Lund, E. (Ed.) (2020). *Advancing Systems Epidemiology in Cancer. Exploring Trajectories of Gene Expression.*
Oslo: Scandinavian University Press.
DOI: https://doi.org/10.18261/9788215041193-2020-09

9. Signals of Death—Post-Diagnostic Single Gene Expression Trajectories in Breast Cancer—A Proof of Concept

Eiliv Lund, Marit Holden, Jean-Christophe Thalabard, Lill-Tove Rasmussen Busund, Igor Snapkov and Lars Holden

Abstract Using the time-dependent dynamics of gene expression from immune cells in blood, we aimed to explore single gene expression trajectories as biomarkers for death after a diagnosis of breast cancer introducing a new statistical method denoted Difference in Time Development Statistics (DTDS). This shows as proof of principle that the gene expression profiles from immune cells in blood differed in the postdiagnostic period are dependent on later vital status.

Keywords gene expression | breast cancer | systems epidemiology | death | statistical method

The gene expression analyses of 394 breast cancer cases and age-matched controls were obtained from the Norwegian Women and Cancer (NOWAC) postgenome biobank (N = 50 000) performed in blood taken 0–8 years after a breast cancer diagnosis. The tube contained a protective buffer that preserved the mRNA in the blood. Cancer diagnosis and cause of death were based on linkage with the Norwegian Cancer Registry. The new statistical method was designed to test the difference in the time development between two strata using a non-parametric representation of the time development of the gene expression and used the area between the curves, i.e. the integral between the cures, as test statistics.

The time-dependent curves or trajectories exerted clearly non-linear changes with rapid transient mostly increasing fold changes, in cases who later died. Survivors had no changes. For cases who died this transient increase was followed by a

regression towards the gene expression profiles of survivors. For 86 genes, the integrated area from 18 months to 8 years post diagnosis was highly significant ($p<0.00001$) among women who died. There were indications of stronger relationship in metastatic cases alone.

INTRODUCTION

In 2017, the number of cancer deaths in Norway exceeded that of cardiovascular deaths for the first time (Norwegian Institute of Public Health, Norway, 2018). While the number of cancer deaths has remained fairly stable over recent years, the number of cardiovascular deaths has decreased rapidly. This points to the urgent need for further improvements in cancer treatment for an ageing population. For women in Norway, breast cancer is the most common invasive cancer, constituting 23% of all cancers diagnosed among women in 2017 (Cancer Registry of Norway, 2018). Although significantly improved, the majority of breast cancer deaths are due to metastasis, not the tumor. One hundred years ago the survival for women with metastatic cancer was only 5% after five years, while today the ten-year survival rate of metastatic breast cancer is 85% (Reddy et al., 2018). In order to further improve cancer diagnosis, personalized treatment is moving forward (Jeibouei et al., 2019). Individualized treatment should be based on predictors for individual outcome. The potential of immune response has become evident through the recent use of immune therapy (Stroncek et al., 2017). Biomarkers in blood or liquid biopsies could be functional genomics i.e. transcriptomics or methylation, or metabolites or proteins.

We proposed the compilation of time trajectories of gene expression in blood from many independent case-control pairs as a potential liquid biopsy in order to study the impact of the immune system on carcinogenesis (Lund et al., 2016). A gene's trajectory corresponds to a curve that represents the changes in gene expression as a function of time, consisting of differences of gene expression between many case-control pairs. Healthy controls establish the level of expression for genes not involved in carcinogenesis, and is assumed to be constant over time. Genes related to the immune system and/or carcinogenesis (expressed in cases) may change over time. Lack of a priori knowledge of the shape of the trajectories demands an agnostic approach (Spitz & Bondy, 2010) including adjustment for multiple testing (Reiner, Yekutieli, & Benjamini, 2003). Gene expression is analyzed as a potential biomarker of carcinogenesis/metastasis, and the *statistical quantity of interest* is the distribution of the gene expression as a function of time after diagnosis.

In a recent study of gene expression profiles in the years after diagnosis stratified on clinical stages significant differences in the overall gene expression profiles were found (Lund submitted PLOS).

The aim of this study is to explore single gene expression trajectories from immune cells in blood over the first years after diagnosis as predictors of later vital status, dead or alive. In order to use the cumulated evidence over time for clinical follow-up a new statistical method, denoted Difference in Time Development Statistics, was developed; see below.

METHODS

This new statistical method, denoted Difference in Time Development Statistics (DTDS), is designed to test differences in time development in a non-parametric manner of two variables or the same variable for two different strata. In this paper, the method is used in order to identify genes where the gene expressions in blood samples have a different time development after diagnosis of breast cancer. The dataset consists of case-control pairs in which the case is diagnosed with the disease and the control is healthy. The data is the difference in log2 gene expression in blood samples between the case and the control. The gene expression profiles that are measured represent an aggregate of the transcriptional activity of all the blood cells at the time of blood collection. The DTDS method will be used on the postdiagnostic or clinical follow-up in the NOWAC postgenome cohort, where each blood sample, regardless of disease status, was collected at a random follow-up time. We will first describe the epidemiological design necessary for studies of the postdiagnostic trajectories, and then describe the statistical concepts.

MATERIAL

The overall NOWAC postgenome biobank

Recruitment for the prospective Norwegian Women and Cancer (NOWAC) study started in 1991 (Lund et al., 2008). Women were randomly sampled from the Norwegian population register in Statistics Norway. The women were mailed a letter of invitation and a questionnaire. Follow-up was based on linkage to the Cancer Registry of Norway and the register of deaths were based on the unique national birth number given to all Norwegian inhabitants. Repeat questionnaires were mailed with intervals of some years. In the years 2002–2006, women were invited to participate in a subcohort, the NOWAC Postgenome cohort study; for further

details see Dumeaux et al., 2008. The main purpose was to establish a biobank suitable for analyses of functional genomics, in particular transcriptomics. Random samples of NOWAC women were drawn in weekly batches of 500 women until 50 000 women had responded positively. Women were invited to fill in another questionnaire and donate a blood sample at a health-care institution such as a GP's office. The blood samples were sent overnight to the institute by special post for biological samples. The tube contained a protective buffer that preserved the mRNA in the blood (PAX gene blood RNA system), allowing frozen storage over time and optimizing sensitivity of the analysis.

