Open access

Eiliv Lund (Ed.)

ADVANCING SYSTEMS EPIDEMIOLOGY IN CANCER

Exploring Trajectories of Gene Expression

SCANDINAVIAN UNIVERSITY PRESS

Advancing Systems Epidemiology in Cancer

Eiliv Lund (Ed.)

Advancing Systems Epidemiology in Cancer

Exploring Trajectories of Gene Expression

Scandinavian University Press

This book was first published in 2020 by Scandinavian University Press.

This book is published with financial support of The Margit and Halfdan Jacobsen trust.

ISBN printed edition (print on demand): 978-82-15-04120-9
ISBN electronic pdf-edition: 978-82-15-04119-3
DOI: 10.18261/9788215041193-2020

Enquiries about this publication may be directed to:
post@universitetsforlaget.no.

www.universitetsforlaget.no

Cover: Scandinavian University Press
Prepress: Teksflyt AS

Contents

Lund, E. (Ed.) (2020). *Advancing Systems Epidemiology in Cancer. Exploring Trajectories of Gene Expression.*
Oslo: Scandinavian University Press.
DOI: https://doi.org/10.18261/9788215041193-2020-00

Preface

In the autumn of 2007, the EU launched its new European Research Council. One of the new funding mechanisms was the "Advanced researcher grants", so-called ERC AdGs. Applications were to be written by one researcher with one novel idea. I spent that same autumn as a visiting researcher at MAP5, Université René Descartes, Paris V. I immediately felt that I had an idea that was unique and novel. I called it Systems Epidemiology. Together with colleagues we discussed the content of the application, particularly the need for novel statistical methods. Back in Tromsø, a small writing group continued the work of searching through all databases in order to be sure that we had a novel idea. With that in mind I submitted the application under the title "Transcriptomics in cancer epidemiology"—TICE—in winter 2008. Six months later we received what was, for us, the sensational news that our application had been accepted—but it had to be a single-researcher project.

The project started in the winter of 2009, a decade ago. Here we will look at the potential of our novelties. The reviewers used a term that inspired us: high risk—high gain. Our intention is to demonstrate that the gain has been high. In 2015 we received a "Proof of concept" ERC grant for the development of TICE findings towards a diagnostic patent for breast cancer based on gene expression profiles. We acknowledge all of the women who gave their time to fill out questionnaires, and donated blood and even tumor and healthy breast tissues biopsies.

I am thankful to all those I have held discussions with over these years, sometimes ending in nothing, sometimes another step forward.

Eiliv Lund
The Cancer Registry of Norway, Oslo
UiT The Arctic University of Norway, Tromsø

Lund, E. (Ed.) (2020). *Advancing Systems Epidemiology in Cancer. Exploring Trajectories of Gene Expression.* Oslo: Scandinavian University Press.
DOI: https://doi.org/10.18261/9788215041193-2020-01

1. Challenges in Systems Epidemiology in Cancer

Eiliv Lund

Abstract Systems epidemiology is a new research discipline that seeks to integrate pathways analyses into observational study designs to improve the understanding of biological processes in human organisms as time-dependent changes or trajectories of functional genomics. This chapter guides the reader on the different aspects of the book. The aim is to improve the understanding of the structures of data from complex study designs, data handling, new statistical methods and the interpretation of the results.

Keywords trajectories | processual research | analogy | carcinogenesis | statistical methods

ABOUT THE BOOK

This chapter gives an overview of challenges to improving systems epidemiology in cancer. As a new research discipline, systems epidemiology has confronted many problems related to design, laboratory work, statistics and biological interpretation. It is important throughout this book that most of the gene expression analyses originate from whole blood, thus representing the body's defense through the immune system. This response is not a copy of the carcinogenic process.

With the description of the human genome in 2001 (Lander et al. 2001, Venter et al. 2001), medical scientists were promised a dramatic paradigmatic shift in future research. In epidemiology at that time, the main focus was on risk estimation. After the Second World War, epidemiology shifted from studies of infectious diseases towards risk factors for cancer, from studies of tuberculosis to long-term effects of carcinogens. The first success in this risk estimation era was the finding that the rapidly increasing incidence of lung cancer in many western countries was due to smoking. It is almost 70 years since Doll and Bradford Hill published their landmark paper linking smoking to lung cancer (Doll and Bradford Hill 1950).

There then first followed a huge number of case-control studies, with a later shift towards large prospective studies covering most aspects of modern lifestyle such as diet, hormonal treatment, physical activity, body mass index, alcohol and so on. Almost all analyses used linear relationships, mostly proportional hazard models and multiplicative models.

In the genomic research era that followed, single nucleotide polymorphisms, (SNPs) became the major interest for studies of individual risk for chronic diseases such as cancer. In traditional epidemiology, an SNP is an ideal exposure, easy to measure and constant throughout life. The genomic research fitted well into the standard statistical methods, only confronted by the multiple test problem (Reiner et al. 2003). Studies of functional genomics, here mRNA (messenger Ribonucleic acid), miRNA (micro Ribonucleic acid) and methylation, had only been part of epidemiology to a minor extent, and then in the same risk context as other exposures. The challenge of performing studies of the time-dependent dynamics of functional genomics was left unresolved. In 2008, systems epidemiology was defined as studies of functional genomics in an epidemiological design (Lund and Dumeaux 2008). The concept was dedicated to a new research discipline different from the traditional risk estimation paradigm of epidemiology. Focus was on the description of changes in functional genomics as part of the study of the carcinogenic process over time (Lund et al 2015). Ten years later, the definition provided by National Cancer Institute, NCI (Workshop to Facilitate Cancer Systems 2019) was research

> directed towards systems modelling based on approaches which account for multiple dimensions, integration over diverse data and changes over time, all needed to better understand contributors to disease and treatment outcomes and provide clues for improved intervention.

The NCI text from the invitation to the meeting Workshop to Facilitate Cancer Systems Epidemiology Research is included as a supplement at the end of this chapter, together with their selected Systems Epidemiology Key references showing an overview of relevant literature.

The introduction of new sampling methods of human biological material, high-throughput technologies and new laboratory analyses have made possible new studies of functional genomic. There are important differences between the time-dependent changes within the functional classes of transcriptomics i.e. mRNA and miRNA, changing rapidly, even during the same day, and methylation changing over years (Guida et al. 2015).

The new research discipline, systems epidemiology, seeks to integrate pathway analyses into observational study designs to improve the understanding of biological processes in the human organism. Systems epidemiology is the observational counterpart to systems biology, which has many definitions, such as

> a discipline that seeks to determine how complex biological systems function by integrating experimentally derived information through mathematical and computing solutions (Imperial College London, Institute of Systems and Synthetic Biology).

One could eliminate both terms and use the common term "systems science" (Green 2006), but this would not emphasize the emerging approaches and designs necessary for the optimal use of new technologies. Systems epidemiology could be seen as a discipline that merges epidemiologic research with biological mechanistic analysis by investigating gene expression patterns related to metabolic pathways.

The aim of this book is to advance systems epidemiology in cancer. After some years of research in systems epidemiology in cancer, several issues have been solved, but in order to advance the following major issues will be proposed as future challenges.

HIGH DIMENSIONAL, TIME-DEPENDENT DYNAMIC CURVES—THE PROCESSUAL CHALLENGE

The conceptual differences between risk-estimating epidemiology and functional genomic epidemiology can best be described through some added definitions and terminology. A more comprehensive description is found in Lund et al. (2015). First, the curves of functions describing time-dependent gene expression or methylation profiles can be named as trajectories. These trajectories have mostly unknown distribution according to lifestyle or outcome. At present there exists no library of gene expression profiles according to different exposures.

The change in approach can be illustrated through the change in the mathematical model.

