NR Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Analysis of gene expression in blood before diagnosis of ovarian cancer

Different statistical methods

Note

Note no.	SAMBA/10/16
Authors	Marit Holden and Lars Holden
Date	March 2016

NR Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

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.

Title	**Analysis of gene expression in blood before diagnosis of ovarian cancer**
Authors	**Marit Holden and Lars Holden**
Date	March 2016
Year	2016
Publication number	SAMBA/10/16

Abstract

The analyses in this note are based on a dataset with gene expression in blood before diagnosis of ovarian cancer. The dataset consists of case-control pairs that are matched on birth year and time of blood sampling, and the data for a pair is the $\log_2$ difference in gene expression between the case and control. For each case-control pair the gene expression is measured once before diagnosis. As the blood samples of the different case-controls pairs are measured at different points in time before diagnosis, we have used the dataset for examining whether the gene expression profile varies with time. We have also used the dataset for examining whether the gene expression profile varies between cases and controls, or between cases with and without spread (metastases), and for predicting whether a case has ovarian cancer with or without spread.

We have used and adapted a method based on hypothesis testing using randomization, that is able to identify small changes that are varying slowly in time and/or among strata, by using a large number of genes in each hypothesis test and predictor. Even though the signals in the data are weak, we concluded that the gene expression profile varies in time, between cases and controls and between cases with and without spread (metastases). The results indicated that there is an increasing variation in the gene expression profiles when approaching the time of diagnosis, while the gene expression profiles far from diagnosis are more stable.

The dataset is quite small, with only 59 case-control pairs with spread and 28 without spread that are distributed over a seven year period before diagnosis. We can therefore not draw any firm conclusion about whether the predictive power of the method used for predicting the metastasis status of the cases is sufficiently good (p-value 0.28, Fisher's test). The best predictive power was observed in a two-year period around year 5 before diagnosis (p-value 0.12, Fisher's test).

Keywords	Gene expression; Ovarian cancer; Blood; Metastasis; Development in time; Hypothesis testing using randomization; Prediction of metastasis status.
Target group	Clinical medicine; Systems epidemiology.
Availability	Open
Project number	ERC-AdG 232997 TICE (Transcriptomics In Cancer Epidemiology). NR project number 220744
Research field	MMH, Bioinformatics
Number of pages	36
© Copyright	Norsk Regnesentral

Table of Content

1 Introduction

The analyses in this note are based on a dataset with gene expression in blood before diagnosis of ovarian cancer. The dataset consists of case-control pairs that are matched on birth year and time of blood sampling, and the data for a pair is the $\log_2$ difference in gene expression between the case and control. For each case-control pair the gene expression is measured once before diagnosis. As the blood samples of the different case-controls pairs are measured at different points in time before diagnosis, we can use the dataset for examining whether the gene expression profile varies with time. We will also use the dataset for examining whether the gene expression profile varies between cases and controls, or between cases with and without spread (metastases), and for predicting whether a case has ovarian cancer with or without spread.

In previous analyses we used the Bioconductor R-package Limma (Linear models for microarrays) for identifying genes that are differentially expressed between cases and controls. No differentially expressed genes were found for any of the examined datasets. Also, when using Limma analysis for identifying genes that are differentially expressed between cases with and without spread, no differentially expressed genes were found. In these analyses information about time to diagnosis was used only when selecting the dataset for each analysis. In addition, we analyzed the data using an approach based on curve groups [1] where information about time to diagnosis is included in the analysis. This approach was used both for comparing the different strata (with and without spread) and for testing whether there is a development in time for any of the strata. No significant results were obtained. See Section 8 (Appendix) for more details about the analyses described above.

In this note we will use and adapt a method that includes time, that is not based on curve groups, and that is able to identify small changes that are varying slowly in time and/or among strata, by using a large number of genes in each hypothesis test and predictor [2]. In Section 2 we present the dataset. Methods are described in Section 3, while results are summarized in Section 4.

2 Dataset

The available dataset consists of data from 87 case-control pairs with time to diagnosis varying between 1 and 2555 days (year 1-7 before diagnosis). Each case belongs to one of the two strata with spread and without spread. More details about the dataset, like the number of case-control pairs in each stratum and the distribution of the case-controls pairs in time, are given in Table 1 and Figure 1. The data used in all analyses are the $\log_2$ differences in gene expression between cases and controls.

Table 1 *Details about the available dataset.*

Number of case-control pairs									
Year before diagnosis		7	6	5	4	3	2	1	Sum
Stratum	Without spread	0	2	4	6	6	7	3	**28**
	With spread	5	5	11	7	13	9	9	**59**
	Sum	**5**	**7**	**15**	**13**	**19**	**16**	**12**	**87**

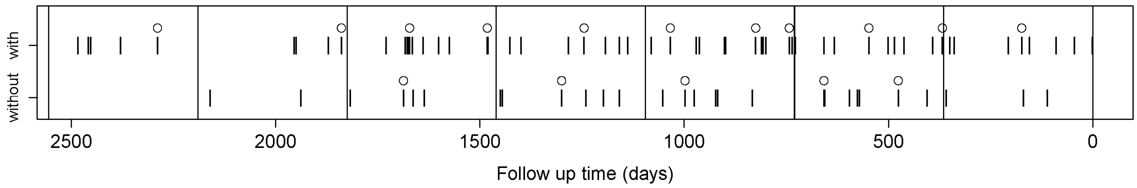


Figure 1 *The distribution of the case-controls pairs in time. Each short vertical line represents a case-control pair. A circle is plotted above every fifth case-control pair. Long vertical lines are plotted to indicate the years. On the y-axis "with" means cases with spread and "without" means cases without spread.*

The dataset has been preprocessed using a procedure that consists of the following steps:

1. Background correct the data using negative control probes.
2. Remove non-present probes, i.e. only probes with detection p-value less than 0.05 in more than 70% of the 90 x 2 = 180 samples remain in the dataset.
3. Transform the data using the variance stabilizing technique described in [3].
4. Quantile normalized the data.
5. Map probes to genes. When several probes map to the same gene, the average expression of the probes is used as expression value for the gene.

A more detailed description of each step is given in [4]. After preprocessing the dataset consists of 9644 genes. All data are from three different plates in the same run, we therefore assume that there are no batch effects in the data. Note that three of the 90 case-control pairs are removed from the dataset as they did not belong to any of the two strata (with spread and without spread). See Section 7 (Appendix) for more details.

3 Methods

The method described here is explained in more detail in [2], and it has been used to analyze a similar dataset based on blood samples from cases with breast cancer [5].

Let $X_{g,c}$ be the log$_2$-expression difference for case-control pair c and gene g. Let $\mu_{g,s,t}$ and $\sigma_{g,s,t}$ be the expectation and standard deviation of $X_{g,c}$, respectively, where s is the stratum and t is the time to diagnosis for $X_{g,c}$. If the distribution of $X_{g,c}$ does not vary in time or between strata, the expectation and variance of $X_{g,C}$ are independent of time and stratum, i.e. $\mu_{g,s,t} = \mu$ and $\sigma_{g,s,t} = \sigma$ for all strata s and time before diagnosis t. Also, if there is no difference between cases and controls, the expectation of $X_{g,C}$ is zero, i.e. $\mu_{g,s,t} = 0$.

3.1 Hypothesis tests for finding signal in the data

For examining whether there are differences between cases and controls, between strata or in time, we will test different hypotheses. For each hypothesis the statistic will be based on either expectation or standard deviation or both. The null distribution of the statistic will be estimated by randomizing the data, and we compute p-values by comparing the statistic for the data to the estimated null distribution.

Let $m_{p,g}$ be the sample mean and $s_{p,g}$ be the sample standard deviations for the gene expression for gene g in time period p. Let $m_{p,g,1}(m_{p,g,0})$ be the sample mean and $s_{p,g,1}$ $(s_{p,g,0})$ be the sample standard deviations for the gene expression for gene g in time period p for stratum 1 (0).

We define the statistics $s_{p,(g)}$, $m^1_{p,(g)}$, $m^2_{p,(g)}$ and $w_{p,(g)}$ as follows[1]:

- $\boldsymbol{s_{p,(g)}}$ is the g'th smallest of $s_{p,g}$ for period p.
- $\boldsymbol{m^1_{p,(g)}}$ is the g'th largest of $|m_{p,g}|$ for period p.
- $\boldsymbol{m^2_{p,(g)}}$ is the g'th largest of $\left|\frac{m_{p,g}}{s_{p,g}}\right|$ for period p.
- $\boldsymbol{w_{p,(g)}}$ is the g'th largest of $|w_{p,g}|$ for period p, where $w_{p,g} = \frac{m_{g,p,1}-m_{g,p,0}}{\sqrt{s^2_{g,p,1}+s^2_{g,p,0}}}$ is the weight for gene g in time period p.

These four statistics are used for testing the following three null hypotheses:

<u>H01: The distribution of $X_{g,c}$ does not depend on the time to diagnosis.</u>

- This means that the expectation and standard deviation of $X_{g,c}$ are the same in all time periods.
- If the null hypothesis is false, the standard deviation for some periods will be lower than the standard deviations for the entire time period for some genes. Also, the absolute value of the expectation for some periods will be higher than the absolute value of the expectation for the entire time period for some genes.
- We test the hypothesis first by using the statistic $s_{p,(g)}$, and then by using the statistic $m^1_{p,(g)}$.
- The null distributions of the statistics are estimated by randomizing the case-control pairs between the periods.

<u>H02: The expectation of $X_{g,c}$ is zero.</u>

- This means that there is no difference between the expectations of the gene expression values for the cases and controls.
- If the null hypothesis is false, the expectation will be different from zero for some periods and genes.
- We test the hypothesis first by using the statistic $m^1_{p,(g)}$, and then by using the statistic $m^2_{p,(g)}$.
- The null distributions of the statistics are estimated by randomizing the case and control in each case-control pair. In practice this is done by keeping the absolute value of all gene expression differences, but simulating their signs.

<u>H03: The expectation of $X_{g,c}$ does not depend on stratum.</u>

- This means that $\mu_{g,1,t} = \mu_{g,0,t}$, i.e. the expectations for the two strata are equal for all genes g and time to diagnosis t.
- If the null hypothesis is false, the difference in expectation will be different from zero for some periods and genes.
- We test the hypothesis by using the statistic $w_{p,(g)}$.

[1] Note that the second and third statistic were not defined in [2].

- The null distribution of the statistic is estimated by randomizing between the two strata within the time period.
- Note that we compute $w_{p,(g)}$ only if there are at least three case-control pairs in period p for each stratum. If this is not the case, we set the p-value to 1 for this period for all genes.

3.2 Predicting metastasis status

Let $m_{p,g,1,-j}(m_{p,g,0,-j})$ be the sample mean and $s_{p,g,1,-\mathrm{j}}$ $(s_{p,g,0,-j})$ be the sample standard deviations for the gene expression for gene g in period p for stratum 1 (0) when sample j is not included.

We define the weights for the genes, $w_{p,g,-j}$, as:

$$w_{p,g,-j} = \frac{m_{p,g,1,-j} - m_{p,g,0,-j}}{\sqrt{s_{p,g,1,-j}^2 + s_{p,g,0,-j}^2}}$$

and compute

$$z_j = \sum_{g=1}^{n} w_{p,(g),-j} x_{(g),j},$$

where (g) is the gene with the g'th largest $|w_{p,g,-j}|$. Large values of $z_j$indicates that case j belongs to group 1. If z_j>c we conclude that case j belongs to group 1, otherwise we conclude that case j belongs to group 0. We may set c=0 if it is not more important to avoid false classification in one group relative to the other and if

$$\sum_{g=1}^{n} w_{p,(g),-j} \frac{m_{p,(g),1,-j} + m_{p,(g),0,-j}}{2} \approx 0,$$

where $m_{p,(g),1,-j}$ and $m_{p,(g),0,-j}$ are the sample means that are used when computing $w_{p,(g),-j}$.

4 Results

Before testing the hypotheses described above and predicting metastasis status, we need to decide how to divide into time periods. We want as short time periods as possible (as the distribution may vary with time), but at the same time we want as many case-control pairs as possible within each time period. As there is trade-off between these two wishes we have tested with some different time periods where the length depends on the number of cases with spread. Periods that contain 25 cases with spread seems to be reasonable both with respect to the number of case-control pairs (25 with spread, 9-17 without spread), and with respect to the length of the time periods (742-987 days except for the five periods that include the case-control pairs in year 7 before diagnosis). We have therefore selected time periods that contain 25 cases with spread. We have defined one time period for each set of 25 cases with spread that are consecutive in time. As there are 59 cases with spread, this resulted in 35 different, and overlapping, time periods.

In all hypothesis tests described in this section the estimated null distribution consists of 1000 samples.

4.1 Comparing periods close to and far from time of diagnosis

We show results for two variants of the dataset, one where we have standardized[2] the data to expectation zero and standard deviation one for each gene, and one without standardizing the data. Figure 2 shows plots of the three statistics that depend on whether the dataset has been standardized, while Figure 3 shows a plot of the statistic $\mathrm{w}_{\mathrm{p},(\mathrm{g})}$ that does not depend on whether the dataset has been standardized.

