	2	22
		0.40	0	20	0	1	30
		0.45	0	18	0	0	33
		0.50	0	16	0	0	35
Second time period (P 11, N 39)							
Covariate	Frequency	Radius	TP	TN	FP	FN	No pred.
231-YY\YN	50	0.10	0	20	0	5	25
132-YY\YN	47	0.15	0	11	1	4	34
312-YY\YN	2	0.20	0	10	2	5	33
312-YN\YY	1	0.25	0	7	1	2	40
		0.30	0	5	0	1	44
		0.35	0	3	0	1	46
		0.40	0	0	0	0	50
		0.45	0	0	0	0	50
		0.50	0	0	0	0	50
Third time period (P 8, N 34)							
Covariate	Frequency	Radius	TP	TN	FP	FN	No pred.
132-YY\YN	41	0.10	1	20	0	5	16
213-YY\YN	1	0.15	0	24	0	6	12
123-YN\YY	2	0.20	0	22	0	5	15
312-YY\YN	40	0.25	0	22	0	5	15
		0.30	0	20	0	5	17
		0.35	0	19	0	3	20
		0.40	0	17	0	1	24
		0.45	0	12	0	1	29
		0.50	0	11	0	0	31

2.6 Some additional analyses

We have repeated the analyses in Section 2.4.1, but with only two time periods (year 1 and 2 are merged to one time period). Small p-values were not obtained for any of the curve groups. We also repeated the prediction analyses in Section 2.5 for two time periods. The obtained prediction results were not good. These negative results may indicate that two periods are not sufficient to discover trends in our dataset. There are only two curve groups, up and down.

With three periods there were six curve groups and the smallest p-values were obtained for curve groups that increased or decreased monotonically.

We have also repeated the analyses in Sections 2.4.1 and 2.5 without log-transforming the data, both with two and with three time periods. With two periods, small p-values were not obtained for any of the curve groups. With three periods, the p-values were similar to those Table 3 and Table 4. Both with two and three time periods the prediction results were far from promising.

3 Data at time of diagnosis and for four years before diagnosis

3.1 Data

The dataset presented in Section 2 is extended with case-control pairs after diagnosis that is represented as an additional year with t=0. The number of genes is reduced from 9060 to 8552. The original case and control data $D^*_{g,p,c}$ (c=1 case and c=2 for control) are available, but we are only using the $\log_2$-difference between case and control. See Table 6 for a summary of the number of observations in each stratum.

Table 6 Number of observations in each year before diagnosis

Stratum\ year	4	3	2	1	0
YScrYSpr	3	5	11	7	10
NScrYSpr	4	5	17	16	8
YScrNSpr	12	22	39	44	34
NScrNSpr	4	10	25	27	13

We merge the data in year 3 and 4 since there are few data in these periods. In the analyses we focus on differentiating between spread and not spread based on a large number of genes.

3.2 Normalization of the data

The tests are performed on data before normalization and with three different normalization methods:

- Normalization method 1: First compute $\log_2$-differences of the gene expression data for the case and control. Then quantile normalize these $\log_2$-differences.
- Normalization method 2: First compute $\log_2$ of the case and control gene expression data. Then quantile normalize the $\log_2$ gene expression data. Finally, compute the difference of the quantile normalized case and control data.
- Normalization method 3: First quantile normalize the gene expressions data for the case and control. Then compute $\log_2$-differences.

Normalization method 1 is the method used in Section2 and we believe normalization 3 is the most common method. We denote the $\log_2$-difference of case and control as $D_{g,p}$ in all four cases. Note that alternative 2 and 3 give very similar results. The next test is similar to the test described in Section 2.4.1 without the last time period.

3.3 Identify curve groups based on four time periods

For each stratum S, gene g and time period t we find the average gene expression, $\overline{D}_{S,g,t}$. All the genes are classified into curve groups based on the order of this average for the four time periods. F.ex. curve group 1234 includes genes where the average gene expression increases in time in the four time periods. We believe most genes are in a class based on coincidence. We denote that a gene has a strong functional form if a Mann-Whitney test on the order of the gene expression values in the time period with the largest and smallest average gives a p-value

smaller than a. Let $M_{a,i}$ be the number of genes in curve group i with a strong functional form. Define $M_a = \sum_i M_{a,i}$.

Then we simulate datasets based on the randomization $D_{g,r(p)}$ and compare the variables M_a and $M_{a,i}$ in the original and simulated datasets. For each simulated dataset, we find the curve group for each gene and the number of genes with a strong functional form in this curve group. Then we compare the number of genes with a strong functional form in the original dataset and the simulated datasets. The p-value is set to K/N+1/(2N) where K is the number of simulations with a larger number of genes with a strong functional form than was observed for the data and N is the number of simulation.

Detailed results for each curve group are presented in the Section 3.5 both for data before normalization and for the different normalization methods. We have selected a=0.01 based on the number of significant values and a reasonable number of genes in the group. The number of genes in the group should at least be 100 and the total number of genes with a strong functional form should be considerably smaller than the total number of genes. We have summarized the main results in Table 7 and Table 8.

Table 7 Number of significant curve groups. The numbers in the 0.05 columns include the numbers in the 0.01 columns.

	Number of significant curve groups									
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr		Sum	
p-value	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05
Not normalized	1	3	0	2	1	3	0	1	2	9
Normalization method 1	1	5	0	0	4	7	0	3	5	15
Normalization method 2	3	5	0	0	5	7	0	5	8	17
Normalization method 3	3	4	0	0	5	7	0	6	8	17

Table 8 Difference in order between the two last time periods for the significant curve groups (p<0.05) illustrated by curve group 1234 has difference 1, curve group 1342 has difference 2 and curve group 3214 has difference 3. There are 12 curve groups with difference 1, 8 curve groups with difference 2 and 4 curve groups with difference 3.

	Difference in order between two last time periods														
Dataset	YScrYSpr			NScrYSpr			YScrNSpr			NScrNSpr			Sum		
Difference in order	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Not normalized	0	1	2	0	0	2	2	1	0	0	0	1	2	2	5
Normalization method 1	0	2	2	0	0	0	4	2	0	0	1	2	4	5	4
Normalization method 2	0	2	2	0	0	0	4	2	0	1	2	2	5	6	4
Normalization method 3	0	2	2	0	0	0	4	2	0	2	2	2	6	6	4

Table 7 shows the number of significant curve groups without normalization and with all three normalization methods. There are a many significant curve groups, but most significant curve groups for normalization methods 2 and 3. Normalization methods 2 and 3 give almost identical numbers for each curve group. There are most significant curve groups in stratum YscrNSpr where there are most observations and then in stratum YScrYSpr where we might expect the strongest signal since it is more homogeneous than the non-screening group and the spread is expected to give a stronger signal.

Naturally, the significant curve groups are almost the same for the different normalization method. More interesting are the numbers shown in Table 8. Gene expressions tend to have a large shift in value between the last two time periods in stratum YScrYSpr (typically 3214 and 2341) and a small shift in value in stratum YScrNSpr (typically 1234 and 4321). Stratum NScrNSpr is believed to be a more heterogeneous group and have values between these extremes. There are no significant values in the stratum NScrYSpr which has the smallest number of observations. But also here we get the smallest p-values for similar curve groups as for stratum YScrYSpr. Hence, it seems that spread implies a shift in value for a quite large number of genes while not spread gives a more continuous development.

3.4 Prediction of spread

We first tested the method described in Section 2.5.1, but as described in that section, it did not give a promising result. It is based on average values from genes with strong functional form in the same curve groups. But we do not know actually whether the values differ between spread and not spread. Then we tested the method described in Section 2.5.2. We have calculated the values of the covariates $\bar{X}_{YY\backslash YN,C,p}$ and $\bar{X}_{YN\backslash YY,C,p}$ described in Section 2.5.2. The subscript YY\YN shows that this covariate is based on genes that have a strong functional form in the YScrYSpr stratum and not a strong functional form in the YScrNSpr stratum. Then we have performed a t-test in order to find out whether the average value is significantly different between the two strata. Results are shown in Table 9-Table 11. For the data before normalization there is no significant difference between the covariates. For normalization 1 (3) there are 4 (2) significant curve groups for the YScrYSpr stratum and no for the YScrNSpr stratum at a 0.01 level with significant difference in average value. Note that this test is performed based on the curve groups and should not be interpreted as ordinary p-values. There is a considerable overlap between these curve groups and the curve groups that had a significant number of genes with strong functional form in the YScrNSpr stratum. However, there is no overlap with the curve groups that are significant for the YScrYSpr stratum. Hence, we try with different covariates here than in the test described in Section 2.5.1. The YScrYSpr stratum has the more extreme values for these curve groups, i.e. smallest values for curve groups ending with 1 (i.e. 4321) and largest values for curve groups ending with 4.

Table 9 Value of covariates based on not normalized gene expression data. P-values are found from a t-test comparing average values for the two strata. We get NA if there are no genes in the curve group.

	Non normalized gene expression data, covariate values					
Dataset	$\bar{X}_{YY\backslash YN,C,p}$			$\bar{X}_{YN\backslash YY,C,p}$		
Curve group	YScrYSpr	YScrNSpr	p-value	YScrYSpr	YScrNSpr	p-value
1234	NA	NA	1	0.56	0.78	0.54
1243	NA	NA	1	NA	NA	1
1423	0.047	0.056	0.88	NA	NA	1
4123	0.22	-0.56	0.25	0.023	0.048	0.91
4132	NA	NA	1	-0.12	0.015	0.67
1432	-0.085	-0.082	0.99	NA	NA	1

1342	0.013	-0.12	0.48	NA	NA	1
1324	0.29	0.14	0.054	0.5	0.63	0.66
3124	0.61	0.42	0.44	0.37	0.49	0.66
3142	NA	NA	1	0.029	0.15	0.41
3412	-0.051	0.026	0.73	-0.03	0.047	0.29
4312	-0.052	0.055	0.68	-0.12	-0.26	0.49
4321	NA	NA	1	-0.27	-0.35	0.72
3421	-0.67	-0.31	0.28	-0.32	-0.3	0.91
3241	-0.5	-0.18	0.13	-0.29	-0.22	0.73
3214	0.64	0.45	0.57	0.24	0.24	0.97
2314	0.47	0.31	0.6	0.34	0.28	0.77
2341	-0.39	-0.13	0.2	-0.3	0.14	0.23
2431	-0.41	-0.18	0.2	-0.11	-0.081	0.82
4231	NA	NA	1	-0.24	-0.28	0.89
4213	0.12	0.16	0.89	0.044	0.011	0.78
2413	0.097	0.074	0.91	0.09	-0.00063	0.32
2143	NA	NA	1	NA	NA	1
2134	0.61	0.33	0.2	0.47	0.64	0.64

Table 10 Value of covariates based on gene expression data normalized with method 1. P-values are found from a t-test comparing average values for the two strata.

	Normalized gene expression data, method 1, covariate values					
Dataset	$\bar{X}_{YY\setminus YN,C,p}$			$\bar{X}_{YN\setminus YY,C,p}$		
Curve group	YScrYSpr	YScrNSpr	p-value	YScrYSpr	YScrNSpr	p-value
1234	0.38	0.17	5.00E-05	0.29	0.41	0.33
1243	0.029	-0.044	0.38	0.045	0.05	0.92
1423	-0.027	-0.011	0.88	0.11	0.085	0.75
4123	0.1	0.083	0.76	0.0026	-0.07	0.43
4132	0.034	-0.33	0.21	-0.094	-0.061	0.76
1432	-0.13	-0.098	0.61	-0.046	0.038	0.35
1342	-0.1	-0.11	0.98	-0.04	-0.0083	0.53
1324	0.48	0.25	0.022	0.25	0.34	0.36
3124	0.47	0.29	0.1	0.21	0.28	0.49
3142	-0.042	-0.053	0.89	-0.031	-0.0039	0.75
3412	-0.1	-0.056	0.43	-0.031	-0.094	0.32
4312	-0.13	-0.038	0.38	-0.099	-0.23	0.21
4321	-0.23	-0.07	0.0098	-0.2	-0.28	0.35
3421	-0.3	-0.15	0.052	-0.18	-0.2	0.79
3241	-0.23	-0.11	0.17	-0.19	-0.17	0.75
3214	0.43	0.29	0.33	0.19	0.22	0.5
2314	0.37	0.19	0.16	0.24	0.22	0.72
2341	-0.27	-0.16	0.21	-0.18	-0.15	0.74

2431	-0.3	-0.17	0.076	-0.16	-0.13	0.66
4231	-0.26	-0.054	0.00031	-0.21	-0.25	0.51
4213	0.14	0.13	0.9	0.02	-0.025	0.55
2413	0.056	0.024	0.67	0.081	-0.044	0.07
2143	0.052	-0.0096	0.41	0.051	0.049	0.99
2134	0.46	0.24	0.0034	0.24	0.33	0.47

Table 11 Value of covariates based on gene expression data normalized with method 3. P-values are found from a t-test comparing average values for the two strata.

	Normalized gene expression data, method 3, covariate values					
Dataset	$\bar{X}_{YY\setminus YN,C,p}$			$\bar{X}_{YN\setminus YY,C,p}$		
Curve group	YScrYSpr	YScrNSpr	p-value	YScrYSpr	YScrNSpr	p-value
1234	0.36	0.15	0.0029	0.37	0.54	0.38
1243	0.061	0.0064	0.61	0.072	0.087	0.82
1423	0.03	-0.034	0.59	0.17	0.051	0.55
4123	0.13	0.099	0.68	0.01	-0.029	0.68
4132	-0.073	-0.21	0.29	-0.049	-0.028	0.82
1432	-0.15	-0.11	0.66	-0.045	-0.067	0.75
1342	-0.15	-0.15	0.98	-0.044	0.013	0.42
1324	0.45	0.22	0.034	0.3	0.42	0.36
3124	0.53	0.32	0.23	0.25	0.32	0.58
3142	-0.086	-0.095	0.92	-0.028	0.029	0.45
3412	-0.088	-0.046	0.48	-0.049	-0.078	0.73
4312	-0.12	-0.02	0.26	-0.069	-0.14	0.32
4321	-0.33	-0.12	0.00032	-0.25	-0.34	0.43
3421	-0.35	-0.18	0.05	-0.22	-0.26	0.73
3241	-0.37	-0.2	0.11	-0.2	-0.21	0.97
3214	0.59	0.38	0.3	0.2	0.24	0.59
2314	0.5	0.29	0.22	0.24	0.21	0.74
2341	-0.37	-0.23	0.23	-0.21	-0.22	0.94
2431	-0.33	-0.18	0.12	-0.21	-0.23	0.9
4231	-0.37	-0.16	0.018	-0.23	-0.3	0.48
4213	0.1	0.11	0.9	0.045	-0.085	0.16
2413	0.051	0.018	0.66	0.06	-0.061	0.16
2143	-0.048	-0.057	0.95	0.043	0.082	0.71
2134	0.49	0.28	0.13	0.29	0.41	0.5

We have predicted spread or not spread using regression based on a partly leave-one-out approach for normalization methods 1 and 3. The result was not sufficiently good for being used for prediction. We believe that if we had performed a t-test or a Mann-Whitney test, this will give a significant difference between the two strata. But the difference will not be sufficiently good for prediction.

