Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Statistical analysis of gene expression in blood before diagnosis of breast cancer

Note

Note no.	SAMBA/07/16
Authors	Marit Holden and Lars Holden
Date	February 2016

© Copyright:
Norsk Regnesentral

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	**Statistical analysis of gene expression in blood before diagnosis of breast cancer**
Authors	**Marit Holden and Lars Holden**
Date	February 2016
Year	2016
Publication number	SAMBA/07/16

Abstract

The analyses in this note are based on a dataset with gene expression in blood before diagnosis of breast 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 breast cancer with or without spread. The dataset consists of two subdatasets, one where the cases participated in the screening program (the screening group) and one where for cases did not participate in the screening program (the clinical group). All analyses have been performed separately for these two subdatasets.

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 dataset is quite small, with only 108 (30) case-control pairs with spread and 272 (57) without spread in the screening (clinical) group, that are distributed over an eight 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. In the screening group we obtained p-value 0.5 for the entire period but 0.03 for the last year before diagnosis. For the clinical group the p-value for the entire period was 0.05. Here the results indicated best prediction 3-4 years before diagnosis. The p-value is equal 0.05 in this time period but this may be due to a small data set).

Keywords	Gene expression; Breast 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 220763
Research field	MMH, Bioinformatics
Number of pages	39
© 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 breast 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 breast cancer with or without spread.

In Section 2 we present the dataset. Methods are described in Section 3, while results are summarized in Section 4.

2 Datasets

The available dataset consists of data from 546 case-control pairs with time to diagnosis varying between 1 and 2920 days (year 1-8 before diagnosis). The dataset consists of two subdatasets, one where the cases participated in the screening program (the screening group) and one where the cases did not participate in the screening program (the clinical group). All analyses have been performed separately for these two subdatasets. 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.

In addition three validation datasets are available. These datasets are denoted CC1, CC2 and CC3, respectively, and the blood sampling was done at the time of the diagnostic biopsy, i.e. time to diagnosis is 0 days for each case-control pair in these datasets. After quality control of the data as described in [1], the three datasets CC1, CC2 and CC3 consisted of 55, 49 and 59 case-control pairs, and 39 426, 48 802 and 47 323 probes, respectively. (In [1], the number of probes were 48 803, 48 803 and 47 323, thus for CC1 we have not used the complete set of probes used in [1].) Some details about the three datasets, like the number of case-control pairs in each stratum, are given in Table 2 a). For more details, see [1].

For all case-control pairs included in the validation datasets CC1, CC2 and CC3, the cases have been diagnosed with cancer. Four small validation datasets, CC0small, CC1small, CC2small and CC3small, where the cases do not have cancer (benign tumors) are also available. Details about these datasets are given in Table 2 b).

Table 1 *Details about the available prospective dataset for the screening (upper panel) and clinical group (lower panel).*

a) Number of case-control pairs in the screening group (includes interval group that consists of cases that are diagnosed with cancer between two screening visits)										
Year before diagnosis		8	7	6	5	4	3	2	1	Sum
Stratum	With spread	0	1	6	15	30	24	20	12	**108**
	Without spread	1	3	10	36	53	59	57	53	**272**
	Insitu	0	0	0	7	19	15	9	16	**66**
	Sum	**1**	**4**	**16**	**58**	**102**	**98**	**86**	**81**	**446**

b) Number of case-control pairs in the clinical group										
Year before diagnosis		8	7	6	5	4	3	2	1	Sum
Stratum	With spread	0	0	1	4	5	2	8	10	**30**
	Without spread	0	1	3	6	14	8	12	13	**57**
	Insitu	0	0	1	1	1	3	2	5	**13**
	Sum	**0**	**1**	**5**	**11**	**20**	**13**	**22**	**28**	**100**

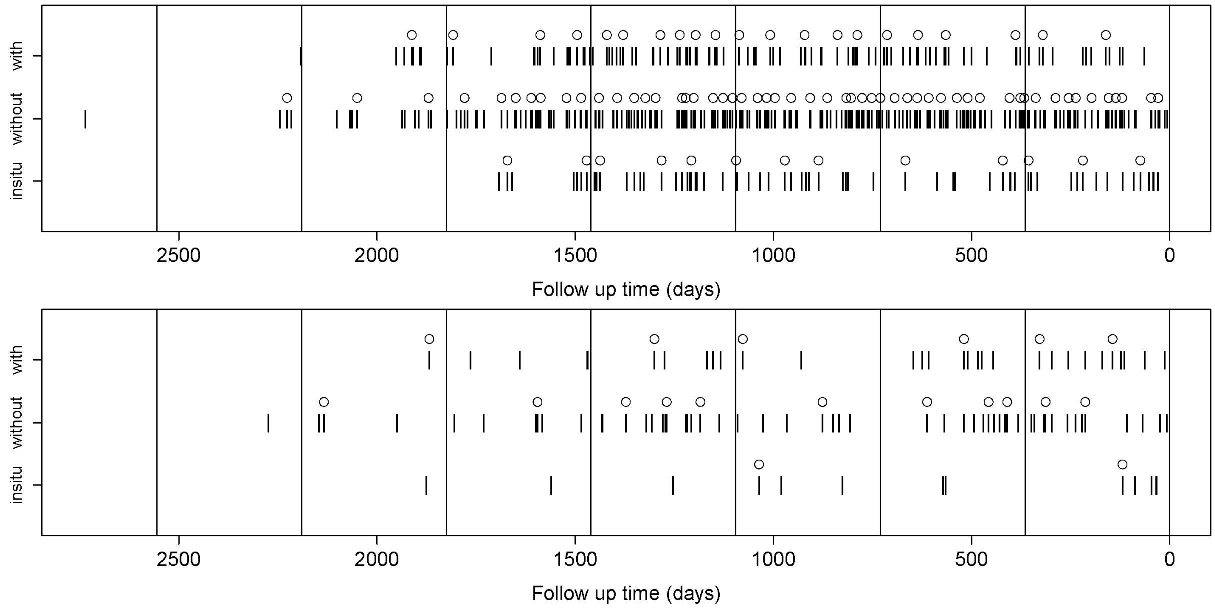


Figure 1 *The distribution of the case-controls pairs in time for the screening (upper panel) and clinical (lower panel) group. Each short vertical line represents a case-control pair. A circle is plotted above every fifth case-control pair. Long vertical lines are plotted to indicate the years. On the 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 546 x 2 = 1092 samples remain in the dataset.
3. Transform the data using the variance stabilizing technique described in [2].
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 [3].

After preprocessing the prospective dataset consists of 8155 genes. The data are from three different runs and there are batch effects between runs[1]. We therefore estimate these batch effects (see Section 10 - Appendix) and include the estimates in the methods that will be used for analyzing the data. Note that some batch effects disappear when we compute $\log_2$ differences in gene expression between cases and controls, while other batch effects do not disappear. See Section 10 (Appendix) for more details.

Table 2 *a) Details about the validation datasets CC1, CC2 and CC3 for the screening and clinical group where the cases are diagnosed with cancer. b) Details about the four small validation datasets, CC0small, CC1small, CC2small and CC3small, where the cases do not have cancer (benign tumors).*

		Stratum	Number of case-control pairs for	
			Screening group	Clinical group
a)	CC1	With Spread	8	4
		Without spread	28	11
	CC2	With Spread	7	4
		Without spread	22	11
	CC3	With Spread	5	7
		Without spread	23	18

	Number of case-control pairs in			
b)	CC0small	CC1small	CC2small	CC3small
	7	4	1	6

After preprocessing the validation datasets CC1, CC2 and CC3 consisted of 10 260, 8 430 and 9 936 genes, respectively. The CC0small, CC1small, CC2small and CC3small datasets are preprocessed so that they contain exactly the same genes as CC3, CC1, CC2 and CC3, respectively.

3 Methods

The method described here is explained in more detail in [4].

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.

[1] The obtained estimated for the batch effects are more different than expected by chance. We demonstrated this by randomizing data between the batches/runs.

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[2]:

- $\boldsymbol{s_{p,(g)}}$ 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:

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

- 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.

H02: The expectation of $X_{g,c}$ is zero.

- 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.

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

- 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 $w_{p,(g)}$.

[2] Note that the second and third statistic were not defined in [4].

- 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 a trade-off between these two wishes we have tested with some different time periods where the length depends on the number of cases without spread. For the clinical group, periods that contain 25 cases without spread seems to be reasonable both with respect to the number of case-control pairs (25 without spread, 9-17 with spread), and with respect to the length of the time periods (605-971 days, except for the four periods that include the case-control pairs in year 6 and 7 before diagnosis). For the screening group periods that contain 50 cases without spread seems to be a reasonable choice (50 case-control pairs without spread, 11-32 with spread) (256-796 days in the each time period except for the period that includes the case-control pair in year 8 before diagnosis). We have therefore selected time periods that contain 25 cases without spread for the clinical group and 50 cases without spread for the screening group. In analyses that only include cases with spread from the screening group, we select time periods that contain 50 cases with spread. When estimating $s_{p,(g)}$ we always includes at least 35 cases without spread as more data are needed to obtain reliable estimates of the standard deviation than the mean. We have defined one time period for each set of 25, 35 or 50 cases with or without spread, that are consecutive in time. The number of periods and number of case-control pairs per period for each subset of data that will be analyzed are summarized in the table below:

Subdatasets	Number of case-control pairs per period (Number of periods)		
	Finding signal in the data		Prediction
	Statistic based on standard deviation	Other statistics	
Screening with spread (108)	50 (59)	50 (59)	
Screening without spread (272)	50 (223)	50 (223)	
Screening with or without spread (380)	50 (223)	50 (223)	50 (223)
Clinical without spread (57)	35 (23)	25 (33)	
Clinical with or without spread (87)	35 (23)	25 (33)	25 (33)

Note that we include no analyses for the case “Clinical with spread” as this subdataset is too small. 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 the data to expectation zero and standard deviation one for each gene, and one without standardizing the data. Figure 2 (screening group) and Figure 3 (clinical group) show plots of the three statistics for both standardized and not standardized data, while Figure 4 shows plots of the statistic $\mathrm{w}_{\mathrm{p,(g)}}$ that does not depend on whether the dataset has been standardized. In these plots we focus on the difference between data close to and far from diagnosis. In Figure 3 we show results from data without spread and all data, instead of with and without spread as in Figure 2, since there are so few case-control pairs with spread in the clinical group.

In Figure 2 and Figure 3 we observe that the shape of the curves in the two plots with standard deviation (Figure 2 a) and Figure 3 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. Note that for both the screening and the clinical group we also observe that $s_{p,(g)}$ is larger far from diagnosis (H01). For the clinical group $s_{p,(g)}$ is larger for all cases both close and far from diagnosis (H03), while $s_{p,(g)}$ is smaller with spread close to diagnosis for the screening group. The difference between the two types of cases (with and without spread) also implies a difference between cases and controls (H02).

All four plots that are based on the expectation, $m^1_{p,(g)}$ and $m^2_{p,(g)}$, are very similar for the screening group (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 case with spread (H03) and close to diagnosis (H01) . For the clinical group, the statistic is quite similar for the two periods and the two datasets in all four plots (Figure 3 b) and c)).

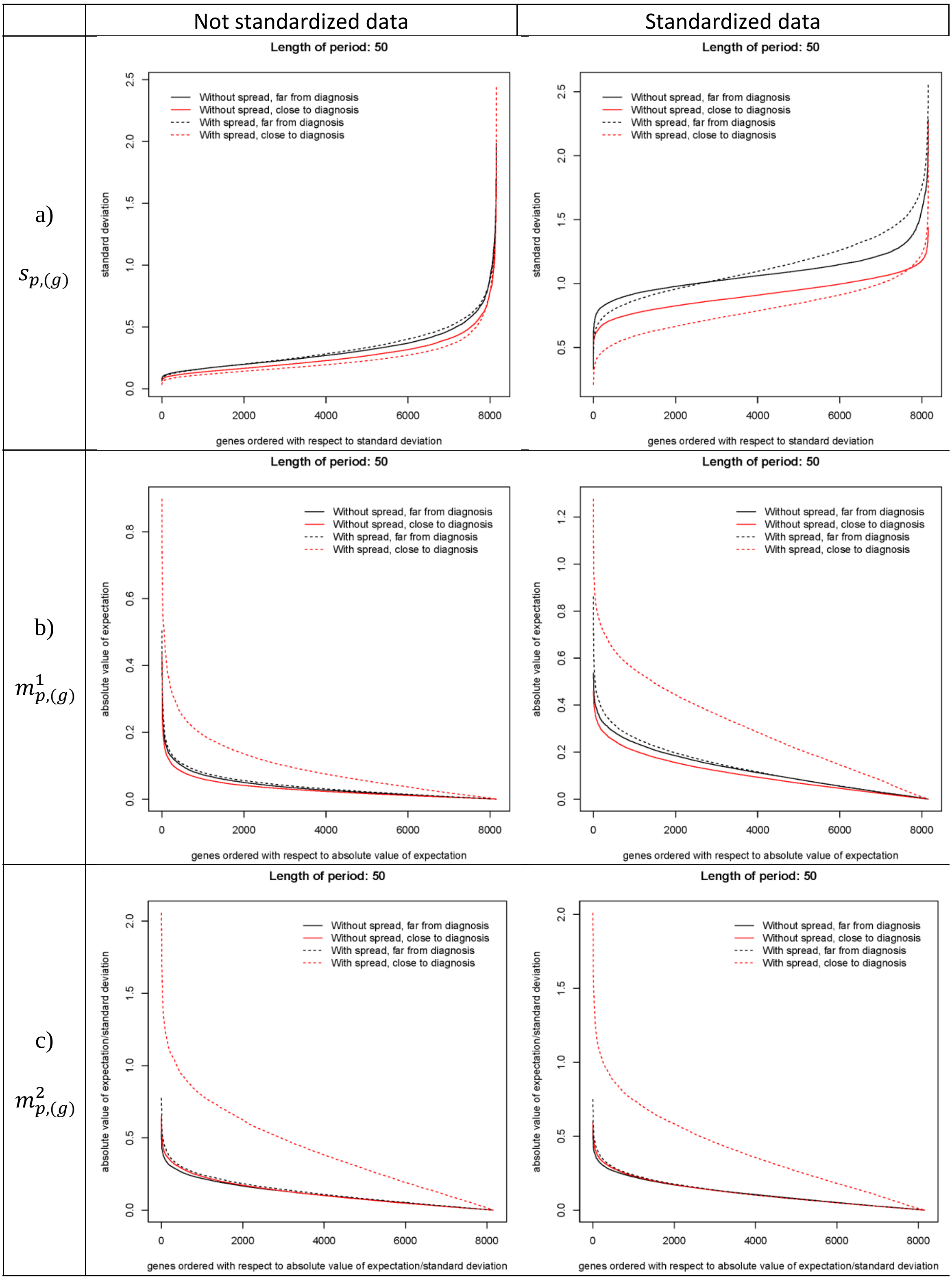


Figure 2 *Plots of the three statistics that depend on whether the dataset has been standardized for data from the screening group. The time period closest to diagnosis is 1-338 days before diagnosis (year 1), while the time period furthest from diagnosis is 1470-2736 days before diagnosis (five months of year 5, year 6, year 7, six months of year 8). The two periods contain 50 case-control pairs where the case is without spread.*