In Figure 2 we observe that the shape of the curves in the two plots with standard deviation (Figure 2 a)) are quite different. In the plot with not standardized data there are many small, and few large standard deviations, while the standard deviations, as expected, are around 1 for the standardized data. All four cases that are based on the expectation are very similar (Figure 2 b) and c)). Independent of whether the data have been standardized or whether the expectations have been divided by the standard deviation, the statistic is largest for the period far from diagnosis (H01) and larger with spread than without spread (H03).

From Figure 2 we also observe that $s_{p,(g)}$ is smaller for the approximately 5000 genes with smallest standard deviation for the stratum with spread in the period far from diagnosis, and that $m_{p,(g)}^1$ and $m_{p,(g)}^2$ are larger for the stratum with spread in the period far from diagnosis. Smaller standard deviations and larger expectations for the stratum with spread in the period far from diagnosis indicate that the gene expression profiles of the cases with and without spread are different (H03). Close to diagnosis the approximately 3000 genes with largest standard deviation are larger for the stratum with spread (H03).

[2] We have also tested with a dataset were we standardized to standard deviation one for each gene under the assumption that the expectation is zero for each gene. This method gave very similar results to the dataset where we standardized to mean zero and standard deviation one. We have therefore not shown results for the alternative standardization method.

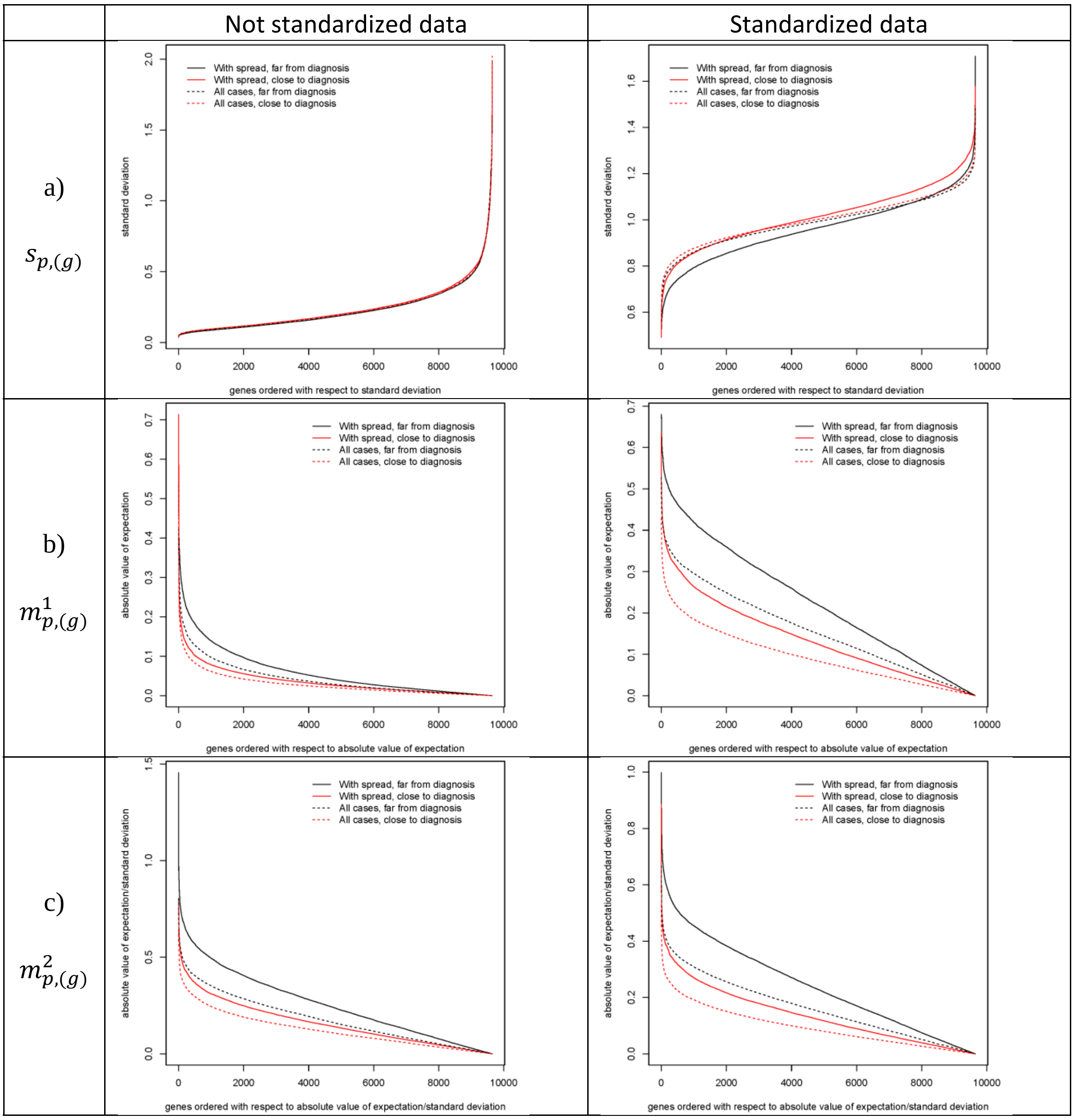


Figure 2 *Plots of the three statistics that depend on whether the dataset has been standardized. The two periods contain 25 case-control pairs where the case is with spread. The time period closest to diagnosis is 1-825 days before diagnosis (year 1, year2, three months of year 3), while the time period furthest from diagnosis is 1245-2484 days before diagnosis (five months of year 4, year 5, year 6, ten months of year 7).*

It is maybe surprising that the expectations ($m^1_{p,(g)}$ and $m^2_{p,(g)}$) are larger, and the standard deviations ($s_{p,(g)}$) are smaller, far from diagnosis than close to diagnosis, i.e. that the gene expression profile of the cases is more similar to the gene expression profile of the controls close to diagnosis than far from diagnosis. One possible explanation could be that the gene expression profiles change several years before diagnosis and is quite stable for several years, but as the point of diagnosis approaches the gene expression profiles change quite rapidly, but at different times before diagnosis for the different cases. Such behavior could lead to smaller standard deviations and larger expectations far from diagnosis than closer to diagnosis. The long sampling period may also give larger variation. This may be more important close to diagnosis.

Figure 3 shows results for the statistic $w_{p,(g)}$ that measures the difference between the gene expression of the cases with and without spread relative to their standard deviations. As the statistics based on expectation, the statistic $w_{p,(g)}$ is largest far from diagnosis. Also, this observation can agree with gene expression profiles that change several years before diagnosis and that are quite stable for several years, also for the cases without spread, but that change quite rapidly as the point of diagnosis approaches (H01).

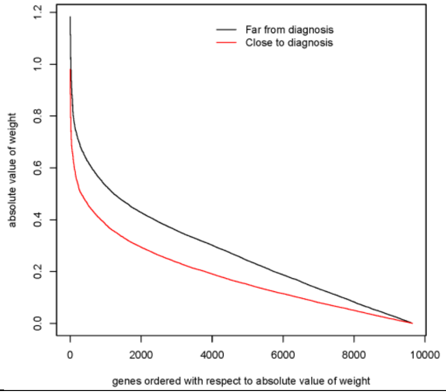
$w_{p,(g)}$	

Figure 3 *Plot of the statistic $w_{p,(g)}$ that does not depend on whether the data are standardized. The two periods contain 25 case-control pairs where the case is with spread.*

Figure 4 and Figure 6 show plots of p-values for four of the hypothesis tests described in Section 3.1 when the dataset with cases with spread is used. Note that in all plots the curves with p-values have been smoothed using a median-filter with window size 11. Figures with results for the entire dataset are shown in Section 9 (Appendix, Figure 11 and Figure 12). Comparing Figure 4 and Figure 6 with Figure 11 and Figure 12, we observe that the results are better, i.e. there are more low p-values, when only cases with spread are included in the dataset. This indicates that the gene expression profiles of the cases with and without spread are different. The plot that is shown in Figure 5 confirms this. For two periods approximately 4 years before diagnosis, most p-values are either below or close to 0.05. The null hypothesis tested is $w_{p,(g)} = 0$. When this hypothesis is rejected, we conclude that the expectations for some genes in period p are different for the two strata. This means that the expectation of $X_{g,c}$, i.e. the log$_2$-expression difference for case-control pair c and gene g, depends on stratum (H03).

Figure 4 shows results for the hypothesis test that is based on the statistic $s_{p,(g)}$ and that is used for testing if the standard deviation in a period is small compared to the standard deviation for all periods. The plots in Figure 4 a) show that the p-values are above 0.05 for all periods. We have used the same amount of data when estimating $s_{p,(g)}$ that is based on estimating the standard deviation, as for the statistics that are based on estimating the mean. As more data are needed to obtain reliable estimates of the standard deviation than the mean, we also performed hypotheses tests based on $s_{p,(g)}$ with fewer periods with more data in each period. The plots in Figure 4 b) show results for periods that each contains 35 case-control pairs where the case is with spread. We observe that in this case the period 4.5 years before diagnosis have low p-values. We also observe that the periods closest to diagnosis have very large p-values. This indicates that the variance in gene expression is larger close to time of diagnosis than further from time of diagnosis, as we also concluded from the results in Figure 2 a). From this we may conclude that for some genes the distribution of $X_{g,c}$ depends on time to diagnosis. This is confirmed by the plots in Figure 6 a), where low p-values are obtained for

$m_{p,(g)}^{1}$ for several of the periods. This indicates that for some genes the expectation of $X_{g,c}$ is larger in some periods than in other periods (H01).

Figure 5 shows results based on the statistic $w_{p,(g)}$ that is used for comparing the expectations of the two strata in the dataset. This statistic is closely connected to the possibility of differentiating between cases with and without spread based on gene expression values and time to diagnosis. The plots in Figure 5 indicate that the cases with and without spread are differentially expressed for some genes far from the time of diagnosis. In these periods it is reasonable to assume that the gene expression profiles differ between cases with and without spread and that they are not changing rapidly with time within the periods (H01 and H03).

In Figure 6 that shows results for mean values, we observe low p-values for 1.5 years and far from diagnosis. This is in accordance with the results that are shown in Figure 2 b) and c) for the data for the statistics $m_{p,(g)}^{1}$ and $m_{p,(g)}^{2}$. The hypotheses tests based on these two statistics that randomize the case and control in each case-control pair, is used for testing whether the expectation of $X_{g,c}$ is zero. The plots in Figure 6 b) and c), show that low p-values are obtained for the period 1.5 years before diagnosis. The best results are obtained when using $m_{p,(g)}^{1}$, that is based on the mean, and not standardized data randomized between periods. From this we conclude that the cases with spread and the controls are differentially expressed for some genes (H02).

4.2 Development in time before diagnosis

In the previous section we observed that the p-values for the different hypotheses tested varied between the different time periods depending on how far the time period was from diagnosis. In this section we illustrate the same results as shown in Figures 4-6 but now focusing on how the p-values vary with time. This is shown in Figure 7 for genes with order 50, 200, 500, 1000 and 2000, respectively. We observe that the standard deviation of the gene expression differences is small and the expectation is large in year 5-7 (Figure 7 a) and b)). Here we randomize the data between the time periods (H01). Also, the expectations of the cases and controls differ significantly in year 5-7 (Figure 7 c) and d)). Here we randomize between case and controls (H02). The expectations of the cases with and without spread differ most in year 4 (Figure 7 e), randomizing between with and without spread) (H03). This is a slightly shorter time period than the period with best prediction results in Figure 8 (see next section).