We have also made predictions using a nearest neighbor approach as in Section 2.5.2. We classified the observations based on the values for the two most significant covariates shown in Figure 12. Note that observations with spread tend to be in the lower right hand side of the figure. Based on these two covariates, we have classified the observations in 3 categories: not spread, spread and uncertain. For each observation we use all observations within a circle with radius r. If all these observations have the same category, we predict that the observation in the middle of the circle also belongs to this category. The results of using this strategy are shown in Table 12. There are 34 observations without spread and 10 with spread. With radius equal to 0.166, we predicted no spread for 12 of the patients and spread for 2 of the patents, i.e. there are prediction for slightly more than 30% of the patients. All these predictions are correct. Note that in this classification we have used all the data for identifying the curve groups and choosing these two covariates. If we want to publish this, we need to use a leave-one-out strategy also in these two steps. Such tests were performed in Section 2.5.2. We do not expect as good results if these steps are included.

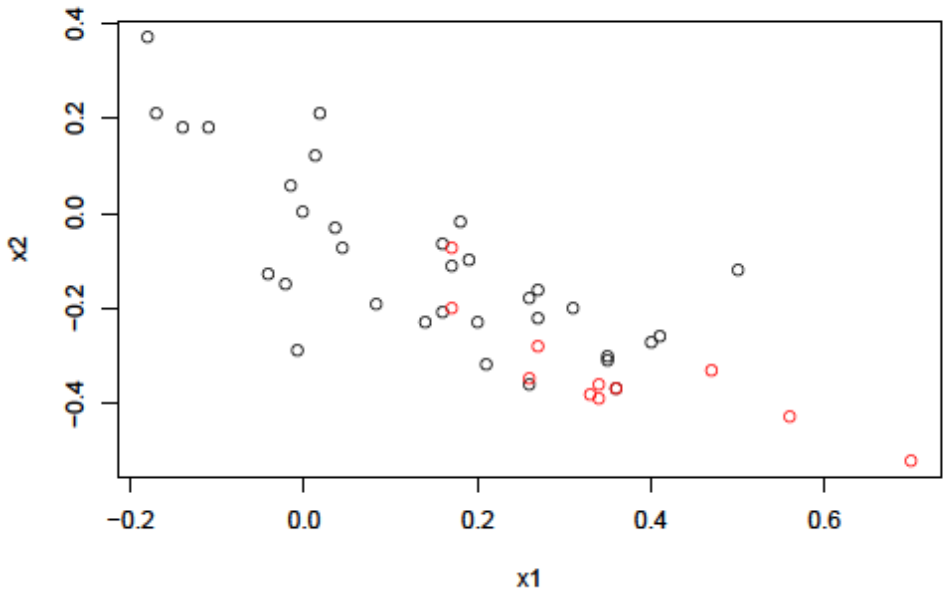

Figure 12 Values for the two significant covariates with normalization 3 for the 44 observations with diagnosis. Observations with spread are shown in red.

Table 12 Prediction of spread or not spread. Classification based on similar values in all points in a circle with the specified radius. The predicted values are divided into the categories: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) and No predictions. No predictions are for observations with no other observations within the circle or there are both observations with and without spread within the circle.

Radius	TP	TN	FP	FN	No pred.
0.033	0	13	2	3	26
0.066	0	15	0	4	25
0.100	0	12	0	3	29
0.133	0	10	0	0	34
0.166	2	12	0	0	30
0.200	2	9	0	0	33

3.5 Detailed results for number of genes in curve groups

Tables with detailed results for number of genes in curve groups for not normalized and 3 different normalization methods.

Table 13 P-values obtained when testing whether there is a significant development in time or not. The gene expression data are not normalized. P-values below 0.01 are highlighted in yellow, while p-values between 0.05 and 0.01 are highlighted in blue. A p-value close to 1 means that there are fewer genes than expected in the curve group. The test for each curve group is based on $\boldsymbol{M}_{a,i}$ and the global test is based on $\boldsymbol{M}_{a}$.

	Not normalized gene expression data							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global	0.068	2111	0.26	560	0.11	2100	0.53	272
1234	0.65	1	0.13	19	0.024	295	0.41	5
1243	1	0	0.69	1	1	0	0.33	5
1423	0.31	4	1	0	1	0	1	0
4123	0.64	1	0.69	1	0.29	8	0.48	3
4132	1	0	1	0	0.5	3	0.27	7
1432	0.27	5	1	0	1	0	0.73	1
1342	0.47	2	0.71	1	1	0	0.48	3
1324	0.28	6	0.068	47	0.098	47	1	0
3124	0.15	16	0.088	35	0.022	268	0.25	10
3142	1	0	1	0	0.72	1	0.19	12
3412	0.072	27	0.63	1	0.3	7	0.7	1
4312	0.056	40	1	0	0.13	30	0.69	1
4321	1	0	1	0	0.094	50	0.24	11
3421	0.39	3	1	0	0.28	8	1	0
3241	0.51	2	0.52	2	0.49	3	0.046	95
3214	0.0045	1383	0.02	328	0.062	90	0.27	9
2314	0.012	502	0.036	95	0.082	64	0.3	7
2341	0.35	4	0.55	2	0.57	2	0.096	39
2431	0.51	2	0.69	1	0.74	1	0.39	5
4231	1	0	1	0	0.15	26	0.12	32
4213	0.044	60	0.29	6	0.17	16	1	0
2413	0.064	50	1	0	0.31	6	1	0
2143	1	0	1	0	1	0	0.17	14
2134	0.43	3	0.11	21	0.0085	1175	0.2	12

Table 14 P-values obtained when testing whether there is a significant development in time or not. The gene expression data are normalized by method 1. P-values below 0.01 are highlighted in yellow, while p-values between 0.05 and 0.01 are highlighted in blue. A p-value close to 1 means that there are fewer genes than expected in the curve group. The test for each curve group is based on $\boldsymbol{M_{a,i}}$ and the global test is based on $\boldsymbol{M_a}$.

	Normalized gene expression data by method 1							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global	0.044	1396	0.58	478	0.0015	2631	0.16	1004
1234	0.27	23	0.65	11	0.0015	569	0.26	38
1243	0.41	14	0.64	11	0.57	14	0.82	7
1423	0.81	5	0.77	7	0.84	7	0.25	29
4123	0.7	7	0.45	17	0.86	6	0.6	14
4132	0.96	2	0.6	11	0.56	15	0.69	10
1432	0.18	31	0.66	10	0.96	3	0.39	22
1342	0.066	71	0.12	57	0.66	11	0.82	8
1324	0.61	9	0.56	14	0.022	174	0.15	52
3124	0.022	143	0.33	24	0.054	99	0.096	77
3142	0.43	14	0.81	6	0.74	9	0.67	9
3412	0.4	14	0.98	2	0.51	16	0.78	7
4312	0.48	12	0.73	9	0.5	16	0.55	13
4321	0.28	22	0.72	9	5,00E-04	464	0.088	86
3421	0.068	78	0.5	14	0.014	218	0.11	54
3241	0.11	60	0.13	51	0.11	67	0.028	111
3214	0.0025	407	0.18	44	0.25	32	0.32	28
2314	0.15	51	0.62	12	0.18	44	0.23	32
2341	0.016	170	0.052	96	0.096	77	0.04	100
2431	0.03	127	0.32	25	0.088	73	0.088	74
4231	0.55	11	0.62	12	0.02	191	0.034	130
4213	0.11	47	0.44	21	0.93	4	0.57	15
2413	0.8	5	0.98	2	0.42	20	0.58	11
2143	0.46	12	0.98	2	0.61	12	0.52	13
2134	0.11	61	0.58	11	0.0025	490	0.088	64

Table 15 P-values obtained when testing whether there is a significant development in time or not. The gene expression data are normalized by method 2. P-values below 0.01 are highlighted in yellow, while p-values between 0.05 and 0.01 are highlighted in blue. A p-value close to 1 means that there are fewer genes than expected in the curve group. The test for each curve group is based on $\boldsymbol{M_{a,i}}$ and the global test is based on $\boldsymbol{M_a}$.

	Normalized gene expression data by method 2							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global	0.016	1648	0.51	533	0.0015	3499	0.11	1176
1234	0.46	13	0.61	13	0.0015	799	0.19	45
1243	0.88	4	0.67	11	0.86	6	0.85	6
1423	0.9	4	0.87	6	1	1	0.68	10
4123	0.53	11	0.71	10	0.91	5	0.85	8
4132	0.84	5	0.81	7	0.81	8	0.96	4
1432	0.31	19	0.66	11	0.93	4	0.5	17
1342	0.14	38	0.19	41	0.88	5	0.94	5
1324	0.28	24	0.58	14	0.018	195	0.29	32
3124	0.012	201	0.28	28	0.086	75	0.044	126
3142	0.76	6	0.87	5	0.68	10	0.96	3
3412	0.83	5	0.88	5	0.2	36	0.57	11
4312	0.5	12	0.96	4	0.44	18	0.66	10
4321	0.18	31	0.58	14	5,00E-04	710	0.084	88
3421	0.12	58	0.57	13	0.0095	374	0.024	134
3241	0.062	88	0.08	65	0.056	97	0.048	80
3214	0.0015	505	0.084	69	0.14	48	0.21	40
2314	0.09	70	0.43	20	0.11	61	0.15	41
2341	0.0015	278	0.052	97	0.12	62	0.036	113
2431	0.0095	183	0.18	39	0.074	77	0.016	197
4231	0.18	31	0.36	23	0.01	311	0.09	79
4213	0.33	19	0.57	16	0.9	5	0.73	10
2413	0.76	6	0.93	4	0.35	22	0.86	5
2143	0.93	3	0.91	4	0.22	30	0.79	7
2134	0.2	34	0.51	14	0.0015	540	0.036	105

Table 16 P-values obtained when testing whether there is a significant development in time or not. The gene expression data are normalized by method 3. P-values below 0.01 are highlighted in yellow, while p-values between 0.05 and 0.01 are highlighted in blue. A p-value close to 1 means that there are fewer genes than expected in the curve group. The test for each curve group is based on $\boldsymbol{M_{a,i}}$ and the global test is based on $\boldsymbol{M_a}$.

	Normalized gene expression data by method 3							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global	0.016	1647	0.51	534	0.0015	3500	0.11	1180
1234	0.43	14	0.61	13	0.0015	793	0.19	45
1243	0.88	4	0.67	11	0.86	6	0.85	6
1423	0.9	4	0.87	6	0.99	2	0.68	10
4123	0.53	11	0.72	10	0.91	5	0.85	8
4132	0.83	5	0.8	7	0.81	8	0.96	4
1432	0.33	18	0.67	11	0.93	4	0.5	17
1342	0.14	38	0.2	40	0.84	6	0.94	5
1324	0.29	23	0.57	14	0.018	194	0.29	32
3124	0.012	200	0.27	29	0.086	76	0.042	127
3142	0.76	6	0.87	5	0.68	10	0.93	4
3412	0.83	5	0.87	5	0.21	35	0.57	11
4312	0.47	13	0.96	4	0.42	19	0.67	10
4321	0.18	31	0.57	14	5,00E-04	709	0.084	88
3421	0.12	57	0.57	13	0.0095	375	0.026	130
3241	0.058	90	0.078	66	0.056	97	0.044	82
3214	0.0015	504	0.084	69	0.14	49	0.21	40
2314	0.086	72	0.43	20	0.1	62	0.16	41
2341	0.0015	278	0.052	97	0.11	63	0.036	114
2431	0.0095	182	0.17	39	0.076	77	0.016	196
4231	0.18	31	0.36	23	0.01	313	0.088	80
4213	0.34	19	0.57	16	0.9	5	0.73	10
2413	0.76	6	0.93	4	0.33	23	0.86	5
2143	0.98	2	0.92	4	0.24	29	0.78	7
2134	0.2	34	0.51	14	0.0015	540	0.034	108

3.6 Detailed results for predictions

The subsection gives the predictions for each case-control pair and a list of significant genes. We study the 44 case-control pairs with Screening and from year 0 (i.e. at time of diagnosis) where 10 have spread and 34 have not spread.

Table 17 These are the genes in the two most significant covariates in Table 9. The covariates are shown in Figure 1 and are used in Table 10. There are 13 genes in the left column and 24 genes in the right column. These genes are selected from the genes in the curve groups in Table 16, where YY,1234 has 14 genes and YY,4321 has 31 genes. We then take out the genes that also are present in YN,1234 and YN,4321 respectively.

$\bar{X}_{YY\setminus YN,1234,p}$	$\bar{X}_{YY\setminus YN,4321,p}$
gene_195	gene_56
gene_1276	gene_272
gene_1300	gene_740
gene_1408	gene_769
gene_2638	gene_903
gene_2708	gene_1146
gene_3519	gene_1240
gene_3834	gene_1362
gene_3867	gene_1405
gene_4194	gene_1869
gene_5108	gene_2463
gene_7145	gene_2579
gene_7753	gene_2794
	gene_3204
	gene_3433
	gene_3938
	gene_4278
	gene_4367
	gene_4444
	gene_4483
	gene_4857
	gene_5385
	gene_6614
	gene_8367

Table 18 These are the PatientID, covariates, spread/no spread and classification results for the 44 case-control pairs with screening and year 0. The classification is as summarized in Table 10 for radius 1.66 with 2 true positives and 12 true negatives.