The present analysis used a *subsample* of the NOWAC postgenome biobank participants. Women who had both filled in a questionnaire in 1996–1998 and had given a Postgenome blood sample were eligible, a total of 31 101 women. Since collection of blood was at random without knowledge of disease status, the procedure gave a uniform distribution of gene expression measurement over time.

In 2010, breast cancer cases diagnosed between 1996 and 2006 were identified through a linkage to the national cancer registry. An age-matched control was drawn at random from the same batch of 500 women. A total of 394 incident breast cancer cases were identified. Those rendered non-eligible were six technical outliers, seven cases with unknown metastases, seven cases with another incidence of breast cancer before blood collection, ten controls diagnosed with cancer before blood donation, and one who emigrated, leaving 363 case-control pairs for the present analyses.

A linkage to the register of vital status in Norway gave a complete follow-up after blood donation until the end of the study on 31.12.2014, or death or emigration. Causes of death according to different strata of metastatic/invasive cancer at time of diagnoses are given below.

In order to update changes in clinical stage or a second breast cancer and to remove controls with an incidence of cancer, another linkage was performed in 2018 with the Cancer Registry of Norway. For six women with metastases and ten cases with another incidence of breast cancer, the updated information was used to change the start of follow-up.

Of the 363 case-control pairs, 85 were omitted since the follow-up time for the cases that are observed before 18 months from diagnosis are heavily influenced by the treatment. We therefore first analyze a dataset of 39 cases who died from cancers and compare them with 239 cases who did not die of cancer, i.e. a total dataset of 278 case-controls. Later, we reduce this to a dataset consisting of 23 cases with metastatic breast cancer who died of breast cancer and 79 cases with metastatic breast cancer who did not die of cancer; see Table 9.1.

Table 9.1. Further classification of the data and specification of the two analyzed datasets with 278 and 102 case-control pairs

Strata	Died of breast cancer	Died from non-breast cancer	Sum died of cancer	Survived	Died, not cancer	Sum, not died of cancer	Sum
Metastatic	32	4	**36**	97	3	**100**	**136**
Invasive	11	5	**16**	205	6	**211**	**227**
Sum	**43**	**9**	**52**	**302**	**9**	**311**	**363**
Dataset one where data before 18 months are excluded							
Metastatic			27			82	**109**
Invasive			12			157	**169**
Sum			**39**			**239**	**278**
Dataset two where data before 18 months are excluded							
Metastatic	**23**			**79**			**102**

STATISTICAL METHODS

The dataset consists of two strata of women with breast cancer in which the cases died or did not die of cancer and the observation time is the time after the last diagnosis. For each gene and stratum, we estimate the differences between cases and controls in gene expression as a smooth function using a moving window in time. We then estimate the differences in the time development between the two strata by calculating the area between the two estimated curves for the smoothed gene expression for the two strata. If there is a systematic difference in the level or the time development of the gene expression between the two strata, this area is large. We will test three hypotheses. The first hypothesis, H0A, concerns individual gene trajectories, while H0B looks at all genes together. We also predict the vital stage, dead or alive, of each case using cross-validation. H0C states that this prediction is independent of vital stage.

H0A: Identify genes with different time development

We first focus on identifying genes with a different time development. Let $X_{c,g}$ be the difference in log2 gene expression for case-control pair $c = 1,2, \ldots, M$ for gene $g = 1,2, \ldots N_g$. Further, let t_c be the time of observation relative to diagnosis for the case in the case-control pair c. We assume $X_{c,g} \sim N(f_{s(c),g}(t_c), \sigma_g)$ where σ_g is the standard deviation and $s(c)$ is the stratum of case c. We estimate the function $f_{s,g}(t)$

by taking an average of the observations $X_{c,g}$ from stratum $s(c)$ in an interval that includes the n nearest observations in time, i.e. the $n/2$ observations with largest t_c but $t_c < t$ and the $n/2$ observations with smallest t_c but $t_c > t$. The number n is a tradeoff between precision and resolution. It should be large enough that the estimate in an interval should not depend on a single data point and at least smaller then $M/4$ in order to get resolution in time. If there is a large difference in the time development between the two strata, the test statistic or area $V_g = |f_{a,g} - f_{b,g}| = \int |f_{a,g}(t) - f_{b,g}(t)| dt$ describing the area between the curves, will be large where the two strata are denoted a and b, respectively. This estimate is the sum of the absolute value of the differences in average gene expression between the two strata in equally spaced time points assuming the controls have similar values. *The integral* is evaluated in a time interval where there are observations from both strata.

We make N_g, independent hypotheses, i.e. one hypothesis for each gene:

H0A: For gene g, the time development of $X_{c,g}$ is independent of the stratum $s(c)$, i.e. $f_{a,g} = f_{b,g}$.

For each gene g, we compare the observed $V_{g,o}$ with the variable V from a simulated distribution where we use a standardization of the same variables $X_{c,g}$ for all the genes simultaneously, but where we randomize the strata $s(c)$ of the cases. We maintain the observations for each gene and the number of observations from each stratum. From the N_u simulations, we estimate the probability distribution $g(v) = P(V > v)$ that is independent of the genes. Based on this, we find a p-value $p_g = p(V > V_{g,o}) = (k + 1)/(N_u N_g + 1)$ for each gene g if k of the $N_u N_g$ simulations have $V > V_{g,o}$. We correct for multiple testing using the (Benjamini & Hochberg, 1995).

We estimated the functions $f_{s,g}$ with a moving average, where the window size is one-quarter of the respective datasets, i.e. 9 and 59 points, respectively. These functions were evaluated in regularly spaced points, making it easy to evaluate the functions when the observations for each stratum changes position in time. The integral was evaluated in the largest interval such that there were data points from the two strata before and after the interval making the interval equal to (547, 2255) days after diagnosis. The method was applied on a dataset with $N_g = 8400$ genes. The analysis is performed for standardized gene expressions for each gene

$$Y_{c,g} = (X_{c,g} - \frac{1}{M} \sum_c X_{c,g}) / \sigma_g$$

where the standard deviation σ_g is taken over the case-controls pairs for each gene. This normalization is necessary in order to compare area between the curves since we want to focus on the differences in time development and not in the mean value and the variance. The results were based on simulations with $N_u = 1000$ realizations.