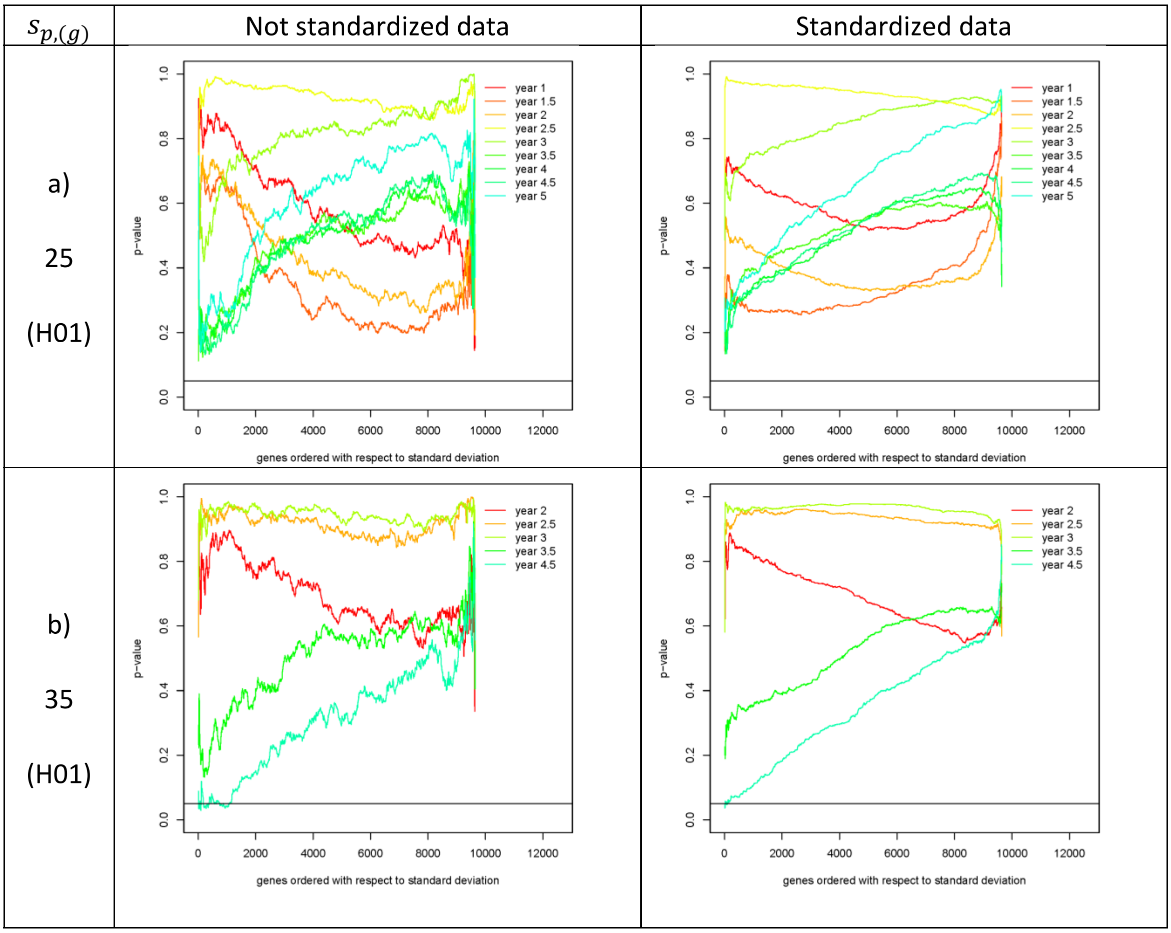


Figure 4 *Plots of p-values for the hypothesis tests based on the statistic* $s_{p,(g)}$ *where the dataset with cases with spread is used. The null distribution is estimated by randomizing the case-control pairs between the periods. a) Plots for periods that contain 25 case-control pairs where the case is with spread. b) Plots for periods that contain 35 case-control pairs where the case is with spread. In each plot there is one curve for every half year with a time period with 25 (35) case-control pairs with spread sufficiently close. The p-value is 0.05 at the black horizontal line.*

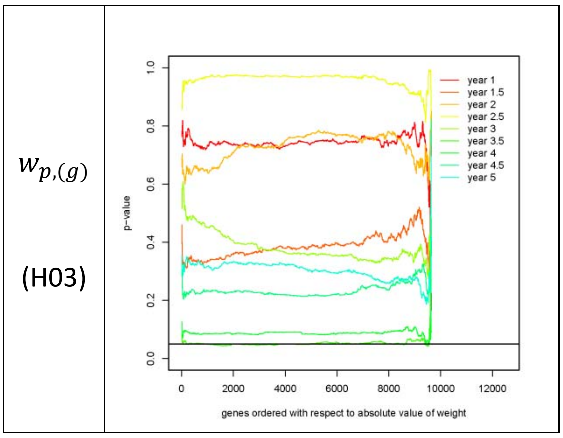


Figure 5 *Plots of p-values for the hypothesis tests based on the statistic* $w_{p,(g)}$ *where the expectations of the two strata in the dataset, with and without spread, are compared. The null distribution is estimated by randomizing the case-control pairs between the strata within the period. In each plot there is one curve for every half year with a time period with 25 case-control pairs with spread sufficiently close. The p-value is 0.05 at the black horizontal line.*

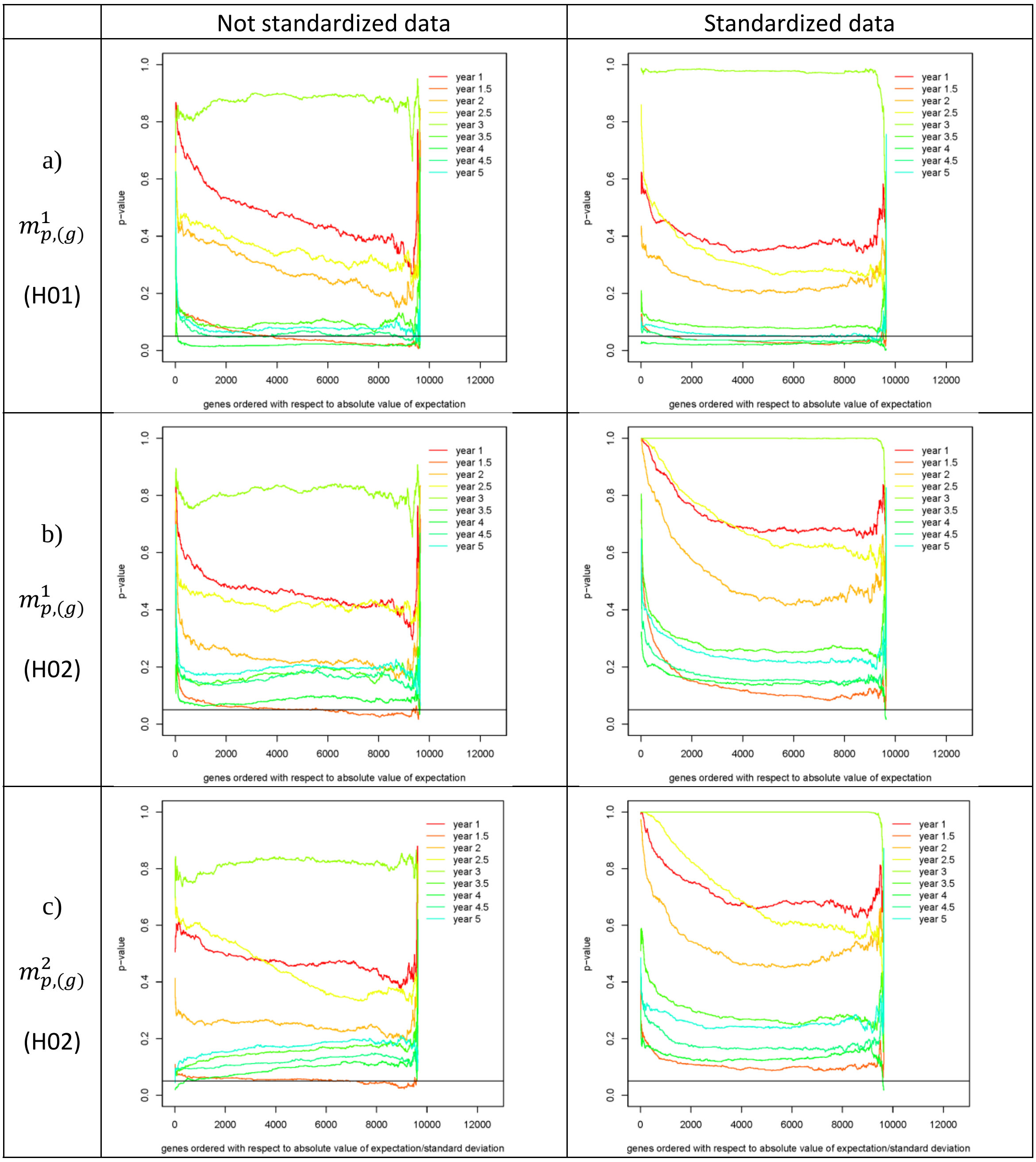
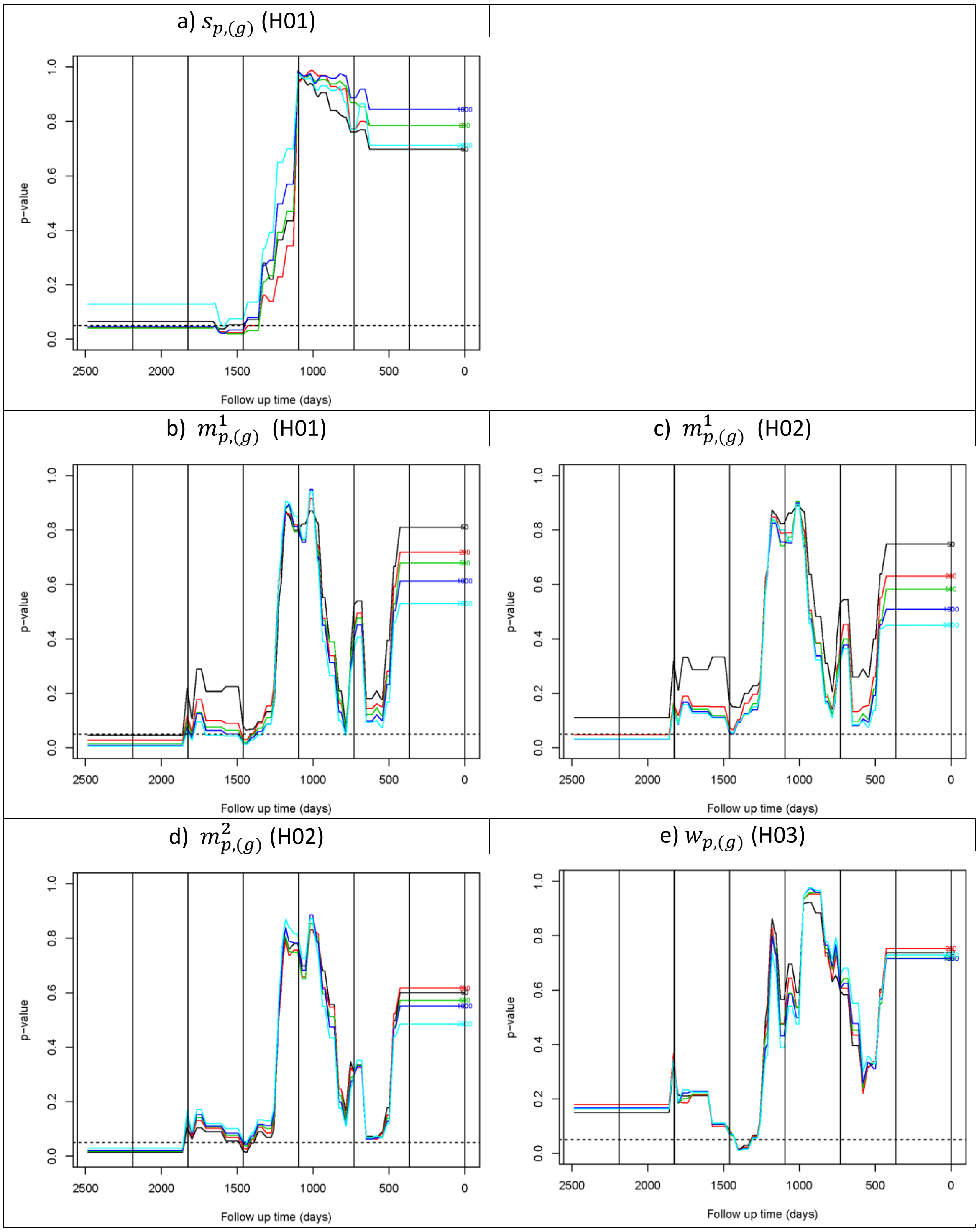


Figure 6 *Plots of p-values for three of the hypothesis tests where the dataset with cases with spread is used. P-values for all genes are included in the plots. a) The hypothesis test is based on the statistic* $m^1_{p,(g)}$ *and the null distribution is estimated by randomizing the case-control pairs between the periods. b) The hypothesis test is based on the statistic* $m^1_{p,(g)}$ *and the null distribution is estimated by randomizing the case and control in each case-control pair. c) The hypothesis test is based on the statistic* $m^2_{p,(g)}$ *and the null distribution is estimated by randomizing the case and control in each case-control pair. In each plot there is one curve for every half year with a time period with 25 case-control pairs with spread sufficiently close. The p-value is 0.05 at the black horizontal line.*



a) $s_{p,(g)}$ (H01)

b) $m^1_{p,(g)}$ (H01)

c) $m^1_{p,(g)}$ (H02)

d) $m^2_{p,(g)}$ (H02)

e) $w_{p,(g)}$ (H03)

Figure 7 *Plots of p-values against time for the hypothesis tests where the not standardized dataset is used. In panels a), b), c) and d) the dataset consists of the cases with spread, while in panel e) the entire dataset is used. In each plot there is one curve for genes with order 50 (black), 200 (red), 500 (green), 1000 (blue) and 2000 (light blue), respectively. P-value for time point t is equal to the p-value for the time period with middle point closest to t (after the p-values has been smoothed using a median-filtered with window size 99). The resulting curve is then smoothed using a mean-filter with a window size of one month. The p-value is 0.05 at the dotted horizontal line. Panel a) corresponds to Figure 4 b), panels b), c) and d) correspond to* Figure 6*a), b) and c), respectively, and panel e) corresponds to Figure 5.*

4.3 Predicting metastasis status of the cases

For predicting the metastasis status of the case in case-control pair j, we used the prediction method described in Section 3.2 with $n = 1000$, i.e. the 1000 genes with highest absolute value of the weights are used for computing the score that is used for prediction. The period selected for predicting the status of the case in case-control pair j is chosen among the 35 periods that contain 25 case-control pairs where the case is with spread, and it is chosen such that case-control pair j is as close to the middle of the time period as possible.

The results of the prediction are shown is Table 2 and Figure 8 a) and b). We observe that 63% of the cases with spread are correctly classified, while 46% of the cases without spread are correctly classified. The numbers of correctly classifies cases is not significantly higher than what is expected by chance (p-value 0.28, Fisher's test) (H03).