No	PasientID	Curve group 1234	Curve group 4321	Spread (1) or not (0)	Classification result
1	252	0.34	-0.39	1	
2	259	0.27	-0.28	1	
3	261	0.47	-0.33	1	
4	262	0.17	-0.2	1	
5	271	0.36	-0.37	1	
6	277	0.33	-0.38	1	
7	285	0.17	-0.072	1	
8	294	0.56	-0.43	1	TP
9	298	0.7	-0.52	1	TP

10	310	0.26	-0.35	1	
11	256	-0.0013	0.0015	0	TN
12	263	0.083	-0.19	0	
13	265	0.26	-0.36	0	
14	266	0.21	-0.32	0	
15	267	-0.021	-0.15	0	TN
16	268	0.19	-0.1	0	
17	269	0.4	-0.27	0	
18	272	0.41	-0.26	0	
19	273	0.26	-0.18	0	
20	274	0.5	-0.12	0	TN
21	275	0.2	-0.23	0	
22	276	-0.17	0.21	0	TN
23	278	0.35	-0.3	0	
24	279	0.16	-0.066	0	
25	280	0.013	0.12	0	TN
26	281	0.044	-0.073	0	
27	283	0.018	0.21	0	TN
28	286	-0.18	0.37	0	TN
29	287	0.14	-0.23	0	
30	288	-0.041	-0.13	0	TN
31	290	-0.0074	-0.29	0	TN
32	293	-0.11	0.18	0	TN
33	300	0.31	-0.2	0	
34	301	0.36	-0.37	0	
35	302	0.16	-0.21	0	
36	303	0.17	-0.11	0	
37	304	0.036	-0.03	0	
38	305	0.27	-0.22	0	
39	307	0.27	-0.16	0	
40	309	-0.14	0.18	0	
41	313	-0.015	0.056	0	TN
42	314	0.18	-0.02	0	TN
43	315	0.35	-0.31	0	
44	316	0.34	-0.36	0	

4 Updated data including test sets hcc2 and hcc3

4.1 Data

The dataset contains many of the same case-control pairs as the dataset presented in Section 3 and also some new pairs, but fewer genes (6952) are included. For making predictions, two test datasets are available. These test dataset are obtained using chips with slightly different designs. Table 19 shows the number of case-control pairs in the dataset. The dataset with hcc1 is used in the estimation and the datasets hcc2 and hcc3 are used as test datasets. We merge year 3-5 in order to get sufficient data in this time period.

Table 19 Number of case-control pairs in each stratum and year.

Stratum\ year	5	4	3	2	1	0 (hcc1 + hcc2 + hcc3)
YScrYSpr	0	3	4	11	7	11 + 7 + 5
NScrYSpr	1	4	5	16	16	8 + 4 + 4
YScrNSpr	0	12	23	39	44	33 + 21 + 26
NScrNSpr	2	4	10	25	23	12 + 10 + 18

4.2 Results

First we show the same tables as in Section 3.5 which are the detailed results with 4 time periods and not normalized and 3 different normalization methods. Note that the results are very similar to the other dataset. It is almost the same curve groups that have a significantly high number of genes.

Table 20 P-values obtained when testing whether there is a significant development in time or not. The gene expression data are not normalized. P-values below 0.01 are highlighted in yellow, while p-values between 0.05 and 0.01 are highlighted in blue. A p-value close to 1 means that there are fewer genes than expected in the curve group. The test for each curve group is based on $\boldsymbol{M_{a,i}}$ and the global test is based on $\boldsymbol{M_a}$.

	Not normalized gene expression data							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global	0.076	1592	0.38	277	0.034	2524	0.29	472
1234	0.42	1	0.16	9	0.0035	1433	1	0
1243	1	0	0.17	6	1	0	1	0
1423	0.26	2	1	0	1	0	1	0
4123	1	0	1	0	1	0	0.17	12
4132	1	0	1	0	1	0	0.084	36
1432	0.27	2	1	0	1	0	1	0
1342	0.19	5	0.49	1	0.53	1	1	0
1324	0.16	6	0.14	12	0.0075	652	1	0
3124	0.31	2	0.25	4	0.11	22	0.14	13
3142	1	0	1	0	1	0	0.14	10
3412	0.14	6	1	0	0.29	4	1	0
4312	0.15	7	1	0	0.19	9	0.54	1
4321	1	0	1	0	0.17	12	0.27	6
3421	1	0	1	0	0.21	8	1	0
3241	1	0	1	0	1	0	0.074	37
3214	0.0025	1166	0.038	125	0.09	26	0.2	8
2314	0.0095	346	0.036	111	0.028	181	1	0
2341	1	0	1	0	0.55	1	1	0
2431	1	0	1	0	1	0	0.54	1
4231	1	0	1	0	1	0	0.024	343
4213	0.082	26	0.3	3	0.56	1	0.56	1
2413	0.076	23	1	0	0.52	1	1	0
2143	1	0	1	0	1	0	0.39	2
2134	1	0	0.19	6	0.03	173	0.4	2

Table 21 P-values obtained when testing whether there is a significant development in time or not. The gene expression data are normalized by method 1. P-values below 0.01 are highlighted in yellow, while p-values between 0.05 and 0.01 are highlighted in blue. A p-value close to 1 means that there are fewer genes than expected in the curve group. The test for each curve group is based on $\boldsymbol{M_{a,i}}$ and the global test is based on $\boldsymbol{M_a}$.

	Normalized gene expression data by method 1							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global	0.036	1120	0.3	602	5,00E-04	2547	0.11	908
1234	0.61	7	0.94	3	0.0015	683	0.17	44
1243	0.86	4	0.42	15	0.62	9	0.84	5
1423	0.87	3	0.75	7	0.99	1	0.68	8
4123	0.28	17	0.54	13	0.94	3	1	1
4132	0.98	1	0.73	7	0.83	5	0.73	7
1432	0.19	26	0.82	6	0.98	2	0.84	5
1342	0.044	95	0.036	94	0.73	7	0.97	2
1324	0.86	3	0.56	11	0.012	193	0.062	84
3124	0.01	129	0.29	24	0.17	39	0.14	42
3142	0.88	3	0.88	3	0.96	2	0.85	4
3412	0.6	7	0.99	1	0.55	11	0.81	5
4312	0.6	9	0.36	18	0.44	15	0.97	2
4321	0.47	10	0.6	9	5,00E-04	659	0.046	91
3421	0.14	34	0.64	9	0.01	175	0.092	49
3241	0.21	27	0.11	43	0.11	52	0.024	105
3214	5,00E-04	361	0.056	65	0.45	15	0.17	35
2314	0.23	24	0.28	23	0.14	42	0.088	53
2341	0.0095	163	0.0075	160	0.21	33	0.028	94
2431	0.016	129	0.31	24	0.14	42	0.068	63
4231	0.73	5	0.46	14	0.0095	218	0.014	168
4213	0.15	38	0.23	30	0.98	1	0.83	5
2413	0.87	3	0.54	10	0.38	17	0.92	3
2143	1	0	0.83	4	0.66	8	0.74	6
2134	0.24	22	0.62	9	0.0015	315	0.21	27

Table 22 P-values obtained when testing whether there is a significant development in time or not. The gene expression data are normalized by method 2. P-values below 0.01 are highlighted in yellow, while p-values between 0.05 and 0.01 are highlighted in blue. A p-value close to 1 means that there are fewer genes than expected in the curve group. The test for each curve group is based on $\boldsymbol{M_{a,i}}$ and the global test is based on $\boldsymbol{M_a}$.

	Normalized gene expression data by method 2							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global	0.014	1553	0.6	396	5,00E-04	2985	0.15	770
1234	0.85	3	0.61	10	5,00E-04	814	0.18	39
1243	0.89	4	0.8	6	0.67	8	0.73	7
1423	0.87	3	0.76	7	0.9	4	0.65	9
4123	0.27	17	0.62	11	0.99	1	0.92	4
4132	0.95	2	0.67	9	0.97	2	0.95	3
1432	0.23	22	0.42	18	0.98	2	0.52	13
1342	0.076	63	0.2	31	0.93	3	0.71	8
1324	0.61	7	0.54	12	0.0065	270	0.15	45
3124	0.01	178	0.48	15	0.2	32	0.092	51
3142	0.95	2	0.62	8	0.65	8	0.61	9
3412	0.65	6	1	0	0.31	19	0.77	6
4312	0.62	9	0.92	4	0.33	19	0.95	3
4321	0.35	13	0.39	17	5,00E-04	718	0.074	73
3421	0.07	56	0.86	5	0.0075	205	0.038	75
3241	0.068	63	0.15	33	0.034	94	0.06	66
3214	5,00E-04	516	0.13	41	0.22	27	0.18	33
2314	0.076	54	0.26	24	0.064	66	0.22	26
2341	0.0035	284	0.082	52	0.092	54	0.046	72
2431	0.012	191	0.22	32	0.16	39	0.048	68
4231	0.39	13	0.2	30	0.0065	313	0.044	106
4213	0.24	26	0.5	14	0.83	5	0.88	5
2413	0.82	4	0.96	2	0.3	20	0.76	6
2143	0.99	1	0.92	3	0.4	15	0.82	5
2134	0.34	16	0.54	12	0.0045	247	0.12	38

Table 23 P-values obtained when testing whether there is a significant development in time or not. The gene expression data are normalized by method 3. P-values below 0.01 are highlighted in yellow, while p-values between 0.05 and 0.01 are highlighted in blue. A p-value close to 1 means that there are fewer genes than expected in the curve group. The test for each curve group is based on $\boldsymbol{M_{a,i}}$ and the global test is based on $\boldsymbol{M_a}$.

	Normalized gene expression data by method 3							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global	0.014	1554	0.59	401	5,00E-04	2981	0.16	767
1234	0.85	3	0.61	10	5,00E-04	814	0.2	38
1243	0.89	4	0.79	6	0.67	8	0.74	7
1423	0.87	3	0.76	7	0.9	4	0.7	8
4123	0.27	17	0.62	11	0.99	1	0.92	4
4132	0.95	2	0.68	9	0.97	2	0.95	3
1432	0.23	22	0.41	18	0.98	2	0.49	14
1342	0.076	63	0.2	32	0.93	3	0.66	9
1324	0.61	7	0.54	12	0.0065	268	0.16	45
3124	0.01	176	0.51	14	0.2	32	0.094	50
3142	0.95	2	0.62	8	0.66	8	0.61	9
3412	0.65	6	1	0	0.29	20	0.77	6
4312	0.61	9	0.92	4	0.33	19	0.91	4
4321	0.33	14	0.39	17	5,00E-04	714	0.072	73
3421	0.068	56	0.8	6	0.0075	206	0.04	75
3241	0.066	64	0.13	35	0.032	96	0.062	65
3214	5,00E-04	517	0.13	41	0.25	25	0.17	34
2314	0.076	53	0.26	24	0.068	65	0.23	26
2341	0.0035	283	0.08	53	0.096	53	0.046	71
2431	0.012	193	0.21	33	0.16	39	0.05	68
4231	0.39	13	0.21	30	0.0055	315	0.044	105
4213	0.24	26	0.5	14	0.83	5	0.88	5
2413	0.82	4	0.96	2	0.32	19	0.83	5
2143	0.99	1	0.92	3	0.42	14	0.81	5
2134	0.34	16	0.54	12	0.0045	249	0.12	38

Then we test the significance of the covariate values for the genes in the different curve groups. Also here there is good correspondence between the two dataset.

Table 24 Value of covariates based on non-normalized gene expression data. P-values are found from a t-test comparing average values for the two strata. We get NA if there are no genes in the curve group.

	Non normalized gene expression data, covariate values							
Dataset	$\bar{X}_{YY\setminus YN,C,p}$				$\bar{X}_{YN\setminus YY,C,p}$			
Curve group	YScrYSpr	YScrNSpr	p-value	nofGenes	YScrYSpr	YScrNSpr	p-value	nofGenes
1234	0.065	0.01	0.21	1	0.41	0.67	0.45	1433
1243	NA	NA	1	0	NA	NA	1	0
1423	0.32	-0.11	0.2	2	NA	NA	1	0
4123	NA	NA	1	0	NA	NA	1	0
4132	NA	NA	1	0	NA	NA	1	0
1432	0.14	-0.13	0.39	2	NA	NA	1	0
1342	-0.054	-0.22	0.64	5	-0.012	-0.2	0.5	1
1324	0.36	0.12	0.12	6	0.35	0.54	0.54	652
3124	0.86	0.51	0.23	2	0.35	0.58	0.49	22
3142	NA	NA	1	0	NA	NA	1	0
3412	0.00057	0.028	0.92	6	-0.15	0.04	0.51	4
4312	0.00027	-0.0071	0.98	7	-0.24	-0.23	0.98	9
4321	NA	NA	1	0	-0.38	-0.26	0.71	12
3421	NA	NA	1	0	-0.33	-0.19	0.63	8
3241	NA	NA	1	0	NA	NA	1	0
3214	0.8	0.46	0.37	1162	0.2	0.21	0.95	22
2314	0.56	0.28	0.4	333	0.23	0.25	0.92	168
2341	NA	NA	1	0	-0.54	0.3	0.19	1
2431	NA	NA	1	0	NA	NA	1	0
4231	NA	NA	1	0	NA	NA	1	0
4213	0.18	0.12	0.85	25	NA	NA	1	0
2413	0.084	0.045	0.89	23	0.27	0.33	0.85	1
2143	NA	NA	1	0	NA	NA	1	0
2134	NA	NA	1	0	0.35	0.6	0.48	173

Table 25 Value of covariates based on normalized gene expression data normalized with method 1. P-values are found from a t-test comparing average values for the two strata.

	Normalized gene expression data, method 1, covariate values							
Dataset	$\bar{X}_{YY\setminus YN,C,p}$				$\bar{X}_{YN\setminus YY,C,p}$			
Curve group	YScrYSpr	YScrNSpr	p-value	nofGenes	YScrYSpr	YScrNSpr	p-value	nofGenes
1234	0.2	-0.036	0.0052	2	0.23	0.34	0.32	678
1243	-0.098	-0.072	0.77	4	-0.0095	0.024	0.49	9
1423	-0.12	-0.054	0.63	3	0.052	0.029	0.86	1
4123	0.17	0.084	0.33	17	-0.1	-0.055	0.71	3
4132	-0.14	-0.77	0.25	1	-0.15	-0.18	0.78	5
1432	-0.19	-0.12	0.25	26	-0.061	-0.058	0.97	2
1342	-0.19	-0.17	0.84	95	-0.083	-0.038	0.48	7
1324	0.17	0.058	0.45	3	0.22	0.29	0.45	193
3124	0.43	0.24	0.087	128	0.17	0.21	0.73	38
3142	-0.087	-0.098	0.86	3	-0.12	0.0074	0.27	2
3412	-0.11	-0.04	0.14	7	-0.043	-0.19	0.036	11
4312	-0.16	-0.09	0.38	9	-0.13	-0.27	0.21	15
4321	-0.21	-0.13	0.14	5	-0.24	-0.3	0.52	654
3421	-0.33	-0.22	0.066	34	-0.21	-0.23	0.73	175
3241	-0.33	-0.2	0.093	27	-0.25	-0.22	0.71	52
3214	0.45	0.26	0.19	360	0.18	0.092	0.22	14
2314	0.37	0.16	0.072	24	0.19	0.13	0.39	42
2341	-0.34	-0.22	0.17	163	-0.27	-0.21	0.53	33
2431	-0.35	-0.2	0.033	129	-0.21	-0.19	0.73	42
4231	-0.32	-0.16	0.03	4	-0.26	-0.28	0.75	217
4213	0.075	0.052	0.8	38	0.061	-0.075	0.42	1
2413	-0.0027	-0.033	0.76	3	0.076	-0.079	0.04	17
2143	NA	NA	1	0	-0.018	0.017	0.67	8
2134	0.37	0.2	0.038	12	0.17	0.28	0.34	305

Table 26 Value of covariates based on normalized gene expression data normalized with method 3. P-values are found from a t-test comparing average values for the two strata.