A classic prospective GWAS (genome-wide association studies) design includes genomics and exposures measured at the start of the study and uses the time elapsed since the beginning of the study, defined as the follow-up time, in survival analyses. The failure time for a case-control pair corresponds to the diagnosis of the case. The main issue adopted is as follows: Given the values of some covari-

ates—genomics and environmental—what is the risk of developing a cancer at some time? Thus, genomics and exposure variables are considered as risk factors for cancer, and the relationship may be expressed in terms of a survival analysis model:

$$P[T|G,E]$$

where T is the failure time, G the genomics measurements, and E the exposures.

The processual analysis of transcriptomics raises a different question: How are transcriptomics affected by the carcinogenic process? Transcriptomics are therefore analyzed as potential biomarkers of the carcinogenic process, and the statistical quantity of interest is the distribution of the gene expression GE as a function of the time to diagnosis T and the exposures E:

$$P[GE|T,E]$$

Ideally, repeated measurements should be available, but for practical reasons a single measurement at time of inclusion may be the only one for each individual. In this case, transcriptome measurements collected from distinct individuals at different times before diagnosis may be considered as consequences of the same carcinogenic process. This point of view is commonly adopted in lab experiments, e.g. when dissections performed at different time points on different animals are analyzed as a longitudinal study. In an epidemiological context the individual variability is expected to be much higher due to the heterogeneity of cancer. This approach relies on the assumption that available information on the outcome allows stratification according to different biological processes, such as positive or negative node status at time of diagnosis.

NOVEL DESIGNS—INTEGRATED SYSTEMS EPIDEMIOLOGY APPROACH

The inclusion of biological material for analyses of transcriptomics both in peripheral blood and in tissues such as breast cancer tumors has created what can be called the beauty of complex designs. This is demonstrated in Chapter 2. At the same time, due to the complexity of the data and the scientific need to store previous work for later replication, a strong platform for data storing and handling has been developed; see Chapter 3. The logistics and ethical issues for sampling biopsies from women is a necessary part of the realization of the new designs for anal-

yses. The procedures for sampling biopsies from breast tissue in healthy women, with related ethical questions, is given in Chapter 4.

For many of the functional genomic analyses, the number of other studies is limited due to the cost of laboratory analyses or limitations in access to biological samples or samples collected under suboptimal conditions. Since different cohorts reflect different populations with different lifestyles, this implies that the estimates should be adjusted for other risk factors. But with little knowledge about the gene expression pattern of most lifestyle factors, this is an uncertain undertaking. As an example, should all users of hormonal replacement therapies be excluded due to unknown effects on gene expression pattern?

Partly to solve the problems of lack of power and mass significance, a novel design has been proposed by us (Lund et al. 2018), named an integrated systems epidemiology approach. The integrated systems epidemiology approach is based on a two-step integrated analysis in the same cohort. The first step is the exploration of hypotheses describing the mathematical relationship between an exposure such as parity and breast cancer incidence in the overall study. This should be done in a large study in order to give unbiased and precise estimates of effects. The results of the explorative research can then be used as hypotheses in the second level of analyses. In a random sub-cohort of the overall cohort, blood sampling can be performed with biological samples with high quality to test the hypotheses. The hypotheses of the relationships between the exposures, the gene expression and outcome can then be tested directly. The random sampling will control the level of lifestyle factors between the full cohort and the sub-cohort. This will be most important with weak or small associations, which could be the reality for most associations between exposures and gene expression. This will reduce the need for adjustment. An example of such an analysis is shown in Lund et al. 2018, demonstrating the statistical strength of the design. Each pregnancy showed a reduction in breast cancer risk of 8% in a linear model. Translated into gene expression as a hypothesis, the testing showed that hundreds of genes showed a similar linear reduction.

STATISTICAL CHALLENGES OF SPARSELY SAMPLED CURVES

For most cancer sites, the incidence rates are lower than 300 per 100 000 person years. Consequently, a cohort study must be rather large, such as the NOWAC postgenome cohort with 50 000 women. The power of the study is improved due to the selection of only one gender. Still, the distribution of new cases, with

breast cancer as an example, will be around 150 cases annually. The age distribution in postmenopausal breast cancer increases only slowly with age. The problem of small data and statistical power in statistical analyses of genomic data is debated in Chapter 5, with emphasis on explorative versus hypothesis-driven research.

Information on changes in gene expression over time will consequently be based on rather few measurements. This gives sparsely sampled curves or trajectories. This has motivated the three new statistical methods described in Chapter 6. Increasing the follow-up time expands the time dimension, but does not increase the incidence rate, i.e. the sparse density remains the same. In a standard cohort design with repeated sampling of biological material with 4-6 years intervals, constructing the trajectories will be almost impossible due to a lack of measurements in between the repeated measurements.

Another challenge is two major issues working mostly together. First, the distribution or curve shapes of the trajectories are mostly unknown in relation to different exposures or outcomes like cancer. To assume linearity would be a simplification. At the same time, the number of genes is up to 20 000, giving a false positive problem. Using FDR (false discovery rate) is a conservative procedure in explorative research. Each explorative analysis or subgroup analysis or stratification generates an immense cloud of results, approaching chaos.

Since a p-value can only be used once, new approaches are important in a situation where the explorative analyses are expected to be followed by predictions in the same material. Today, there are limited numbers of large cohort studies with large-scale analyses i.e. analyses of thousands of individuals for test-retest analyses. One proposed solution is leave-one-out procedures. These stratifications reduce the statistical power dramatically. The use of the curve group analysis in Lund et al. 2016 Chapter 8 is an example of using different measures at each level of the analysis; first mean and standard deviations for grouping of curves, then a significance test using p-values.

FROM MICE TO WOMEN—SEVEN LEVELS OF ANALOGICAL FALLACIES

"Analogy is a comparison between things that have similar features, often used to help explain a principle or idea" (Cambridge Dictionary). The use of analogies is common in daily life and in science. However, the validity of the analogies should always be considered; see Chapter 7. There are no standard criteria for valid transfer over species or for the use of analogies. The interpretation of pathways or

single genes is traditionally proposed by reviewers expressing a belief in the validity for humans of the genetic findings from mostly animal or in vitro cell experiments. The lack of consistency has been described in an analysis of single genes in the epidemiological study of pregnancies and risk of breast cancer in Lund et al. 2018.

The rapidly growing interest and potential for analyses of functional genomics in human studies, systems epidemiology, confronts researchers with interpretations of statistical associations based on biological knowledge. There are differences in human versus animal or in vitro experiments, between the observational study designs in epidemiology and the experimental designs in basic biological research, and in the interpretations of findings especially concerning the comparability or validity of transferring information from one biological species to another—humans. The transfer is often based on analogical thinking, often with little knowledge about the nature of these biological and methodological differences. Analogy was one of Bradford Hill's original criteria of causality (Bradford Hill 1965, Häfler 2005, Fedak et al. 2015), but was dismissed after some years and lost its meaning due to the potential for fallacious thinking, and it became difficult to assume that results from one study could be generalized to other research areas.

Today, epidemiologists working with analyses of functional genomics import knowledge of function from databases with information on basic biology. This might be experiments on mice, cell lines or humans. Our concern is that the information collected for systems biology cannot necessarily be used for interpretations of the biology of statistical associations. In order to avoid analogical fallacies, we should aim to classify more of the information in databases according to human parameters. What is the sum total of the conditions around each experiment compared to human observations?

Taking gene functions from mice to humans or vice versa is a journey through many levels of analogical fallacies. The critical view on the transfer of knowledge over the borders of species has grown over recent years as a consequence of the rapidly increasing use of annotation of gene functions in the interpretation of results from clinical or epidemiological studies.

EXPLORING THE TRAJECTORIES OF GENE EXPRESSION

Examples of statistical methods for the analyses of trajectories are shown in Chapters 8 and 9. The chapters illustrate two new methods for longitudinal analyses of gene expression trajectories before and after a diagnosis of breast cancer.