H0B: Identify difference in gene development for all genes simultaneously

We will also make a weaker hypothesis where we analyze all the genes simultaneously:

H0B: For all genes, the time development of $X_{c,g}$ is independent of the stratum $s(c)$, i.e. for all genes $f_{a,g} = f_{b,g}$.

Note that we only make one hypothesis here. We perform the same N_u simulations as in the hypothesis test for H0A, but we use the test statistics $V_{(1),o} > V_{(2),o} > \cdots$ which is the $V_{g,o}$ variables for all the genes that are sorted in decreasing order. From the simulation, we find the probability for the ordered variables $g_m(v) = P(V_{(m)} > v)$ for $m = 1,2, ...,$ and the p-value for the hypothesis test $p_{(m)} = P(V_{(m)} > V_{(m),o} = (k + 1)/(N_u + 1)$ if k of the N_u simulations have $V_{(m)} > V_{(m),o}$. Here, we have many highly correlated test statistics $V_{(m),o}$ for $m = 1,2, ...,$ for testing the same hypothesis H0B. The integer m is chosen by the user. $m = 1$ means that we are only interested in the most extreme gene and $m = 10$ means that we are interested in the 10 most extreme genes. This method is most interesting for $3 < m < 50$, i.e. where no single gene is significant, but several/many genes have deviating values and where we avoid the multiple testing problem. Ideally, m should be decided before the data is analyzed, but this is not as critical as when alternative test statistics are independent of each other.

H0C: Prediction of strata

It is also possible to use the same technique in order to predict the stratum of a case. The idea is to find out if the observations $X_{c,g}$ for $g = 1,2, ..., N_g$ is closest to $f_{a,g}(t_c)$ or $f_{b,g}(t_c)$ for the genes with lowest p-values in the hypothesis test H0A above. Our ambition is only to find the quality of the prediction, not to make a diagnosis for each case. Hence, we make the following hypothesis:

H0C: The prediction $P_{a,c}$, that the case c belongs to stratum a, is independent of the stratum $s(c)$.

We test the hypothesis using cross-validation. The case-control pairs are divided into the D_1, D_2, ... D_{Nd} groups, which are described in more detail further down. For each of the pairs $c \in D_k$ we find

$$A_{c,a} = \sum_{g=1,s(c)=a}^{N_g} w_g (X_{c,g} - f_{a,g,k}(t_c))^2$$

where $f_{a,g,k}(t_c)$ is the estimated gene expression for gene g and stratum a at time t_c, i.e. the time of observation $X_{c,g}$ based on all the data except the data in D_k. This is

based on the assumption that $X_{c,g} \sim \mathrm{N}(f_{s(c),g,\mathrm{k}}(t_c), \sigma_{\mathrm{g}})$. The weight w_g may be set equal to $1/\sigma_g^2$, possibly modified based on the correlation between the gene expressions for different genes and how significant this gene is for the prediction. How important gene g is for the prediction is estimated from $p_{g,k}$, the p-value for hypothesis test HOA estimated from all the data except D_k. The prediction that the observation $X_{c,g}$ is from stratum a is then

$$P_{c,a} = \frac{A_{c,a}}{A_{c,a} + A_{c,b}}$$

This model gives probabilities that are approximately uniform in (0,1), see Figure 9.1. If we had assumed $X_{c,g} \sim N(f_{s(c),g,k}(t_c), \sigma_g)$ independently for each gene g, then most $P_{c,a}$ would be close to 0 or 1, which does not correspond to our ignorance in the classification. We use the test statistics

$$S_o = \sum_{c \in a} \left|1 - P_{c,a}\right| + \sum_{c \in b} \left|P_{c,a}\right|$$

which is the L_1 distance between the prediction for stratum a and the indicator for stratum a. We may randomize $P_{c,a}$ between the observations and find a distribution for S. The p-value for the hypothesis test H0C is found from the distribution for S, i.e. $p = P(S < S_o)$.

We used cross-validation and therefore needed to divide the dataset into separate groups. The 39 case-control pairs where the case died of cancer and 239 case-control pairs where the case did not die of cancer were divided into 13 groups, D_1, D_2, ... D_{13}. The data in each stratum was divided into three time periods for each of the two strata with an (almost) equal number of observations. Each of the 13 groups had (almost) the same number of observations from each stratum in each of the three time periods. For each group k we find the values $p_{g,k}$ from the hypothesis test H0A based on all the data except the data in D_k based on 1000 randomizations of the strata $s(c)$.