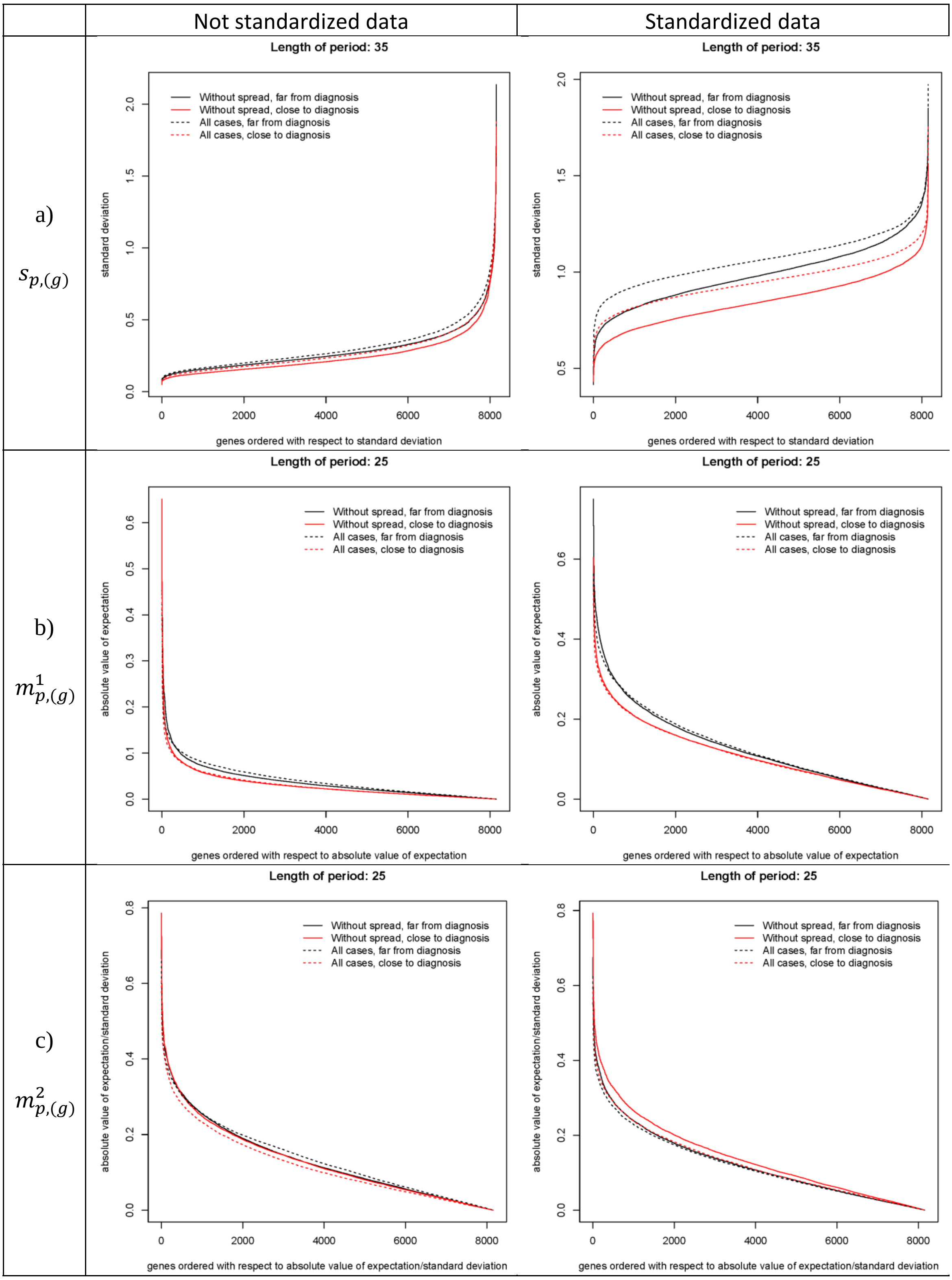


Figure 3 *Plots of the three statistics that depend on whether the dataset has been standardized for data from the clinical group. The time period closest to diagnosis is 1-612 days before diagnosis (year 1, eight months of year 2), while the time period furthest from diagnosis is 1090-2274 days before diagnosis (year 4, year 5, year 6, three months of year 7). The two periods contain 25 ($m^1_{p,(g)}$ and $m^2_{p,(g)}$) or 35 ($s_{p,(g)}$) case-control pairs where the case is without spread. All cases means cases with or without spread (i.e. not insitu).*

Figure 4 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. For the screening group, the statistic $w_{p,(g)}$ is largest close to diagnosis, while for the clinical group the difference is smaller and in the opposite direction, i.e. largest far from diagnosis (H01, H03). The difference in the screening group between without and with spread may be due to the difference in expectation shown in Figure 2 b) and c).

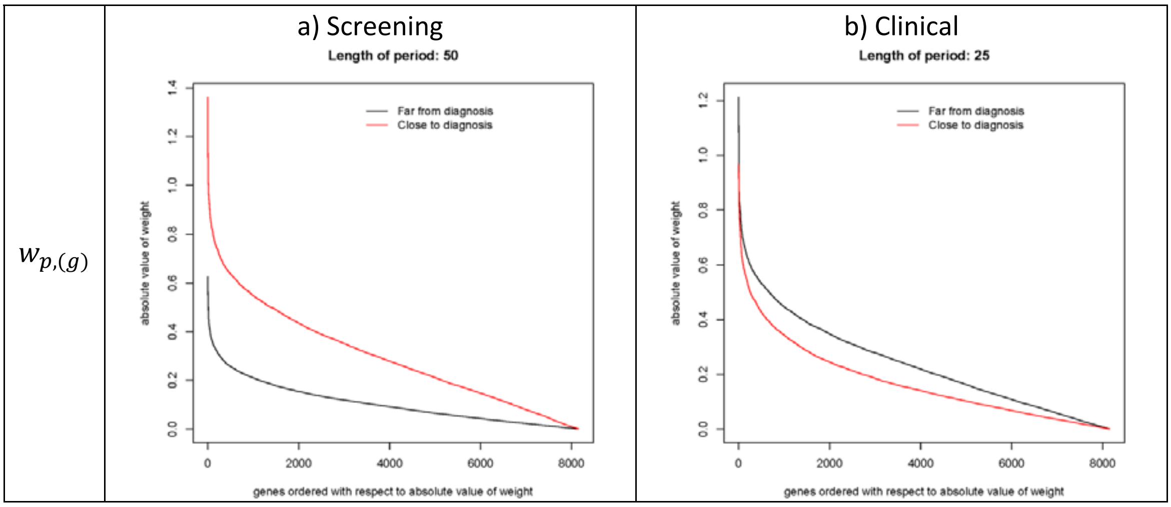


Figure 4 *Plot of the statistic $w_{p,(g)}$ that does not depend on whether the data are standardized. a) The dataset consists of case-control pairs where the case belongs to the screening group. b) The dataset consists of case-control pairs where the case belongs to the clinical group. The two periods contain 50 (a) or 25 (b) case-control pairs where the case is without spread.*

Figure 5 – Figure 10 show plots of p-values for all hypothesis tests described in Section 3.1. Note that in all the plots, the curves with p-values have been smoothed using a median-filter with window size 11.

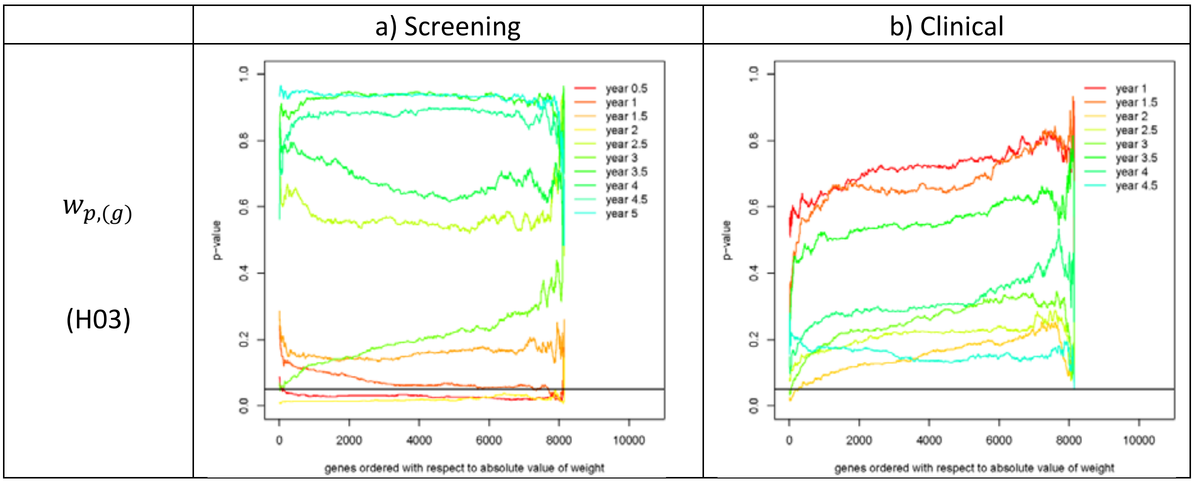


Figure 5 *Plots of p-values for the hypothesis tests based on the statistic $w_{p,(g)}$ where the expectations of the two strata in the dataset, with and without spread, are compared. The null distribution is estimated by randomizing the case-control pairs between the strata within the period. a) P-values for the screening group. b) P-values for the clinical group. In each plot there is one curve for every half year with a time period with 25 (clinical group) or 50 (screening group) case-control pairs sufficiently close. The p-value is 0.05 at the black horizontal line.*