TWO THEORIES OF IMMUNE DYSFUNCTION AND RESISTANCE DURING CARCINOGENESIS

Over recent decades, many theories of carcinogenesis have been proposed. Common to most of them is the lack of integration of knowledge from many scientific research disciplines such as epidemiology, biostatistics, immunology, basic biology and clinics. Consequently, cancer could be viewed as an intracellular process or as an exposure-driven process. Models have been built on mathematical modelling of incidence rates or hallmarks of cancer. On the other hand, lay people talk about the disease using the metaphor of "the war on cancer", in French "la dernière lutte" or in Norwegian "sin livs kamp". The essence of these metaphors has been neglected both historically and today, despite over 120 years of diverse experiences. The first observations of the potential effect of the immune system on cancer was back in 1895, when injections of serum from infested mice led to the reduction of tumor masses in humans (Kaplon and Dieu-Nosjean 2018). In the interwar period, a number of experiments were performed and a toxin was even produced as a treatment (Kucerova and Cervinkova 2016). Unfortunately, many of the experiments went wrong, with death of the cancer patient due to infection rather than due to the cancer. After the Second World War this became an obscure idea as cancer was considered as a multistage process that was in some ways deterministic (Tomasetti and Vogelstein 2015, Perduca et al. 2019).

In systems epidemiology the fundamental change is the methodological position, from a study design with information only on the driving forces of carcinogenesis (Lund 2011) to the implementation of measures of the immune response, mainly through gene expression analyses in peripheral blood.

While the research on carcinogens has been both prioritized and highly valuable, any understanding of the carcinogenic process must involve measures of the immune defense. Immune evasion as a concept was introduced a few years ago (Hanahan and Weinberg 2011), but at that time with limited knowledge about its mechanisms. However, the idea of carcinogenesis as a war between the tumor cells and the immune system should take into account the lack of consistency between the gene expression in the tumor and in peripheral blood. Instead, these two biological tissues represent these two forces.

Further support for this theory is the increasing importance of infectious agents as risk factors for cancers, such as helicobacter pylori and stomach cancer, HPV (Human papillomavirus) and cervical (and now also pharyngeal) cancer, hepatitis C, liver cancer and Burkitt's lymphoma. Recently there have been indications of an association between leukemia among children and viruses, demonstrating the importance of the immune system for development of several cancer sites. Vacci-

nation against HPV clearly demonstrates the potential of using the immune system for prevention, and possibly also for treatment.

The assembled evidence supports the hypothesis or model of cancer development as balancing act or war between two opponents. Most breast cancers remain invasive and will only kill women slowly due to ulcerations and infections. Metastases, however, will kill.

Understanding the balancing act between the two dimensions—the tumor gene expression profiles representing the carcinogenic process as the aggressor, and the immune cells' gene expression representing the defense—might be vital for the future success of cancer research.

In Chapter 10, the old carcinogenic theory of aviatism is therefore discussed. The atavistic hypothesis postulates that white blood cells responses to cells under threat continue to run their core functionalities, preserving its most vital functions. The findings are essentially in accordance with the atavistic hypothesis, showing that the gene expression of white blood cells under threat from a cancer run their core functionalities, preserving their most vital functions. Chapter 11 is a model of carcinogenesis based on the overall findings in TICE in relation to the complexities of the evolution of the immune system, tumor tissue gene expression, the carcinogens, and the immune response as seen in peripheral blood.

WHEN SYSTEMS EPIDEMIOLOGY MERGES WITH SYSTEMS BIOLOGY

So far, information on functional genomics runs from reductionist experiments with animals or cells to epidemiological findings. Huge repositories have collected all information from experiments or from systems biology. In epidemiology this information is used to explain the findings from explorative epidemiological studies with analogical fallacies as a potential source of poor validity of interpretations. However, recently the process was reversed. After several studies of mice in relation to systemic dysfunction and plasticity of the macroenvironment in mice models, the authors turned to open access data from a case-control study of breast cancer in the NOWAC postgenome biobank. The human information was used as a validation of the mice results (Allen BM et al., Nature Medicine). This opens up for complicated study designs combining mice experiments with systems epidemiology studies.

THE FUTURE

The list of challenges could have been extended with several others, such as changing technologies and multilevel omics. Both issues would have added highly specific technical and statistical methods, considered to be outside the scope of explaining systems epidemiology. The focus on transcriptomics was chosen as it most clearly defines the challenges of time-dependent functional genomic studies.

EPILOGUE: EXPLORATIVE RESEARCH AS SAILING IN UNCHARTED WATERS

The ten years of our ERC project TICE has a metaphor in the North Pole expedition of Frithjof Nansen—ten years in uncharted waters. Everybody could dream of reaching the pole, but nobody had managed to survive. Nansen had extensive experience of travel in the North and had one observation—perhaps he was the only one who understood the consequences: that a small part of the French ship *Jeanette,* which was crushed by the ice, took three years to travel from East Siberia, where it had sunk, across the Arctic Ocean, ending up in Greenland. He then went to a famous Norwegian shipwright, who constructed the ship *Fram*. The vessel was constructed like a nutshell in order to withstand the pressure of the ice and serve as a base during his years of drifting. But he ignored all polar people who told him that the Arctic Ocean was shallow, perhaps no more than 100–200 meters deep, and that his boat would be destroyed against a rock or an island. Nansen found that the Polar sea was 3000 meters deep. As he departed for an expedition that would last 4–5 years, he claimed that reaching the North Pole would be nice, but the real endeavor was to explore the Arctic Ocean. He returned without reaching the pole, but with a wealth of knowledge that others were able to use.

REFERENCES

Allen BM, Hiam KJ, Burnett CE, Venida A, DeBarge R, Tenvooren I et al. Systemic dysfunction and plasticity of the immune macroenvironment in cancer models. Nature Medicine 2020

Bradford Hill A. The Environment and Disease: Association or Causation? Proc R Soc Med 1965 May; 58: 295–300. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1898525/

Cambridge Dictionary [Internet]. Accessed: 15.11.2019. Available from: https://dictionary.cambridge.org/dictionary/english/analogy

Doll R, Bradford Hill A. Smoking and Carcinoma of the Lung. Br Med J. 1950 Sep 30; 2(4682): 739–748. Available from: https://www.bmj.com/content/2/4682/739

Fedak KM, Bernal A, Capshaw ZA, Gross S. Applying the Bradford Hill criteria in the 21st century: how data integration has changed casual interference in molecular epidemiology. Emerg Themes Epidemiol. 2015; 12: 14. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4589117/

Green LW. Public health asks of systems science: to advance our evidence-based practice, can you help us get more practice-based evidence? Am J Public Health. 2006 Mar; 96(3): 406–409. Available from: https://ajph.aphapublications.org/doi/full/10.2105/AJPH.2005.066035?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed&

Guida F, Sandanger TM, Castagné R, Campanella G, Polidoro S, Palli D, et al. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. Hum Mol Genet. 2015 Apr 15; 24(8): 2349–2359. Available from: https://academic.oup.com/hmg/article/24/8/2349/652785

Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011 Mar 4; 144(5): 646–674. Available from: https://www.sciencedirect.com/science/article/pii/S0092867411001279?via%3Dihub

Holden M, Holden L, Olsen KS, Lund E. 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. 2017; 7: 11–28. Available from: https://www.dovepress.com/local-in-time-statistics-for-detecting-weak-gene-expression-signals-in-peer-reviewed-article-AGG

Höfler N. The Bradford Hill considerations on causality: a counterfactual perspective. Emerg Themes Epidemiol. 2005 Nov 3; 2: 11. Available from: https://ete-online.biomedcentral.com/articles/10.1186/1742-7622-2-11

Imperial College London, Institute of Systems and Synthetic Biology [Internet]. Accessed: 15.11.2019. Available from: http://www.imperial.ac.uk/systems-biology/about-the-institute/

Kaplon H, Dieu-Nosjean MC. Quelle avenir pour les lymphocytes B infiltrant les tumeures solides. MedSci (Paris). 2018 Jan; 34(1): 72–78. Available from: https://www.medecinesciences.org/articles/medsci/pdf/2018/01/medsci20183401p72.pdf

Kucerova P, Cervinkova M. Spontaneous regression of tumour and the role of microbila infection – possibilities for cancer treatment. Anticancer Drugs. 2016 Apr; 27(4): 269–277. Available from: https://insights.ovid.com/article/00001813-201604000-00001

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15; 409(6822): 860–921. Available from: https://www.nature.com/articles/35057062

Lund E. An exposure driven functional model of carcinogenesis. Med Hypotheses. 2011 Aug; 77(2): 195–198. Available from: https://www.sciencedirect.com/science/article/pii/S0306987711001721?via%3Dihub

Lund E, Dumeaux V. Systems epidemiology in cancer. Cancer Epidemiol Biomarkers prev. 2008 Nov; 17(11): 2954–2957. Available from: https://cebp.aacrjournals.org/content/17/11/2954.long

Lund E, Holden L, Bøvestad H, Plancade S, Mode N, Günther CC et al. 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 Med Res Methodol. 2016 Mar 5; 16: 28.