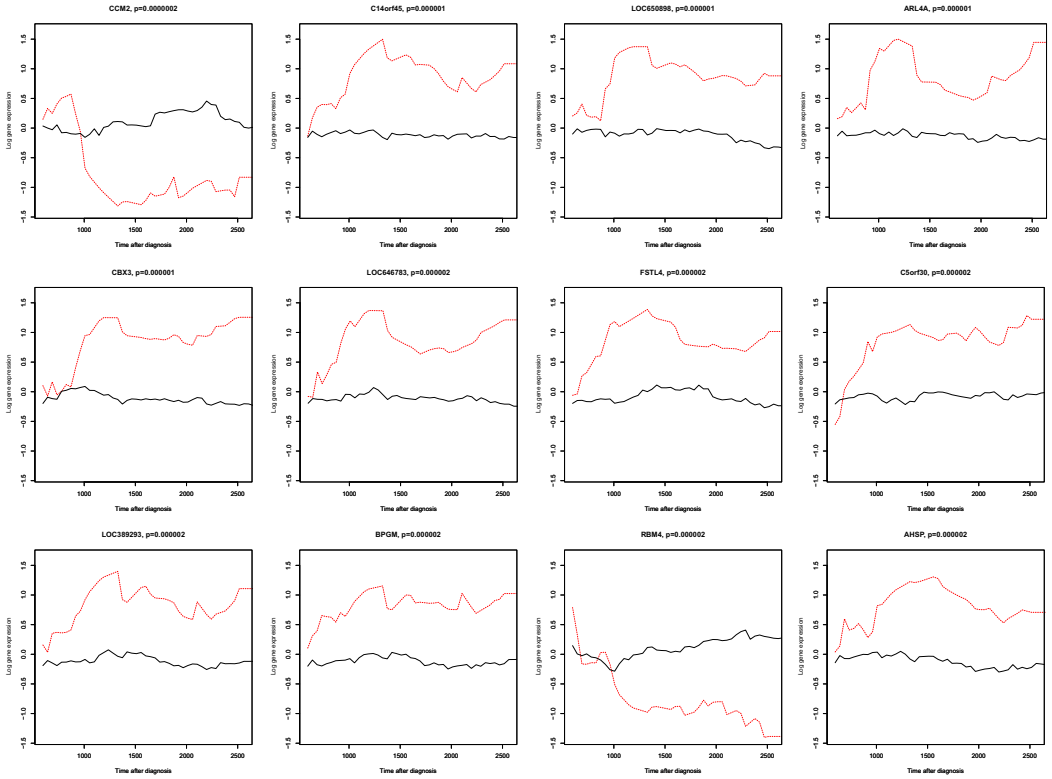

Figure 9.1. Log2 gene expression from the 12 case-control curves with the smallest p-value of the 8400 genes. The 12 p-values less than 0.00001. The figure uses normalized data as is used in the test statistics. The black continuous and the red dashed curves are the log2 gene expression from the case-controls who survived and died, respectively.

RESULTS

The data used in all the analyses are the differences in log2 gene expression between cases and controls in the period after diagnosis that are shown in Table 9.1, i.e. 278 case-control pairs.

Testing H0A

Results from testing the H0A hypothesis are shown in Table 9.2. The function of the top 10 is shown in Chapter 10.

Table 9.2. The 50 genes with smallest p-value from the 39+239 dataset with metastatic and invasive cancer. The columns show the name, p-value, q-value and area between the smooth curves between the cases who survived and died.

Gene	p-value	q-value	$\int\|f_{a,g}(t) - f_{b,g}(t)\|dt$
CCM2	2.38e-07	0.0016	1997
C14orf45	7.14e-07	0.0016	1883
LOC650898	1.07e-06	0.0016	1869
ARL4A	1.07e-06	0.0016	1867
CBX3	1.36e-06	0.0016	1842
LOC646783	1.90e-06	0.0016	1823
FSTL4	1.90e-06	0.0016	1820
C5orf30	1.90e-06	0.0016	1816
LOC389293	1.90e-06	0.0016	1812
BPGM	2.07e-06	0.0016	1798
RBM4	2.17e-06	0.0016	1789
AHSP	2.28e-06	0.0016	1788
CA1	3.21e-06	0.0020	1775
RP11-529I10.4	3.33e-06	0.0020	1766
ISCA1L	3.69e-06	0.0021	1757
NCBP1	4.42e-06	0.0023	1739
DARC	8.09e-06	0.0040	1697
HPS1	9.43e-06	0.0043	1686
TSTA3	9.65e-06	0.0043	1686
PDSS1	1.16e-05	0.0046	1668
STOM	1.19e-05	0.0046	1667
DHX29	1.21e-05	0.0046	1666
RBBP4	1.41e-05	0.0051	1658
RNF11	1.51e-05	0.0051	1653
FZD1	1.51e-05	0.0051	1652
RIPK4	1.75e-05	0.0053	1643
RBM28	1.81e-05	0.0053	1639
XK	1.88e-05	0.0053	1636
KIAA0174	1.92e-05	0.0053	1633
LOC646508	1.92e-05	0.0053	1633

Gene	p-value	q-value	$\int\lvert f_{a,g}(t) - f_{b,g}(t)\rvert dt$
GYPB	1.94e-05	0.0053	1632
MGC13057	2.06e-05	0.0053	1627
LOC649604	2.06e-05	0.0053	1627
BNIP3L	2.28e-05	0.0055	1618
TRIM10	2.29e-05	0.0055	1616
SLC14A1	2.36e-05	0.0055	1615
C14orf124	2.41e-05	0.0055	1615
EWSR1	2.88e-05	0.0062	1603
TRAK2	2.89e-05	0.0062	1603
SELK	3.34e-05	0.0070	1592
HMBS	3.39e-05	0.0070	1590
NUDT1	3.67e-05	0.0071	1585
SRRD	3.79e-05	0.0071	1583
WDR89	3.81e-05	0.0071	1583
NR1D1	3.85e-05	0.0071	1581
SLITRK1	3.91e-05	0.0071	1579
HEMGN	3.96e-05	0.0071	1577
DNAJB1	4.24e-05	0.0074	1570
LOC649044	4.31e-05	0.0074	1569
PPIA	4.66e-05	0.0075	1563

The q-values are the FDR-corrected p-values. From 8400 genes, 733 genes had q-values below the given threshold for hypothesis test H0A (97 with $q<0.01$ and 636 with $q<0.05$). The result shows that many genes have a different time development between the two strata. The reduced dataset with only metastatic breast cancer is too small to get significant results on this test. Figure 9.2 shows the functions $f_{died,g}(t)$ and $f_{survived,g}(t)$ separately for each of the 12 most significant genes of the 8400 genes ($p<0.000001$). The test statistics is the area between the pair of curves.