Table 2 *Number of correctly and wrongly classified cases.*

Number of correctly and wrongly classified cases				
With spread		Without spread		
FN	TP	FP	TN	P-value (Fisher's test)
22	37	15	13	0.28

To examine whether the probability of correctly classifying the status of the cases varies with time (H01), we plotted the prediction results against time in Figure 8 a) and b). This shows that the probability of correct classification is somewhat higher in a two-year period around year 5 before diagnosis. For this period the p-value obtained using Fisher's test is equal to 0.12. This is in accordance with the results shown in Figure 5 and Figure 7 e) for the statistic $w_{p,(g)}$, where we observe that the cases with and without spread are differentially expressed for some genes in some periods that are neither very close to nor very far from the time of diagnosis (H01, H03).

In Figure 10 a) we examine how the score is influenced by n, i.e. the number of genes included in the score, for the period with best prediction results (around year 5 before diagnosis). The score for the cases with spread should be positive, while the scores for the cases without spread should be negative. We observe that the score stabilizes when the number of genes increases. It is difficult to conclude how many genes to include in the score to optimize the power of the predictor, but at least 50 genes seem to be needed. To find out more about how sensitive the predictor is to the choice of n, we have repeated the analyses with $n = 50$ for the ovarian dataset and the two-year period around year 5 before diagnosis. The results of these additional analyses are shown in Figure 13 a) and Table 3 in Section 9 (Appendix). We observe that the results are similar to the results obtained with $n = 1000$, indicating that the predictor is not very sensitive to the number of genes included in the score.

It is difficult to draw any firm conclusions from the prediction results as the available dataset is too small, and the signals in the data are too weak.

4.4 Comparing prediction results for ovarian and breast cancer

In [5] we predicted metastasis status for the cases of two other prospective datasets where the cases were diagnosed with breast cancer, one where the cases participated in the screening program (the screening group, 380 case-control pairs) and one where the cases did not participate in the screening program (the clinical group, 87 case-control pairs). In Figure 8 we compare the prediction results for the three datasets (panel b - ovarian cancer group; panel c - breast cancer, clinical group; and panel d - breast cancer, screening group). We observe that the prediction results are best around year 5 before diagnosis for the ovarian cancer, around year 3 for the breast cancer, clinical group and around year 1 for the breast cancer, screening group. Note that for all three datasets and all time points the fraction of correctly classified cases are quite similar for the cases with and without spread indicating that a limit of 0 (c=0, see Section 3.2) for the score is a reasonable choice.

We have defined a set of genes for each of the three groups based on data for case-control pairs from the period with best prediction results: i) One for the ovarian cancer group for the period around 4 years and 6 months before diagnosis; ii) One for the breast cancer, clinical group around 2 years and 6 months before diagnosis; and iii) One set for the breast cancer, screening group around 6 months before diagnosis. In each set of genes we select the 1000 genes that with the largest $|w_{p,g}|$ (absolute value of weight for gene g in time period p). In Figure 9 we illustrate how the score based on each of the set of genes develops over time. As expected we observe that the difference between the cases with and without spread is largest around year 5, 3 and 1 before diagnosis, respectively. For each of the three sets of genes we have also examined how the score develops over time for all three datasets, not only the dataset that was used for selecting the genes. The results are shown in Figure 14 – Figure 16 in Section 9 (Appendix). For each of the three datasets, we observe that the differences between scores for the cases with and without spread are small when the set of genes is selected based on another dataset.

For the ovarian-cancer dataset we concluded that the predictor is not very sensitive to the number of genes included in the score, and we want to examine whether the same conclusion can be drawn also for other datasets. Figure 10 shows how the score is influenced by n, i.e. the number of genes included in the score, for the period with best prediction results for each of four different datasets. These four datasets are the three prospective ovarian and breast cancer datasets, and a validation dataset where the cases have no follow up time, i.e. zero days to diagnosis. This validation dataset is denoted the CC3 dataset, and the cases have breast cancer and they participated in the screening program. We observe that for all datasets there is a distinct difference in the score between cases with and without spread. The score stabilizes when the number of genes increases. When we repeat the analyses with $n = 50$, we obtain results that are similar to the results obtained with $n = 1000$, indicating that the predictor is not very sensitive to the number of genes included in the score for any of the datasets. See Figure 13 in Section 9 (Appendix).

Note that in Figure 10 c) the genes are selected from the screening group and applied for the CC3 dataset. For Figure 10 a), b) and d), however, the genes are selected based on the same dataset as we use when computing the scores.

5 Conclusion

For examining whether there are differences between cases and controls, between strata or in time, we have tested different hypotheses. For each hypothesis the statistic has been based on either expectation or standard deviation or both. The null distribution of the statistic has been estimated by randomizing the data, and we computed p-values by comparing the statistic for the data to the estimated null distribution.

Even though the signals in the data are weak, we conclude that the gene expression profile varies in time (H01), between cases and controls (H02) and between cases with and without spread (metastases) (H03). The results indicate that there is an increasing variation in the gene expression profiles when approaching the time of diagnosis, while the gene expression profiles far from diagnosis are more stable. We find the same results in different tests. All the tests are based on the same data and it is natural that this results in the same conclusions. We use several tests since each test illustrates slightly different properties of the same phenomenon. We also compared the prediction properties for ovarian cancer and breast cancer from a clinical and a screening group. We find that the prediction is best around year 5 before diagnosis for the ovarian-cancer dataset, year 3 for the breast-cancer dataset, clinical group, and year 1 for the breast-cancer dataset, screening group.

The dataset is quite small, with only 59 case-control pairs with spread and 28 without spread, that are distributed over a seven year period before diagnosis. We can therefore not draw any firm conclusion about whether the predictive power of the method used for predicting the metastasis status of the cases is sufficiently good (p-value 0.28, Fisher's test). The best predictive power was observed in a two-year period around year 5 before diagnosis (p-value 0.12, Fisher's test).

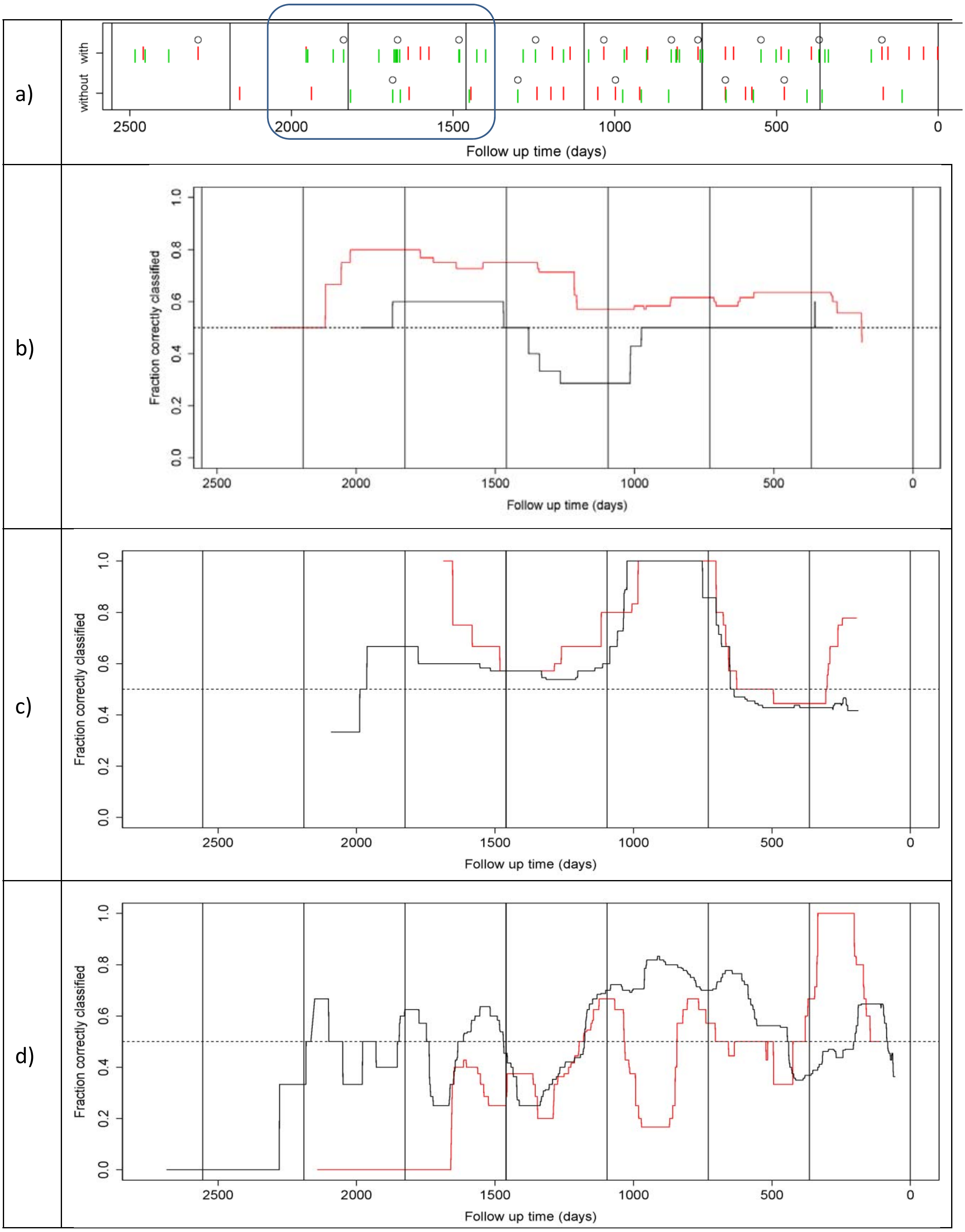


Figure 8 *a) Correctly (green) or wrongly (red) classified cases with ovarian cancer plotted against follow up time. A circle is plotted above every fifth case. Long vertical lines are plotted to indicate the years. On the y-axis "with" means cases with spread and "without" means cases without spread. The blue rectangle highlights the time period with best classification result. The p-value for Fisher's test for this period is 0.12. b) Fraction of correctly classified cases with ovarian cancer with (red) and without (black) spread over time. c) Fraction of correctly classified cases from the clinical group with breast cancer with (red) and without (black) spread over time. d) Fraction of correctly classified cases from the screening group with breast cancer with (red) and without (black) spread over time. The fraction for each point in time is computed using a moving window of one year. The resulting curve is then smoothed using a median-filter using a window size of one year.*