	Normalized gene expression data, method 3, covariate values							
Dataset	$\bar{X}_{YY\setminus YN,C,p}$				$\bar{X}_{YN\setminus YY,C,p}$			
Curve group	YScrYSpr	YScrNSpr	p-value	nofGenes	YScrYSpr	YScrNSpr	p-value	nofGenes
1234	0.41	0.17	0.013	3	0.33	0.53	0.24	814
1243	0.07	-0.0096	0.28	4	0.032	0.08	0.5	8
1423	0.032	-0.0052	0.61	3	0.14	0.16	0.91	4
4123	0.19	0.12	0.36	17	0.082	0.15	0.86	1
4132	-0.057	-0.44	0.16	2	-0.033	-0.051	0.89	2
1432	-0.13	-0.069	0.38	22	-0.0072	0.11	0.15	2
1342	-0.19	-0.17	0.92	63	-0.037	0.14	0.07	3
1324	0.37	0.19	0.049	5	0.31	0.44	0.37	266
3124	0.53	0.3	0.16	175	0.23	0.32	0.47	31
3142	-0.13	-0.083	0.55	2	-0.063	0.0093	0.24	8
3412	-0.11	-0.049	0.23	6	-0.031	-0.13	0.21	20
4312	-0.071	-0.021	0.51	9	-0.044	-0.14	0.14	19
4321	-0.35	-0.13	0.0012	9	-0.24	-0.34	0.37	709
3421	-0.33	-0.14	0.027	53	-0.19	-0.25	0.52	203
3241	-0.34	-0.18	0.11	63	-0.21	-0.2	0.87	95
3214	0.6	0.35	0.18	515	0.18	0.16	0.74	23
2314	0.5	0.27	0.15	53	0.24	0.2	0.69	65
2341	-0.37	-0.21	0.14	279	-0.21	-0.19	0.89	49
2431	-0.34	-0.16	0.061	191	-0.21	-0.23	0.86	37
4231	-0.33	-0.11	0.0051	11	-0.25	-0.32	0.47	313
4213	0.12	0.14	0.85	26	0.024	-0.088	0.11	5
2413	0.048	0.0062	0.58	4	0.077	-0.066	0.11	19
2143	0.027	-0.026	0.77	1	0.02	0.12	0.23	14
2134	0.39	0.24	0.21	15	0.23	0.4	0.28	248

We now continue with the two most significant curve groups from normalization method 3, i.e. curve groups 4321 and 4231 with only 9 and 11 genes. In the test described in Section 3.6, we selected curve groups 1234 with 13 genes and 4321 with 31 genes. We do not know whether these genes are classified in another curve group or if they are not present in this dataset. In this dataset there are only 3 genes in the curve group 1234 making it unsuitable for further study. Results are shown in Figure 13 and Table 27.

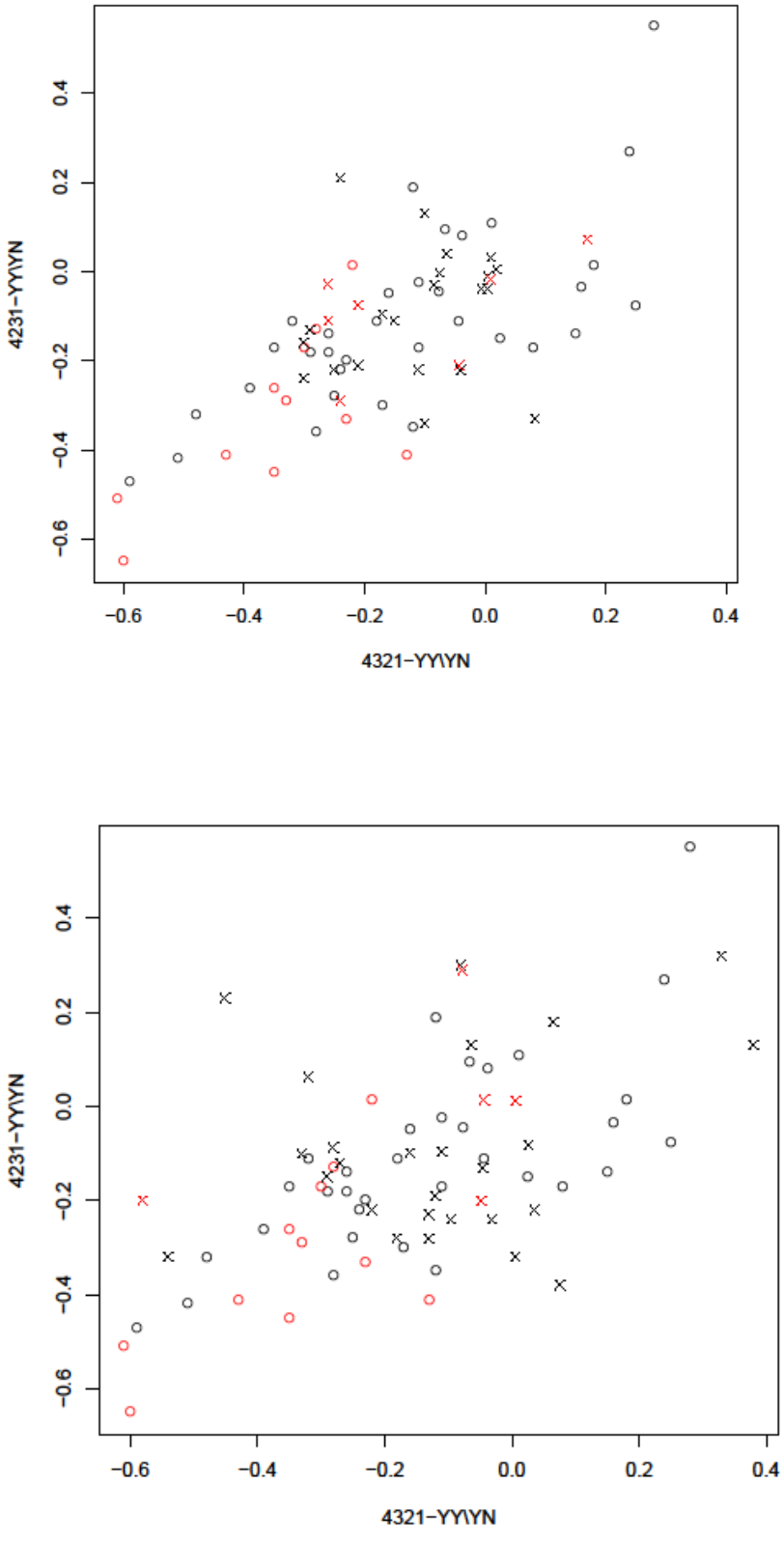

Figure 13 Values for the two significant covariates with normalization method 3. Circles indicate the dataset that is used to identify the covariates and x is the test dataset Hcc2 in the upper figure and test dataset Hcc3 in the lower figure. Observations with spread shown in red.

Table 27 Prediction of spread. Classification based on similar values in all points in a circle with the specified radius. The predicted values are divided into the categories: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) and No predictions. No predictions are for observations with no other observations within the circle or there are both observations with and without spread within the circle.

Hcc2 (P 7, N 21)						Hcc3 (P 5, N 26)					
Radius	TP	TN	FP	FN	No pred.	Radius	TP	TN	FP	FN	No pred.
0.10	1	9	2	2	14	0.10	0	10	2	2	17
0.15	0	6	0	3	19	0.15	0	13	2	2	14
0.20	0	4	0	3	21	0.20	0	10	0	2	19
0.25	0	4	0	2	22	0.25	0	11	0	2	18
0.30	0	6	0	1	21	0.30	0	7	0	2	22
0.35	0	4	0	0	24	0.35	0	5	0	1	25
0.40	0	3	0	1	24	0.40	0	5	0	1	25
0.45	0	3	0	1	24	0.45	0	3	0	1	27
0.50	0	2	0	0	26	0.50	0	3	0	1	27

The genes in the most significant covariates with normalization method 3 from Table 26 is shown in Table 28. This is similar to Table 12.

Table 28 These are the genes for the two most significant covariates in Table 26. The covariates are shown in Figure 13 and used in Table 27. There are 9 genes in the left column and 11 genes in the right column. These genes are select from genes in Table 23, where YY\YN,4321 has 9 genes and YY\YN,4231 has 11 genes.

$\bar{X}_{YY\backslash YN,4321,p}$	$\bar{X}_{YY\backslash YN,4231,p}$
gene_630	gene_1173
gene_1004	gene_1376
gene_1985	gene_2148
gene_2080	gene_2692
gene_3509	gene_3109
gene_3573	gene_3276
gene_5377	gene_3604
gene_6048	gene_3695
gene_6807	gene_4549
	gene_6445
	gene_6800

Single predictions are shown in Table 29 and Table 30. We have used the shortest distance in order to get many predictions, but then we also get many FN. It is not possible to avoid the FN by increasing the radius without also removing all the TN predictions.

Table 29 These are the PatientID, covariates, spread/no spread and classification of the 28 case-control pairs with screening and year 0 in dataset hcc2 with radius 0.1.

No	PasientID	Curve group 1234	Curve group 4321	Spread (1) or not (0)	Classification result
1	hcc2_15	-0.21	-0.076	1	-
2	hcc2_22	-0.043	-0.21	1	FN
3	hcc2_3	-0.26	-0.028	1	TP
4	hcc2_31	0.01	-0.018	1	FN
5	hcc2_4	-0.24	-0.29	1	-
6	hcc2_42	-0.26	-0.11	1	-
7	hcc2_6	0.17	0.072	1	FN
8	hcc2_10	-0.3	-0.16	0	-
9	hcc2_11	0.02	0.0059	0	TN
10	hcc2_12	-0.04	-0.22	0	TN
11	hcc2_13	0.0043	-0.038	0	TN
12	hcc2_14	0.011	0.032	0	TN
13	hcc2_16	-0.24	0.21	0	-
14	hcc2_17	0.083	-0.33	0	-
15	hcc2_19	-0.21	-0.21	0	-
16	hcc2_24	-0.29	-0.13	0	-
17	hcc2_26	-0.25	-0.22	0	-
18	hcc2_27	-0.1	-0.34	0	-
19	hcc2_29	-0.17	-0.095	0	TN
20	hcc2_30	-0.084	-0.031	0	TN
21	hcc2_32	-0.0055	-0.038	0	TN
22	hcc2_33	-0.3	-0.24	0	-
23	hcc2_34	0.0045	-0.011	0	TN
24	hcc2_35	-0.1	0.13	0	TN
25	hcc2_40	-0.075	-0.0027	0	TN
26	hcc2_7	-0.063	0.041	0	TN
27	hcc2_8	-0.11	-0.22	0	TN
28	hcc2_9	-0.15	-0.11	0	TN

Table 30 These are the PatientID, covariates, spread/no spread and classification of the 28 case-control pairs with screening and year 0 in dataset hcc3 with distance 0.1.

No	PasientID	Curve group 1234	Curve group 4321	Spread (1) or not (0)	Classification result
1	hcc3_2	-0.58	-0.2	1	-
2	hcc3_27	-0.043	0.014	1	FN
3	hcc3_5	-0.078	0.29	1	-
4	hcc3_50	-0.048	-0.2	1	FN
5	hcc3_9	0.006	0.013	1	FN
6	hcc3_1	-0.13	-0.28	0	TN
7	hcc3_12	-0.16	-0.099	0	TN
8	hcc3_15	0.38	0.13	0	-
9	hcc3_19	0.036	-0.22	0	TN
10	hcc3_20	-0.095	-0.24	0	TN
11	hcc3_25	0.33	0.32	0	-
12	hcc3_28	-0.11	-0.096	0	TN
13	hcc3_29	-0.32	0.063	0	-
14	hcc3_3	0.066	0.18	0	TN
15	hcc3_30	-0.27	-0.12	0	-
16	hcc3_31	0.026	-0.082	0	TN
17	hcc3_32	-0.063	0.13	0	TN
18	hcc3_33	-0.031	-0.24	0	-
19	hcc3_34	0.0057	-0.32	0	-
20	hcc3_35	-0.28	-0.087	0	-
21	hcc3_39	-0.13	-0.23	0	TN
22	hcc3_4	0.076	-0.38	0	-
23	hcc3_40	-0.08	0.3	0	-
24	hcc3_43	-0.12	-0.19	0	TN
25	hcc3_44	-0.45	0.23	0	-
26	hcc3_46	-0.33	-0.1	0	-
27	hcc3_47	-0.045	-0.13	0	TN
28	hcc3_49	-0.29	-0.15	0	-
29	hcc3_52	-0.54	-0.32	0	TN
30	hcc3_7	-0.18	-0.28	0	-
31	hcc3_8	-0.22	-0.22	0	-
38	hcc3_12	-0.16	-0.099	0	TN
39	hcc3_15	0.38	0.13	0	-
40	hcc3_19	0.036	-0.22	0	TN
41	hcc3_20	-0.095	-0.24	0	TN
42	hcc3_25	0.33	0.32	0	-
43	hcc3_28	-0.11	-0.096	0	TN
44	hcc3_29	-0.32	0.063	0	-

5 Updated data including test sets hcc2 and hcc3 and insitu data

5.1 Data

The dataset is the dataset in Section 4 extended with insitu data and where a less strict filtering criterion has been used (8130 genes remain after filtering). As for the dataset in Section 4, we merge year 3-5 in order to get sufficient data in this time period.

Table 31 Number of case-control pairs in each stratum and year[2].

Year	5	4	3	2	1	0
Stratum	Invasive pros + Insitu pros					Invasive hcc1 + hcc2 + hcc3
YScrYSpr	0 + 0	3 + 0	4 + 0	11 + 0	7 + 0	11 + 7 + 5
NScrYSpr	1 + 0	4 + 0	5 + 0	16 + 0	16 + 0	8 + 4 + 4
YScrNSpr	0 + 1	12 + 4	23 + 10	39 + 9	44 + 14	33 + 21 + 26
NScrNSpr	2 + 0	4 + 1	10 + 2	25 + 2	23 + 6	12 + 10 + 18

The training dataset consists of the Invasive pros and Invasive hcc1 data, while there are three different test sets:

- Insitu pros,
- Invasive hcc2 and
- Invasive hcc3.