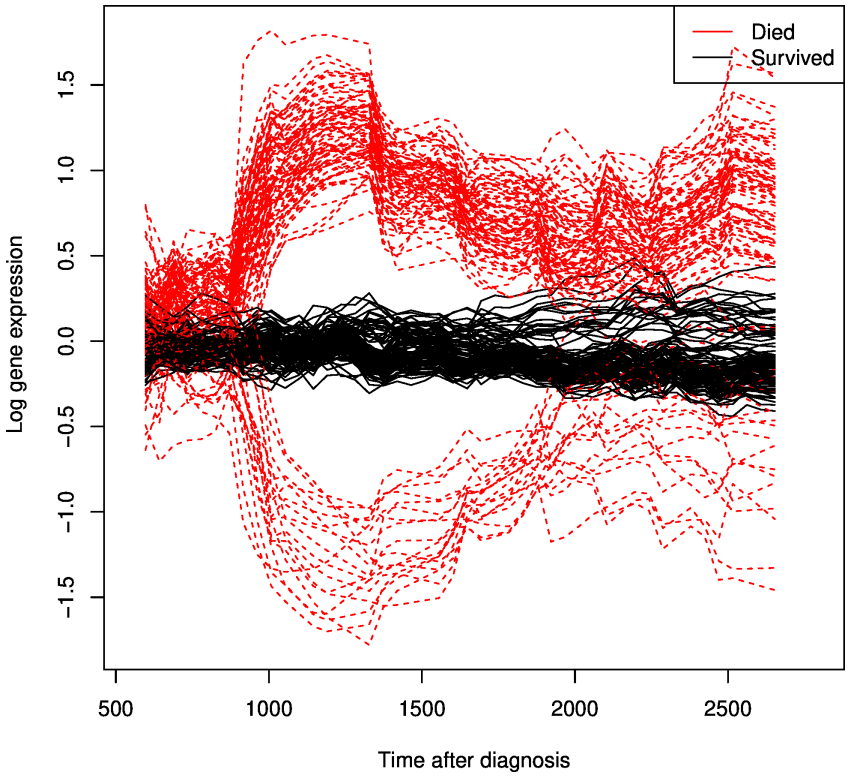

Figure 9.2. Log2 gene expression from the case-controls curves for the 12 genes with smallest p-value of the 8400 genes. The black continuous and the red dashed curves are the log2 gene expression from the case-controls who survived and died, respectively.

As shown for most genes, the gene expression increases. Noticeably, $f_{survived,g}(t)$ is almost constant and close to 0 in the entire period while $f_{died,g}(t)$ is closer to 1 or −1 in the period (1000,1500) days and then for many genes closer to 0 after 1500 days. The normalization (1) implies that the data for each gene have average value 0 and standard deviation 1 in order to compare data between genes. Since the stratum that survived is much larger, it is natural that the average of these curves is smoother and close to 0. The statistical test shows that deviation in averages value between the strata is significant for many genes. The p-value depends on whether there is a systematic difference in level or time variation of the gene expression, not the size of the difference in average value between the strata since this is removed in the standardization.

H0B: identify difference in gene development for all genes simultaneously

We also want to test all the genes simultaneously. Since we only make one test, it is easier to reject the hypothesis for a smaller dataset. First, we test hypothesis H0B on the dataset with 39 and 239 case-control pairs. There is only one hypothesis, but we have many different test statistics, one for each of m ordered $V_{(m)}$ test statistics for the area between the two curves. The different test statistics indicate whether there is a strong difference in the time development in one or a few genes compared to a smaller difference in many genes. The test for each of the ordered variables is highly correlated. Table 9.3 shows the p-values from the H0B.

Table 9.3. The p-values from hypothesis test H0B for each of the ordered variables. The upper row is from the 39+239 dataset with metastatic and invasive cancer and the lower line is from the 23+79 dataset with metastatic cancer. "<0.001" means that we have not observed any simulated values above the observed value from the data. The lower row shows the p-values from hypothesis test H0B for the reduced dataset on metastatic breast cancer.

Ordered variables	1	5	10	25	50	100	500	1000
39+239	0.002	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.002
23+79	0.003	0.061	0.043	0.036	0.037	0.034	0.045	0.052

Notice that we get very significant results and that many genes have a different time development between the two strata.

This test is also performed on the reduced dataset with only metastatic breast cancers. There are only 23 and 79 case-control pairs in the two strata (Table 9.1), those with metastases who died of breast cancer and those who did not die of cancer, respectively. We still get significant p-values, but much larger values than in the larger data set with both metastatic and invasive cancer; see Table 9.4. The differently ordered variables are highly correlated and give typically p-values between 3–6%.

H0C: Prediction of strata

We also want to test whether it is possible to predict the stratum of each case by testing hypothesis H0C. The 13 different datasets leaving out one of the groups D_k give a slightly different rankings of the importance of the different genes. The mean correlation between the rankings of the genes for these 13 datasets is 0.85.

Table 9.4 shows that there is a large overlap in the most important genes in the 14 different datasets when we include the ranking using all the data. On average, four of the five genes with the lowest p-value when using the entire dataset were among the 10 smallest p-values in the reduced datasets. We have marked the five genes with the smallest p-values when using all the genes with colors. Notice that many of the same genes have small p-values for the different subsets.

Table 9.4. Ranking of the 10 most important genes when we leave out D_k from the dataset. The lowest line is the most important genes when we use all the data.

D_k	Ranking of the most important genes for each of datasets
1	CCM2, C14orf45, LOC650898, BPGM, FSTL4, AHSP, CA1, C5orf30, LOC389293, ISCA1L
2	CCM2, LOC650898, C5orf30, C14orf45, BPGM, RBM4, LOC389293, RP11-529I10.4, ARL4A, CA1
3	CCM2, LOC646783, C5orf30, CBX3, FSTL4, RBBP4, LOC650898, RBM4, AHSP, PPIA
4	FSTL4, ARL4A, CBX3, LOC650898, C14orf45, LOC389293, DARC, CCM2, LOC646783, ISCA1L
5	CCM2, LOC650898, C14orf45, CBX3, ARL4A, C5orf30, BPGM, AHSP, RBM4
6	CCM2, C5orf30, C14orf45, ARL4A, CBX3, BPGM, FSTL4, LOC650898, RBM4, AHSP
7	LOC646783, ARL4A, NCBP1, CBX3, CCM2, C5orf30, LOC650898, LOC649604, C14orf45, LOC389293
8	CCM2, CBX3, C14orf45, BPGM, ARL4A, NCBP1, RP11-529I10.4, LOC646783, LOC389293, CA1
9	CCM2, FSTL4, TSTA3, ARL4A, C14orf45, KIAA0174, AHSP, LOC389293, RP11-529I10.4, ISCA1L
10	C14orf45, LOC389293, LOC646783, ARL4A, LOC650898, FSTL4, ISCA1L, RP11-529I10.4, CBX3, FZD1
11	CCM2, C14orf45, RBM4, C5orf30, LOC389293, ISCA1L, LOC650898, LOC646783, PDSS1, CA1
12	CCM2, ARL4A, CBX3, LOC650898, NCBP1, C14orf45, RP11-529I10.4, LOC389293, LOC646783, CA1
13	CCM2, C5orf30, FSTL4, LOC650898, DMD, CBX3, RBM4, ARL4A, CXCR4, LOC646783
all	CCM2, C14orf45, ARL4A, LOC650898, CBX3, C5orf30, FSTL4, LOC389293, LOC646783, BPGM

We have tested different predictions methods, i.e. different choices of the weights $w_{g,k}$. The different choices give highly correlated probabilities. We have found out that $w_{g,k} = 1/p_{g,k}$ for the 50 genes g with smallest $p_{g,k}$ value for each group is a quite robust choice. Figure 9.3 shows the predicted probabilities for each of the 278 case-control pairs after time of follow-up. Ideally, we wanted all the 39 red and yellow circles to be equal to 1 and the remaining circles equal to 0.