Available from https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0129-z

Lund E, Nakamura A, Snapkov I, Thalabard JC, Olsen KS, Holden L, et al. Each pregnancy linearly changes immune gene expression in the blood of healthy women compared with breast cancer patients. Clin Epidemiol. 2018 Aug 6; 10: 931–940. Available from: https://www.dovepress.com/each-pregnancy-linearly-changes-immune-gene-expression-in-the-blood-of-peer-reviewed-article-CLEP

Lund E, Plancade S, Nuel G, Bøvelstad H, Thalabard JC. A processual model for functional analyses of carcinogenesis in the prospective cohort design. Med Hypotheses. 2015 Oct; 85(4): 494–497. Available from: https://www.sciencedirect.com/science/article/pii/S0306987715002704?via%3Dihub

Perduca V, Alexandrov LB, Kelly-Irving M, Delpierre C, Omichessan H, Little MP, et al. Stem cell replication, somatic mutations and role of randomness in the development of cancer. Eur J Epidemiol. 2019 May; 34(5): 439–445. Available from: https://link.springer.com/article/10.1007%2Fs10654-018-0477-6

Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics. 2003 Feb 12; 19(3): 368–375. Available from: https://academic.oup.com/bioinformatics/article/19/3/368/258230

Tomasetti C, Vogelstein B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science. 2015 Jan 2; 347(6217): 78–81. Available from: https://science.sciencemag.org/content/347/6217/78.long

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG et al. The sequence of the human genome. Science. 2001 Feb 16; 291(5507): 1304–1351. Erratum in: Science. 2001 Jun 5;292(5523):1838. Available from: https://science.sciencemag.org/content/291/5507/1304.long

Workshop to Facilitate Cancer Systems, Epidemiology Research, February 28 – March 1, 2019, Bethesda MD. National Cancer Institute, Division of Cancer Control & population Sciences, Biology and Genomics Research Program. Accessed: 15.11.2019. Available from: https://epi.grants.cancer.gov/events/systems-epidemiology/?cid=eb_govdel_en_egrp_newsletter_jan19

Lund, E. (Ed.) (2020). *Advancing Systems Epidemiology in Cancer. Exploring Trajectories of Gene Expression.* Oslo: Scandinavian University Press.
DOI: https://doi.org/10.18261/9788215041193-2020-02

2. The Beauty of Complex Designs

Jo Inge Arnes and Lars Ailo Bongo

Abstract The increasing use of omics data in epidemiology enables many novel study designs, but also introduces challenges for data analysis. We describe the possibilities for systems epidemiological designs in the Norwegian Women and Cancer (NOWAC) study and show how the complexity of NOWAC enables many beautiful new study designs. We discuss the challenges of implementing designs and analyzing data. Finally, we propose a systems architecture for swift design and exploration of epidemiological studies.

Keywords Systems epidemiology | Norwegian Women and Cancer | study designs | hypothesis exploration | computer systems

INTRODUCTION

Analytical observational epidemiology was, and primarily still is, about disease risk estimation. In the past, most studies used simple case-control designs with data from questionnaires, registers, and health records. The analyses relied on Cox and classical survival analysis methods. Because case-control designs are prone to selection and recall bias, prospective cohorts with nested designs are increasingly used, but typically still focus on risk estimation. However, there is a shift in epidemiology towards more basic research in which we study how diseases affect biological systems at a biomolecular level over time – for example, to understand the dynamics of human carcinogenesis.

This shift was motivated by the sequencing of the human genome, officially completed in April 2003 (The Human Genome Project), which led to the incorporation of genetic variants into epidemiological studies, primarily single nucleotide polymorphisms (SNPs). SNPs are ideal as exposures because they do not change over a lifetime. Hence, risks can be estimated using classical statistical methods. There are also many hospital and research biobanks with samples usable for SNP

analyses, such as biobanks incorporated in the European Prospective Investigation into Cancer and Nutrition (EPIC) (Bingham and Riboli 2004). In the ensuing decade, considerable resources were spent on genome-wide association studies (GWAS), but the studies repeatedly failed to find robust, replicable associations between SNPs and common diseases (Lund and Dumeaux 2008). The focus, therefore, shifted to functional genomics to find biological markers associated with environmental exposures, lifestyle, age, or disease.

In 2008, Lund and Dumeaux (Lund and Dumeaux 2008) introduced systems epidemiology and proposed the globolomic design. Systems epidemiology incorporates functional genomics and observes how diseases affect human biological systems over time. The globolomic design extends the existing prospective design by integrating functional genomics analyses from blood and tissue. In 2015, Lund, with collaborators, introduced a processual approach to systems epidemiology (Lund et al. 2015). The processual approach differs from traditional risk-related research in that we view disease as a multi-stage process and use functional genomics to observe disease-associated changes over time. In connection with the new direction in epidemiology, there was a need for new statistical methods. An example is a statistical method for longitudinal gene expression analysis using the concept of curve groups (Lund et al. 2016, Chapter 8), developed in cooperation with the Norwegian Computing Center.

Omics (Vailati-Riboni et al. 2017) plays an essential part in systems epidemiology. The different omics are, unlike genes, affected by exposures and diseases. By integrating omics in nested case-control studies, we can find altered levels of gene expressions or methylation that are biological markers of the disease. For example, studies have discovered changes in pre-diagnostic DNA methylation associated with breast and lung cancer risk (Baglietto et al. 2017, Fasanelli et al. 2015, van Veldhoven et al. 2015). Other studies have found changes in the inflammatory transcriptome in adults related to early-life socioeconomic status (Castagne et al. 2016). We can also use other types of biological data that contain changes associated with a disease, including epigenetics, gene expressions, proteins, and metabolites. Finally, we can combine different types of omics and observe them together in a multi-omics approach (Hasin et al. 2017).

In systems epidemiology, we observe how diseases affect human biological systems at the molecular level over time in order to gain more knowledge about the mechanisms involved throughout the natural history of a disease. The development of cancer, for example, is a multi-stage process (Foulds L 1958, Grizzi and Chiriva-Internati 2006). The omics may be affected differently at different stages of the process. Thus, the temporal aspects are essential – for example, the time to

diagnosis. Systems epidemiology can help to bridge the gap between epidemiology and research in biological sciences. The study findings can provide input into research on molecular level biological systems, which can enhance our understanding of diseases, e.g. through pathway analysis (Garca-Campos et al. 2015). We can, therefore, see systems epidemiology as a shift in epidemiology from applied research towards basic research. The emphasis on the dynamic nature of biological systems and processes in systems epidemiology can be seen as a counterpart to systems biology, which is a discipline that seeks to determine how complex biological systems function by integrating experimentally derived information through mathematical and computing solutions (Institute of Systems and Synthetic Biology).

We can integrate systems epidemiological designs into existing prospective studies if the studies include omics and relevant questionnaire data. The Norwegian Women and Cancer study is an example of a complex prospective study with extensive data from questionnaires and registers, nested studies, different types of preserved biological samples, and omics data.