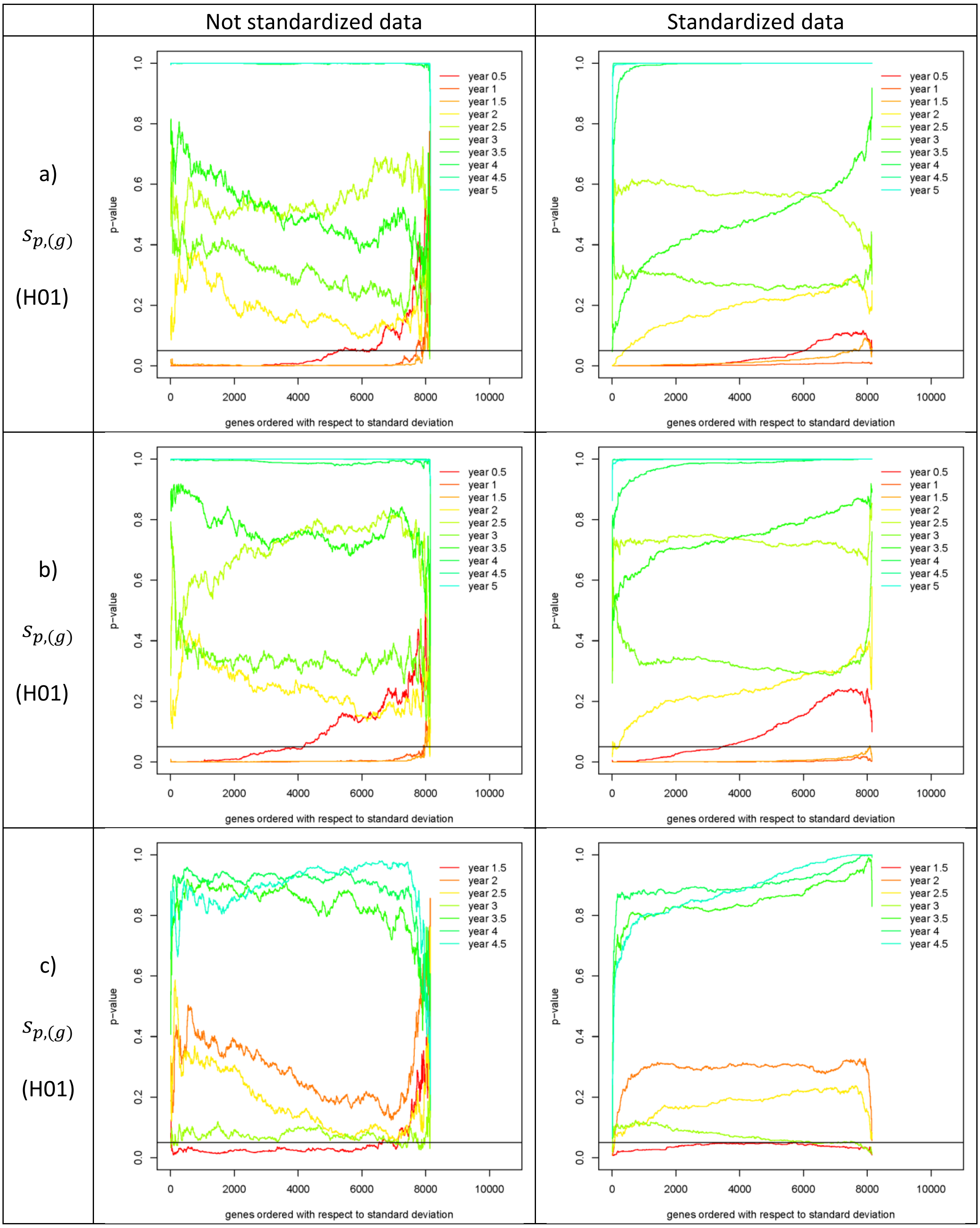


Figure 6 *Plots of p-values for the hypothesis tests based on the statistic* $s_{p,(g)}$ *where the following datasets with cases from the screening group are used: a) Cases with or without spread; b) Cases without spread; and c) Cases with spread. The null distribution is estimated by randomizing the case-control pairs between the periods. In each plot there is one curve for every half year with a time period with 50 case-control pairs sufficiently close. The p-value is 0.05 at the black horizontal line.*

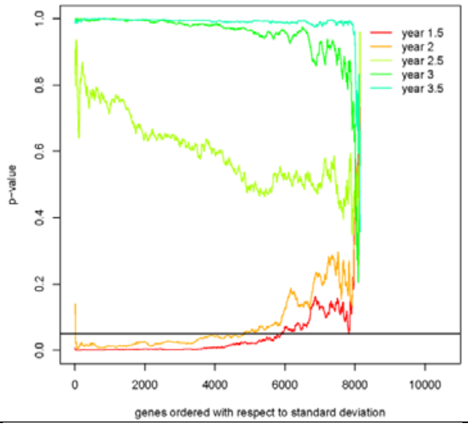
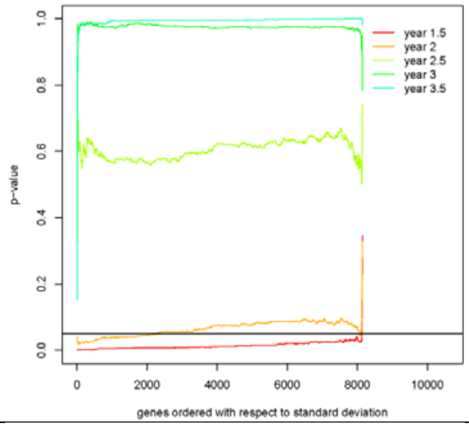
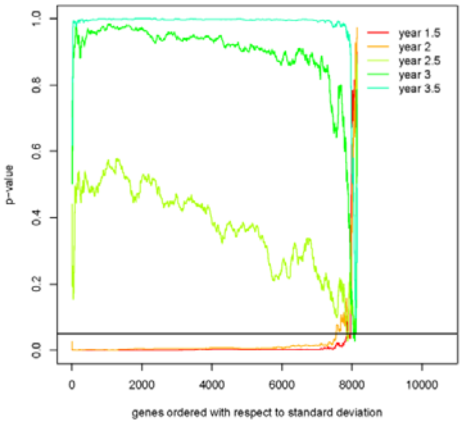
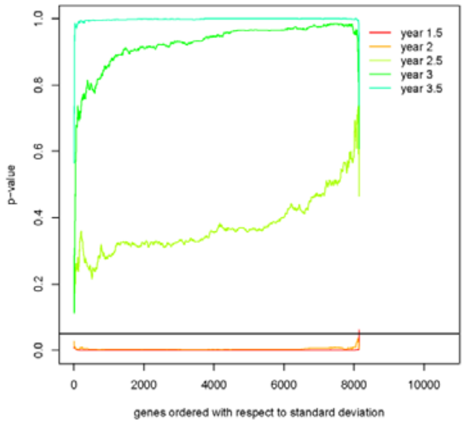
$s_{p,(g)}$	Not standardized data	Standardized data
a) (H01)		
b) (H01)		

Figure 7 *Plots of p-values for the hypothesis tests based on the statistic* $s_{p,(g)}$ *where the following datasets with cases from the clinical group are used: a) cases with or without spread; b) cases without spread. The null distribution is estimated by randomizing the case-control pairs between the periods. In each plot there is one curve for every half year with a time period with 35 case-control pairs sufficiently close. The p-value is 0.05 at the black horizontal line.*

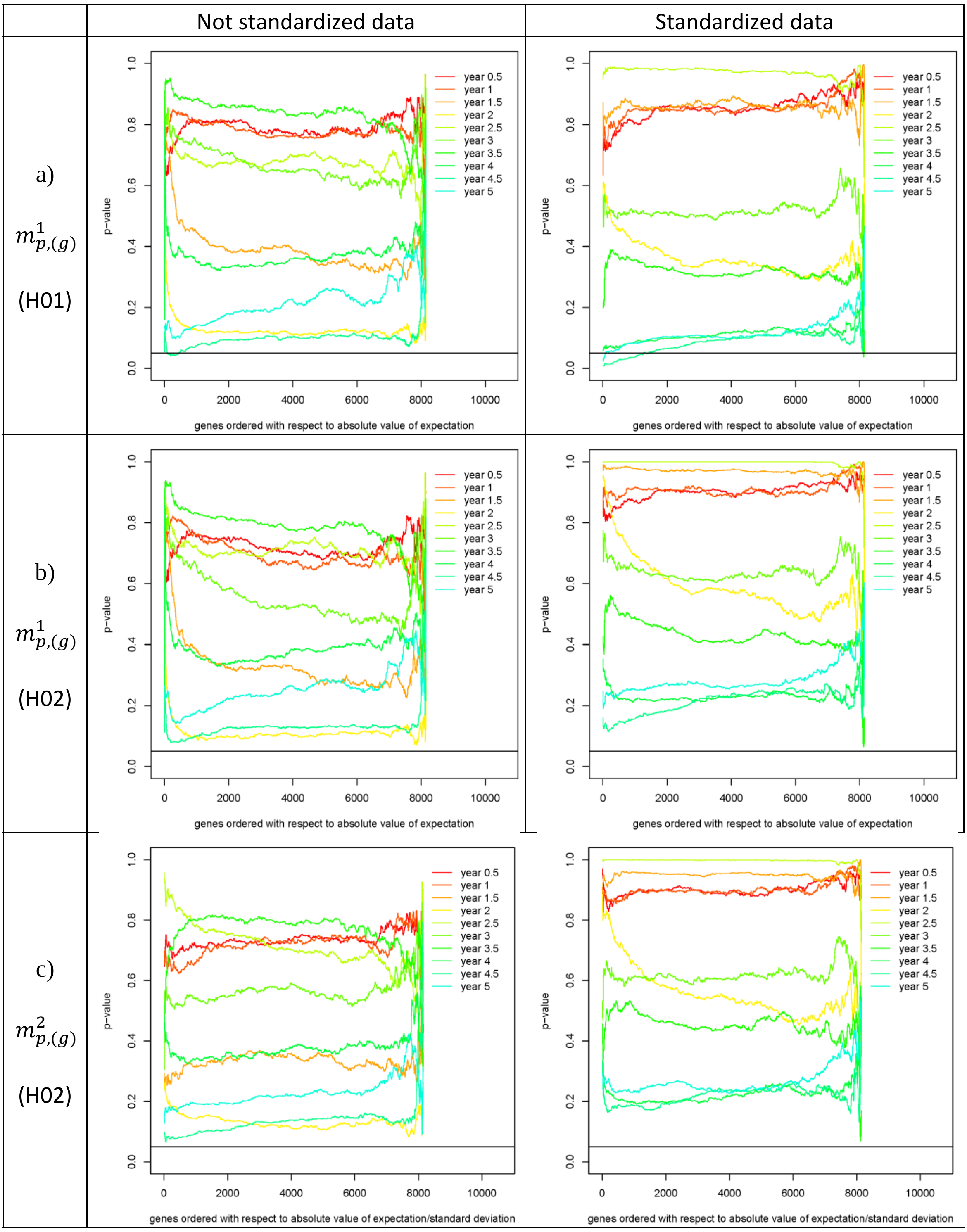


Figure 8 *Plots of p-values for three of the hypothesis tests where the dataset with cases with and without spread in the screening group is used. a) The hypothesis test is based on the statistic $m^1_{p,(g)}$ and the null distribution is estimated by randomizing the case-control pairs between the periods. 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 50 case-control pairs sufficiently close. The p-value is 0.05 at the black horizontal line.*

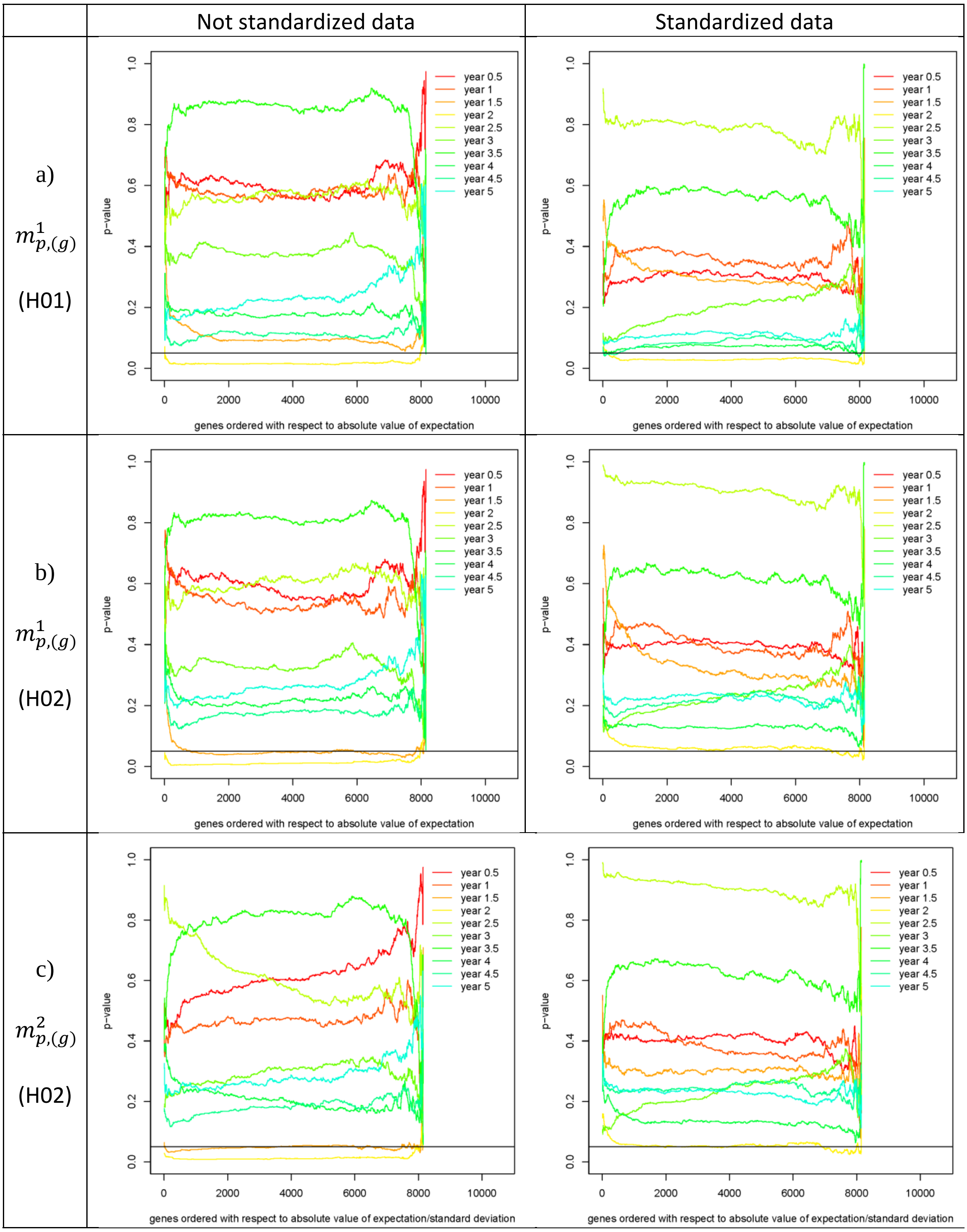


Figure 9 *Plots of p-values for three of the hypothesis tests where the dataset with cases without spread in the screening group is used. a) The hypothesis test is based on the statistic* $m^{1}_{p,(g)}$ *and the null distribution is estimated by randomizing the case-control pairs between the periods. 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 50 case-control pairs sufficiently close. The p-value is 0.05 at the black horizontal line.*