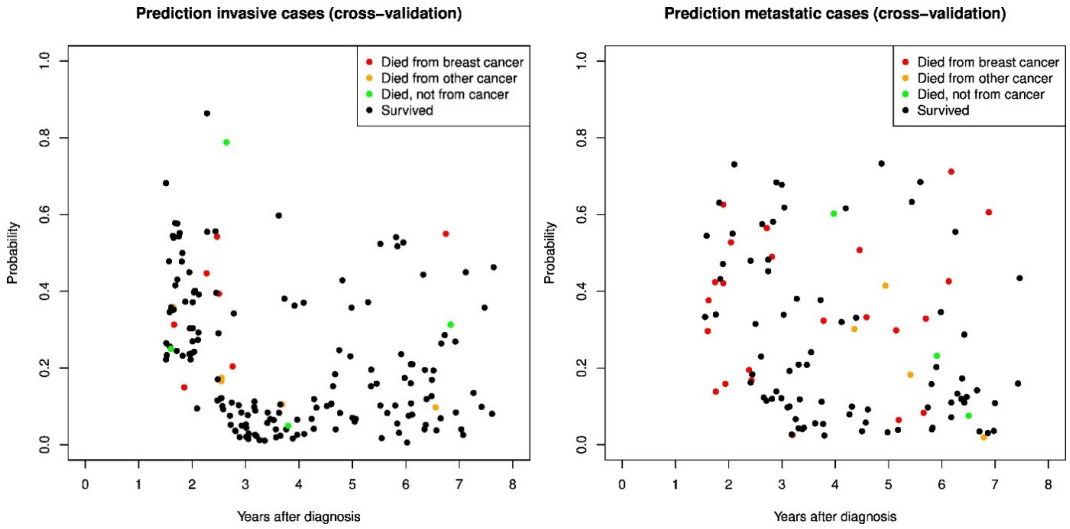

Figure 9.3. The prediction of dying from cancer for the cases who died of breast cancer, died of other types of cancers and died, but not from cancer, and cases who survived for each of the 278 case-control pairs. The figure to the left shows predictions for cases with invasive cancer, while the figure to the right shows prediction for cases with metastatic cancer.

Based on these variables, we find $P_{c,a}$, S_o and the p-value $p = P(S < S_o)$ based on the 10 000 randomization of $P_{c,a}$ We find the p-value less than 0.004 indicating that the prediction is far from random. Table 9.5 gives another presentation of the prediction based on whether $P_{c,a} > 0.3$ or not.

Table 9.5. Prediction for each of the 278 cases based on a threshold equal 0.3

	Sum	$P_{a,c} > 0.3$	$P_{a,c} < 0.3$
Cases who died of cancer	39	22	17
Cases who did not die of cancer	239	75	164

Increasing the thresholds from 0.3 will decrease both the number of true positive and the number of false positive. The threshold 0.3 is chosen as a balance between false positive and false negative. This gives a sensitivity equal 0.56 and specificity equal 0.69. The mean prediction value for the 39 cases who died is 0.32 and the mean prediction value for the 239 cases who survived is 0.23. The predictions are also shown in the boxplot in Figure 9.4.

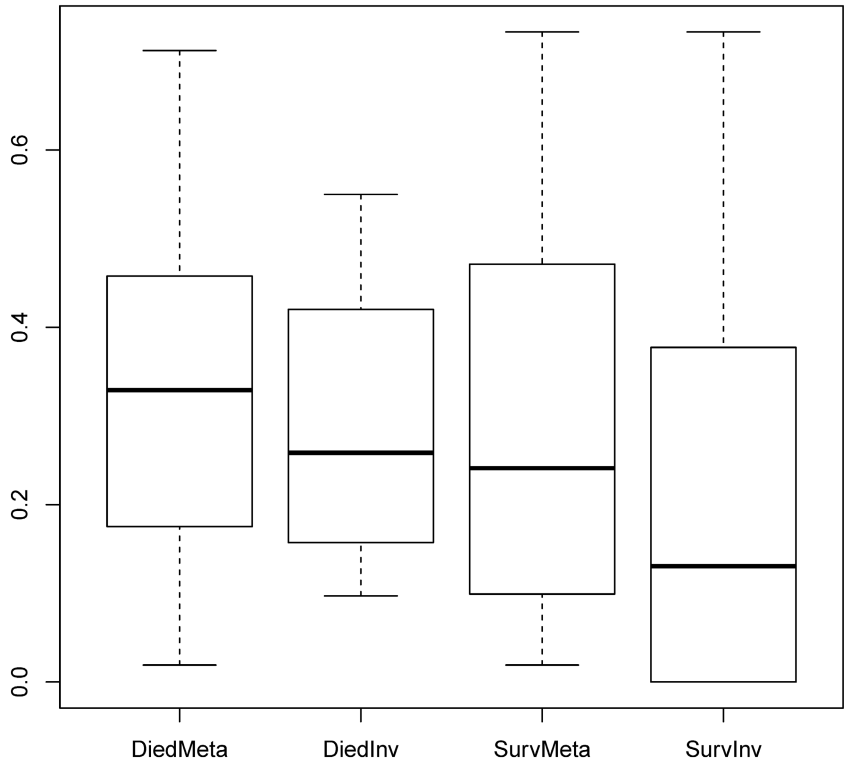

Figure 9.4. Boxplot of prediction of death from cross-validation of cases after 18 months from diagnosis. Horizontal lines describe 0.25, 0.5 and 0.75 quantiles. The number of cases and mean in the four categories are metastatic cancer where case died, no: 27, mean 0.33, invasive cancer where case died, no: 12, mean 0.29, metastatic cancer where case survives, no: 82, mean: 0.29, invasive cancer where case survives no: 157, mean: 0.22.