However, many opportunities remain unexplored due to the time-consuming and expensive steps required to conduct a full systems epidemiological project. We could reduce the problem by making it possible to quickly design studies and explore potential hypotheses at an early stage, before starting thorough research projects.

In this paper, we show that many novel systems epidemiological studies are possible by utilizing existing data from population-based prospective cohort studies. We also propose a computer systems architecture enabling the swift design of studies and exploration of hypotheses.

COMPLEX DESIGNS

Systems epidemiological study designs can be nested within existing cohort studies, such as the Norwegian Women and Cancer (NOWAC) study. The novel studies thus become part of a larger, complex design. Here, we describe the NOWAC study and data types, and we show that the existing cohort enables many novel study design possibilities. We give a stepwise example of a systems epidemiological design process. We also provide examples of two other variations of study designs to show that there are several ways to design studies. Lastly, in this section we discuss the potential for realizing more of the potential for designing studies and exploring hypotheses.

NORWEGIAN WOMEN AND CANCER STUDY

In this paper, we use the Norwegian Women and Cancer (NOWAC) Study (Lund et al. 2008) to describe the systems epidemiological design process. NOWAC is a population-based prospective cohort study approved by the Regional Committee for Medical Research Ethics and the Norwegian Data Inspectorate (P REK NORD 141/2008 Biobanken KVINNER OG KREFT). It was initially designed for breast cancer research and has later been used to research other types of cancer. The cohort includes 172 556 Norwegian women born between 1926–1965 (Gram et al. 2013). Invitations to the study were sent by mail in different batches for different time periods (The Norwegian Women and Cancer Study, NOWAC). Most of the women were recruited between 1991–1997 (179 387 invited, 102 540 recruited) and 2003–2006 (130 577 invited, 63 232 recruited) (Lund et al. 2008). All of the invited women had been randomly drawn from the Norwegian Central Person Register. Each woman in the study has participated in surveys with questionnaires covering a wide range of topics, from smoking, alcohol, diet, and physical activity to the use of oral contraceptives and hormonal replacement therapy, reproductive history, and diseases in the family.

The women have answered follow-up surveys with intervals of between four to six years, resulting in a total of one to four answered questionnaires per woman. The latest follow-up was in 2017. NOWAC periodically updates data with information from the Norwegian Cancer Registry and the Cause of Death Registry.

There are also blood and tissue samples. The number of women in NOWAC born 1943–1957 is about one-third of all Norwegian women born in those years, and between 2003–2006, the NOWAC postgenome cohort study (Dumeaux et al. 2008) collected blood samples from about 50 000 of these participants. At the time of blood sampling, the participants filled out an accompanying two-page questionnaire. The samples were collected using the PAXgene™ Blood RNA System (PreAnalytiX GmbH, CH–8634 Hombrechtikon, Switzerland) with buffers specially designed for the conservation of RNA (Barnung et al. 2018).

Other types of samples also exist for a smaller portion of the women, such as biopsies from both malignant tumors (Dumeaux V 2017) and healthy tissue (Chapter 4). NOWAC produced its first microarray-based gene expression dataset in 2009 and later miRNA, DNA methylation, metabolomics, and RNA-Seq datasets (Fjukstad 2019).

The samples have been preserved with the future in mind. Assessment of the mRNA quality in whole blood samples after 15 years has been reassuring (data not shown). We are still early in the post-genomic era, and the omics field is rapidly evolving. In the future, new or improved types of assays will be developed. We can

then use the preserved samples together with these assays. Also, tissue and blood samples can be analyzed in new ways as new areas of interest emerge in cancer research. For example, the immune system's role in cancer is promising (de Visser et al. 2006). In the future, other areas may attract attention.

Systems epidemiology's use of biological samples from human participants has a number of advantages compared to the alternatives. In biomedical research, for example, it is common to conduct experiments either on live laboratory animals (in vivo) or in Petri dishes and test tubes (in vitro). It is reasonable to assume that there are relevant differences between humans and laboratory mice that must be taken into account when studying human diseases (Breschi et al. 2017, Mestas and Hughes 2004). In their daily lives, humans experience very different exposures compared to laboratory mice. Systems epidemiological designs make it possible to investigate gene expression profiles resulting from the complex real-life situations of the participants, with hundreds of different exposures that interact with genetic predispositions to cancer (Lund and Dumeaux 2008).

A prospective study, such as NOWAC, will often start as a cross-sectional study in which data collection is done at a defined time. The study will usually involve surveys about the past and data originally collected for other purposes. Cross-sections of the cohort can be made, but the temporality desired in a prospective study is still missing. For each following year, some percentage of the participants will be affected by cancer or another disease, which forms the basis for the prospective aspect of the study. Additionally, the cohort needs to be followed up. Follow-ups of a cohort can involve mailing follow-up questionnaires, updating data from disease and cause-of-death registers, and possibly blood and tissue sampling.

The NOWAC study was designed as a prospective cohort study from the beginning. The aim of the study was initially to research hormonal contraceptives and breast cancer risk, but the surveys included questions covering a far broader scope. This is the reason why NOWAC can be used to research many other cancers and risk factors. In addition to the original study, there are different nested studies within NOWAC. These are mostly case-control studies. An advantage of nesting case-control studies in prospective cohorts is the reduction of recall and selection bias. Other study designs can be nested, as well. Some studies exist that only use the controls from a nested case-control study.

We can use the data in NOWAC for many novel epidemiological studies (Figure 2.1). Before any diagnosis, most participants have answered multiple surveys and donated blood samples. Data from the surveys give an insight into the participants' prior exposures and risk factors related to lifestyle, family history, socioeconomic status, and health status. This information is supplemented with data from passive

follow-up based on cancer and death register data, and active follow-up based on collaboration with 11 major Norwegian hospitals and the Norwegian Breast Cancer Group (NBCG). Blood samples were collected and stored in a way that makes new functional genomics analyses possible. Because the blood was collected before diagnosis, the time between blood sampling and diagnosis varies for different cases. In addition to the pre-diagnostic blood samples, some post-diagnostic samples were collected as well. NOWAC also includes tissue samples from hospital biobanks for many of the participants that developed cancer. The study even has four hundred biopsies from healthy women; see Chapter 4. The blood and tissue samples are analyzed using several omics technologies. All this data can be combined in many different ways, enabling many system epidemiology studies, which we will show in the following section.

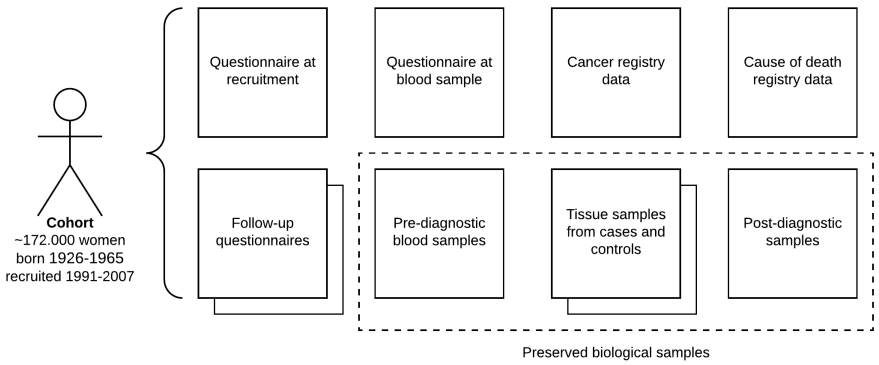


Figure 2.1. NOWAC cohort overview; biological samples and data types.

DESIGNING SYSTEMS EPIDEMIOLOGICAL STUDIES

Here, we describe how novel systems epidemiological studies can be designed using data from NOWAC. We first describe limitations of the data material before moving on to the many possible combinations of data that exist. We then provide an example of the design process.

Limitations

Before we describe the many possibilities in a prospective cohort, we first discuss the limitations. One type of limitation is when the data material does not contain

the necessary information. A trivial example is that a cohort without male participants probably does not have the data needed for prostate cancer research.