5.2 Results for four time periods

5.2.1 Identifying curve groups from the training dataset

Curve groups are identified from the training dataset. See Section 5.5.1 for detailed results. The results are similar to those presented in Table 20-Table 23 in Section 4.2.

5.2.2 Prediction of spread using training and test datasets

For each time period the two most significant of the 48 covariates $\bar{X}_{YY\backslash YN,C,p}$ and $\bar{X}_{YN\backslash YY,C,p}$ where identified from the training dataset. These two covariates were used when predicting the diagnosis (spread or not spread) of the cases in each of the three test sets (only for cases diagnosed at screening). No sufficiently good prediction results were obtained. This is illustrated in Figure 14 for the test set Insitu pros. We observe that the insitu cases are not well separated from the cases with spread in the training dataset. The test sets Invasive hcc2 and Invasive hcc3 contain cases both with and without spread. Figure 15 shows that it is not possible to predict the diagnosis (spread or not spread) of the cases using the two selected covariates.

[2] In addition the dataset consisted of Insitu hcc1 (2 pairs), Insitu hcc2 (6 pairs) and Insitu hcc3 (3 pairs) data. These data were included when the entire dataset was normalized, but they will not be included in any other parts of the data analyses.

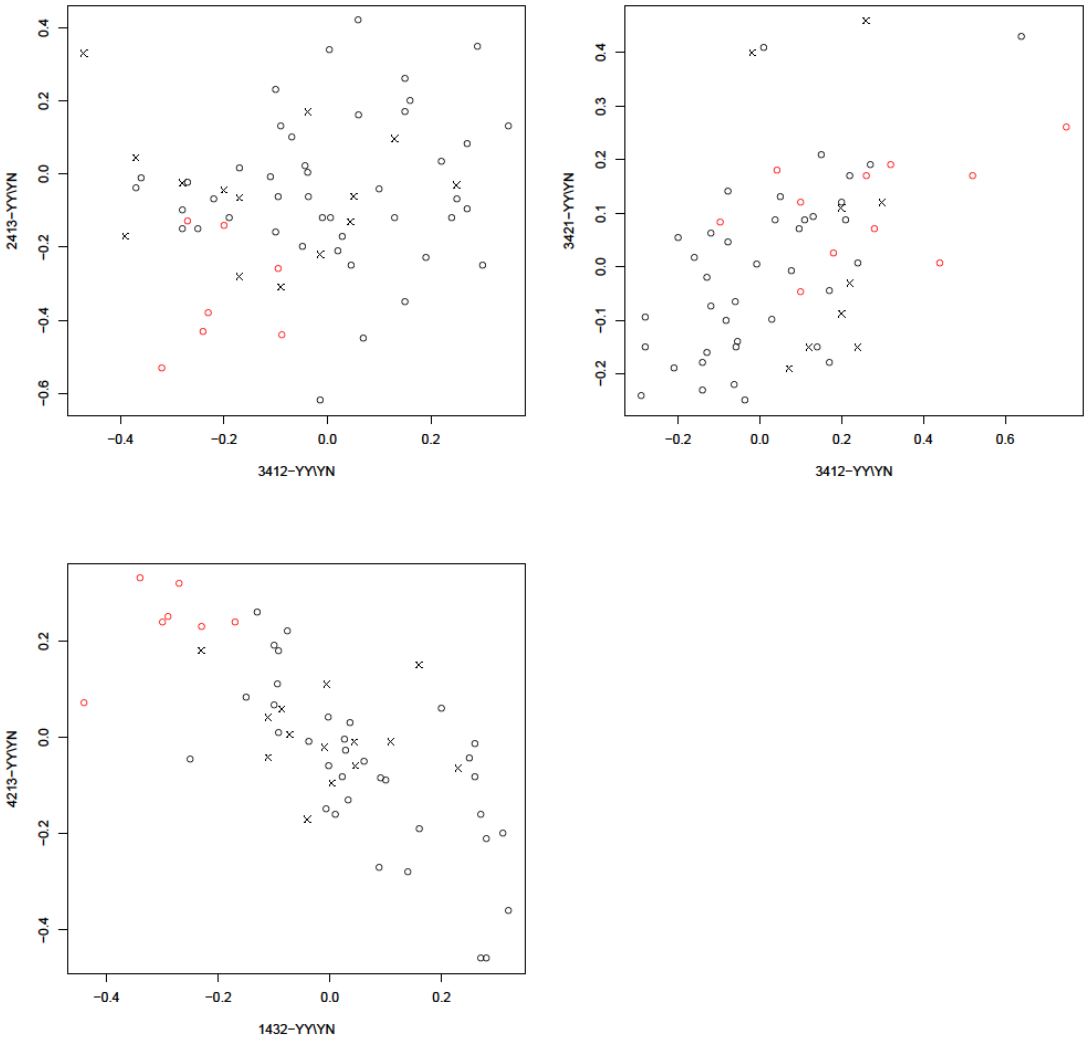

Figure 14 Values for the two significant covariates with normalization method 3. Circles indicate the dataset that is used to select the covariates and x is the test dataset Insitu pros in period 1 (upper left panel), period 2 (upper right panel) and period 3 (lower left panel). Observations with spread shown in red.

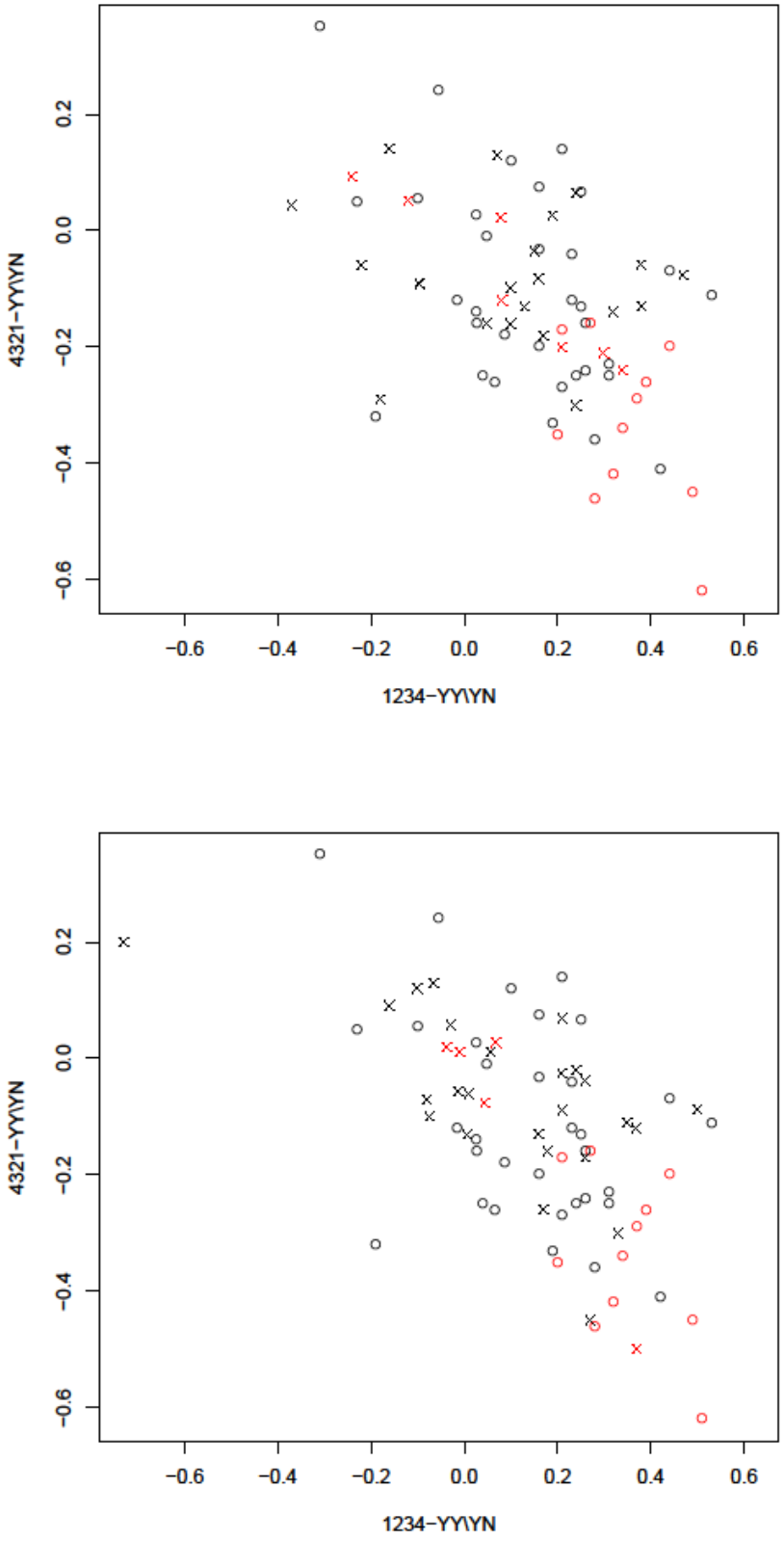

Figure 15 Values for the two significant covariates with normalization method 3. Circles indicate the dataset that is used to select the covariate and x is the test dataset Invasive Hcc2 (upper panel) and test dataset Invasive Hcc3 (lower panel). Observations with spread shown in red.

5.2.3 Prediction of spread based on leave-one-out

The test sets have either no data before diagnosis or they contain only case-control pairs without spread. The training set however contains both data before diagnosis and case-control pairs with and without spread. It is therefore interesting to examine whether diagnosis can be predicted for data in the training set. Prediction results were not sufficiently good (results not shown).

5.3 Results for three time periods

A hypothesis is that some of the observed gene expression differences between cases and controls at time of diagnosis can be caused by stress reactions following the breast cancer diagnosis. If this is the case, it is interesting to perform the analyses described in Section 5.2, but where data at the time of diagnosis are excluded. This means that we get three instead of four time periods. Then it is also possible to compare with the analysis performed in Section 2.5 where we had about 10% more genes and the same or almost the same case-control pairs.

The following two sections show that we get less significant results using the dataset analyzed in Section 5 than in Section 2. As far as we know the case-control pairs are almost the same. Hence the difference is probably due to the selection of genes. We do not known how many of the genes that are the same in the two datasets.

5.3.1 Identifying curve groups from the training dataset

Curve groups are identified from the Invasive pros dataset (part of the training dataset). This is similar to Section 2.4.1. The detailed results shown in Section 5.5.2 shows slightly less significant results in this dataset compared to the dataset analyzed in Section 2 by comparing Table 38 with Table 4.

5.3.2 Prediction of spread

Here we do the same analyses as in Section 5.2.2 and 5.2.3, but based on three instead of four time periods (data from period 0, i.e. time of diagnosis, are excluded). Results are shown in Figure 16 (similar to Figure 11) and Table 32 (similar to Table 5). The figure shows that we are slightly less able to differentiate between the two strata (spread/not spread) with the dataset used in this section. But when we perform a leave-one-out analysis and perform a prediction on each case-control pair shown in Table 32, the difference between the two datasets is large. In the dataset in this section (Section 5) we are not able to make prediction of practical value.

Table 32 Classification based on similar values in all points in a circle with the specified radius. The predicted values are divided into the categories: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) and No predictions. No predictions are for observations with no other observations within the circle or there are both observations with and without spread within the circle.

First time period (P 7, N 44)							
Most significant covariates		Radius	TP	TN	FP	FN	No pred.
Covariate	Frequency	0.10	0	33	0	6	12
231-YY\YN	51	0.15	0	38	0	6	7
213-YY\YN	49	0.20	0	36	0	7	8
123-YY\YN	2	0.25	0	34	0	5	12
		0.30	0	31	0	4	16
		0.35	0	29	0	4	18
		0.40	0	28	0	4	19
		0.45	0	26	0	3	22
		0.50	0	24	0	3	24
Second time period (P 11, N 39)							
Covariate	Frequency	Radius	TP	TN	FP	FN	No pred.
231-YY\YN	49	0.10	0	23	3	7	17
132-YY\YN	34	0.15	0	18	2	8	22
213-YY\YN	2	0.20	0	11	1	7	31
132-YN\YY	15	0.25	0	8	1	7	34
		0.30	0	6	0	3	41
		0.35	0	2	0	2	46
		0.40	0	0	0	2	48
		0.45	0	0	0	1	49
		0.50	0	0	0	1	49
Third time period (P 7, N 35)							
Covariate	Frequency	Radius	TP	TN	FP	FN	No pred.
132-YY\YN	41	0.10	0	28	1	5	8
321-YY\YN	3	0.15	0	26	0	5	10
123-YN\YY	15	0.20	0	23	0	4	15
312-YY\YN	25	0.25	0	21	0	4	17
		0.30	0	19	0	2	21
		0.35	0	16	0	2	24
		0.40	0	14	0	2	26
		0.45	0	10	0	2	30
		0.50	0	5	0	2	35