Notice that the invasive cases who died and the metastatic cases who survived have a relatively similar prediction which is between the prediction of the metastatic cases who died and the invasive cases who survived.

DISCUSSION

We have shown that the trajectories of gene expression after diagnosis of breast cancer were mostly significantly upregulated for hundreds of genes in the years after a diagnosis of metastatic breast cancer compared to invasive cancer, as shown in Figure 9.4. These signals may be considered as signals of an upcoming death due

to cancer. Fewer genes were downregulated. After some years, most upregulated genes levelled off while downregulated genes slowly returned to the normal expression level. Among women with invasive breast cancer, no significant trajectories were found. These results were based on a new statistical approach using the differences in the area between the trajectories of gene expression between diseased and healthy women.

For practical and economic reasons, only one single measurement at time of inclusion was available for each individual in the NOWAC postgenome cohort. Hence, the processual approach relies on the assumption that the gene expression in distinct individuals at different times before or after diagnosis is a consequence of the same carcinogenic process (Lund & Plancade, 2012). Semi-parametric models with time-varying covariates, e.g. the Cox model (Cox, 1972), cannot be estimated from a prospective design including only one unique measurement at time of inclusion, unless covariates are assumed to be constant over time. Consequently, this assumption would not allow us to address changes in gene expression over time.

The DTDS is a further development of the LITS method (Holden, Holden, Olsen, & Lund, 2017), where we used a moving window and summary statistics for all genes for each of the stratum and time period. The genes that were significant in each time interval varied between the intervals, making the LITS method not suitable for identifying genes with different time development. In contrast, the DTDS method is able to identify genes with different time development. Both methods use the same method for simulation and randomization of gene expressions between the case-control pairs with cases from the different strata.

The distribution of measurements of gene expression must follow a constant function, i.e. with measurements spaced over the time interval. Most cohort studies have repeated measurements, but usually they are collected for all participants with several years of spacing and can be used as repeated measurements only.

We cannot predict the outcome for single individuals, only on a group level. The results can be looked upon as a proof of concept for the idea that gene expression measured repeatedly over time after diagnosis can be used as a predictive test for the vital outcome.

Little is known about the changes in gene expression in the blood in the period after a breast cancer diagnosis, i.e. the time period after the primary treatment (Lund & Plancade, 2012). In the stratified analysis, both invasive and metastatic cases were compared to healthy women without known cancers. The consistent and highly significant differences between the two strata adds information that can be used toward a new hypothesis of metastatic breast cancer and its high

lethality. For hundreds of genes, the integrated area between the two curves for each stratum accumulates during follow-up, indicating ongoing dysregulation of important genes. These strong changes in gene expression from the immune cells can be viewed as signals of upcoming death. The intention here was to explore the unknown trajectories of gene expression after diagnosis of breast cancer. The interpretation of each gene was outside the scope of our exploration. Still, some hypotheses can be put forward.

Human model of carcinogenesis—interpretation of highly expressed genes

No unifying theory exists for human carcinogenesis; the number of proposals is many (Vineis, Schatzkin, & Potter, 2010). To date, most mechanistic or pathways analyses have been experimental in-vitro or animal studies. With the increasing knowledge about human carcinogenesis in tumor tissues or in blood at time of diagnosis, some disturbing facts about the validity of the animal models for human carcinogenesis have been brought up. First, the biology of mice and men is comparatively different (Mak, Evaniew, & Ghert, 2014; Anisimov, Ukraintseva, & Yashin, 2005), and a controversial *Nature* editorial ("Of men, not mice", 2013) advocated the need for human functional studies. Similarly, the translational value of mouse models in oncology drug development was recently questioned (Gould, Junittila, & de Sauvage, 2015). While cancer can be manufactured in mice quite easily, these models do not necessarily apply to humans (Mak et al., 2014). Consequently, an increasing number of studies use functional genomics as biomarkers, looking both at the exposure relationship and the outcome. While interesting, this approach lacks the distinct focus on the time-dependent process of carcinogenesis. Few, if any, prospective studies have been designed for longitudinal analyses of functional genomics related to the processes of carcinogenesis and metastasis.

Table 9.6. Annotated functions of the most significant genes from Table 9.2

CCM2	Regulate angiogenesis and formation of new blood vessels
C14orf45	Gene responsible for cilia orientation. One paper shows as low-expressed gene associated with poor survival in BC (higher number of cylia is necessary for improved migration of breast cancer cells)
ARL4A	Increase cell migration
CBX3	Shown to be overexpressed in BC and associated with low survival, might block differentiation and promote self-renewal of cancer stem cells

FSTL4	Shown to be involved in BC cell migration in mice. Was discussed in relation to late distant metastases in BC here without any conclusions (Mittempergher et al., 2013)
C5orf30	Known to be expressed in BC and especially in lymph-node metastases. Promote inflammation and hypothesized to reduce immune response against cancer cells
RBM4	Known tumor suppressor in BC

The interpretation of these genes points towards important changes in genes known to be affected at breast cancer, and in addition some more general ones.

During the different laboratory steps, several decisions had to be taken on level of noise and the use of specific distribution of noise. Further, since a gene maybe not expressed in all individuals, the percentage of cases or controls with sufficient signals had to be decided. The stronger the criteria moving towards hundred percent, the harder the exclusion.

The strength of the study is the unique biobank created with the purpose of gene expression analysis in peripheral blood. This gave a unique opportunity to study the immune response since the mRNA in blood came from immune cells. This opened for the view that the carcinogenic process not only included exposures to carcinogens, but also has an important counterforce in the immune system. This has been known for more than a hundred years, and today documented through the new immune therapies.

The design has been population-based with a complete follow-up on cancer incidence, emigration and death based on linkage to national registers using the unique national birth number given to all residents in Norway from 1960. In addition, we had access to updated information on metastases and second breast cancers in the time between inclusion and blood donation. This somewhat reduced the noise from carcinogenic processes hidden at the time of diagnosis.

CONCLUSION

In this systems epidemiology approach, we have given a proof of concept for the use of gene expression as an individualized biomarker of prognosis related to death or not. The design of NOWAC is population-based and the results should be validated in a more specific clinical setting. With improved technology and individual repeated measurements gene expression followed over time could offer a unique opportunity for personalized treatment of metastatic breast cancer.