When it comes to questionnaire data, it is important to be aware that not all groups respond to surveys to the same extent. The validity of studies concerning high alcohol consumption can be problematic because people who suffer from alcoholism answer questionnaires to a lesser extent than others. Consequently, data on this group may be insufficient. However, studies involving other groups can still be valid. The validity of the questionnaire items can also be of concern—have the participants understood the questions? Furthermore, the types of data obtainable from samples are limited by the technology used for collection and cold storage. To conserve RNA in blood, we must use technologies such as PAXgene or similar.

The size of the cohort is another limiting factor. In studies involving subgroups, statistical power can often become a problem due to too few participants. One way of counteracting the problem is through international collaborations. The European Prospective Investigation into Cancer and Nutrition (EPIC) (Bingham and Riboli 2004) is one such international collaboration. EPIC is one of the largest prospective cohort studies in the world. It has 521 000 participants and has been followed for almost fifteen years. The cohort is composed of other cohorts from ten European countries, including NOWAC.

A significant problem internationally is the follow-up of mortality and disease. In Norway and the other Nordic countries, follow-up is easier thanks to public register data. All Nordic countries have a central person register, cause-of-death register, disease registers, and other public registers. Although not perfect in every respect, the Nordic registers have long been celebrated as a 'gold mine' for research (van der Wel et al. 2019).

The many possible studies

When we design a study, there are many types of choices that we can make depending on the research hypothesis. The different types of choices comprise a high number of possible studies when combined.

Figure 2.2 shows the intersection of seven different types of choices as separate dimensions. There are many options for each dimension, and the intersection of the dimensions results in an ample decision space where each combination is a potential study design. In the following, we describe the different choice dimensions.

The first dimension (1) concerns choices related to the study design's time aspect, which is an integral part of most epidemiological study designs. In system epidemiological designs, we define a timeline dimension explicitly. We can divide the timeline into the time before diagnosis, time of diagnosis, or time after diagnosis. For some samples, such as biopsies taken at diagnosis, the time will coincide with the time of diagnosis, but we can combine this with other samples taken before or after diagnosis. We can also further divide the timeline into intervals, e.g. 0–1 years before diagnosis, 2–3 years before diagnosis, and 3–5 years before diagnosis, which is useful for statistical analyses.

The second dimension (2) is the exposures and risks dimension. Many different types of exposures can increase the risk of a condition. In NOWAC's prospective questionnaires, we find information about each participant's risk factors, such as lifestyle, use of medication, conditions in the family, number of births, and much more. Additionally, genetic variants can be viewed as risk factors that can be identified by analyzing blood samples.

The third dimension (3) is the different types of measurements and assays that we can choose. In the NOWAC context, each assay is an omics or multi-omics assay – for example methylation, gene expressions, and metabolomics.

However, there are more than three dimensions. Instead of adding more axes, we label the remaining dimensions with lower case letters a–d on a cube (see label 4 in the figure). Each cube in the figure will have these four additional choice dimensions, which differentiate the many possible studies.

The fourth dimension (4a) represents the possible diagnoses that can be studied. In NOWAC, we have information about various diagnoses from the Norwegian Cancer Registry and the Cause of Death Registry.

The fifth dimension (4b) is the participant selection dimension. This dimension concerns the criteria for choosing and grouping participants for the study. A typical example is a case-control study in which we select cases from the cohort based on criteria that we choose. We then choose controls nested in the cohort matched on the cases. The criteria that we use to match controls to cases can vary from study to study, while selecting controls with the same sex and similar age since the case is quite common. There will usually be far more possible controls than cases available for selection in a study. A ratio of about a thousand to one is not uncommon. The statistical power is dependent on the number of available cases and the number of controls drawn for each case.

The sixth dimension (4c) is the sample type dimension. Usually, it matters where the analyzed sample was acquired from; it can be a blood sample, a tissue sample, or a sample of specific types of immune cells. We can compare results from differ-

ent sample types from each participant, such as comparing methylation levels in peripheral blood and tumor tissue.

The seventh dimension (4d) applies to stratification and de-confounding. The purpose is to adjust for underlying factors that skew the results, and we usually use exposure and risk factor data for this. An example of how we can adjust for smoking exposure when analyzing biomarkers for lung cancer is given in a later description of a three-level study design.

We have now described the many available choices that exist when designing studies. Each dimension consists of many options, and the number of possible studies becomes very large when we combine different dimensions. The reason for the high number of combinations is that the number of options for each dimension must be multiplied together. The total number of combinations then becomes: *(The number of ways to arrange the timeline) * (The number of exposures) * (The number of available measurements and assays, e.g. for single or multi-omics) * (The number of available diagnoses) * (The number of ways to select participants) * (All sample types and relevant combinations) * (The de-confounding and stratification factors)*

After we have chosen the study parameters from the described dimensions, we will have a clearer understanding of the selection of data we need for a study. The next step is to apply the data selection to systems epidemiological designs.

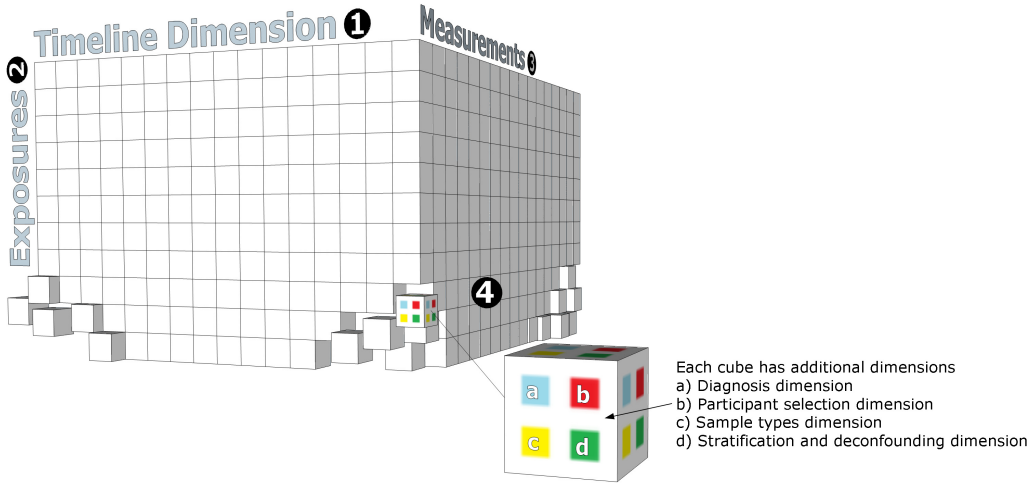


Figure 2.2. The different dimensions that can be combined for each study design.

Applying data to systems epidemiological designs

After deciding on the parameters and data for our study, we apply the data within a systems epidemiological design. We now give a stepwise example of a systems epidemiological design process using existing data from a prospective cohort study with omics data, such as NOWAC.

In systems epidemiology, imagine that we organize our sample data points along several axes, where one is the timeline (Figure 2.3). We usually split the timeline into the time before diagnosis, of diagnosis, and after diagnosis. It is also possible to split the timeline by an event other than the diagnosis. The decision on how to split the timeline was described earlier as one of the dimensions from which we choose our study parameters.

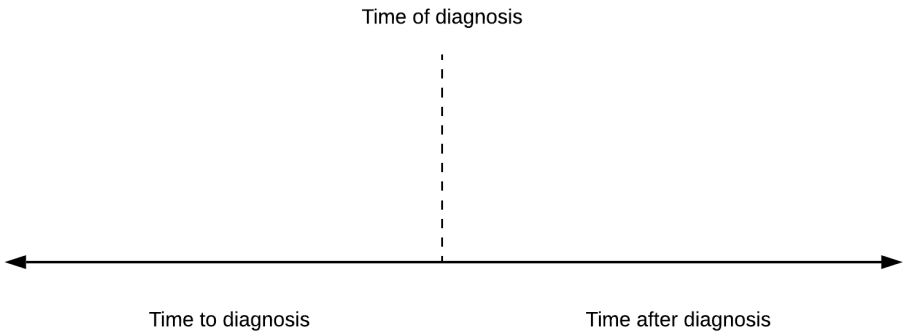


Figure 2.3. Time to diagnosis, time of diagnosis, and time after diagnosis.