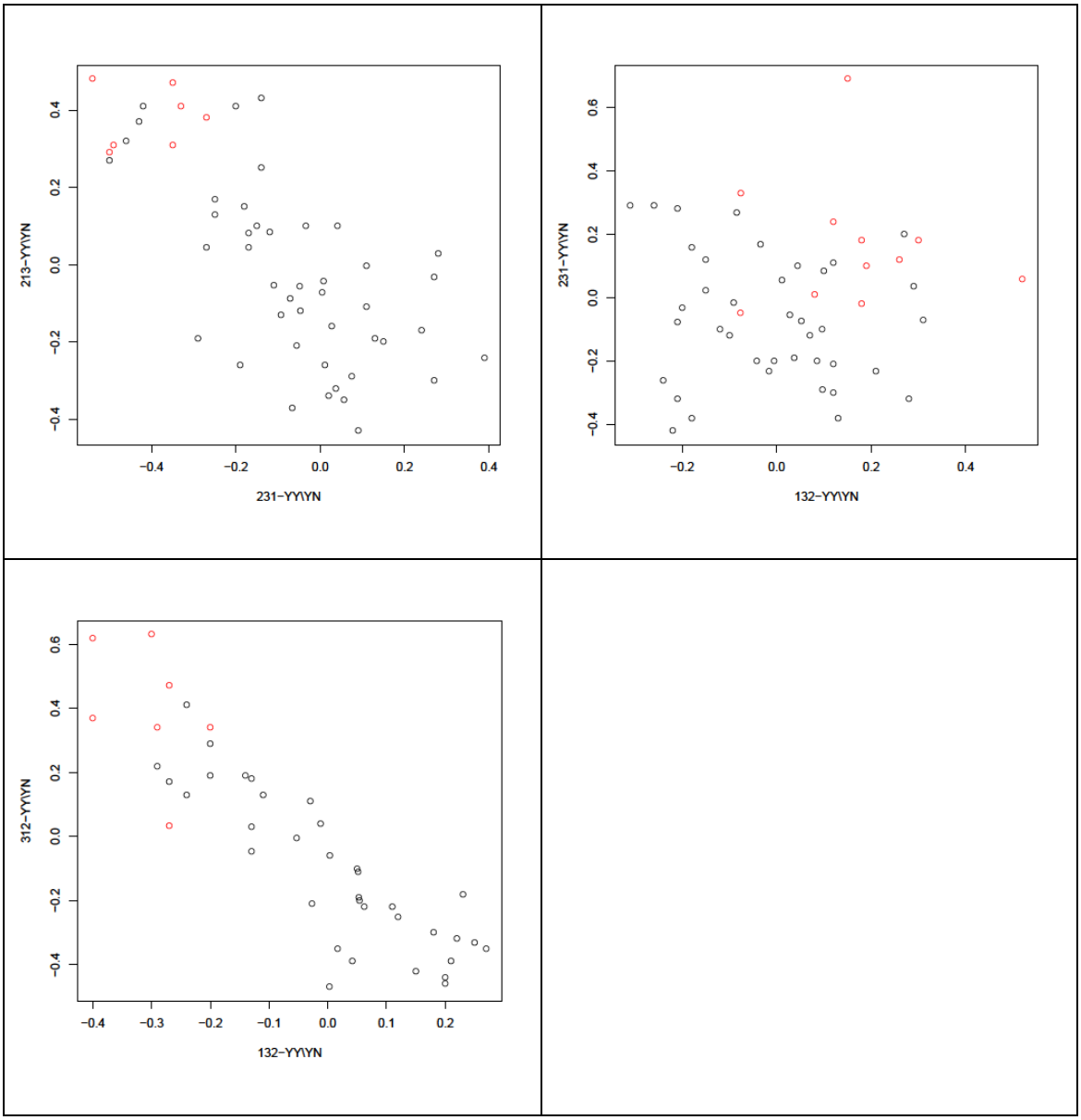

Figure 16 Case control pairs with spread (red) and not spread (black) for the most important covariates for time period 1 (upper left), 2 (upper right) and 3 (lower left), respectively.

5.4 Further work

There are many options for further work. We should analyze why this dataset gives less significant results than the dataset analyzed in Section 2. This is probably due to the selection of genes and hence this selection should be studied more carefully.

It is possible to compare data at time of diagnosis with data before diagnosis to identify potential stress related genes.

5.5 Detailed results

5.5.1 Identifying curve groups (four periods)

Table 33 Estimation of number of genes in the different curve groups for not normalized data.

	Not normalized gene expression data							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global	0.084	1758	0.37	365	0.032	3264	0.33	512
1234	0.44	1	0.064	41	0.0035	1967	1	0
1243	1	0	0.066	45	0.56	1	1	0
1423	0.33	2	1	0	1	0	1	0
4123	1	0	1	0	1	0	0.17	14
4132	1	0	1	0	1	0	0.082	43
1432	0.3	2	1	0	1	0	1	0
1342	0.22	5	0.16	11	0.56	1	1	0
1324	0.12	12	0.13	13	0.0055	797	1	0
3124	0.25	3	0.28	4	0.13	25	0.12	24
3142	1	0	1	0	1	0	0.18	11
3412	0.19	6	1	0	0.29	4	1	0
4312	0.21	7	1	0	0.2	9	0.58	1
4321	1	0	1	0	0.18	12	0.28	7
3421	1	0	1	0	0.22	8	1	0
3241	1	0	1	0	1	0	0.098	38
3214	0.0045	1261	0.046	125	0.12	28	0.17	14
2314	0.012	407	0.054	114	0.028	199	1	0
2341	1	0	0.29	3	0.57	1	1	0
2431	1	0	1	0	1	0	0.6	1
4231	1	0	1	0	1	0	0.028	347
4213	0.11	26	0.35	3	0.57	1	0.28	5
2413	0.088	24	1	0	0.55	1	1	0
2143	1	0	1	0	1	0	0.45	2
2134	0.32	2	0.21	6	0.034	210	0.3	5

Table 34 Estimation of number of genes in the different curve groups for data normalized using method 1

	Normalized gene expression data by method 1							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global	0.042	1242	0.18	877	0.0025	2753	0.12	1054
1234	0.56	8	0.99	1	5.00E-04	726	0.22	45
1243	0.64	10	0.21	33	0.62	11	0.79	6
1423	0.88	3	0.75	7	0.99	1	0.9	4
4123	0.36	15	0.48	18	0.94	3	0.92	4
4132	0.99	1	0.84	5	0.91	4	0.72	8
1432	0.23	23	0.81	6	0.95	3	0.91	4
1342	0.038	115	0.016	183	0.85	5	0.95	3
1324	0.64	7	0.53	12	0.0095	208	0.074	97
3124	0.026	124	0.39	21	0.18	43	0.18	40
3142	0.84	4	0.75	6	0.97	2	0.72	7
3412	0.68	6	0.9	3	0.45	17	0.94	3
4312	0.62	10	0.24	30	0.36	21	0.99	1
4321	0.47	11	0.61	10	5.00E-04	735	0.072	98
3421	0.094	56	0.6	11	0.012	198	0.15	42
3241	0.18	33	0.074	60	0.16	48	0.024	137
3214	5.00E-04	418	0.034	101	0.38	22	0.19	41
2314	0.16	37	0.16	39	0.2	39	0.092	63
2341	0.016	144	0.0045	217	0.2	39	0.026	128
2431	0.034	110	0.42	19	0.17	48	0.096	59
4231	0.65	7	0.77	6	0.0075	217	0.014	220
4213	0.15	45	0.098	66	0.94	3	0.92	4
2413	0.83	4	0.58	10	0.52	14	0.97	2
2143	0.98	1	0.68	7	0.94	3	0.79	6
2134	0.1	50	0.79	6	0.0045	343	0.22	32

Table 35 Estimation of number of genes in the different curve groups for data normalized using method 2.

	Normalized gene expression data by method 2							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global	0.0045	1743	0.6	483	0.0015	3304	0.18	873
1234	0.73	6	0.74	9	5.00E-04	900	0.25	39
1243	0.87	6	0.77	8	0.7	10	0.74	9
1423	0.93	3	0.74	9	0.85	7	0.59	13
4123	0.31	17	0.55	16	0.99	2	0.85	7
4132	0.78	6	0.67	11	0.97	3	0.96	4
1432	0.22	24	0.45	19	0.96	4	0.36	22
1342	0.07	70	0.2	38	0.95	4	0.67	11
1324	0.5	11	0.59	13	0.0045	297	0.17	46
3124	0.0085	216	0.53	16	0.17	43	0.098	55
3142	0.98	2	0.58	11	0.82	7	0.7	9
3412	0.89	4	1	1	0.32	24	0.87	6
4312	0.65	11	0.93	5	0.37	21	0.89	6
4321	0.31	18	0.47	17	5.00E-04	760	0.084	76
3421	0.08	61	0.86	7	0.01	223	0.052	77
3241	0.074	66	0.14	41	0.042	113	0.054	74
3214	5.00E-04	562	0.094	55	0.23	34	0.18	39
2314	0.078	59	0.21	33	0.078	83	0.17	39
2341	5.00E-04	313	0.076	65	0.12	55	0.046	80
2431	0.0075	206	0.24	34	0.14	48	0.05	74
4231	0.35	17	0.2	35	0.0025	334	0.036	114
4213	0.21	32	0.53	16	0.82	7	0.83	8
2413	0.85	5	0.9	5	0.28	27	0.5	14
2143	0.99	1	0.9	5	0.49	16	0.83	7
2134	0.23	27	0.58	14	0.01	282	0.13	44

Table 36 Estimation of number of genes in the different curve groups for data normalized using method 3

	Normalized gene expression data by method 3							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global	0.0045	1736	0.59	486	0.0015	3304	0.18	871
1234	0.73	6	0.74	9	5.00E-04	898	0.24	39
1243	0.87	6	0.78	8	0.71	10	0.74	9
1423	0.93	3	0.71	10	0.85	7	0.59	13
4123	0.31	17	0.53	17	1	1	0.81	8
4132	0.77	6	0.66	11	0.97	3	0.93	5
1432	0.22	24	0.48	18	0.96	4	0.36	22
1342	0.07	70	0.2	39	0.92	5	0.67	11
1324	0.49	11	0.59	13	0.0045	298	0.18	46
3124	0.0085	214	0.52	16	0.18	41	0.098	55
3142	0.98	2	0.58	11	0.83	7	0.71	9
3412	0.89	4	1	1	0.31	24	0.91	5
4312	0.66	11	0.93	5	0.39	20	0.89	6
4321	0.27	20	0.47	17	5.00E-04	765	0.084	76
3421	0.08	60	0.85	7	0.01	219	0.054	77
3241	0.076	65	0.14	41	0.044	112	0.054	73
3214	5.00E-04	560	0.096	55	0.23	34	0.2	37
2314	0.078	60	0.21	33	0.082	80	0.17	39
2341	5.00E-04	311	0.076	66	0.12	56	0.05	79
2431	0.0075	204	0.23	35	0.15	48	0.052	74
4231	0.37	16	0.2	35	0.0025	336	0.036	113
4213	0.21	32	0.54	16	0.81	7	0.83	8
2413	0.85	5	0.89	5	0.27	27	0.48	15
2143	1	1	0.9	5	0.49	16	0.79	8
2134	0.22	28	0.62	13	0.0095	286	0.13	44

5.5.2 Identifying curve groups (three periods)

These tables correspond to Table 3 and Table 4 (normalization method 1) for the dataset analyzed in Section 2.4.1. Note that the results are slightly less significant.

Table 37 P-values obtained when testing whether there is a significant development in time or not. The gene expression data are not normalized. P-values below 0.01 are highlighted in yellow, while p-values between 0.05 and 0.01 are highlighted in blue. A p-value close to 1 means that there are fewer genes than expected in the curve group. The test for each curve group is based on $\boldsymbol{M_{a,i}}$ and the global test is based on $\boldsymbol{M_a}$.

	Not normalized gene expression data							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global	0.21	4608	0.59	1948	0.76	1683	0.36	3332
123	0.54	47	0.17	571	0.17	948	0.96	5
132	0.31	191	0.46	98	0.32	252	0.99	1
312	0.87	5	0.85	11	0.81	19	0.066	2207
321	0.058	2597	0.16	963	0.49	131	0.50	125
231	0.090	1764	0.29	283	0.35	260	1.00	1
213	0.90	4	0.69	22	0.58	73	0.13	993

Table 38 P-values obtained when testing whether there is a significant development in time or not. The gene expression data are normalized. P-values below 0.01 are highlighted in yellow, while p-values between 0.05 and 0.01 are highlighted in blue. A p-value close to 1 means that there are fewer genes than expected in the curve group. The test for each curve group is based on $\boldsymbol{M_{a,i}}$ and the global test is based on $\boldsymbol{M_a}$.

	Normalized gene expression data by method 1							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global	0.31	3148	0.23	3609	0.56	3001	0.96	1953
123	0.060	946	0.018	1262	0.20	743	0.91	252
132	0.49	398	0.89	240	0.88	272	0.62	408
312	0.70	281	0.66	362	0.92	244	0.62	406
321	0.036	1081	0.07	994	0.15	820	0.91	259
231	0.83	231	0.43	449	0.52	441	0.71	311
213	0.86	211	0.71	302	0.47	481	0.70	317

Table 39 P-values obtained when testing whether there is a significant development in time or not. The gene expression data are normalized. P-values below 0.01 are highlighted in yellow, while p-values between 0.05 and 0.01 are highlighted in blue. A p-value close to 1 means that there are fewer genes than expected in the curve group. The test for each curve group is based on $\boldsymbol{M_{a,i}}$ and the global test is based on $\boldsymbol{M_a}$.

	Normalized gene expression data by method 2							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global	0.20	3390	0.87	2215	0.38	3286	0.88	2302
123	0.05	951	0.59	402	0.17	756	0.67	396
132	0.41	421	0.88	277	0.80	309	0.48	482
312	0.38	434	0.66	389	0.81	312	0.61	413
321	0.038	1066	0.66	372	0.12	891	0.87	306
231	0.72	285	0.62	359	0.33	547	0.62	353
213	0.84	233	0.48	416	0.44	471	0.61	352

Table 40 P-values obtained when testing whether there is a significant development in time or not. The gene expression data are normalized. P-values below 0.01 are highlighted in yellow, while p-values between 0.05 and 0.01 are highlighted in blue. A p-value close to 1 means that there are fewer genes than expected in the curve group. The test for each curve group is based on $\boldsymbol{M_{a,i}}$ and the global test is based on $\boldsymbol{M_a}$.

	Normalized gene expression data by method 3							
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr	
# Simulations	1000		1000		1000		1000	
Curve group	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$	p-value	$M_{a,i}$
Global	0.20	3394	0.87	2221	0.38	3297	0.88	2305
123	0.050	959	0.58	407	0.16	761	0.66	399
132	0.42	417	0.89	273	0.80	311	0.47	484
312	0.38	432	0.66	389	0.80	315	0.61	414
321	0.038	1064	0.65	373	0.12	892	0.87	308
231	0.71	287	0.61	363	0.33	544	0.63	351
213	0.84	235	0.48	416	0.44	474	0.62	349

6 Data for eight years before diagnosis

We have extended the analysis in Section 2.4 and 2.5 with more data, now extending back to 8 years, but with slightly fewer genes, 8952 instead of 9060, see Table 1. As shown in Table 41 we divide the data into 4 periods by merging years 3-4 and 5-8. This gives 24 curve groups that is a feasible number. It also seems reasonable with longer time periods many years before diagnosis. We have used the same methods as in Section 2.4, with a=0.1 and with a=0.01.

In Section 6.1 we analyze the number of significant genes in each curve group. There are not more significant numbers than we expect from the null model. We have also performed exactly the same test as in Section 2.4, i.e. we neglect data earlier than year 4, merge year 3 and 4, and use a=0.1 and a=0.01. The results in Section 2.4 were slightly significant, but when we repeat the analysis with more data, and with a slightly different list of genes, there are no significant results.