DISCLAIMER

1. Some of the data in this article are from the Cancer Registry of Norway. The Cancer Registry of Norway is not responsible for the analysis or interpretation of the data presented.
2. Microarray service was provided by the Genomics Core Facility, Norwegian University of Science and technology, and NMC—a national technology platform supported by the functional genomics program (FUGE) of the Research Council of Norway.

ACKNOWLEDGEMENTS

We are impressed by and thankful to the women who donated blood for this cancer research project. Bente Augdal, Merete Albertsen, and Knut Hansen were responsible for all infrastructure and administrative issues. This study was supported by a grant from the European Research Council (ERC-AdG 232997 TICE) and a donation from Halfdan Jacobsen og frues legat (The Norwegian Cancer Society)

The funders had no role in the design of the study; in the collection, analyses and interpretation of the data; in the writing of the manuscript; or in the decision to submit for publication.

AUTHORS' CONTRIBUTIONS

EL is PI of the NOWAC Study and initiated the methodological collaboration; LH and MH developed the statistical methods. KSO addressed the gene function. J-CT and L-TB added clinical information. All authors have participated in the discussions and have read and approved the final manuscript.

REFERENCES

Anisimov, V.N., Ukraintseva, S.V., & Yashin, A.I. (2005). Cancer in rodents: does it tell us about cancer in humans? *Nature Reviews. Cancer*, 5(10), 807–819. doi: https://doi.org/10.1038/nrc1715

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 289–300.

Cancer Registry of Norway. (2019). *Cancer in Norway 2018 – Cancer incidence, mortality, survival and prevalence in Norway.* Oslo, Norway: Cancer Registry of Norway. Retrieved from: https://www.kreftregisteret.no/globalassets/cancer-in-norway/2018/cin-2018.pdf

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2), 187–220.

Dumeaux, V., Børresen-Dale, A.L., Frantzen, J.O., Kumle, M., Kristensen, V.N., Lund, E. (2008). Gene expression analyses in breast cancer epidemiology: the Norwegian Women and Cancer postgenome cohort study. *Breast Cancer Research*, 10(1), R13. doi: https://doi.org/10.1186/bcr1859

Gould, S.E., Junttila, M.R., & de Sauvage, F.J. (2015). Translational value of mouse models in oncology drug development. *Nature Medicine*, 21(5), 431–439. doi: https://doi.org/10.1038/nm.3853

Holden, M., Holden, L., Olsen, K.S., & Lund, E. (2017). Local in Time Statistics for detecting weak gene expression signals in blood – illustrated for prediction of metastases in breast cancer in the NOWAC Post-genome Cohort. *Advances in Genomics and Genetics*, 7, 11–28. doi: https://doi.org/10.2147/AGG.S130004

Jeibouei, S., Akbari, M.E., Kalbasi, A., Aref, A.R., Ajoudanian, M., Rezvani, A., Zali, H. (2019). Personalized medicine in breast cancer: pharmacogenomics approaches. *Pharmacogenomics and Personalized Medicine*,12, 59–73. doi: https://doi.org/10.2147/PGPM.S167886

Lund, E., Dumeaux, V., Braaten, T., Hjartåker, A., Engeset, D., Skeie, G., Kumle, M. (2008). Cohort profile: The Norwegian Women and Cancer Study--NOWAC--Kvinner og kreft. *International Journal of Epidemiology*, 37(1), 36–41. doi: https://doi.org/10.1093/ije/dym137

Lund, E., & Plancade, S. (2012). Transcriptional output in a prospective design conditionally on follow-up and exposure: the multistage model of cancer. *International Journal of Molecular Epidemiology and Genetics*, 3(2), 107–114.

Lund, E., Holden, L., Bøvelstad, H., Plancade, S., Mode, N., Günther, C.C., ... Holden, M. (2016). 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. *BMC Medical Research Methodology*, 16, 28. doi: https://doi.org/10.1186/s12874-016-0129-z

Mak, I.W., Evaniew, N., & Ghert, M. (2014). Lost in translation: animal models and clinical trials in cancer treatment. *American Journal of Translational Research*, 6(2), 114–118.

Mittempergher, L., Saghatchian, M., Wolf, D.M., Michiels, S., Canisius, S., Dessen, P., …van't Veer, L.J. (2013). A gene signature for late distant metastasis in breast cancer identifies a potential mechanism of late recurrences. *Molecular Oncology*, 7(5), 987–999. doi: https://doi.org/10.1016/j.molonc.2013.07.006

Norwegian Institute of Public Health (n.d.). Causes of death & Life expectancy. [Internet]. Accessed: 10.10.2019. Retrieved from: http://www.fhi.no/en/hn/cause-of-death-and-life-expectancy/

Of men, not mice [Editorial]. (2013). *Nature Medicine*, 19(4), 379. Retrieved from: https://www.nature.com/articles/nm.3163

Reddy, S.M., Barcenas, C.H., Sinha, A.K., Hsu, L., Moulder, S.L., Tripathy, D., … Valero, V. (2018). Long-term survival outcomes of triple-receptor negative breast cancer survivors who are disease free at 5 years and relationship with low hormone receptor positivity. *British Journal of Cancer,* 118(1), 17–23. doi: https://doi.org/10.1038/bjc.2017.379

Reiner, A., Yekutieli, D., & Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics (Oxford, England)*, 19(3), 368–375. doi: https://doi.org/10.1093/bioinformatics/btf877

Spitz, M.R., & Bondy, M.L. (2010). The evolving discipline of molecular epidemiology of cancer. *Carcinogenesis*, 31(1), 127–134. doi: https://doi.org/10.1093/carcin/bgp246

Stroncek, D.F., Butterfield, L.H., Cannarile, M.A., Dhodapkar, M.V., Greten, T.F., Grivel, J.C., … Seliger, B. (2017). Systematic evaluation of immune regulation and modulation. *Journal for Immunotherapy of Cancer*, 5, 21. doi: https://doi.org/10.1186/s40425-017-0223-8

Vineis, P., Schatzkin, A., Potter, J.D. (2010). Models of carcinogenesis: an overview. *Carcinogenesis*, 31(10), 1703–1709. doi: https://doi.org/10.1093/carcin/bgq087