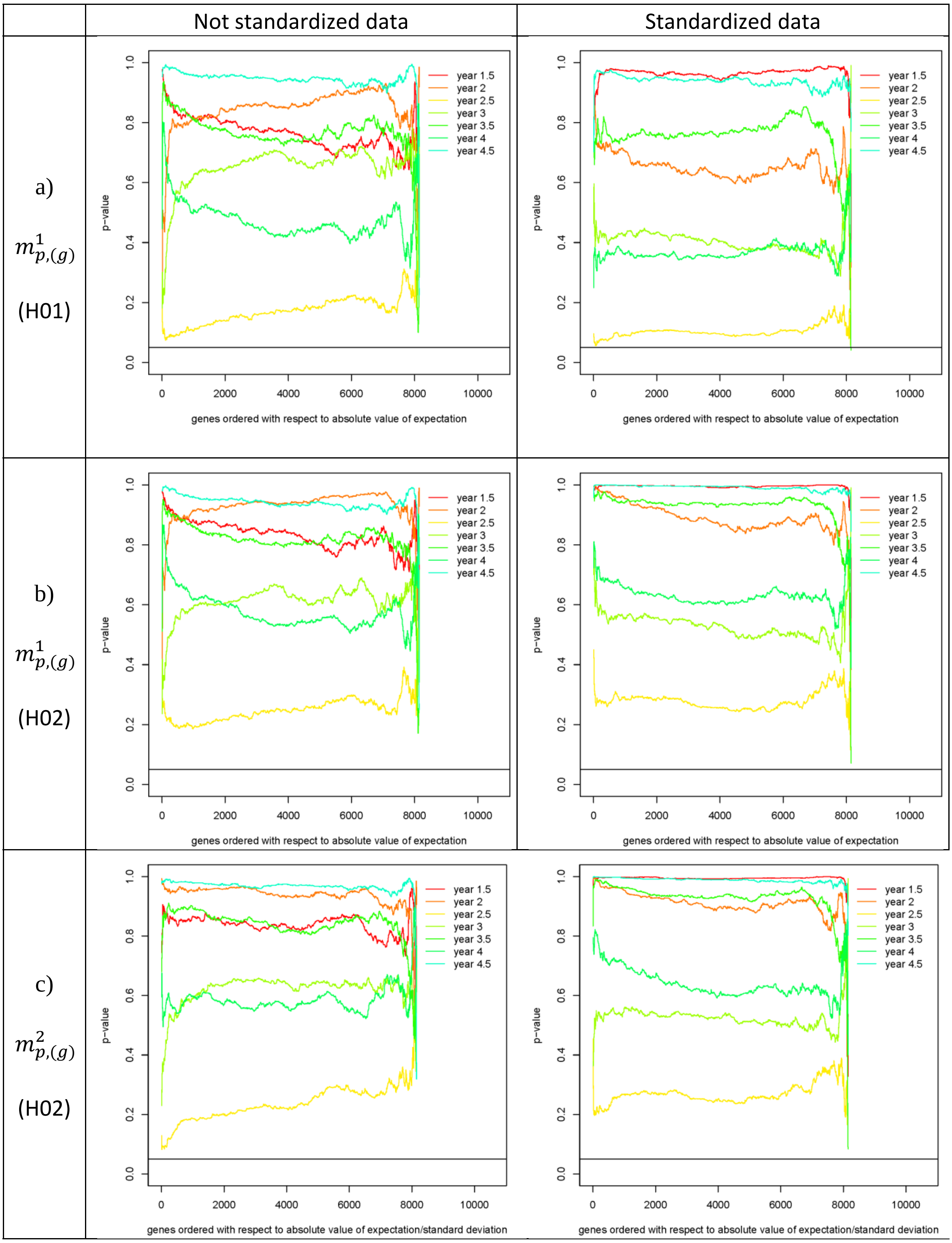


Figure 10 *Plots of p-values for three of the hypothesis tests where the dataset with cases with spread in the screening group is used. a) The hypothesis test is based on the statistic* $m^{1}_{p,(g)}$ *and the null distribution is estimated by randomizing the case-control pairs between the periods. 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 50 case-control pairs sufficiently close. The p-value is 0.05 at the black horizontal line.*

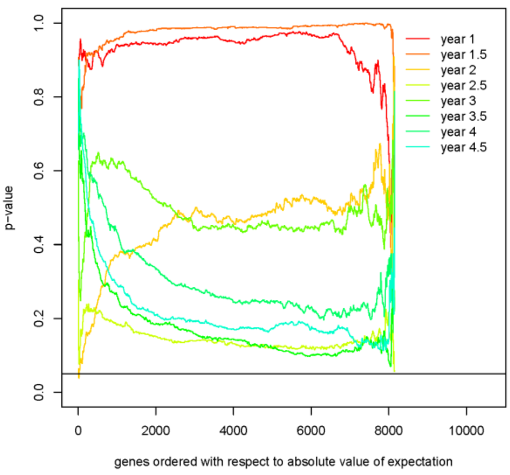
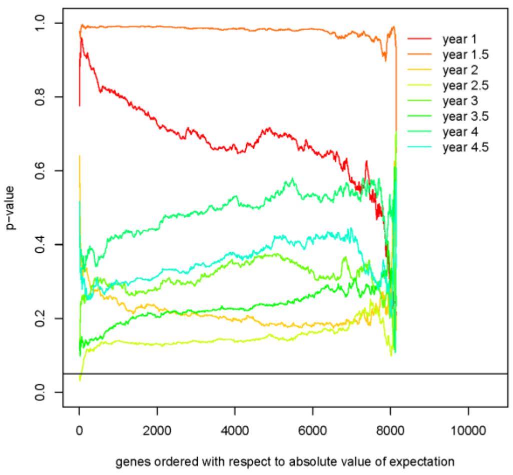
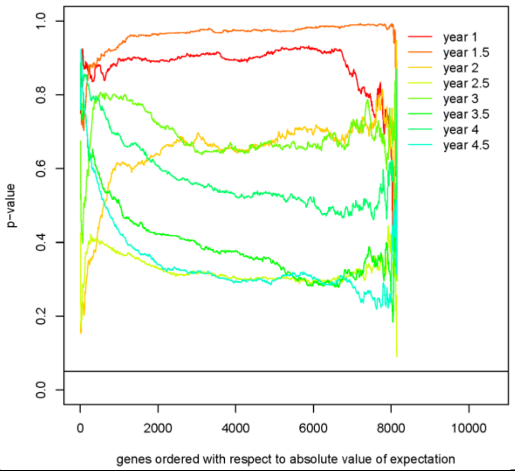
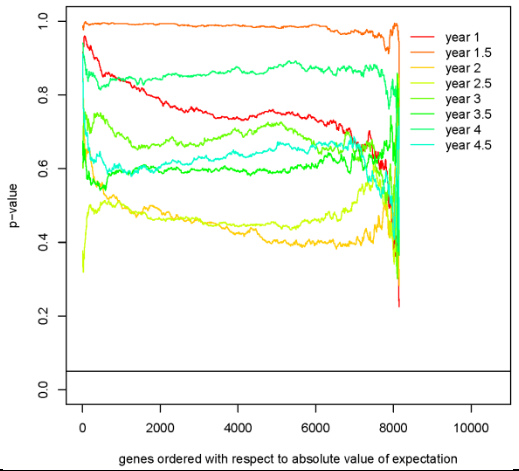
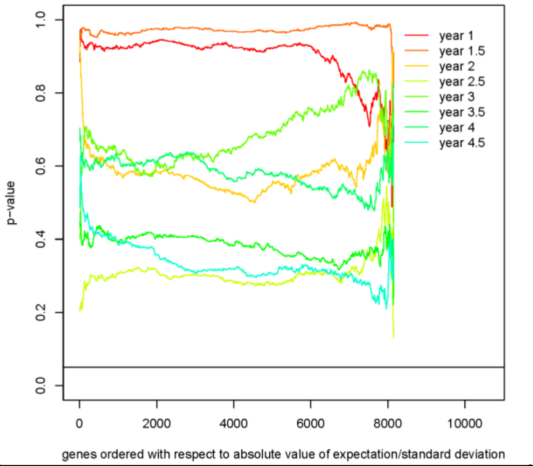
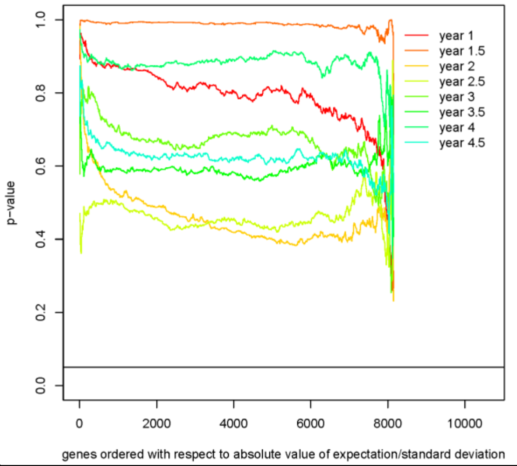
	Not standardized data	Standardized data
a) $m^1_{p,(g)}$ (H01)		
b) $m^1_{p,(g)}$ (H02)		
c) $m^2_{p,(g)}$ (H02)		

Figure 11 *Plots of p-values for three of the hypothesis tests where the dataset with cases with and without spread in the clinical group is used. 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 in 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 sufficiently close. The p-value is 0.05 at the black horizontal line.*

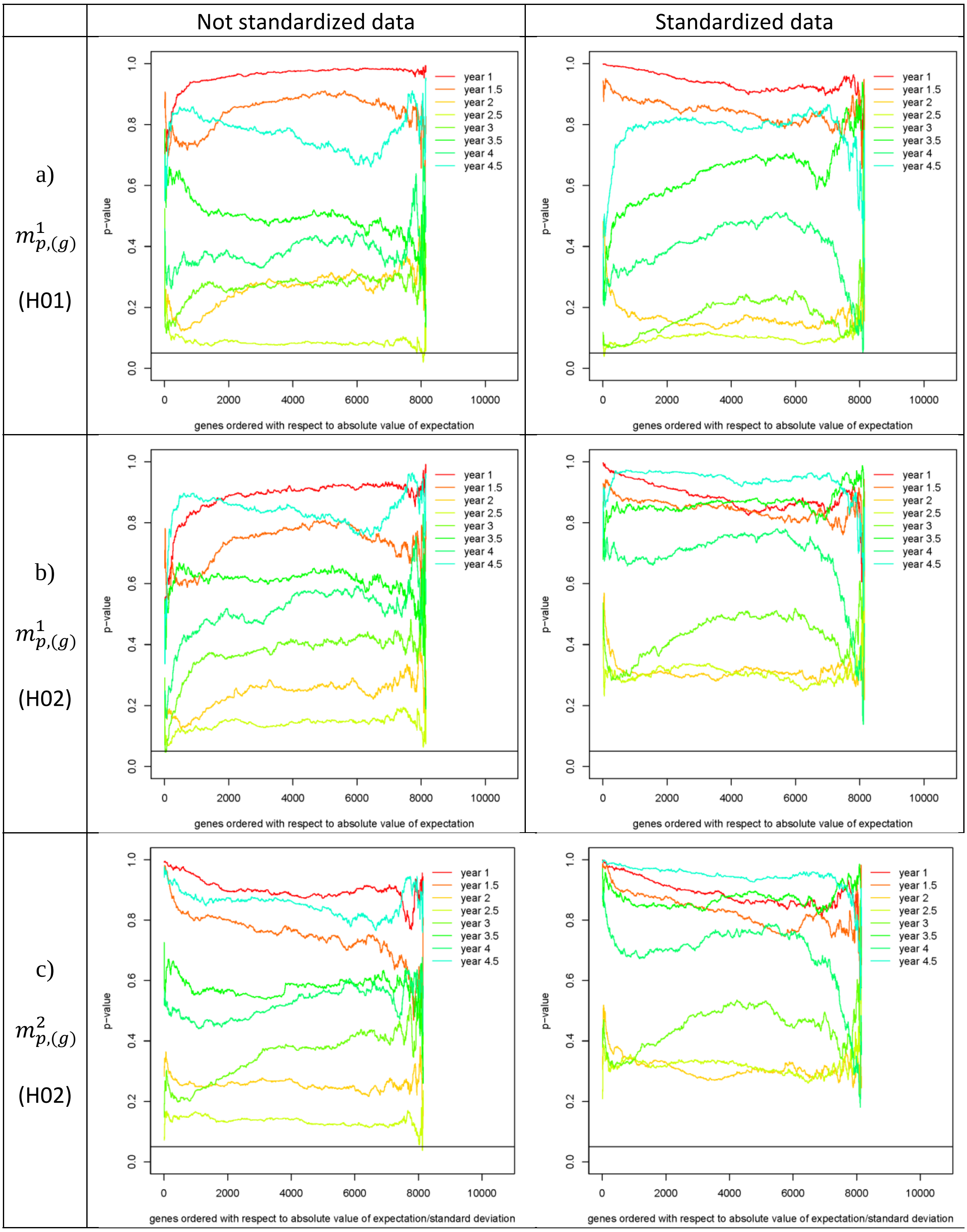


Figure 12 *Plots of p-values for three of the hypothesis tests where the dataset with cases without spread in the clinical group is used. a) The hypothesis test is based on the statistic* $m^1_{p,(g)}$ *and the null distribution is estimated by randomizing the case-control pairs between the periods. 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 sufficiently close. The p-value is 0.05 at the black horizontal line.*

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 in year 1-2 before diagnosis for the screening group (H03). This corresponds to the result in Figure 4 and the difference in expectation shown in Figure 2 b) and c). There are few p-values below 0.05 for the clinical group, but the lowest p-values are obtained around year 2-4 before diagnosis.