In Section 6.2 we test a method for selecting covariates that are different for the two strata YScrYSpr and YScrNSpr. Also here there are not more significant numbers than we expect from the null model.

Table 41 Number of case-control pairs in each stratum and year.

Year	8	7	6	5	4	3	2	1		Sum
Period	4				3		2	1		all
Stratum	Invasive (Insitu)									years
YScrYSpr	0	1	5	9	17	15	11	7		65
NScrYSpr	0	0	2	10	18	11	17	15		73
YScrNSpr	0 (0)	3 (0)	5 (0)	26 (5)	48 (16)	45 (14)	43 (9)	43 (14)		213 (58)
NScrNSpr	1 (0)	1 (0)	8 (1)	16 (3)	19 (4)	22 (4)	26 (2)	23 (7)		116 (21)
Sum strata	1 (0)	5 (0)	20 (1)	61 (8)	102 (20)	93 (18)	97 (11)	88 (21)		467 (79)

Note that the data are produced in three different runs (run1, run2 and run3). We assume that the effect of including a case-control pair in run 2 (or run 3) instead of run 1 is that all intensities are multiplied by the same constant. This constant disappears when we use data that are $\log_2$-differences between case and control data, i.e. for the $\log_2$-difference data there are no systematic differences between the three runs. The assumption about the effect of the run also means that it is possible to use all three normalization methods, even if the data are obtained in different runs.

Note that the p-values have not been adjusted for multiple testing.

6.1 Identifying curve groups

The number of significant curve groups are given in Table 42 and Table 43 below. There are not more significant values than we expect from the null model.

Table 42 Number of significant curve groups with four time periods. (Detailed results are found in M:\bioinf\Prj\EilivLund\Datasett5\Res4periods*\resNorm*TestForEachStratum.txt – files not included.)

# Simulations=1000	Number of significant curve groups 4 time periods, 24 curve groups									
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr		Sum	
p-value	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05
	a=0.1									
Not normalized	1	0	0	0	1	2	1	1	3	3
Norm. method 1	0	2	0	0	1	1	0	0	1	3
Norm. method 2	0	0	0	0	1	1	0	0	1	1
Norm. method 3	0	0	0	0	1	1	0	0	1	1
	a=0.01									
Not normalized	1	2	0	0	3	2	0	2	4	6
Norm. method 1	0	2	0	0	0	2	0	0	0	4
Norm. method 2	0	1	0	0	0	2	0	0	0	3
Norm. method 3	0	1	0	0	0	2	0	0	0	3

Table 43 Number of significant curve groups with three time periods. (Detailed results are found in M:\bioinf\Prj\EilivLund\Datasett5\Res3periods*\resNorm*TestForEachStratum.txt – files not included.)

# Simulations=1000	Number of significant curve groups 3 time periods, 6 curve groups									
Dataset	YScrYSpr		NScrYSpr		YScrNSpr		NScrNSpr		Sum	
p-value	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05
	a=0.1									
Not normalized	0	0	0	0	0	0	1	0	1	0
Norm. method 1	0	0	0	0	0	0	0	0	0	0
Norm. method 2	0	0	0	0	0	0	0	0	0	0
Norm. method 3	0	0	0	0	0	0	0	0	0	0
	a=0.01									
Not normalized	0	0	0	0	0	0	0	1	0	1
Norm. method 1	0	0	0	0	0	0	0	0	0	0
Norm. method 2	0	0	0	0	0	0	0	0	0	0
Norm. method 3	0	0	0	0	0	0	0	0	0	0

6.2 Selecting covariates with different means for spread and not spread

We have first randomized the data between the YScrYSpr and YScrNSpr strata and in time. This randomization and the p-value computation are similar to what is described for only one stratum in Section 2.4.1 where we analyzed the number of genes in the different curve groups. We have kept the time periods and strata for each case-control pair, but randomized the

values between the pairs for the different time periods and strata. We have compared the absolute value of the differences between the two strata in the averages over the time period t and case-control pairs p of the covariates $\bar{X}_{YY\backslash YN,C,p}$ and $\bar{X}_{YN\backslash YY,C,p}$, respectively, i.e.

$$Y_{YY\backslash YN,C,t} = \left|\text{mean}_{p\in\{YY,t\}}\left(\bar{X}_{YY\backslash YN,C,p}\right) - \text{mean}_{p\in\{YN,t\}}\left(\bar{X}_{YY\backslash YN,C,p}\right)\right|$$

and correspondingly for $\bar{X}_{YN\backslash YY,C,p}$. The covariates $\bar{X}_{YY\backslash YN,C,p}$ and $\bar{X}_{YN\backslash YY,C,p}$, are described in Section 2.5.2. The covariate $\bar{X}_{YY\backslash YN,C,p}$ is the average values for the genes for case-control pair p that are significant for curve group C for YScrYSpr and not YScrNSpr. Similarly, $\bar{X}_{YN\backslash YY,C,p}$ is the average over the genes that are significant for YScrNSpr and not YScrYSpr. The variable $Y_{YY\backslash YN,C,t}$ from the data is compared with the distribution of the same variable from the simulations. Except for the absolute value the variable is symmetric around 0 in the simulated data. Based on all the simulated data, we have found a p-value for each of the two covariates for each curve group and each time period. The global test for each period compares

$$\max_C\{Y_{YY\backslash YN,C,t}\}=\max_C\{\left|\text{mean}_{p\in\{YY,t\}}\left(\bar{X}_{YY\backslash YN,C,p}\right) - \text{mean}_{p\in\{YN,t\}}\left(\bar{X}_{YY\backslash YN,C,p}\right)\right|\}$$

and

$$\max_C\{Y_{YN\backslash YY,C,t}\}= \max_C\{\left|\text{mean}_{p\in\{YY,t\}}\left(\bar{X}_{YN\backslash YY,C,p}\right) - \text{mean}_{p\in\{YN,t\}}\left(\bar{X}_{YN\backslash YY,C,p}\right)\right|\},$$

respectively, with the distribution of the corresponding expressions obtained from the simulated data. Summary of the results of the tests are given in Table 44 and Table 45. For the tests with both three and four periods, a=0.1 and a=0.01 was chosen when identifying the genes with strong functional form.

Table 44 Number of significant covariates with four time periods. . (Detailed results are found in M:\bioinf\Prj\EilivLund\Datasett5\Res4periods\TestCovariates*\resS?norm?.txt – files not included.)

# Simulations=1000	Number of significant covariates 4 time periods, 24 curve groups, 48 tests per time period									
Period	4		3		2		1		Sum	
p-value	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05
	a=0.1									
Not normalized	0	2	0	0	0	1	0	1	0	4
Norm. method 1	0	0	0	0	0	0	1	7	1	7
Norm. method 2	0	0	0	0	0	0	1	4	1	4
Norm. method 3	0	0	0	0	0	0	1	5	1	5
	a=0.01									
Not normalized	1	1	0	1	1	1	2	0	4	3
Norm. method 1	0	0	0	0	0	0	2	3	2	3
Norm. method 2	0	1	0	0	0	0	1	3	1	4
Norm. method 3	0	1	0	0	0	0	1	4	1	5

Table 45 Number of significant covariates with three time periods. . (Detailed results are found in M:\bioinf\Prj\EilivLund\Datasett5\Res3periods\TestCovariates*\resS?norm?.txt – files not included.)

# Simulations=1000	Number of significant covariates 3 time periods, 6 curve groups, 12 test per time period							
Period	3		2		1		Sum	
p-value	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05
	a=0.1							
Not normalized	0	0	0	0	0	1	0	1
Norm. method 1	0	0	0	0	0	2	0	2
Norm. method 2	0	0	0	0	0	2	0	2
Norm. method 3	0	0	0	0	0	1	0	1
	a=0.01							
Not normalized	0	0	0	0	0	1	0	1
Norm. method 1	0	0	0	0	0	0	0	0
Norm. method 2	0	0	0	0	0	2	0	2
Norm. method 3	0	0	0	0	0	2	0	2

6.3 Prediction of spread based on leave-one-out

This is not included in this version.

6.4 Strata defined from HRT

We repeat the analysis in Sections 6.1- 6.2 with a different division into strata. The four strata are now defined based on HRT, i.e. Ycase means use of HRT while Ncase means not use of HRT. The main conclusion is that no significant curve groups are identified. Except for normalization method 1 in time period 2, there is not more significant values than expected from the null model.

Table 46 Number of case-control pairs in each HRT stratum and year.

Year	5-8	3-4	2	1		Sum
Period	4	3	2	1		all
Stratum	Invasive (Insitu)					years
YcaseYctrl	7 (0)	13 (2)	5 (0)	2 (1)		27 (3)
NcaseYctrl	10 (2)	23 (5)	8 (1)	12 (2)		53 (10)
YcaseNctrl	15 (0)	44 (6)	22 (3)	23 (5)		104 (14)
NcaseNctrl	55 (7)	115 (25)	62 (7)	51 (13)		283 (52)
Sum strata	87 (9)	195 (38)	97 (11)	88 (21)		467 (79)

Except for stratum YcaseYctrl, there is a sufficient amount of data for all time periods in each stratum.

6.4.1 Identifying curve groups for HRT

No significant curve groups were identified (Table 47 and Table 48).

Table 47 Number of significant curve groups for HRT with four time periods. (Detailed results are found in M:\bioinf\Prj\EilivLund\Datasett5\HRT\Res4periods*\resNorm*TestForEachStratum.txt – files not included.)

# Simulations=1000	Number of significant curve groups with four periods for HRT 24 curve groups for each strata									
Dataset	YcaseYctrl		NcaseYctrl		YcaseNctrl		NcaseNctrl		Sum	
p-value	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05
	a=0.1									
Not normalized	0	1	1	2	0	2	0	2	1	7
Norm. method 1	1	0	0	0	1	2	0	0	2	2
Norm. method 2	1	1	0	0	0	0	0	0	1	1
Norm. method 3	1	1	0	0	0	0	0	0	1	1
	a=0.01									
Not normalized	0	0	0	5	0	2	0	0	0	7
Norm. method 1	0	0	0	0	0	2	0	0	0	2
Norm. method 2	0	0	0	0	0	0	0	0	0	0
Norm. method 3	0	0	0	0	0	0	0	0	0	0

Table 48 Number of significant curve groups for HRT with three time periods. (Detailed results are found in M:\bioinf\Prj\EilivLund\Datasett5\HRT\Res3periods*\resNorm*TestForEachStratum.txt – files not included.)

# Simulations=1000	Number of significant curve groups with three periods for HRT 6 curve groups for each strata									
Dataset	YcaseYctrl		NcaseYctrl		YcaseNctrl		NcaseNctrl		Sum	
p-value	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05
	a=0.1									
Not normalized	0	0	0	1	0	0	0	0	0	1
Norm. method 1	0	0	0	0	0	0	0	0	0	0
Norm. method 2	0	0	0	0	0	0	0	0	0	0
Norm. method 3	0	0	0	0	0	0	0	0	0	0
	a=0.01									
Not normalized	0	0	0	0	0	0	0	0	0	0
Norm. method 1	0	0	0	0	0	0	0	0	0	0
Norm. method 2	0	0	0	0	0	0	0	0	0	0
Norm. method 3	0	0	0	0	0	0	0	0	0	0

6.4.2 Selecting covariates with different means for HRT

For each covariate (see Section 6.2) and time period, we compare the means of the case-control pairs in two different strata. We will only compare two strata where either the controls

or the cases have the same value for HRT (Yes or No) in the two strata. As there is not a sufficient amount of data for all time periods for YcaseYctrl, this stratum is not included in any comparisons. We are therefore left with two comparisons: YcaseNctrl versus NcaseNctrl, and NcaseYctrl versus NcaseNctrl. We also compare YcaseNctrl versus NcaseYctrl.

With four time periods we perform 192 (48 tests x 4 periods) tests for each normalization method. Notice, however that these tests are not completely independent. If there is no signal in the data we expect a p-value below 0.05 in 10 tests. For YcaseNctrl and NcaseNctrl we observe that there are 26 (12 with a=0.01) tests with p-values below 0.05 for normalization method 1 and time period 2 (Table 49). Otherwise, there are not more significant values than expected under the null model.

With three time periods we perform 36 (12 tests x 3 periods) tests. If there is no signal in the data we expect a p-value below 0.05 in 2 tests. Here, there are not more significant values than expected under the null model.

Table 49 Number of significant covariates for HRT with four time periods. (Detailed results are found in M:\bioinf\Prj\EilivLund\Datasett5\HRT\Res4periods\TestCovariates*\resS?norm?.txt – files not included.)

# Simulations=1000	Number of significant covariates with four periods for HRT 24 curve groups and 48 test per time period									
Period	4		3		2		1		Sum	
p-value	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05
	YcaseNctrl and NcaseNctrl – a=0.1									
Not normalized	0	2	1	0	0	1	0	2	1	5
Norm. method 1	0	3	0	0	9	17	1	5	10	25
Norm. method 2	0	0	0	0	0	2	0	0	0	2
Norm. method 3	0	0	0	0	0	2	0	0	0	2
	NcaseYctrl and NcaseNctrl – a=0.1									
Not normalized	1	2	0	1	1	2	1	1	3	6
Norm. method 1	0	1	0	0	1	3	0	0	1	4
Norm. method 2	0	0	0	0	0	0	0	0	0	0
Norm. method 3	0	0	0	0	0	0	0	0	0	0
	YcaseNctrl and NcaseYctrl – a=0.1									
Not normalized	1	1	1	1	2	0	0	2	4	4
Norm. method 1	0	0	0	1	0	1	0	0	0	2
Norm. method 2	0	0	0	1	0	2	0	3	0	6
Norm. method 3	0	0	0	1	0	2	0	2	0	5
	YcaseNctrl and NcaseNctrl – a=0.01									
Not normalized	0	1	0	2	0	2	0	1	0	6
Norm. method 1	0	1	0	0	3	9	0	1	3	11
Norm. method 2	0	2	0	3	0	3	0	0	0	8
Norm. method 3	0	1	0	3	0	3	0	0	0	7
	NcaseYctrl and NcaseNctrl – a=0.01									

Not normalized	0	1	0	2	0	1	0	1	0	5
Norm. method 1	0	1	0	1	1	4	0	0	1	6
Norm. method 2	0	2	0	1	0	0	1	0	1	3
Norm. method 3	0	2	0	1	0	0	1	0	1	3
	YcaseNctrl and NcaseYctrl – a=0.01									
Not normalized	0	0	0	0	0	0	0	0	0	0
Norm. method 1	1	0	0	2	1	0	0	1	2	3
Norm. method 2	0	1	0	2	0	1	0	1	0	5
Norm. method 3	0	1	0	2	0	1	0	1	0	5

Table 50 Number of significant covariates for HRT with three time periods. (Detailed results are found in M:\bioinf\Prj\EilivLund\Datasett5\HRT\Res3periods\TestCovariates*\resS?norm?.txt – files not included.)