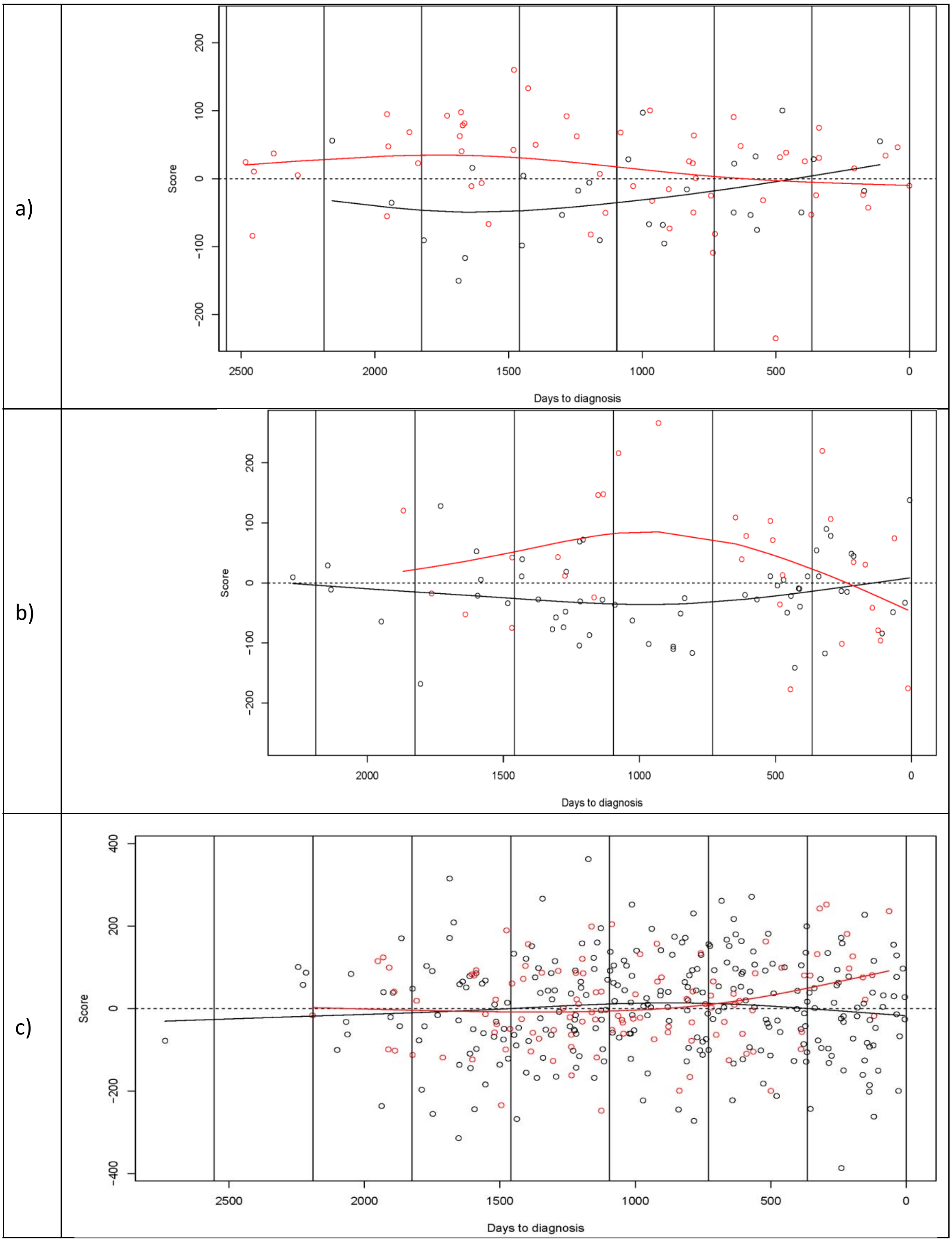
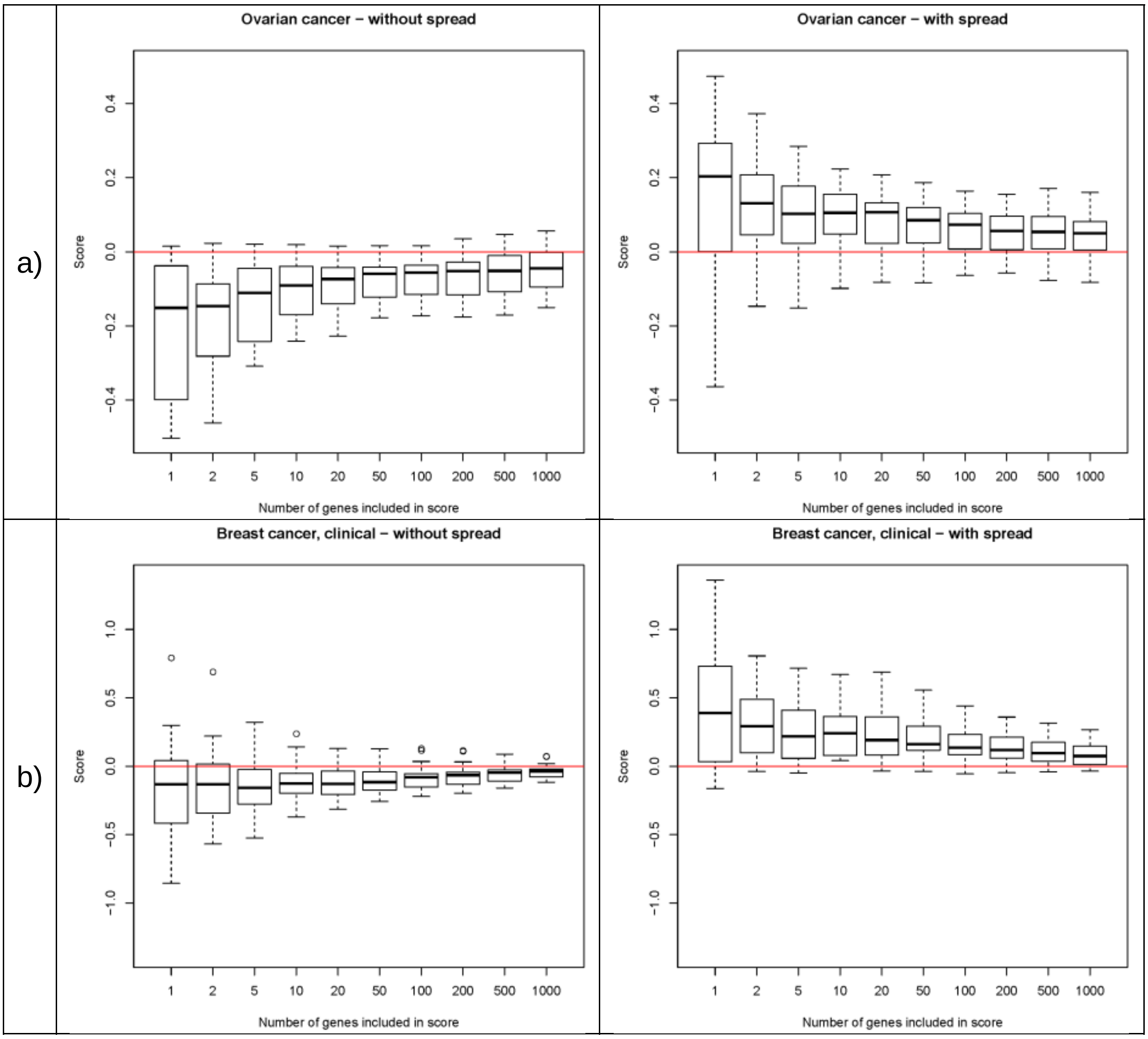


Figure 9 *Plots of scores for each case-control pair against time (days to diagnosis). The score is plotted in red (black) if the case is with (without) spread. The score is computed using the weights of 1000 genes that are selected based on data around 4 years and 6 months (a), 2 years and 6 months (b) and 6 months (c), respectively, before diagnosis. For illustrational purposes, curves have been estimated from the scores using splines and plotted in the same color as the individual scores. a) Scores for case-control pairs where the case has <u>ovarian cancer</u>. b) Scores for case-control pairs where the case has <u>breast cancer</u> and belongs to the <u>clinical group</u>. c) Scores for case-control pairs where the case has <u>breast cancer</u> and belongs to the <u>screening group</u>.*



The figure continues on the next page.

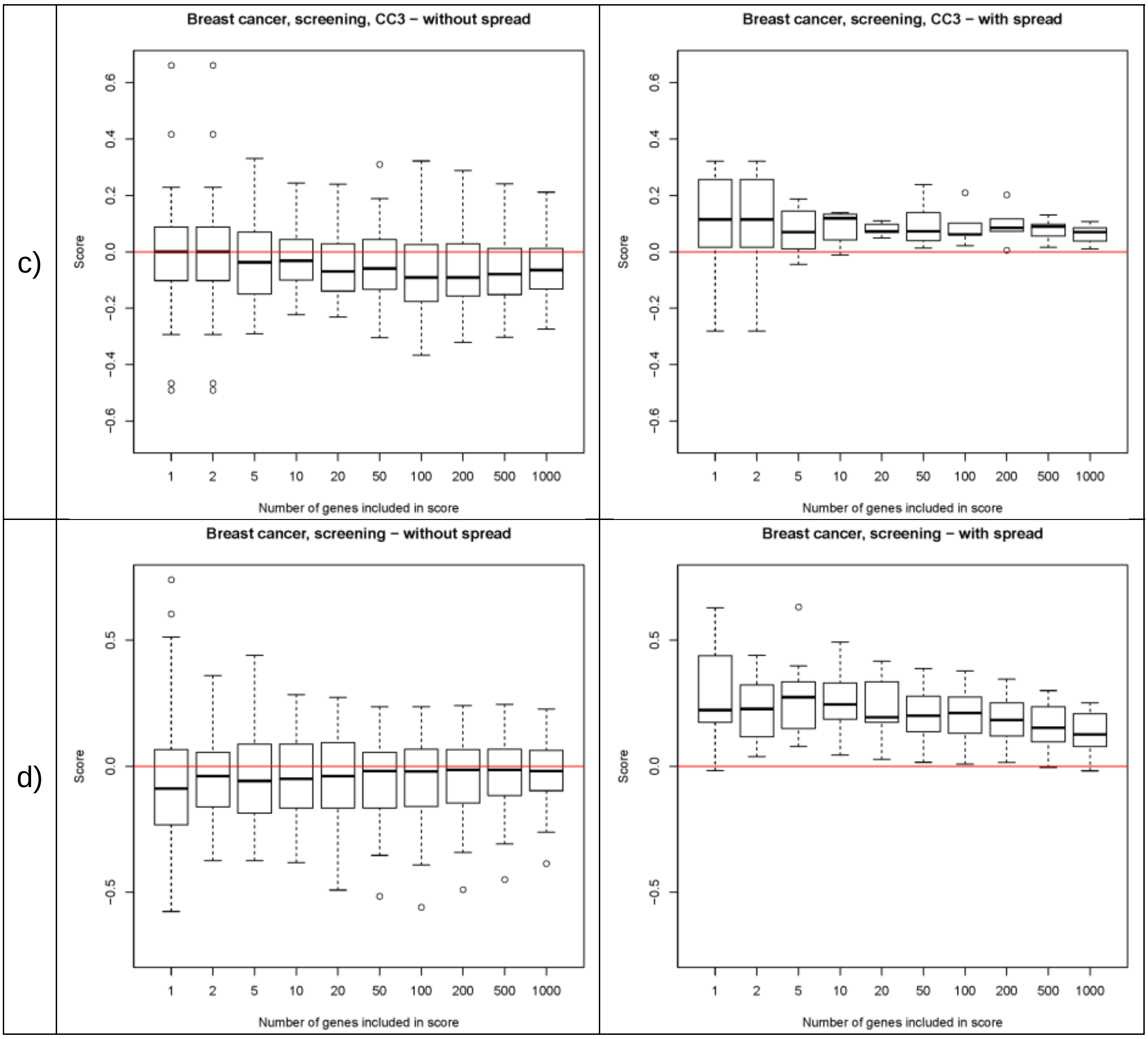


Figure 10 *Boxplots illustrating how the score used in the predictor depend on the number of genes included in the score. Note that the score has been normalized by dividing with the number of genes included in the score. Note also that the score for the cases with spread should be positive, while the scores for the cases without spread should be negative. a) Scores for case control around 4 years and 6 months from the ovarian cancer dataset. The genes are selected based on the same data. b) Scores for case control pairs around 2 years and 6 months from the breast cancer dataset, clinical group. The genes included in the score are selected based on the same data. c) Scores for case control pairs from the CC3 validation dataset, breast cancer, screening group. The genes are selected based on the data from the breast cancer dataset, screening group around 6 months before diagnosis. d) Scores for case control pairs around 6 months from the breast cancer dataset, screening group. The genes included in the score are selected based on the same data.*

6 References

[1] Eiliv Lund, Lars Holden, Hege Bøvelstad, Sandra Plancade, Nicolle Mode, Clara-Cecilie Günther, Gregory Nuel, Jean-Christophe Thalabard and Marit Holden. A new statistical method for curve group analysis of longitudinal gene expression data illustrated for breast cancer in the NOWAC postgenome cohort as a proof of principle. Accepted for publication in BMC Medical Research Methodology, 2016.

[2] Lars Holden. Time development of gene expression. NR note SAMBA/35/15, 2015.

[3] Lin SM, Du P, Huber W, et al. Model-based variance-stabilizing transformation for Illuminamicroarray data. Nucleic Acids Res 2008;36:e11.

[4] Marit Holden, Clara-Cecilie Günther and Lars Holden. Verification of a blood-based test for breast cancer (BLOBREC): Distinguishing breast-cancer patients from population-based controls. NR note SAMBA/33/15, 2015.

[5] Marit Holden and Lars Holden. Statistical analysis of gene expression in blood before diagnosis of breast cancer. NR note SAMBA/07/16, 2016.

[6] Clara-Cecilie Günther, Marit Holden, Lars Holden. Preprocessing of gene-expression data related to breast cancer diagnosis. NR note SAMBA/35/14, 2014.

7 Appendix – Details about the data

Computation of number of days to diagnosis: Number of days to diagnosis is computed as the «DIAGNOSEDATO1» for the case minus the «Nedfrysing_dato» for the case. For one of the cases the «DIAGNOSEDATO1» was before the «Nedfrysing_dato». For this case we set the number of days to diagnosis equal to one.

We define the following strata:

- **Borderline** consists of the 19 case-control pairs with «BORDERLINE_OVARIE1» equal to 1 for the cases (and METASTASE1=NA, 9 or 0 so that the pair with a case with METASTASE1=4 and BORDERLINE_OVARIE1=1 is not included in the borderline stratum).
- **Without spread** consists of the 9 case-control pairs with «METASTASE1» equal to 0 and «BORDERLINE_OVARIE1» equal to NA for the cases.
- **With spread** consists of the 59 case-control pairs with «METASTASE1» larger than 0 and different from 9 (and «BORDERLINE_OVARIE1» equal to NA for the cases so that the pair with a case with METASTASE1=4 and BORDERLINE_OVARIE1=1 is not included in the with spread stratum).

Using these definitions, 3 out of 90 case-control pairs are not included in any stratum; one because METASTASE1=4 and BORDERLINE_OVARIE1=1, and two because METASTASE1=9 and BORDERLINE_OVARIE1=NA. The table below shows how the 87 case-control pairs are distributed over the three strata and seven years before diagnosis:

	Year before diagnosis	7	6	5	4	3	2	1	Sum
Stratum	Borderline	0	1	2	4	4	7	1	**19**
	Without spread	0	1	2	2	2	0	2	**9**
	With spread	5	5	11	7	13	9	9	**59**
	Sum	**5**	**7**	**15**	**13**	**19**	**16**	**12**	**87**

As there are few case-control pairs in the borderline and without spread strata, and these strata are expected to have quite similar gene expression profiles, we include the borderline stratum in the without spread stratum.