Figure 6 and Figure 7 show results for the hypothesis tests that are based on the statistic $s_{p,(g)}$ that is used for testing if the standard deviation in a period is small compared to the standard deviation for all periods. We observe that in all the tests low p-values are obtained close to time of diagnosis. From this we may conclude that for some genes the distribution of $X_{g,c}$ depends on time to diagnosis (H01).

Figure 8 – Figure 12 show results for mean values and we only observe low p-values in Figure 9 with case-controls from the screening group with spread[3]. This is the largest homogeneous group implying higher power of the hypothesis tests. In this case results are significant around two years before diagnosis for the statistic a) $\mathrm{m}^1_{\mathrm{p,(g)}}$ when randomizing between periods (H01), b) $\mathrm{m}^1_{\mathrm{p,(g)}}$ when randomizing between the case and control (H02) and c) $\mathrm{m}^2_{\mathrm{p,(g)}}$ when randomizing between the case and control in each case-control pair (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 Figure 5 – Figure 10, but now focusing on how the p-values vary with time. This is shown in Figure 13 – Figure 15. Figure 13 shows results for the screening group with spread, Figure 14 for the screening group without spread and Figure 15 for the clinical group without spread.

In all the tests shown in Figure 13 – Figure 15 there are significantly low $s_{p,(g)}$ values the last two years before diagnosis, (H01). This corresponds to what we observed in Figure 2 a) and Figure 3 a). In Figure 14 there are significantly high values for the two expectations $m^1_{p,(g)}$ and $m^2_{p,(g)}$ around two years before diagnosis (H01) for the screening group without spread. This is similar to what we found in Figure 9. In Figure 15 there are significantly high values for the expectations $m^1_{p,(g)}$ around two-three years before diagnosis (H01) for the clinical group without spread. This is similar to what we found in Figure 12. For the weights $w_{p,(g)}$ there are significantly low p-values the last year before diagnosis in the screening group and around two-three years before diagnosis for the clinical group.

[3] The other figures are for the clinical group: Figure 11 with all cases and Figure 12 with cases without spread, and for the screening group: Figure 8 with all cases and Figure 10 with cases with spread.

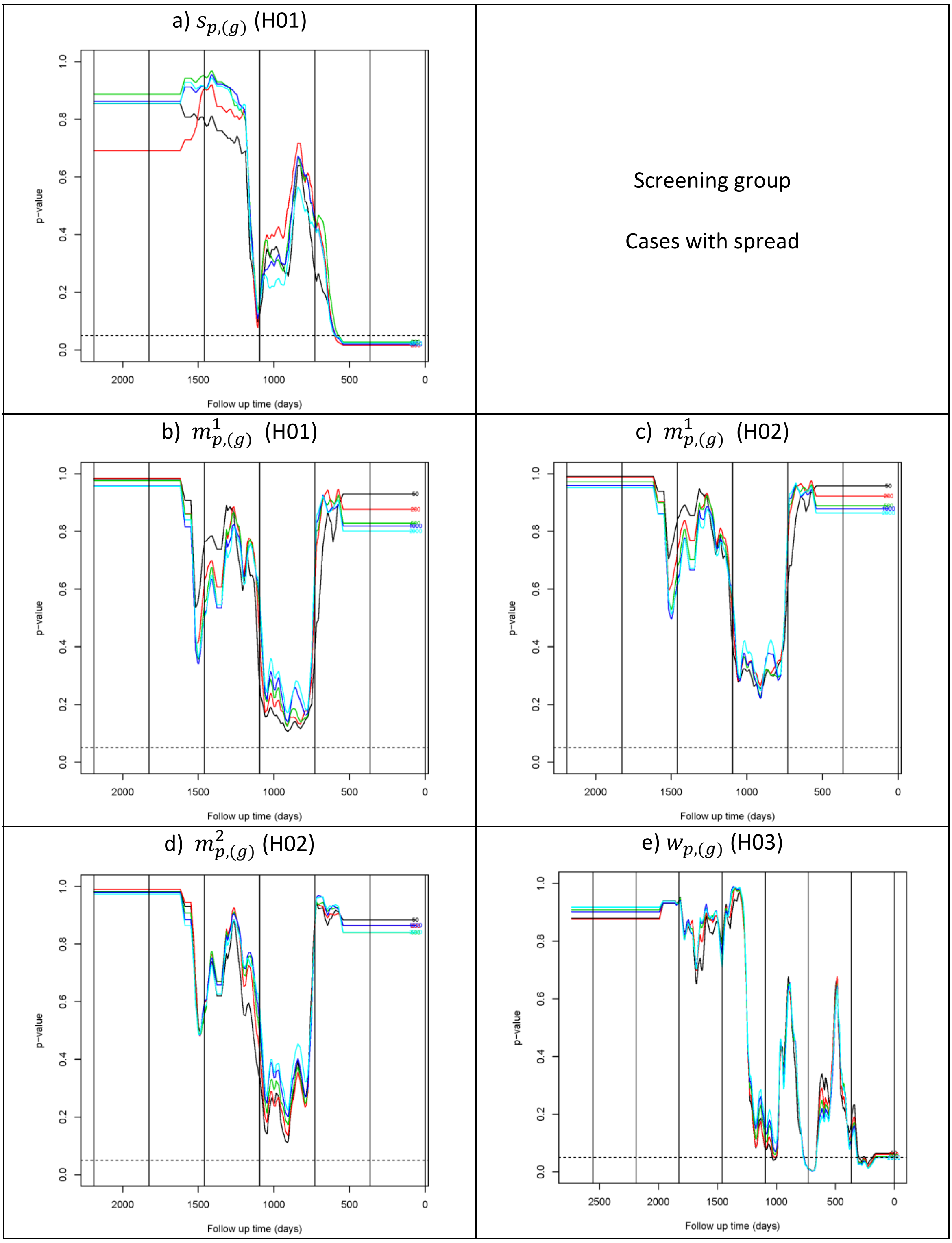
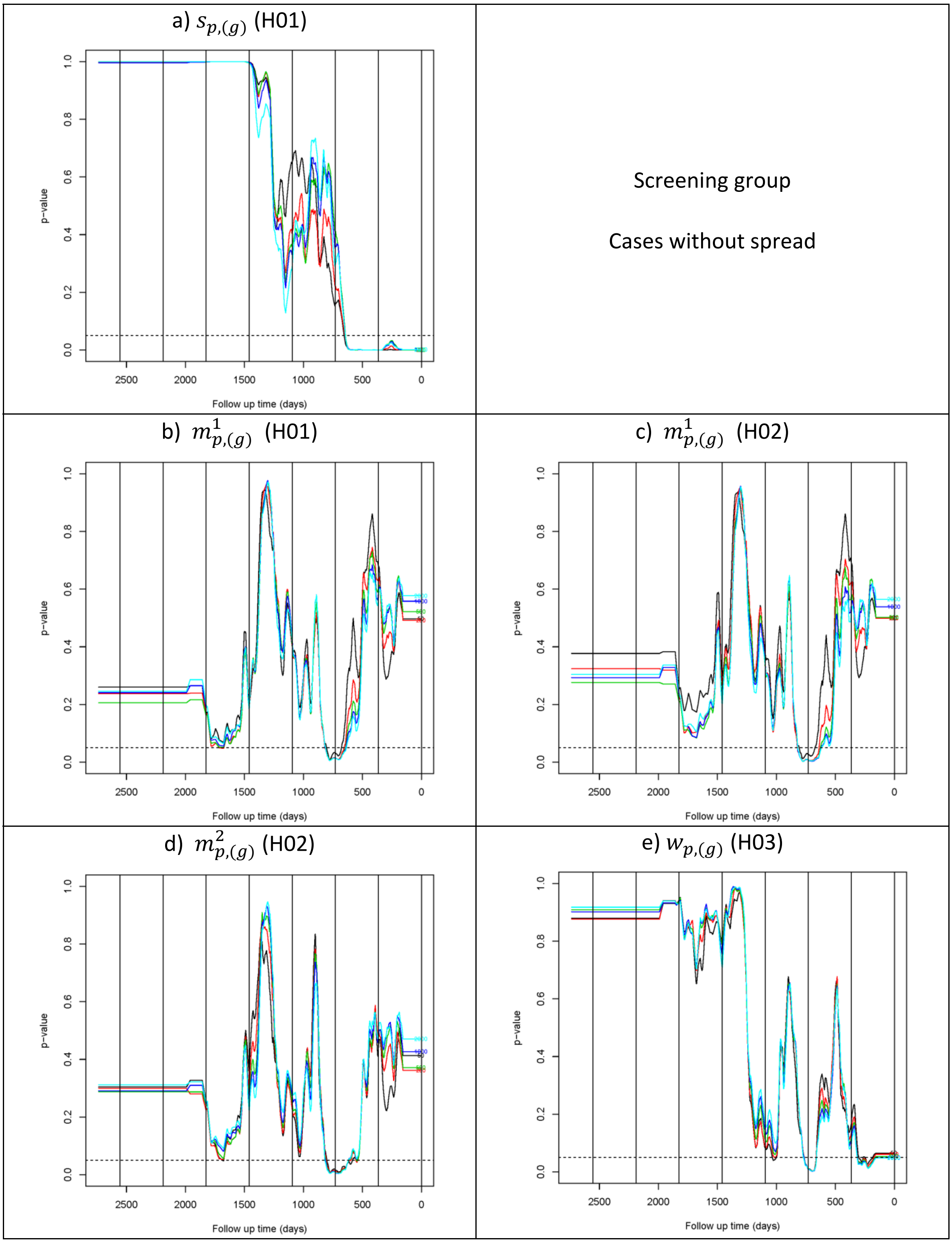


Figure 13 *Plots of p-values against time for the hypothesis tests where the not standardized dataset for the screening group is used. In panels a), b), c) and d) the dataset consists of the cases with spread, while in panel e) the entire dataset is used. In each plot there is one curve for genes with 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 4b), panels b), c) and d) correspond to Figure 5 a), b) and c), respectively, and panel e) corresponds to Figure 5a).*


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

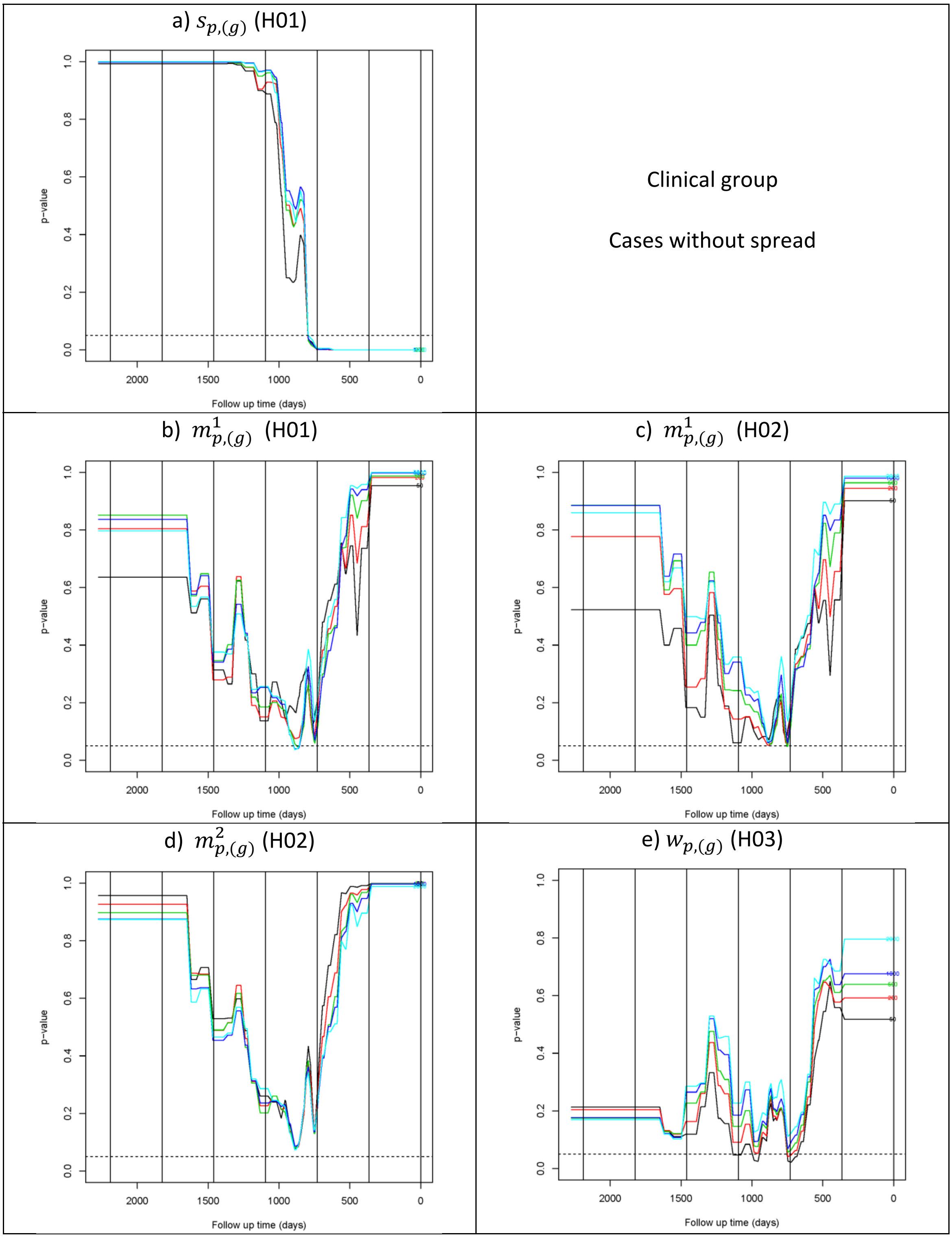
b) $m^1_{p,(g)}$ (H01)

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

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

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

Figure 14 *Plots of p-values against time for the hypothesis tests where the not standardized dataset for the screening group is used. In panels a), b), c) and d) the dataset consists of the cases without 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 6 b), panels b), c) and d) correspond to Figure 9 a), b) and c), respectively, and panel e) corresponds to Figure 5a). Note that panel e) in this figure is the same as panel e)* in Figure 13.