# Simulations=1000	Number of significant covariates with three periods for HRT 6 curve groups, 12 tests per time period							
Period	3		2		1		Sum	
p-value	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05
	YcaseNctrl and NcaseNctrl – a=0.1							
Not normalized	0	0	0	1	0	1	0	2
Norm. method 1	0	0	0	0	1	3	1	3
Norm. method 2	0	0	0	0	0	0	0	0
Norm. method 3	0	0	0	0	0	0	0	0
	NcaseYctrl and NcaseNctrl – a=0.1							
Not normalized	0	0	0	0	0	0	0	0
Norm. method 1	0	0	0	0	0	0	0	0
Norm. method 2	0	0	0	0	0	0	0	0
Norm. method 3	0	0	0	0	0	0	0	0
	YcaseNctrl and NcaseYctrl – a=0.1							
Not normalized	0	0	0	0	0	0	0	0
Norm. method 1	0	0	0	0	0	0	0	0
Norm. method 2	0	0	0	0	0	0	0	0
Norm. method 3	0	0	0	0	0	0	0	0
	YcaseNctrl and NcaseNctrl – a=0.01							
Not normalized	0	0	0	0	0	0	0	0
Norm. method 1	0	0	0	0	0	1	0	1
Norm. method 2	0	1	1	0	0	0	1	1
Norm. method 3	0	1	1	0	0	0	1	1
	NcaseYctrl and NcaseNctrl – a=0.01							
Not normalized	0	0	0	0	0	0	0	0
Norm. method 1	0	0	0	0	0	0	0	0
Norm. method 2	0	0	0	0	0	0	0	0
Norm. method 3	0	0	0	0	0	0	0	0

	YcaseNctrl and NcaseYctrl – a=0.01							
Not normalized	0	1	0	1	0	0	0	2
Norm. method 1	0	0	0	0	0	0	0	0
Norm. method 2	0	1	1	0	0	0	1	1
Norm. method 3	0	1	0	1	0	0	0	2

6.5 Strata defined from smoke

We repeat the analysis in Sections 6.1- 6.2 with a different division into strata. The four strata are now defined based on smoke. The main conclusion is that no significant curve groups or groups of covariates are identified. The results are very similar to the previous section with HRT.

Table 51 Number of case-control pairs in each smoke stratum and year.

Year	5-8	3-4	2	1		Sum
Period	4	3	2	1		all
Stratum	Invasive (Insitu)					years
YcaseYctrl	4 (0)	16 (3)	6 (0)	7 (2)		33 (5)
NcaseYctrl	17 (1)	42 (9)	14 (2)	16 (7)		89 (19)
YcaseNctrl	14 (1)	39 (6)	20 (2)	14 (2)		87 (11)
NcaseNctrl	52 (7)	98 (20)	57 (7)	51 (10)		258 (44)
Sum strata	87 (9)	195 (38)	97 (11)	88 (21)		467 (79)

Except for stratum YcaseYctrl, there is a sufficient amount of data for all time periods in each stratum.

6.5.1 Identifying curve groups for smoke

No significant curve groups are identified (Table 52 and Table 53).

Table 52 Number of significant curve groups for smoke with four time periods. (Detailed results are found in M:\bioinf\Prj\EilivLund\Datasett5\SMOKE\Res4periods\resNorm*TestForEachStratum.txt – files not included.)

# Simulations=1000	Number of significant curve groups with four periods for smoke 24 curve groups for each strata									
Dataset	YcaseYctrl		NcaseYctrl		YcaseNctrl		NcaseNctrl		Sum	
p-value	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05
	a=0.1									
Not normalized	0	0	0	0	0	0	0	3	0	3
Norm. method 1	0	0	2	1	0	0	0	0	2	1
Norm. method 2	0	0	1	1	0	0	0	1	1	2
Norm. method 3	0	0	1	1	0	0	0	1	1	2
	a=0.01									
Not normalized	0	0	0	2	0	0	0	6	0	8

Norm. method 1	0	0	2	1	0	0	0	0	2	1
Norm. method 2	0	0	0	1	0	0	0	0	0	1
Norm. method 3	0	0	0	1	0	0	0	0	0	1

Table 53 Number of significant curve groups for smoke with three time periods. (Detailed results are found in M:\bioinf\Prj\EilivLund\Datasett5\SMOKE\Res3periods\resNorm*TestForEachStratum.txt – files not included.)

# Simulations=1000	Number of significant curve groups with three periods for smoke 6 curve groups for each strata									
Dataset	YcaseYctrl		NcaseYctrl		YcaseNctrl		NcaseNctrl		Sum	
p-value	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05
	a=0.1									
Not normalized	0	0	0	1	0	0	0	1	0	2
Norm. method 1	0	0	0	0	0	0	0	0	0	0
Norm. method 2	0	0	0	0	0	0	0	0	0	0
Norm. method 3	0	0	0	0	0	0	0	0	0	0
	a=0.01									
Not normalized	0	0	0	0	0	0	0	0	0	0
Norm. method 1	0	0	0	0	0	0	0	0	0	0
Norm. method 2	0	0	0	0	0	0	0	0	0	0
Norm. method 3	0	0	0	0	0	0	0	0	0	0

6.5.2 Selecting covariates with different means for smoke

For each covariate (see Section 6.2) and time period, we compare the means of the case-control pairs in two different strata. We will only compare two strata where either the controls or the cases have the same value for smoke (Yes or No) in the two strata. As there is not is a sufficient amount of data for all time periods for YcaseYctrl, this stratum is not included in any comparisons. We are therefore left with two comparisons: YcaseNctrl versus NcaseNctrl, and NcaseYctrl versus NcaseNctrl. We also compare YcaseNctrl versus NcaseYctrl.

With four time periods we perform 192 (48 tests x 4 periods) tests for each normalization method. If there is no signal in the data we expect a p-value below 0.05 in 10 tests. There are not more significant values than expected under the null model.

With three time periods we perform 36 (12 tests x 3 periods) tests. If there is no signal in the data we expect a p-value below 0.05 in 2 tests. There are not more significant values than expected under the null model.

Table 54 Number of significant covariates for smoke with four time periods. (Detailed results are found in M:\bioinf\Prj\EilivLund\Datasett5\SMOKE\Res4periods\TestCovariates*resS?norm?.txt – files not included.)

# Simulations=1000	Number of significant covariates with four periods for smoke 24 curve groups, 48 test per time period									
Period	4		3		2		1		Sum	
p-value	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05
	YcaseNctrl and NcaseNctrl – a=0.1									
Not normalized	1	4	1	1	1	0	0	2	3	7
Norm. method 1	0	0	0	0	0	0	0	0	0	0
Norm. method 2	0	0	0	2	0	0	0	0	0	2
Norm. method 3	0	0	0	1	0	0	0	0	0	1
	NcaseYctrl and NcaseNctrl – a=0.1									
Not normalized	1	6	0	1	1	5	1	2	3	14
Norm. method 1	0	0	2	9	0	2	1	7	3	18
Norm. method 2	0	0	1	3	0	3	1	8	2	14
Norm. method 3	0	0	1	4	0	3	1	8	2	15
	YcaseNctrl and NcaseYctrl – a=0.1									
Not normalized	0	2	0	2	0	4	0	0	0	8
Norm. method 1	0	1	0	3	2	1	0	1	2	6
Norm. method 2	0	1	0	2	0	3	0	1	0	7
Norm. method 3	0	1	0	2	0	3	0	0	0	6
	YcaseNctrl and NcaseNctrl – a=0.01									
Not normalized	1	2	0	1	0	1	1	0	2	4
Norm. method 1	1	3	0	2	0	3	2	1	3	9
Norm. method 2	1	1	0	1	1	1	1	3	3	6
Norm. method 3	1	1	0	2	1	0	1	1	3	4
	NcaseYctrl and NcaseNctrl – a=0.01									
Not normalized	1	2	0	0	1	1	1	2	3	5
Norm. method 1	0	0	0	4	0	1	1	3	1	8
Norm. method 2	0	0	0	2	0	2	1	4	1	8
Norm. method 3	0	0	0	2	0	1	1	6	1	9
	YcaseNctrl and NcaseYctrl – a=0.01									
Not normalized	1	1	0	3	0	2	0	2	1	8
Norm. method 1	1	2	0	6	1	1	1	2	3	11
Norm. method 2	1	3	0	3	1	1	1	0	3	7
Norm. method 3	2	2	0	3	1	0	1	0	4	5

Table 55 Number of significant covariates for smoke with three time periods. (Detailed results are found in M:\bioinf\Prj\EilivLund\Datasett5\SMOKE\Res3periods\TestCovariates*\resS?norm?.txt – files not included.)

# Simulations=1000	Number of significant covariates with three periods for smoke 6 curve groups, 12 tests per time period							
Period	3		2		1		Sum	
p-value	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05	< 0.01	0.01-0.05
	YcaseNctrl and NcaseNctrl – a=0.1							
Not normalized	0	1	0	0	0	1	0	2
Norm. method 1	1	0	0	1	0	0	1	1
Norm. method 2	1	0	0	0	0	0	1	0
Norm. method 3	1	0	0	0	0	0	1	0
	NcaseYctrl and NcaseNctrl – a=0.1							
Not normalized	0	0	0	0	0	1	0	1
Norm. method 1	0	0	0	0	0	0	0	0
Norm. method 2	0	1	1	0	0	0	1	1
Norm. method 3	0	1	1	0	0	0	1	1
	YcaseNctrl and NcaseYctrl – a=0.1							
Not normalized	0	0	0	0	0	2	0	2
Norm. method 1	1	1	0	0	0	0	1	1
Norm. method 2	1	0	0	1	0	1	1	2
Norm. method 3	1	0	0	1	0	1	1	2
	YcaseNctrl and NcaseNctrl – a=0.01							
Not normalized	0	1	0	1	0	0	0	2
Norm. method 1	0	0	0	0	0	0	0	0
Norm. method 2	0	0	0	0	0	0	0	0
Norm. method 3	0	0	0	0	0	0	0	0
	NcaseYctrl and NcaseNctrl – a=0.01							
Not normalized	0	0	0	0	0	0	0	0
Norm. method 1	0	1	0	0	0	0	0	1
Norm. method 2	0	0	0	0	0	0	0	0
Norm. method 3	0	0	0	0	0	0	0	0
	YcaseNctrl and NcaseYctrl – a=0.01							
Not normalized	0	1	0	1	0	0	0	2
Norm. method 1	0	1	0	0	0	0	0	1
Norm. method 2	0	0	0	0	0	0	0	0
Norm. method 3	0	0	0	0	0	0	0	0

7 Conclusion

This note describes methods for and results of analyzing a large and impressive dataset of $\log_2$-transformed gene expression values in blood cells related to breast cancer. In the analyses we are looking for very weak signals that differentiate between spread and not spread of breast cancer based on average expression values of many genes. Under the null model we do not expect to identify any signal in the gene expression data. For being able to discover weak signals, we focus on groups of genes that have a particular functional form the last years before diagnosis.

The developed methods have been tested on five different version of the dataset as the data are continuously updated when new information becomes available (for example when new individuals are diagnosed with cancer or the quality of the data is improved) and because different subsets of the dataset have been selected dependent of what information we wanted to include in the analyses. This, and slightly different choices in the preprocessing steps, resulted in different subsets of genes selected for the different versions of the dataset.

The dataset analyzed in Section 2 have a significantly high number of genes that increase or decrease monotonically in gene expression the years before diagnosis in the stratum where we a priori expect it is most likely to observe a signal. We expect a more homogeneous dataset for persons participating in a screening program and expect a stronger signal from patients with spread. However, the signal is still weak. Using information from the identified groups of genes when predicting spread or not spread, we were able to identify about 1/3 of the cases without spread and no or few false negatives based on two covariates.

In Section 3 the dataset is extended with year of diagnosis, but the number of genes is reduced. Also here we are able to identify a significantly high number of genes in some curve groups (monotonically increasing gene expression). We get most significant results using quantile normalization method 3, i.e. when normalizing the gene expression data for the case and control data separately and then compute the $\log_2$-differences. Also, the stratum with screening and spread have largest change in gene expression the last year before diagnosis. Our prediction of spread or not is about the same as for the dataset of Section 2.

In Section 4 and 5 we extended the dataset with several new case-control pairs at the time of diagnosis. These data were produced using chips with a slightly different design. The dataset in Section 4 (5) have 20% (10%) less genes than the dataset in Section 2. The dataset in Section 5 is also extended with 38 case-control pairs where the cancer of the cases is defined as in situ and where gene expression is measured before diagnosis. The results are quite similar to the results in Section 3 with a significant number of genes for some curve groups and we are able to predict no spread for a group a patients within the training dataset. The new datasets are used as test sets where we estimate the parameters from the dataset before diagnosis and one of the datasets at the time of diagnosis. However, the prediction properties in the test sets are weaker than in the leave-one-out study in the other datasets. It is difficult to explain this difference. It may be due to the change in genes selected for the analyses, the difference in the test set up used in the test sets or that we for these dataset find other covariates (compared to Section 2) as the most significant covariates, or a combination of these three effects.

In Section 6 we extend the dataset with new case-control pair up to 8 years before diagnosis and there are 10% less genes than in the dataset in Section 2. Here we are not able to find a significant number of genes in any of the curve groups. When testing the effects of HRT and smoking we did not obtain any significant results.

We have observed that the results are sensitive to the subset of genes selected. Later, we will examine this further to find the procedure for selecting genes to be included in the statistical analyses that is best suited for our dataset.

From the results described in Section 2.2 we conclude that it is important to normalize the data before further analysis. However, normalizing the data may also remove trends we are looking for, and we have observed that the results presented are sensitive to the choice of normalization method. Therefore different normalization methods should be tested and evaluated to decide which method is best suited in our case.

The preliminary statistical methods will be further developed later, and they will also be tested on a dataset with improved quality where more optimal preprocessing procedures and normalization methods are used. This work will include performing a more theoretical study on how to analyze weak signals in many genes in contrast to stronger signals in few genes. This study may result in development of new statistical techniques that are particularly suited for the unique dataset we are analyzing. The different methods developed will also be extended to incorporate different kinds of additional information like HRT status (Section 6.4), smoking habits (Section 6.5), or the possibility of stress reactions connected to a breast cancer diagnosis (Section 5.3).