8 Appendix – Previous analyses

8.1 Comparing cases and controls

We used the Bioconductor R-package Limma (Linear models for microarrays) for identifying genes that are differentially expressed between cases and controls. No differentially expressed genes were found for any of the five examined datasets. In each of the five analyses information about time to diagnosis was used only when selecting the dataset for the analysis.

The 10 most significantly differentially expressed genes – cases from year 1-7 before diagnosis							
With spread				Without spread			
Gene	logFC	p-value	FDR q-val.	Gene	logFC	p-value	FDR q-val.
KCTD12	-0.26	0.0001	0.4	HVCN1	0.11	0.0005	1
LOC100128269	-0.12	0.0001	0.4	TMEM97	-0.08	0.0010	1
CYSLTR1	-0.15	0.0001	0.4	ZNF638	0.06	0.0010	1
LOC100131253	-0.11	0.0002	0.4	UBAP2	-0.06	0.0012	1
CD93	-0.21	0.0002	0.4	LOC728758	-0.06	0.0020	1
NR4A2	-0.07	0.0003	0.4	DAAM1	-0.07	0.0029	1
TMEM154	-0.21	0.0004	0.5	LOC100128775	-0.25	0.0034	1
LOC440043	-0.07	0.0005	0.5	LOC100131253	-0.13	0.0036	1
RPL8	0.08	0.0005	0.5	MACF1	-0.08	0.0051	1
TAOK1	-0.20	0.0007	0.6	ZFYVE27	-0.06	0.0053	1
The 10 most significantly differentially expressed genes – cases from year 1-2 before diagnosis							
With spread				Without spread			
Gene	logFC	p-value	FDR q-val.	Gene	logFC	p-value	FDR q-val.
ABCA1	-0.29	0.0019	1	LOC144438	-0.19	0.0013	1
PLCXD1	0.12	0.0035	1	PKN2	-0.23	0.0021	1
RORC	0.07	0.0056	1	FKBP1A	0.11	0.0036	1
LOC649143	0.9	0.0058	1	FAM101B	-0.30	0.0037	1
KCTD12	-0.33	0.0068	1	ZRSR2	0.13	0.0043	1
ZDHHC11	0.09	0.0072	1	SMG7	-0.20	0.0048	1
PLA2G7	-0.10	0.0092	1	TADA2B	-0.15	0.0062	1
MAPRE2	0.10	0.0096	1	ODF3B	0.13	0.0074	1
FAM73B	0.08	0.0119	1	LOC221442	-0.11	0.0077	1
TCTN1	-0.08	0.0127	1	MUT	-0.12	0.0090	1
The 10 most significantly differentially expressed genes – cases from year 1 before diagnosis							
With spread				Without spread: Small dataset - not analyzed			
Gene	logFC	p-value	FDR q-val.				
LOC642817	0.27	0.0003	1				
TMEM154	-0.36	0.0022	1				
PKP4	0.15	0.0026	1				
CCDC90A	0.11	0.0036	1				
SNHG5	0.69	0.0038	1				
RAB6B	0.11	0.0054	1				
GSTT1	-0.25	0.0055	1				
C5orf4	0.54	0.0064	1				
LOC100131967	-0.17	0.0069	1				
PHCA	-0.29	0.0069	1				

8.2 Comparing cases with and without spread

When using Limma analysis for identifying genes that are differentially expressed between cases with and without spread, no differentially expressed genes were found. In each of these two analyses information about time to diagnosis was used only when selecting the dataset for the analysis.

The 10 most significantly differentially expressed genes							
Cases from year 1-7 before diagnosis				Cases from year 1-2 before diagnosis			
Gene	logFC	p-value	FDR q-val.	Gene	logFC	p-value	FDR q-val.
HVCN1	-0.17	0.0001	0.8	KIF5C	0.12	0.0013	1
LOC730324	0.16	0.0002	1	RAD1	0.13	0.0042	1
TSHZ3	-0.14	0.0010	1	C15orf57	-0.17	0.0048	1
SCPEP1	-0.19	0.0011	1	WASH3P	0.12	0.0049	1
ZNF638	-0.06	0.0018	1	ABHD6	-0.10	0.0050	1
RILPL2	-0.17	0.0020	1	ZDHHC11	0.13	0.0051	1
NUP188	0.06	0.0022	1	SMG7	0.27	0.0061	1
UBAP2	0.07	0.0024	1	RORC	0.10	0.0062	1
MCCC1	0.09	0.0025	1	LOC643336	-0.45	0.0081	1
OAF	-0.17	0.0029	1	BTD	0.10	0.0087	1

8.3 Curve group analysis

We analyzed the data using an approach based on curve groups [1] where information about time to diagnosis is included in the analysis. This approach was used both for comparing the different strata (with and without spread) and for testing whether there is a development in time for any of the strata. No significant results were obtained. More detailed results are given below. Note that in these analyses we used a slightly different method for preprocessing the data (see [6]) than the one described in Section 2.

For the curve group analyses we divided year 1-7 into three time periods using the following two partitions:

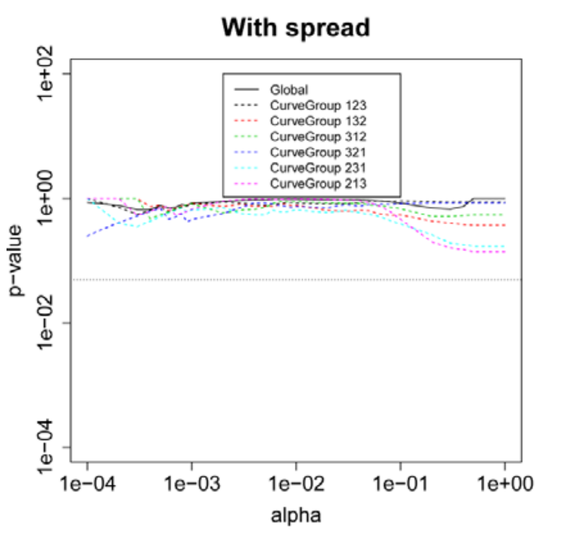
Year before diagnosis (time period)	3-5(3)	2(2)	1(1)		5-7(3)	3-4(2)	1-2(1)	
	Partition 1				Partition 2			
Stratum				Sum				Sum
Without spread	16	7	3	**26**	6	12	10	**28**
With spread	31	9	9	**49**	21	20	18	**59**
Sum	**47**	**16**	**12**	**75**	**27**	**32**	**28**	**87**

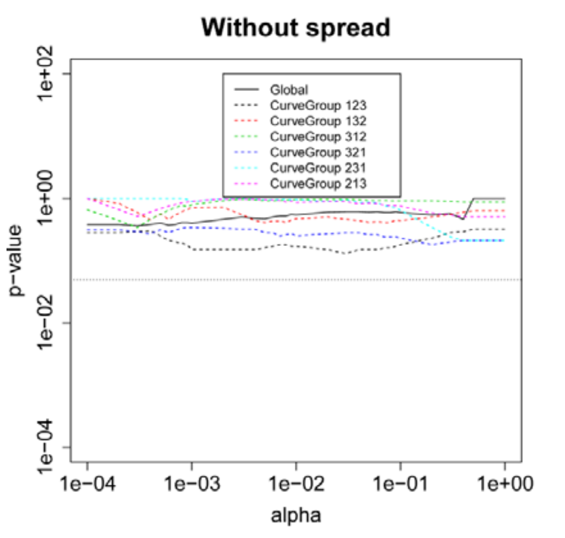
When preprocessing the data used for partition 1, we kept probes that where present for at least 2% of the individuals, i.e. at least 3 of the 75 x 2 = 150 individuals. This resulted in a dataset with 11337 genes (17213 probes after filtering, 34438 probes before filtering). When preprocessing the data used for partition 2, we kept probes that where present for at least 5% of the individuals, i.e. at least 9 of the 87 x 2 = 174 individuals. This resulted in a dataset with 10375 genes (15296 probes after filtering, 34438 probes before filtering). See [1] for a description of the method for curve group analysis.

Testing for development in time for each stratum

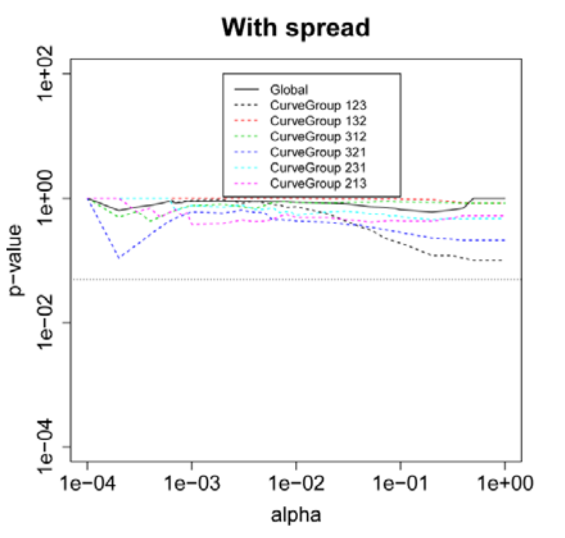
P-values obtained when testing whether there are more genes in a curve group than expected

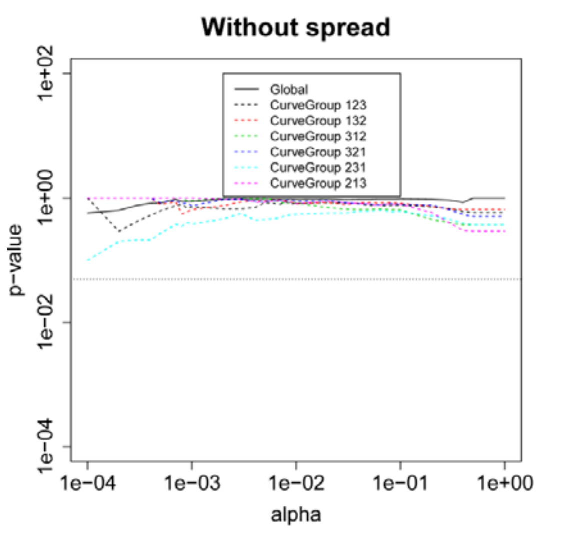
Partition 1





Partition 2





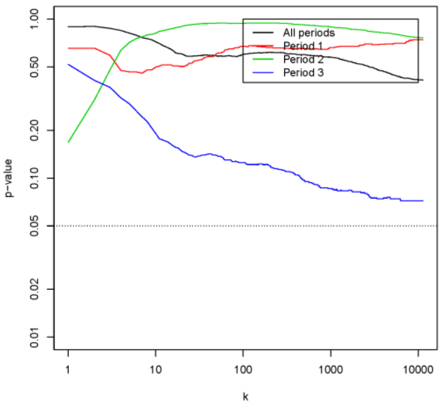
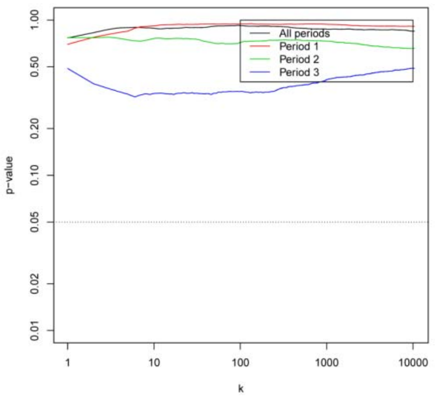
	p-values obtained when testing whether there are more genes in the curve groups than what is expected by chance			
	Partition 1		Partition 2	
Curve group	With spread	Without spread	With spread	Without spread
Global	0.97	0.54	0.86	0.96
123	0.84	0.17	0.73	0.82
132	0.78	0.47	1.00	0.80
312	0.89	0.96	0.86	0.83
321	0.74	0.25	0.43	0.89
231	0.65	0.96	0.53	0.54
213	0.99	0.86	0.49	0.81

	Number of genes in each curve group (expected number of genes)			
	Partition 1		Partition 2	
Curve group	With spread	Without spread	With spread	Without spread
Global	180 (524)	507 (626)	212 (497)	196 (446)
123	29 (81)	187 (140)	30 (83)	32 (76)
132	36 (99)	84 (112)	11 (82)	28 (68)
312	29 (100)	40 (109)	23 (88)	26 (66)
321	33 (81)	146 (131)	54 (83)	28 (79)
231	38 (83)	22 (65)	45 (79)	50 (80)
213	15 (81)	28 (69)	49 (81)	32 (77)

Comparing cases with and without spread

P-values obtained when testing whether the variables $Z_{p,c,s}$ are different in the two strata												
	Partition 1 - genes selected based on stratum						Partition 2 - genes selected based on stratum					
	with spread			without spread			with spread			without spread		
Period t	3	2	1	3	2	1	3	2	1	3	2	1
N1	21	20	18	21	20	18	21	20	18	21	20	18
N2	6	12	10	6	12	10	6	12	10	6	12	10
Curve group c												
123	0.35	0.97	0.82	0.10	0.62	0.77	0.26	0.20	0.74	0.17	0.32	0.73
132	0.68	0.34	0.39	0.43	0.74	0.55	0.90	0.76	0.92	0.31	0.39	0.60
312	0.17	0.67	0.39	0.83	0.64	0.81	0.51	0.47	0.72	0.18	0.18	0.34
321	0.49	0.87	0.69	0.11	0.57	0.84	0.27	0.65	0.70	0.52	0.44	0.97
231	0.20	0.78	0.61	0.31	0.80	0.83	0.40	0.85	0.94	0.60	0.96	0.61
213	0.09	0.64	0.77	0.26	0.70	0.92	0.27	0.91	0.99	0.32	0.81	0.50

N1 is the number of case-control pairs in the stratum «With spread» in the time period, while N2 is the number of case-control pairs in the stratum «Without spread» in the time period. See [1] for a definition and explanation of the variables $Z_{p,c,s}$ and the statistic L_k.