Each sample in our data has a temporal distance to the time of diagnosis (Figure 2.4). We therefore place the data points on the timeline relative to how long before or after diagnosis the sample was collected. The second axis is a value axis. The values of the data points can be the raw measured values, such as the expression levels for a gene, but they are often the results of a function that takes one or more measured values as parameters. For example, the vertical position of the data point may represent the difference between cases and controls (Formula 2.1).

$$f(x_{case}, x_{ctrl}) = \log_2(x_{case}) - \log_2(x_{ctrl})$$

Formula 2.1. In the formula, *x* is a case-control pair's expression levels for a gene or other omics value.

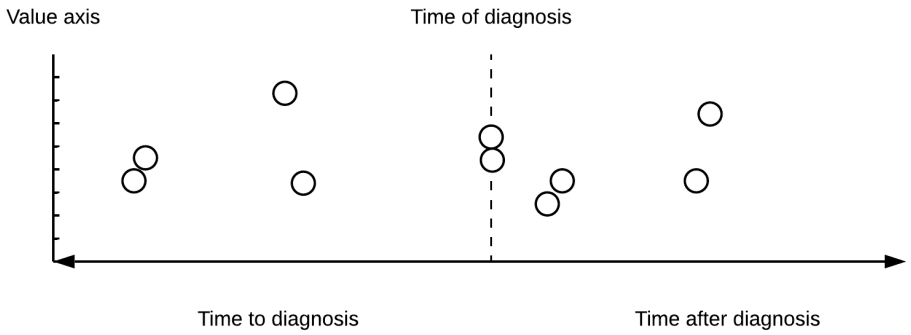


Figure 2.4. Sample data points positioned by distance from diagnosis. The value axis does not have to be linear; it can be logarithmic or other.

Next, we can group data points into strata that we are interested in comparing (Figure 2.5). By observing data points at a group level, we can envision a curve or trajectory for each stratum (Figure 2.6). If we compare the trajectories and find significant differences between the strata, this could potentially be of importance not only for future research on differential diagnosis or prognosis, but also for understanding which biological systems are involved.

It is not mandatory to stratify by grouping data points as described. If the data point values come from a function that represents a comparison of different samples, then this too is a type of stratification. When using Formula 1 for data point values, the height of the curve is a case-control comparison. Consequently, multiple levels of stratification can be achieved through a combination of grouping and use of functions.

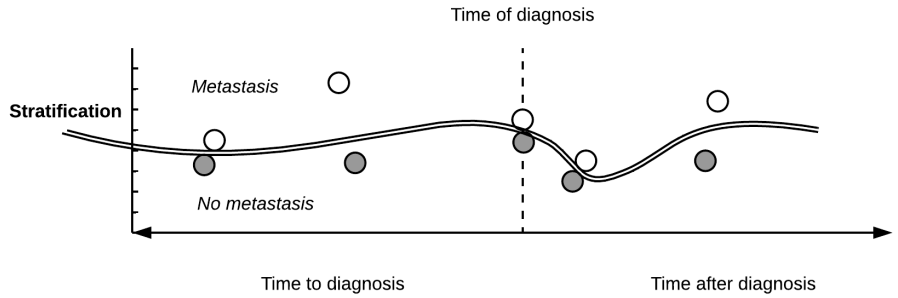


Figure 2.5. Stratification of data points. In this example, the white-filled circles represent women with metastasis, and the grey-filled circles represent women without.

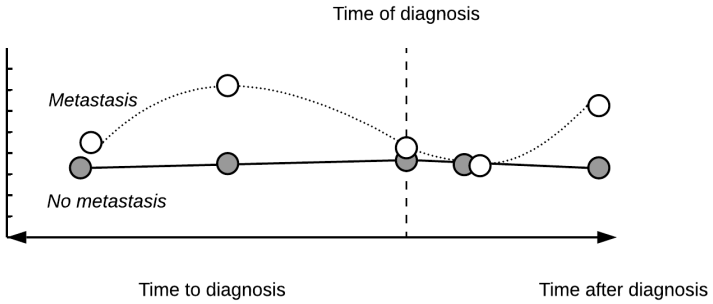


Figure 2.6. An illustration of estimated curves or trajectories for each stratum. The curves for the two strata are different.

Because the measured values are from biological processes that interact as part of a system, it is interesting to compare the curves of many types of values simultaneously (Figure 2.7). The figure shows three curves per stratum, one for each type of gene expression.

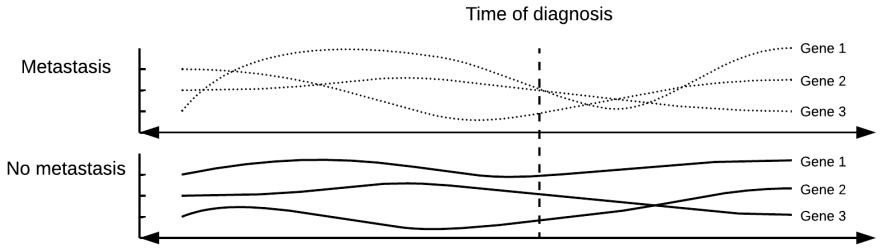


Figure 2.7. For each biological sample, we can measure the levels of many different expressed genes. For each, we can imagine a separate curve per strata. In the illustration, only the expression levels for "Gene 1" differ between the two strata. Note that we are not restricted to gene expressions. Other omics can be used.

However, the reality is more challenging than illustrated in Figure 2.7. For example, we can measure the expression levels for 19 950 protein-coding genes from each blood sample and present each expressed gene as a separate curve along the timeline. Curves for other omics can be included as well, such as methylation. The results can thus consist of thousands of intersecting curves per stratum, which is too much information to be presented as an overview of the data. Therefore, we must use other techniques for analyzing the data. Many methods exist for analyzing high-dimensional omics data. Usually we use methods related to clustering or dimensionality reduction techniques for high-dimensional data (Breschi et al.

2017). Examples of dimensionality reduction techniques include principal component analysis (PCA), multidimensional scaling (MDS), and t-distributed stochastic neighbor embedding (tSNE). An alternative approach is to map the omics data to a biological context, e.g. we can map gene expressions to where they occur in biological pathways. We are also interested in including the temporal aspect as part of the data analysis, which is a hallmark of systems epidemiology.

We have now described how studies can be designed by applying existing cohort data, for example, a combination of questionnaire data and high-dimensional molecular data from NOWAC. The steps in the design process described in this section can be summarized as:

- Establish an axis for the time to diagnosis (or another event) and an axis for values
- Define strata
  - For example, cases with spread or without spread
- Calculate data point values and position them in the coordinate system
  - The basis for the values is analyzed samples, taken from different participants at different times. Pre-diagnostic samples acquired from the cases will usually have different distances to the time of diagnosis
  - The data point values can be the raw measured values from samples, but more often we use derived values from computations and statistical methods that include values from case-control pairs
- Imagine curves for each similar type of data point belonging to the same stratum
  - For example, all data points for a specific mRNA that involve cases with spread belong to the same curve
- For high-dimensional data, there will be too many curves to comprehend, and advanced clustering or dimensionality reduction techniques are thus needed
- Compare the strata to find differences
  - Statistical methods, data explorations, and visualizations

TWO ALTERNATIVE TYPES OF STUDY DESIGN

In the previous section we based the studies on comparing cases and controls, but there are other possibilities. Here we describe two design variations.