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

Clinical group

Cases without spread

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

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

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

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

Figure 15 *Plots of p-values against time for the hypothesis tests where the not standardized dataset for the* <u>*clinical group*</u> *is used. In panels a), b), c) and d) the dataset consists of the* <u>*cases without spread*</u>*, while in panel e) the* <u>*entire dataset*</u> *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 7 b), panels b), c) and d) correspond to Figure 12a), b) and c), respectively, and panel e) corresponds to Figure 5b).*

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 periods that contain 50 (25) case-control pairs from the screening (clinical) group where the case is without 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 3 and Figure 16. For the screening (clinical) group we observe that 44% (67%) of the cases with spread are correctly classified, while 56% (54%) of the cases without spread are correctly classified. For the screening group the numbers of correctly classifies cases is not significantly higher than what is expected by chance (p-value 0.56, Fisher's test) (H03), while for the clinical group the numbers of correctly classifies cases is significantly higher than expected (p-value 0.049, Fisher's test).

Table 3 *Number of correctly and wrongly classified cases. a) Results for the screening group. b) Results for the clinical group. c) Results for the screening group when insitu cases are included amongst the cases without spread.*

		Number of correctly and wrongly classified cases				
		With spread		Without spread		
		FN	TP	FP	TN	P-value (Fisher's test)
a) Screening group	All years	61	47	119	153	0.561
	Year 1	2	10	26	27	0.030
b) Clinical group	All years	10	20	26	31	0.049
	Year 3-4	2	5	6	16	0.051
c) Screening group, including insitu	All years	61	47	119+37	153+29	0.722
	Year 1	2	10	26+13	27+3	0.072

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 16. For the screening group we observe that the probability of correct classification is much higher in year 1 before diagnosis. For this period the p-value obtained using Fisher's test is equal to 0.030. This is in accordance with the results shown in Figure 5 a), Figure 13 e) and Figure 14 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 close to the time of diagnosis (H01, H03). For the clinical group we observe that the probability of correct classification is much higher in year 3 before diagnosis. For year 3 and 4 the p-value obtained using Fisher's test is equal to 0.051, while for year 3 it is 0.00 (year 3 contains only 10 case-control pairs where two are with spread). This is in accordance with the results shown in Figure 5 b) and Figure 15 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 close to year 3 before time of diagnosis (H01, H03).

We also predicted the insitu cases using the weights and genes selected based on the cases with and without spread. An insitu case is correctly classified if it is classified as without spread. For the 13 insitu cases in the clinical group 6 were correctly classified, while 29 of the

66 cases in the screening group were correctly classified. From Table 3 we observe that we obtain poorer results when the insitu cases are included amongst the cases without spread.

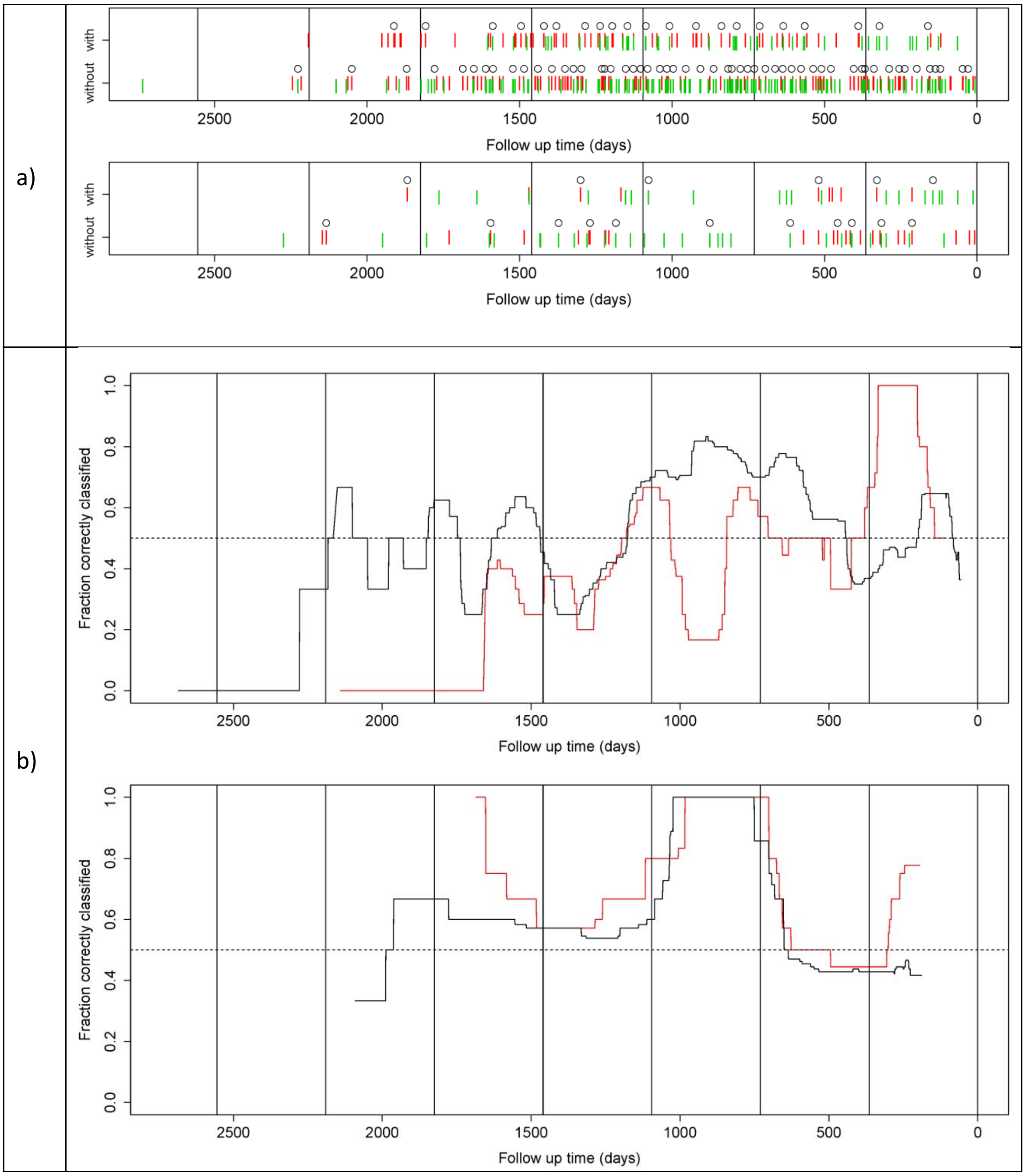


Figure 16 *a) Correctly (green) or wrongly (red) classified cases plotted against follow up time for the screening (upper panel) and the clinical group (lower panel). A circle is plotted above every fifth case. Long vertical lines are plotted to indicate the years. On the y-axis "with" means cases with spread and "without" means cases without spread. All cases in the clinical group are correctly classified in a period around year 3 before diagnosis. b) Fraction of correctly classified cases with (red) and without (black) spread over time for the screening (upper panel) and the clinical group (lower panel). 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.*

4.4 Prediction of metastasis status for the validation datasets

A set of 1000 genes with corresponding weights are selected based on data from the screening group in a period around 6 months before diagnosis. Table 4 shows prediction results obtained when using scores based on these 1000 genes for the case-control pairs of the CC1, CC2 and CC3 datasets where the case belongs to the screening group. We observe that the predictive power obtained for CC3 (p-value 0.005, Fisher test) is much better than for CC1 (p-value 0.7, Fisher test) and CC2 (p-value 0.5, Fisher test). Figure 17 shows the scores for CC1, CC2 and CC3. (Positive score: classified as with spread; Negative score: classified as without spread).

We used a Fisher test for testing whether the results for the cases with spread for CC1, CC2 and CC3 (left FN and TP column in Table 4) are more different than expected by chance. The p-value obtained (0.038) indicates that CC1, CC2 and CC3 are more different than expected, which again indicates that there are differences between the three datasets. Previous analyses of the three datasets indicate that the quality of the CC3 dataset is clearly better than that of the CC1 dataset, which again is clearly better than the quality of the CC2 dataset. Our results strengthen this conclusion. Also, the CC3 dataset has been run on Illumina HumanHT-12 version 4, while the CC1 and CC2 datasets have been run on Illumina HumanAWG-6 version 3.

Table 4 *Prediction results for the case-control pairs of the CC1, CC2 and CC3 datasets where the case belongs to the screening group. The genes used for computing the scores used in the prediction rule are selected based on case-control pairs from the prospective dataset where the case belongs to the screening group. Case-control pairs in a period around 6 months before diagnosis were used.*

Dataset	FN	TP	FP	TN	p-value	FN	TP	FP	TN	p-value
CC1	6	2	7	21	0.6625	10	10	21	53	0.0618
CC2	4	3	8	15	0.5145					
CC3	0	5	6	17	0.0047*					

**p-value<0.05.*

Prediction results for the case-control pairs of the CC1, CC2 and CC3 datasets where the case belongs to the clinical group are shown in Section 8 (Appendix). For the clinical group no p-values obtained in Fisher tests for the prediction results were significant.

We have also predicted the metastasis status for the 18 cases in the CC0small, CC1small, CC2 small and CC3 small datasets, where the cases do not have cancer (benign tumors). The results are presented in Table 5. We observe that 13 (72%) of the cases are correctly classified, a higher number of correctly classified cases than expected by chance (p-value 0.048, binomial test).

Table 5 *Prediction results for the case-control pairs of the CC0small, CC1small, CC2small and CC3small datasets where the cases do not have cancer (benign tumors). The genes used for computing the scores used in the prediction rule are selected based on case-control pairs from the prospective dataset where the case belongs to the screening group. Case-control pairs in a period around 6 months before diagnosis were used.*

Number of	CC0small	CC1small	CC2small	CC3small	Sum
correctly classified as without spread	6	3	1	3	13
wrongly classified as with spread	1	1	0	3	5

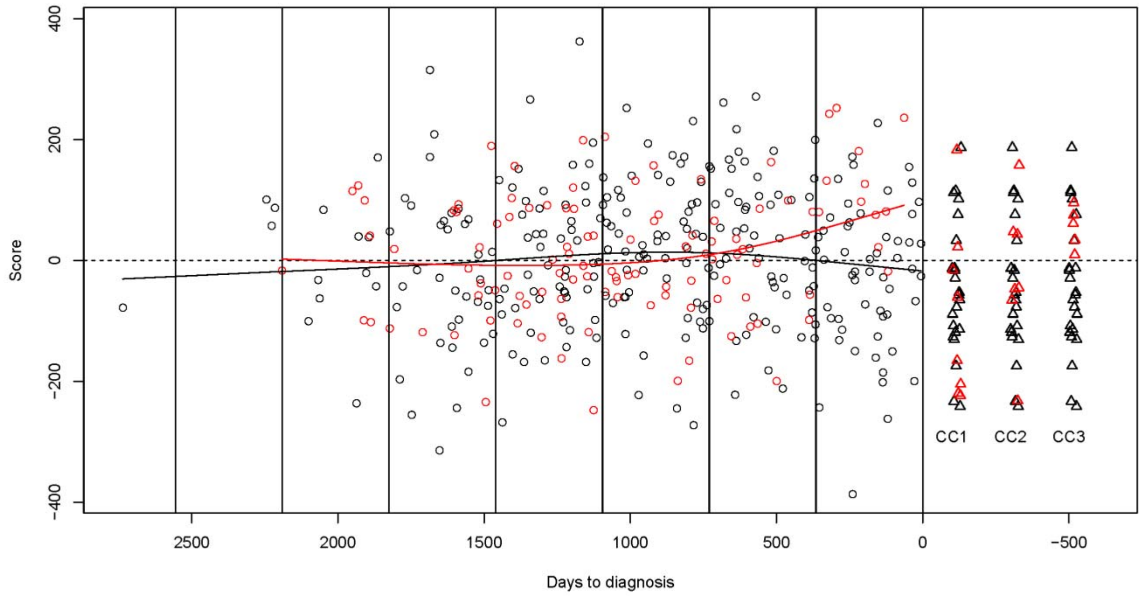


Figure 17 *Plots of scores for each case-control pair of the screening group against time (days to diagnosis). The score is plotted in red (black) if the case is with (without) spread and is computed using the weights of 1000 genes that are selected based on data around 6 months before diagnosis. For illustrational purposes, curves have been estimated from the scores of the prospective dataset (circles) using splines and plotted in the same color as the individual scores. Scores for case-control pairs in the screening group of the CC1, CC2 and CC3 datasets are plotted as triangles to the right of time of diagnosis (days to diagnosis = 0).*

Note that only 826 of the 1000 genes selected for the screening group are included in all three validation datasets (CC1, CC2, CC3)[4]. This means that less than 1000 genes are included in the scores computed for the case-control pairs in CC1, CC2 and CC3. Prediction results for each of CC1, CC2 and CC3, where cases participated in the screening program, do not change if we use the common set of 826 genes as if we use the set of 990, 910 and 883 genes, respectively. Figure 18 shows the scores for the case-control pairs from the screening group of the prospective dataset. We observe that the scores computed from the set of 826 and 1000 genes, respectively, are very similar. We conclude that the prediction results do not seem to be sensitive to exclusion of some genes from the set of genes selected for computing the score.