The p-value in a hypothesis test with test statistic L_k, plotted against the number of genes k used in the calculation of L_k	
Partition 1	Partition 2
	

9 Appendix – Additional figures and tables

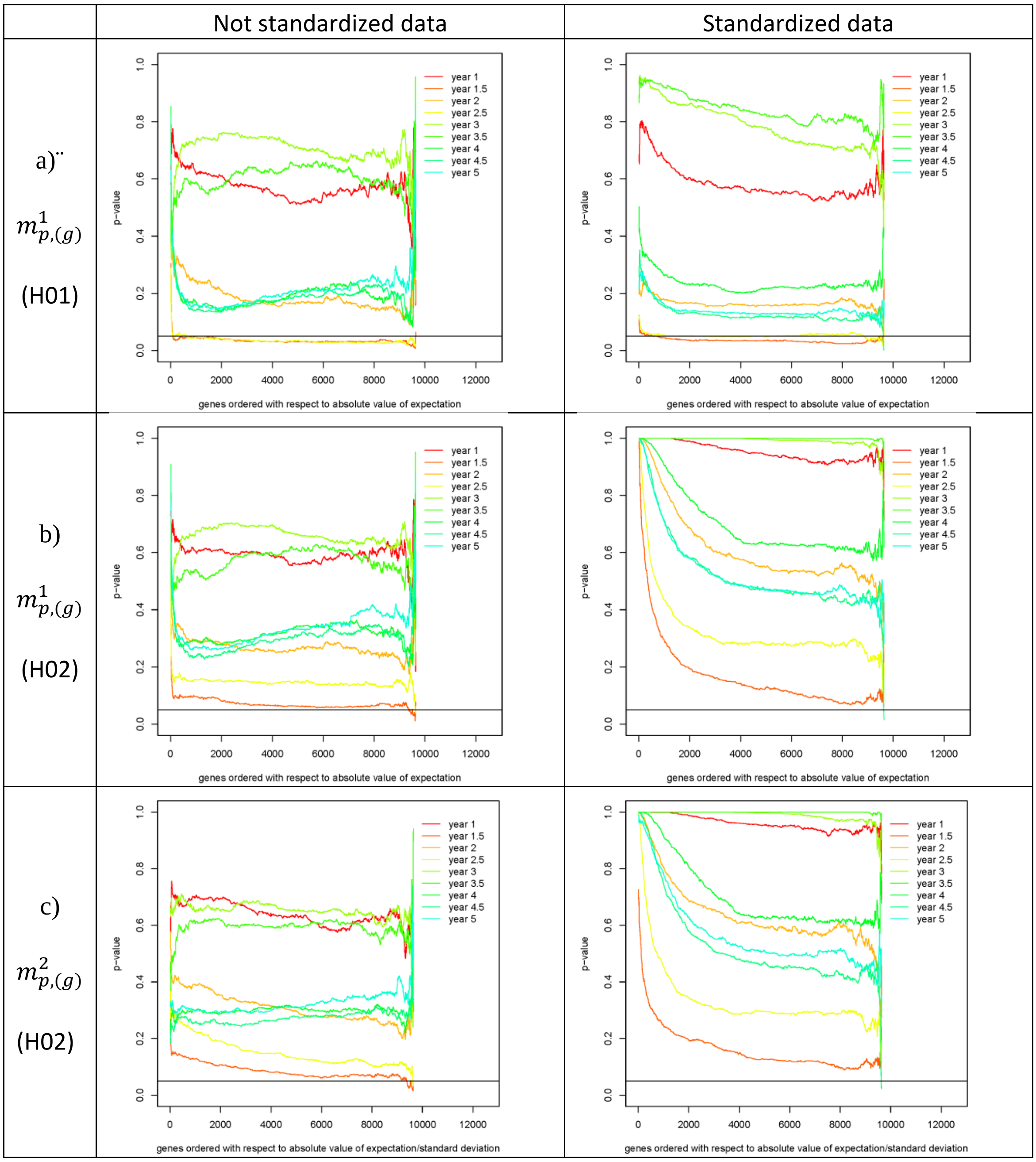
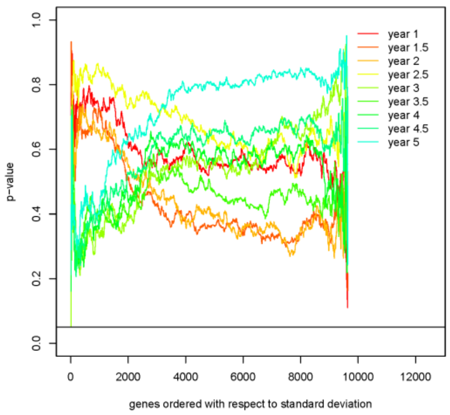
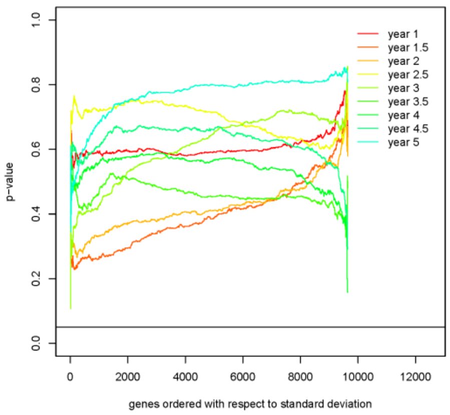
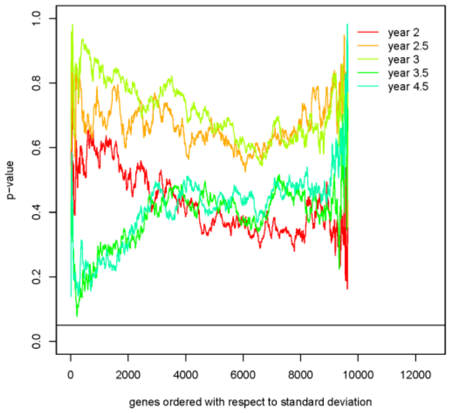
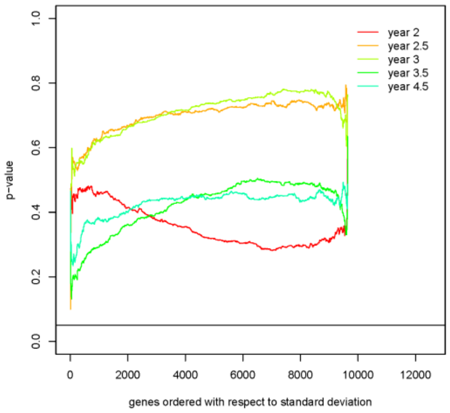


Figure 11 *Plots of p-values for three of the hypothesis tests where entire dataset is used. P-values for all genes are included in the plots. a) The hypothesis test is based on the statistic $m^1_{p,(g)}$ and the null distribution is estimated by randomizing the case-control pairs between the periods. b) The hypothesis test is based on the statistic $m^1_{p,(g)}$ and the null distribution is estimated by randomizing the case and control in each case-control pair. c) The hypothesis test is based on the statistic $m^2_{p,(g)}$ and the null distribution is estimated by randomizing the case and control in each case-control pair. In each plot there is one curve for every half year with a time period with 25 case-control pairs with spread sufficiently close. The p-value is 0.05 at the black horizontal line.*

<table>
<tr><th>$s_{p,(g)}$</th><th>Not standardized data</th><th>Standardized data</th></tr>
<tr><td>a)
25
(H01)</td><td>


</td><td>


</td></tr>
<tr><td>b)
35
(H01)</td><td>


</td><td>


</td></tr>
</table>

Figure 12 *Plots of p-values for the hypothesis tests based on the statistic* $s_{p,(g)}$ *where the entire dataset is used. The null distribution is estimated by randomizing the case-control pairs between the periods. a) Plots for periods that contain 25 case-control pairs where the case is with spread. b) Plots for periods that contain 35 case-control pairs where the case is with spread. In each plot there is one curve for every half year with a time period with 25 (35) case-control pairs with spread sufficiently close. The p-value is 0.05 at the black horizontal line.*

Table 3 *Number of correctly and wrongly classified cases from the ovarian-cancer dataset when 50 genes are included in the score. For the two-year period around year 5 before diagnosis, the p-value obtained using Fisher's test is equal to 0.12.*

Number of correctly and wrongly classified cases				
With spread		Without spread		
FN	TP	FP	TN	P-value (Fisher's test)
23	36	15	13	0.33

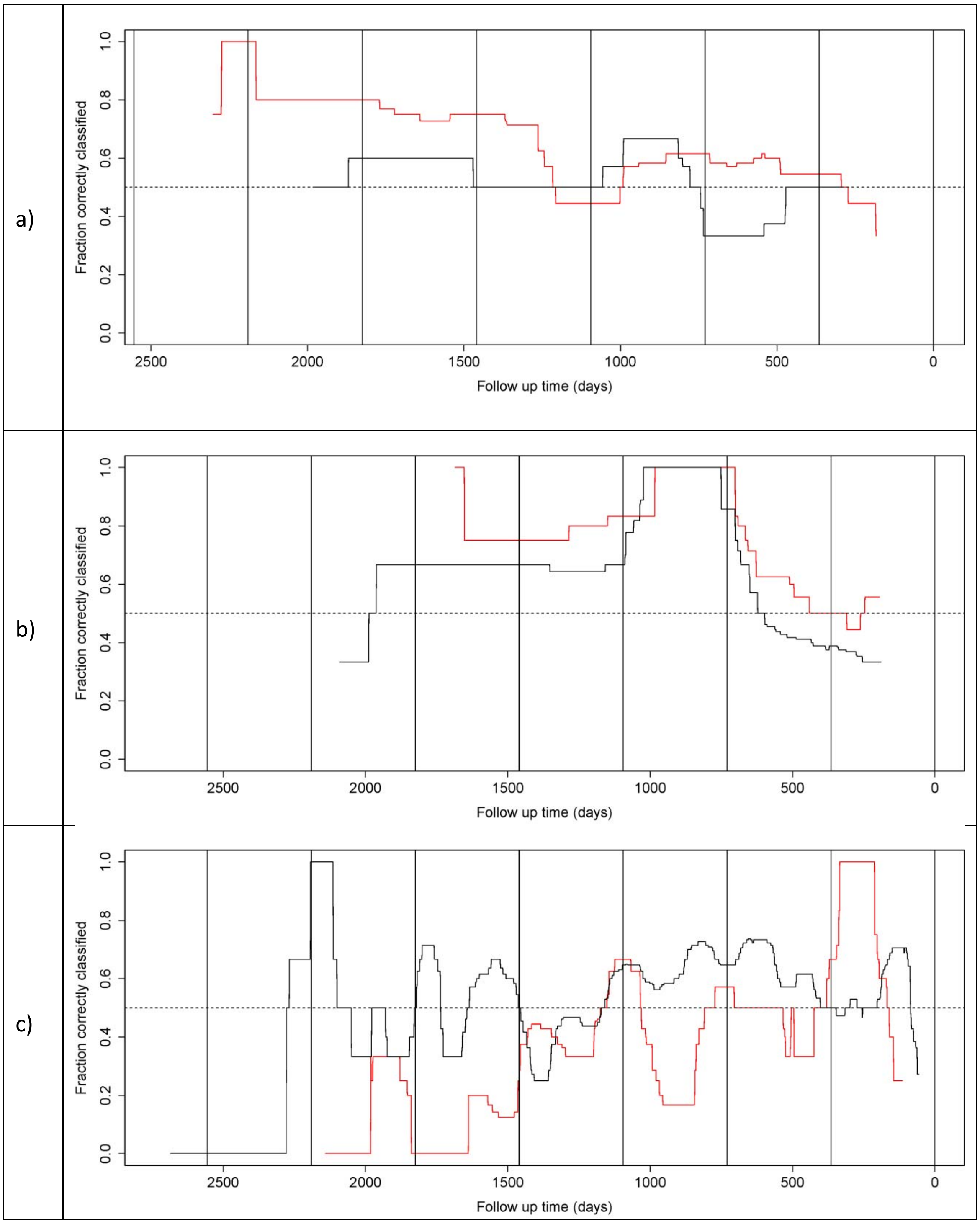


Figure 13 *Prediction when 50 genes are included in the score. a) Fraction of correctly classified cases with ovarian cancer with (red) and without (black) spread over time. b) Fraction of correctly classified cases from the clinical group with breast cancer with (red) and without (black) spread over time. c) Fraction of correctly classified cases from the screening group with breast cancer with (red) and without (black) spread over time. The fraction for each point in time is computed using a moving window of one year. The resulting curve is then smoothed using a median-filter using a window size of one year. Fifty genes are included in the scores that are used in the predictors.*