In Figure 19 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 for each of three different datasets. These three datasets are the two prospective datasets (clinical and screening group), and the part of the CC3 dataset where the cases participated in the screening program. The score for the cases with spread should be positive, while the scores for the cases without spread should be negative. 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. 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 in Table 3 – 4 and Figure 16 – 17 with $n = 50$. The results of these additional analyses are shown in Section 9 (Appendix). We

[4] The CC1, CC2 and CC3 datasets include 990, 910 and 883, respectively, of the 1000 genes that are selected for the screening group.

observe that 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.

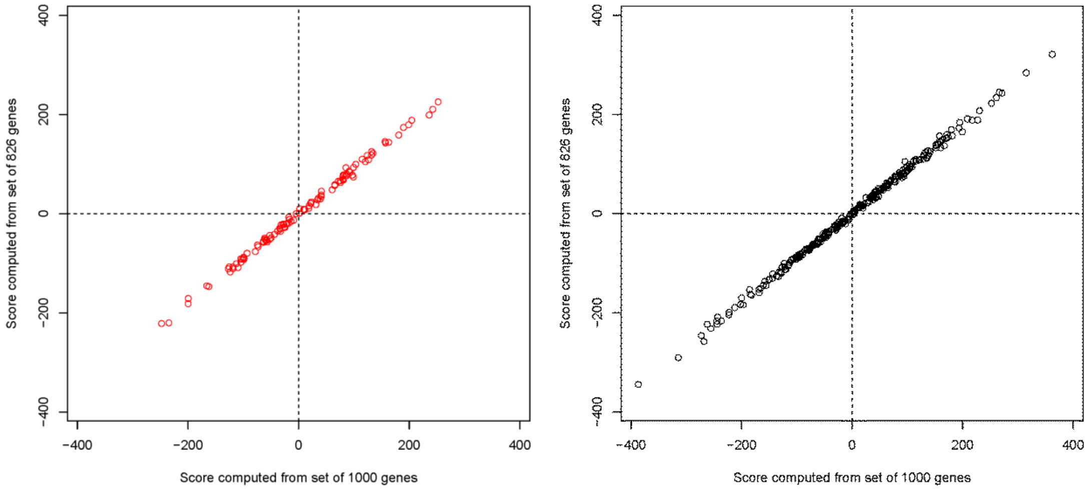


Figure 18 *Scores for the case-control pairs from the screening group of the prospective dataset. A set of 1000 genes with corresponding weights are selected based on data in a period around 6 months before diagnosis. The scores computed from the set of 1000 genes are plotted against scores computed from set of 826 of the 1000 genes that are included in all three validation datasets (CC1, CC2, CC3). a) Scores for the case-control pairs where the case is with spread. b) Scores for the case-control pairs where the case is without spread.*

Note that in Figure 19 a) the genes are selected from the screening group and applied for the CC3 dataset. For Figure 19 b) and c), however, the genes are selected based on the same dataset as we use when computing the scores. The purpose of Figure 19 is to examine how the score in the predictor is influenced by the number of genes included, and this is interesting for all three datasets independent of which dataset that was used for selecting the genes. Note however that in Table 3 and Table 4, where we show p-values computed for prediction results for the three different datasets used in Figure 19, we have used different data when selecting genes and predicting metastasis status.

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). 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.

The dataset is quite small, with only 108 (30) case-control pairs with spread and 272 (57) without spread in the screening (clinical) group, that are distributed over a eigth 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. In the screening group we obtained p-value 0.5 for the entire period but 0.03 for the last year before diagnosis. For the clinical group the p-value for the entire period was 0.05. Here the results indicated best prediction 3-4 years before diagnosis. The p-value is equal 0.05 in this time period but this may be due to a small data set.

6 References

[1] Vanessa Dumeaux, Josie Ursini-Siegel, Arnar Flatberg, Hans E. Fjosne, Jan-Ole Frantzen, Marit Muri Holmen, Enno Rodegerdts, Ellen Schlichting and Eiliv Lund. Peripheral blood cells inform on the presence of breast cancer: A population-based case.control study. Int. J. Cancer: 136, 656.667 (2015).

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

[3] 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.

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

[5] W. Evan Johnson and Cheng Li. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostat: 8 (1), 118-127 (2007). doi: 10.1093/biostatistics/kxj037

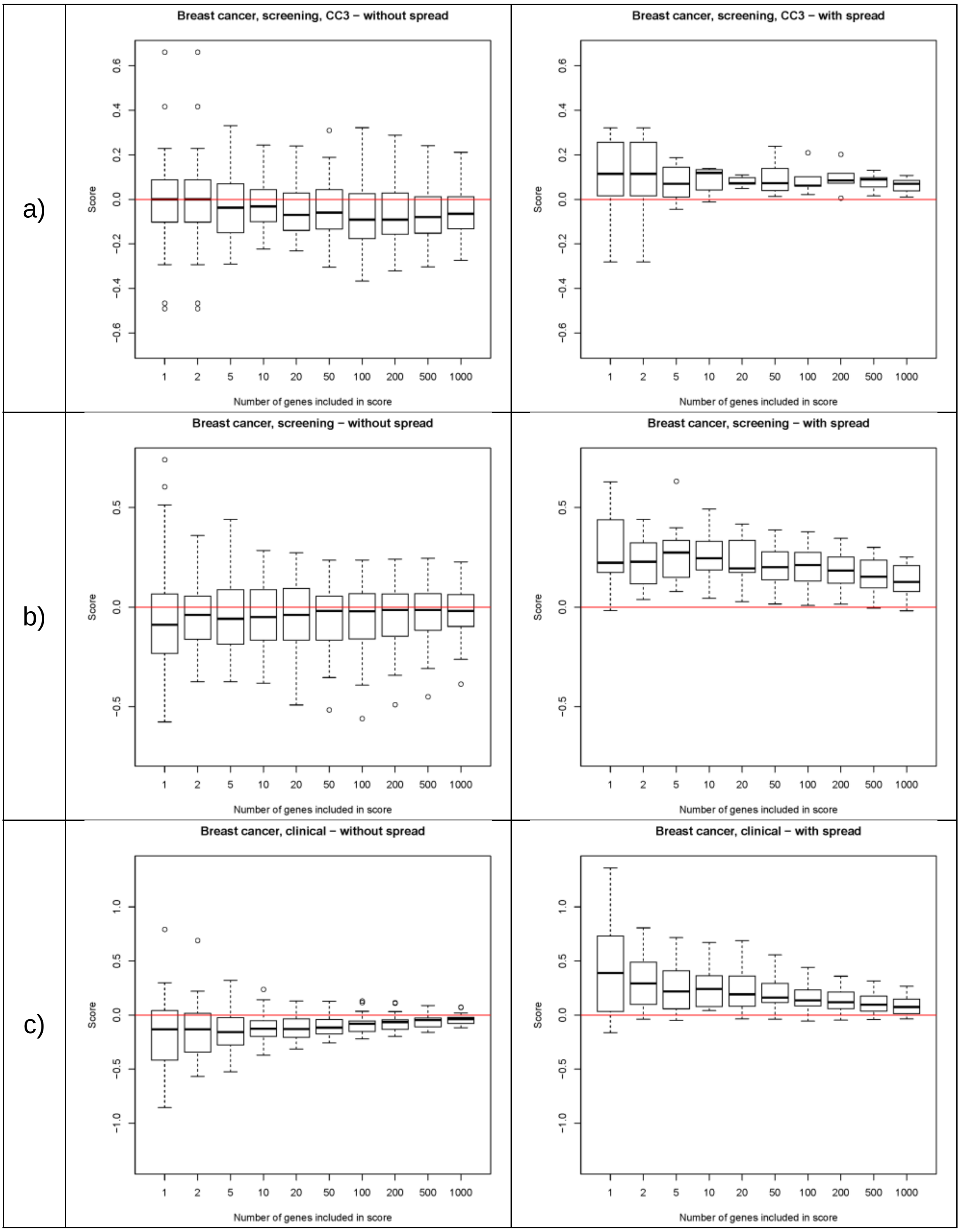


Figure 19 *Boxplots illustrating how the score used in the predictor depend on the number of genes included in the score. Note that the score has been normalized by dividing with the number of genes included in the score. Note also that the score for the cases with spread should be positive, while the scores for the cases without spread should be negative. a) Scores for case control 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. b) 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. c) 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.*

7 Appendix – Interval group included in clinical group

In all analyses described so far, the case-control pairs where the case belongs to the interval group have been included in the screening group. In this Appendix we show the results of some analyses where the interval group instead has been included in the screening group. 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 6. Results of predicting metastasis status of the cases are shown in Table 7 and Figure 20. We observe that the results are not as good as the results shown in Section 4.3.

Table 6 *Details about the available dataset for the screening (upper panel) and clinical group (lower panel).*

a) Number of case-control pairs in the screening group										
Year before diagnosis		8	7	6	5	4	3	2	1	Sum
Stratum	With spread	0	1	5	9	17	15	11	6	
	Without spread	1	3	5	27	48	43	42	43	
	Insitu	0	0	0	5	16	13	8	14	
	Sum									

b) Number of case-control pairs in the clinical group (includes interval group)										
Year before diagnosis		8	7	6	5	4	3	2	1	Sum
Stratum	With spread	0	0	2	10	18	11	17	16	
	Without spread	0	1	8	15	19	24	27	23	
	Insitu	0	0	1	3	4	5	3	7	
	Sum									

Table 7 *Number of correctly and wrongly classified cases. a) Results for the screening group. b) Results for the clinical group.*

		Number of correctly and wrongly classified cases				
		With spread		Without spread		
		FN	TP	FP	TN	P-value (Fisher’s test)
a) Screening group	All years	34	30	93	119	0.389
	Year 1	1	5	22	21	0.148
b) Clinical group	All years	38	36	63	54	0.802
	Year 3-4	16	23	23	20	0.832

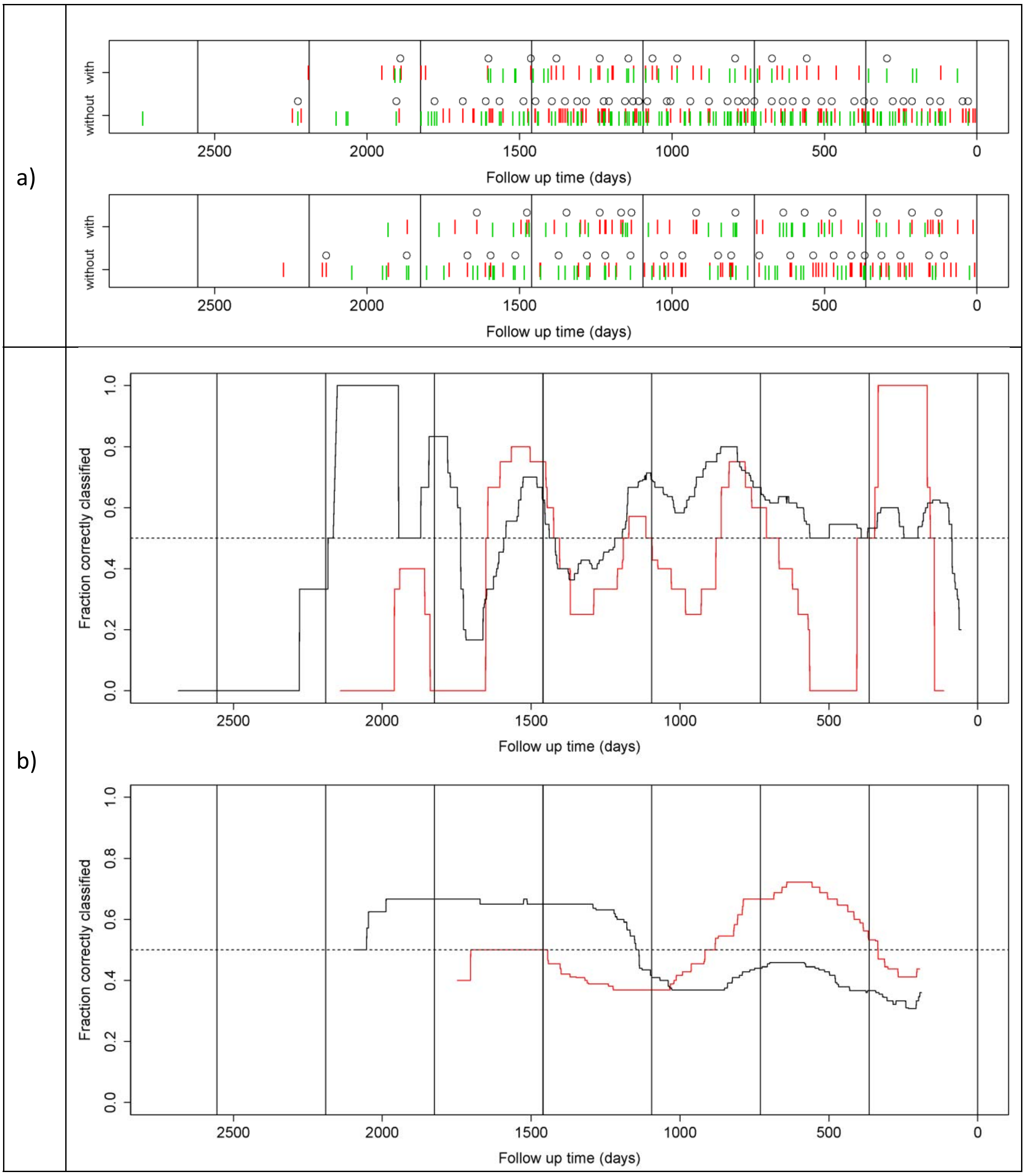


Figure 20 *a) Correctly (green) or wrongly (red) classified cases plotted against follow up time for the screening (upper panel) and the clinical group (lower panel). A circle is plotted above every fifth case. Long vertical lines are plotted to indicate the years. On the y-axis "with" means cases with spread and "without" means cases without spread. b) Fraction of correctly classified cases with (red) and without (black) spread over time for the screening (upper panel) and the clinical group (lower panel). 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.*

8 Appendix – Prediction results for the clinical group

Table 8 *Prediction results for the case-control pairs of the CC1, CC2 and CC3 datasets where the case belongs to the clinical group. The genes used for computing the scores used in the prediction rule are selected based on case-control pairs from the prospective dataset where the case belongs to the clinical group. Case-control pairs in a period around 2 years and 6 months before diagnosis were used.*

Gene selected from prospective dataset (period)	Dataset	FN	TP	FP	TN	p-value	FN	TP	FP	TN	p-value
Screening (around six months before diagnosis)	CC1	2	2	2	9	0.27	9	6	12	28	0.35
	CC2	2	2	4	7	0.54					
	CC3	5	2	6	12	0.75					
Clinical (around two years and six months before diagnosis)	CC1	3	1	7	4	0.97	9	6	25	15	0.96
	CC2	0	4	9	2	0.52					
	CC3	6	1	9	9	0.99					
Clinical (around one year before diagnosis)	CC1	2	2	10	1	0.99	7	8	28	12	0.93
	CC2	2	2	6	5	0.77					
	CC3	3	4	12	6	0.82					

9 Appendix – Including 50 genes in score

Table 9 *Number of correctly and wrongly classified cases when 50 genes are included in the score. a) Results for the screening group. b) Results for the clinical group*

		Number of correctly and wrongly classified cases				
		With spread		Without spread		
		FN	TP	FP	TN	P-value (Fisher's test)
a) Screening group	All years	63	45	118	154	0.662
	Year 1	3	9	24	29	0.061
b) Clinical group	All years	10	20	25	32	0.036
	Year 3-4	1	6	5	17	0.006

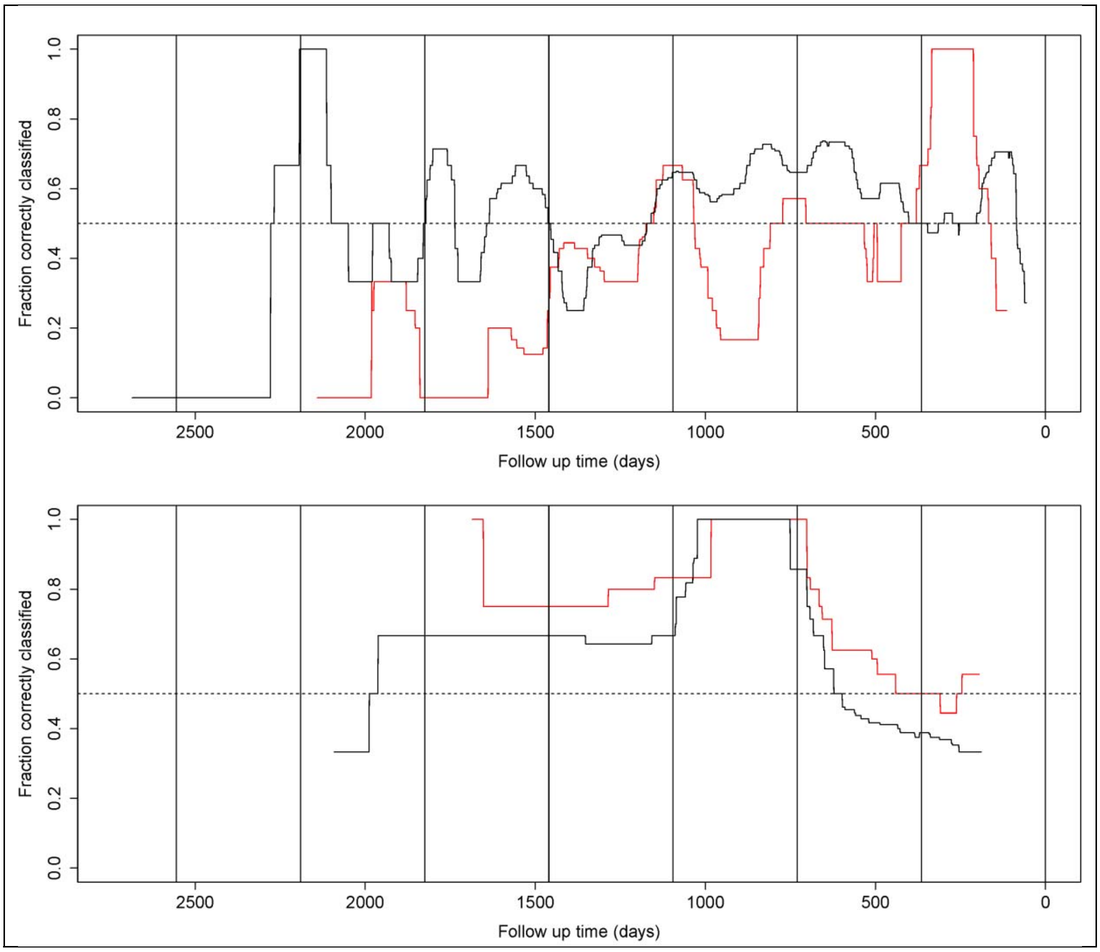


Figure 21 *Fraction of correctly classified cases with (red) and without (black) spread over time for the screening (upper panel) and the clinical group (lower panel) when 50 genes are included in the score. The fraction for each point in time is computed using a moving window of one year. The resulting curve is then smoothed using a median-filter using a window size of one year.*

Table 10 *Prediction results when 50 genes are included in the score for the case-control pairs of the CC1, CC2 and CC3 datasets where the case belongs to the screening group. The genes used for computing the scores used in the prediction rule are selected based on case-control pairs from the prospective dataset where the case belongs to the screening group. Case-control pairs in a period around 6 months before diagnosis were used.*

Dataset	FN	TP	FP	TN	p-value	FN	TP	FP	TN	p-value
CC1	6	2	10	18	0.8384	10	10	26	48	0.170
CC2	4	3	8	15	0.5145					
CC3	0	5	8	15	0.0131*					

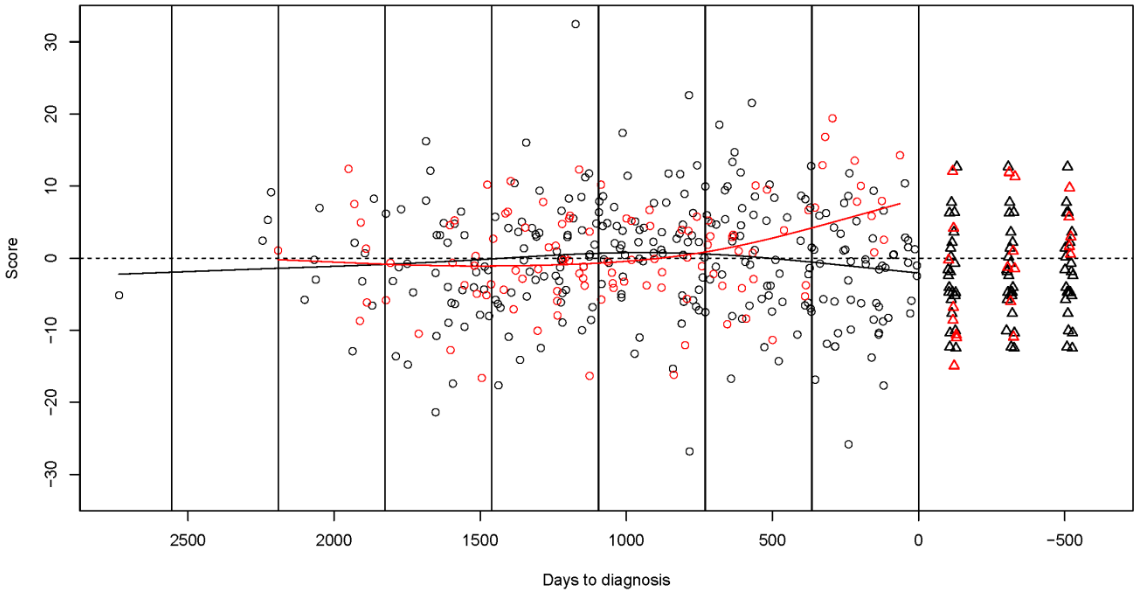


Figure 22 *Plots of scores for each case-control pair of the screening group against time (days to diagnosis) when 50 genes are included in the score. The score is plotted in red (black) if the case is with (without) spread and is computed using the weights of 1000 genes that are selected based on data around 6 months before diagnosis. For illustrational purposes, curves have been estimated from the scores of the prospective dataset (circles) using splines and plotted in the same color as the individual scores. Scores for case-control pairs in the screening group of the CC1, CC2 and CC3 datasets are plotted as triangles to the right of time of diagnosis (days to diagnosis = 0).*

10 Appendix – Adjusting for the batch effect

Here we give a short description of the ComBat method developed by Johnson & Li [5] for estimating the batch effects and how to use these estimates for adjusting for the batch effects when computing sample means and standard deviations.

The gene expression value Y_{ijg} for gene g and sample j from batch i is modelled as

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg} \text{ and } \varepsilon_{ijg} \sim \text{Normal}(0, \sigma^2),$$

where

- α_g is the overall gene expression,
- X is a design matrix for sample conditions,
- β_g is the vector of regression coefficients corresponding to X,
- γ_{ig} is the additive batch effect and
- δ_{ig} is the multiplicative batch effect.

The batch-adjusted data Y^*_{ijg} can then be computed as

$$Y^*_{ijg} = \frac{Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g - \hat{\gamma}_{ig}}{\hat{\delta}_{ig}} + \hat{\alpha}_g + X\hat{\beta}_g.$$

The estimates of the parameters α_g, β_g, γ_{ig} and δ_{ig} are computed using an empirical Bayes method[5]. A more detailed explanation is found in [6].

Both the expectation and the variance of a gene for the cases can vary both with time and stratum. We can therefore not use the ComBat method described above for batch-adjusting the dataset that consists of $\log_2$ differences in gene expression between cases and controls. Instead we will use ComBat to estimate the batch effects $\hat{\gamma}_{ig}$ and $\hat{\delta}_{ig}$ from a dataset that includes only the $\log_2$ gene expressions for the controls.

Log_2 gene expression data that are adjusted for the additive batch effect γ_{ig}, but not for the multiplicative batch effect δ_{ig}, can then be computed as

$$Y'_{ijg} = Y_{ijg} - \hat{\gamma}_{ig} = \mu_{Gg} + \hat{\delta}_{ig}\varepsilon_{ijg} \text{ where } \varepsilon_{ijg} \sim \text{Normal}(0, {\sigma_G}^2) \text{ for group G.}$$

For case-control pair c from batch i with sample j_1 as control (from group G1) and sample j_2 as case (from group G2) we have for the $\log_2$-expression difference $X_{g,c}$

$$X_{g,c} = Y_{ij_2g} - Y_{ij_1g} = Y'_{ij_2g} - Y'_{ij_1g} = \mu_g + \hat{\delta}_{ig}\varepsilon_{g,c} \text{ where } \varepsilon_{g,c} \sim \text{Normal}(0, \sigma^2).$$

We observe that $X_{g,c}$ is adjusted for the additive batch effect γ_{ig}, but not for the multiplicative batch effect δ_{ig}.

We compute the estimate of μ_g, $\hat{\mu}_g$, as the weighted average of $X_{g,c}$, where the weights are $\frac{1}{\hat{\delta}_{ig}}$, and we compute the estimate of σ^2, $\hat{\sigma}^2$, as $\frac{1}{n-1}\sum_{c=1}^{n}\left(\frac{X_{g,c} - \hat{\mu}_g}{\hat{\delta}_{ig}}\right)^2$. We will compare estimated sample means and standard deviations between genes. For each gene we therefore multiply the estimates of $\hat{\delta}_{ig}$ by a constant so that for this gene $\frac{1}{B}\sum_{i=1}^{B}\hat{\delta}_{ig} = 1$, where B is the number of batches / runs.

[5] Note that in the implementation of the method, the batch-adjusted data Y^*_{ijg} are computed as $Y^*_{ijg} = \frac{\frac{Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g}{\hat{\sigma}} - \hat{\gamma}'_{ig}}{\hat{\delta}_{ig}}\hat{\sigma} + \hat{\alpha}_g + X\hat{\beta}_g$, where $\hat{\gamma}'_{ig} = \frac{\hat{\gamma}_{ig}}{\hat{\sigma}}$ is the parameter that is estimated instead of $\hat{\gamma}_{ig}$.