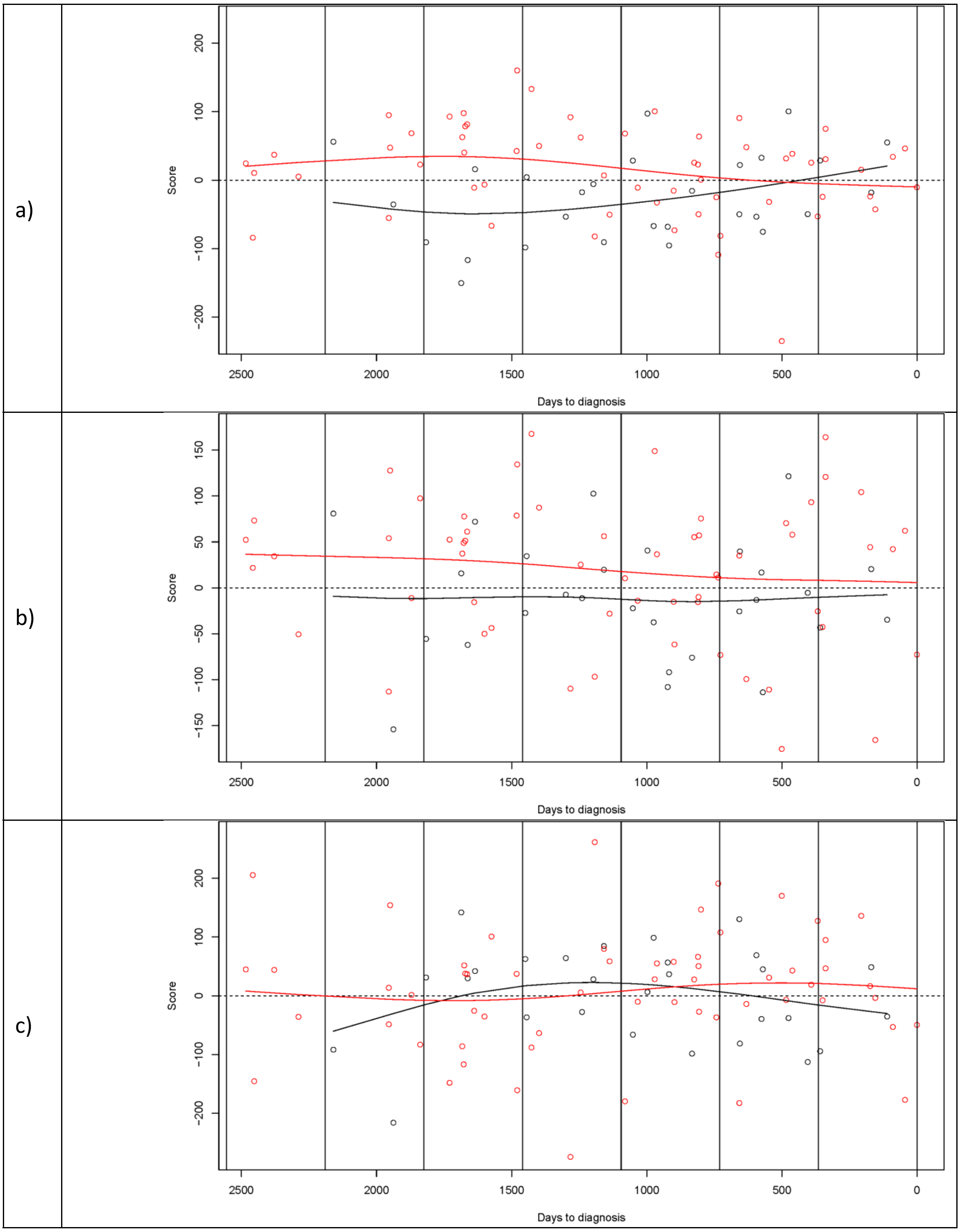


Figure 14 *Plots of scores for each case-control pair against time (days to diagnosis) for the ovarian cancer dataset. The score is plotted in red (black) if the case is with (without) spread. The score is computed using the weights of 1000 genes that are selected based on data around 4 years and 6 months (a), 2 years and 6 months (b) and 6 months (c), respectively, before diagnosis. For illustrational purposes, curves have been estimated from the scores using splines and plotted in the same color as the individual scores. a) Genes selected using the ovarian cancer dataset. b) Genes selected using the clinical group from the breast cancer dataset. c) Genes selected using the screening group from the breast cancer dataset.*

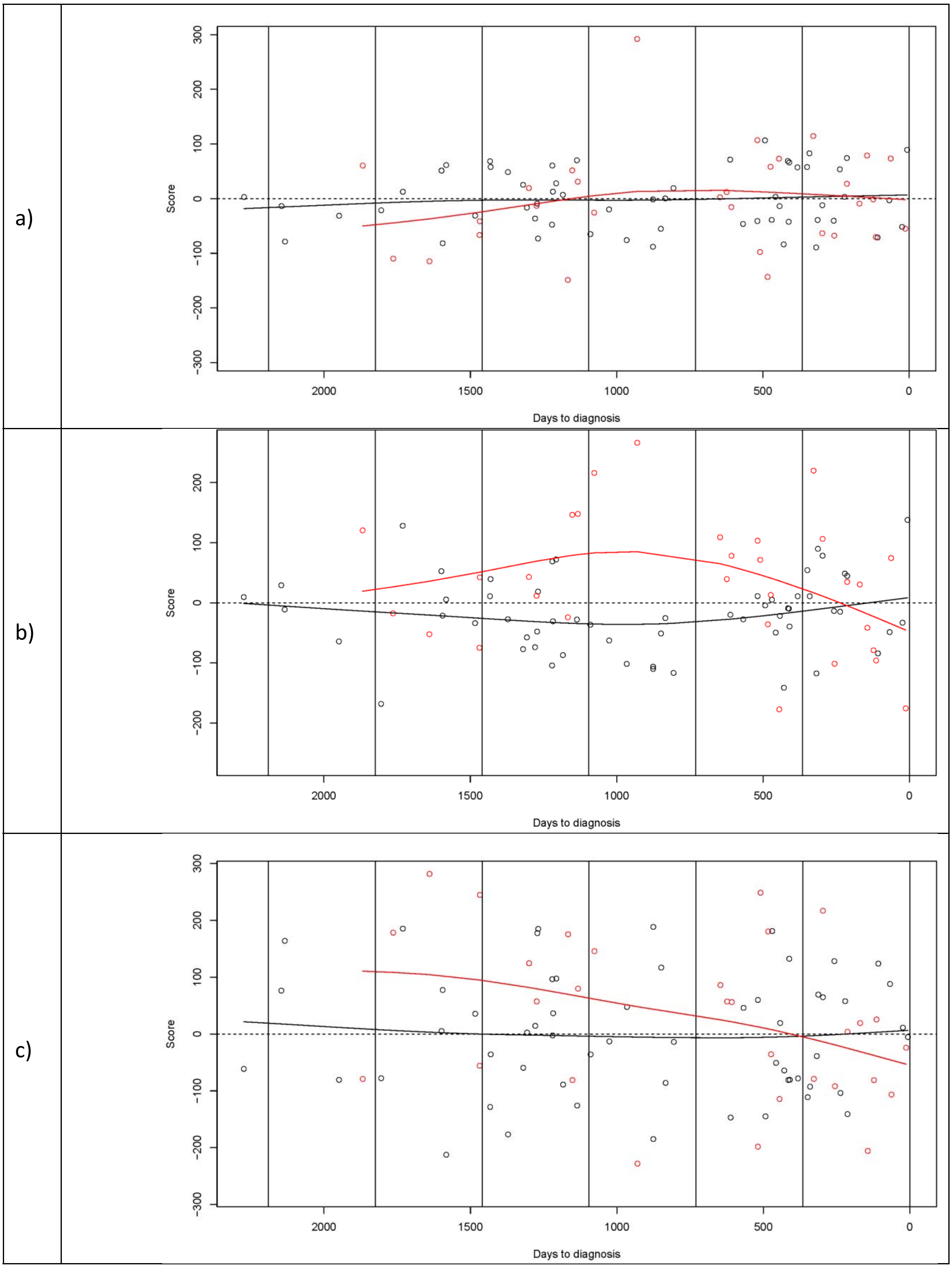


Figure 15 *Plots of scores for each case-control pair against time (days to diagnosis) for the clinical group from the breast cancer dataset. The score is plotted in red (black) if the case is with (without) spread. The score is computed using the weights of 1000 genes that are selected based on data around 4 years and 6 months (a), 2 years and 6 months (b) and 6 months (c), respectively, before diagnosis. For illustrational purposes, curves have been estimated from the scores using splines and plotted in the same color as the individual scores. a) Genes selected using the ovarian cancer dataset. b) Genes selected using the clinical group from the breast cancer dataset. c) Genes selected using the screening group from the breast cancer dataset.*

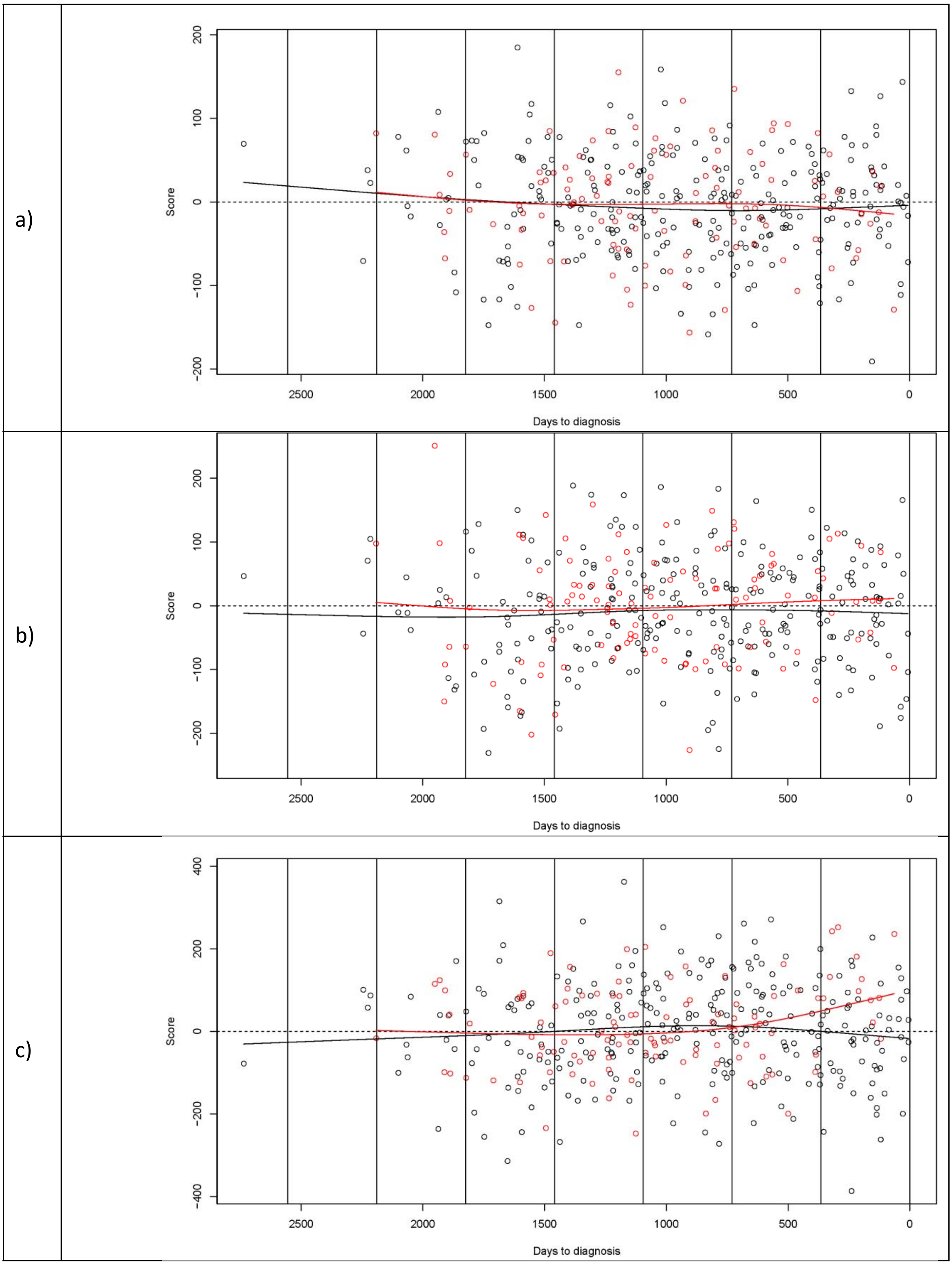


Figure 16 *Plots of scores for each case-control pair against time (days to diagnosis) for the screening group from the breast cancer dataset. The score is plotted in red (black) if the case is with (without) spread. The score is computed using the weights of 1000 genes that are selected based on data around 4 years and 6 months (a), 2 years and 6 months (b) and 6 months (c), respectively, before diagnosis. For illustrational purposes, curves have been estimated from the scores using splines and plotted in the same color as the individual scores. a) Genes selected using the ovarian cancer dataset. b) Genes selected using the clinical group from the breast cancer dataset. c) Genes selected using the screening group from the breast cancer dataset.*