The NOWAC study has tissue samples that we can analyze and compare to peripheral blood. That is, we compare samples from different locations in the same person instead of between cases and controls. NOWAC includes case-control pairs for which diagnostic blood and tissue samples exist both for cases and matching

controls, which means that women allowed health-care professionals to take biopsies of healthy tissue for research purposes. For these participants, we can design studies that compare tissue and blood samples and also include the case-control aspect (Dumeaux et al. 2017).

Figure 2.8 It is also possible to define study designs with more than one level of nesting. For example, we can create a three-level design comprised of the cohort, a nested case-control study, and a cross-sectional study that only includes the controls (Figure 2.8). The following case exemplifies this type of design: For some diseases, such as lung cancer, a large percent of the cases has a history of smoking exposure. As a result, it can be hard to separate the early biological effects of cancer from the effects of smoking. We can solve this problem by first finding biomarkers for smoking exposure in the controls. In the cross-sectional study, the controls are stratified based on exposure data from the cohort's prospective questionnaires. The gene expressions are then analyzed to find the biological markers of smoking. In the parent case-control study, the findings can be used for de-confounding purposes to prevent smoking markers from being misinterpreted as cancer markers. A study similar to this has been conducted by (Baiju et al. 2020) as part of the Id-Lung project. The same type of design was used by to demonstrate altered gene expression levels in the NOWAC cohort associated with coffee consumption (Barnung et al. 2018).

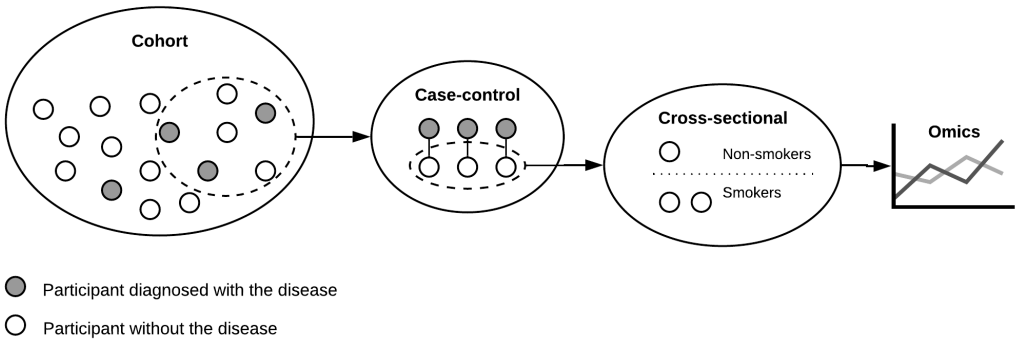


Figure 2.8. An illustration of a three-level design. Case-control pairs are selected from the prospective cohort. The cross-sectional study selects controls from the case-control study. The controls are stratified by exposure, which in this case is smoking status. The smoking statuses are calculated from the cohort study's questionnaires, and the biological samples are also from the cohort. The gene expression data is part of the case-control study. The cross-sectional study analyzes the gene expressions to find exposure markers.

TOWARDS REALIZING THE POTENTIAL

We have shown that it is possible to combine data in numerous ways to design many different studies. Unfortunately, a lot of time and resources are needed to carry out full epidemiological studies. Consequently, many opportunities that lie in the prospective cohorts may be left unrealized.

If, instead, we had carried out lightweight studies in a simple way in advance wherein we could quickly explore potential hypotheses, then we could have had a better starting point when deciding whether it would be worth going ahead with larger projects.

To realize more of the potential that lies in the NOWAC data and similar studies, we suggest that a computer system should be created that supports the rapid design of studies, analysis of data, and exploration of hypotheses. In the following sections, we propose a computer systems architecture for this purpose.

COMPUTER SYSTEMS ARCHITECTURE

In systems epidemiology, we design complex studies with many types of data, including high-dimensional molecular data. Computer systems are essential for managing data and performing computations. In the previous section we discussed the possibility of a computer system helping to realize more of the potential in cohort data by enabling the users to explore different hypotheses quickly. However, no such unified system presently exists for systems epidemiology.

Here, we propose a systems architecture that enables the swift design of studies, analysis of data, and exploration of hypotheses. The aim is to explore different hypotheses quickly at a preliminary stage of research, or explained with a metaphor: "We wish to explore the data by swimming and delving into it." (Lund 2019, personal communication)

There exists a range of software tools and systems that are used in systems epidemiology. Examples are tools that are concerned with processing omics data in pipelines, data management, or reproducibility in science. Fjukstad et al. 2018 (Chapter 3) used a combination of such tools to organize data storage and documentation and to standardize the analysis of gene expression data in NOWAC. Various unrelated tools and scripts for statistical analyses of omics also exist. None of these tools and systems constitute a unified system for the swift design of studies, epidemiological analysis, and exploration of hypotheses. We present a high-level, conceptual architecture for this missing system.

Figure 2.9 shows a conceptual view of the proposed system's architecture. The system is illustrated as having a pipelined architecture in which one part's output becomes the next part's input. The arrows between the parts represent the flow of data. Each part may be composed of loosely coupled subsystems.

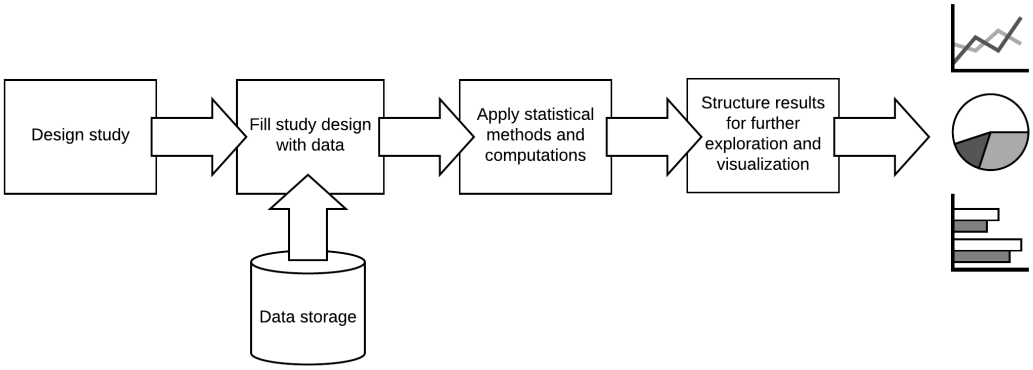


Figure 2.9. A high-level conceptual view of a computer system for systems epidemiology.

In addition to designing each part of the system, we must design good abstractions for the interfaces between them. We can view most of the system's parts as separate black boxes; the outside does not know the details of how the part functions on the inside. The outside can only interact with it through limited interfaces and is not permitted to manipulate its inner state and workings directly. An abstraction is a well-defined view or model that only includes what is relevant and excludes all that is irrelevant. The art is to define abstractions that are correct for use, flexible and general enough to include relevant variations, yet simple and coherent. We commonly prefer interfaces and data structures with these properties. We implement them by using the available features for declaring data types, functions, and schemas in our programming languages, software frameworks, and environments. The conscious use of abstractions when designing systems is an important tool for avoiding accidental complexity, and it provides the system with clean and simple-to-understand façades (Kleppman 2017). Abstractions also help to clearly separate the system's different concerns and make it more flexible to changes.

First, we provide an example use case describing the system from the researcher's point of view. Next, we discuss the five main parts of the system. We additionally touch upon the importance of reproducibility in science.

Example use case: Design a study in an interactive notebook

In this section, we describe how the researcher can use the system through an interactive notebook. Interactive notebooks are increasingly popular in data science and scientific computing. The notebooks enable researchers to create dynamic documents containing a mix of text and runnable code fragments. We use the notebooks as interactive development environments and share them with others. Two examples of notebook environments are R Notebook (Chapter 3.2 in Xie et al. 2019) and (The Jupyter Notebook). We provide a casual use case (Cockburn 2000) describing a notebook approach to designing studies.

A researcher wants to design a study in order to explore a hypothesis. The researcher has already opened a notebook and loaded the required packages belonging to the system. The researcher types in and runs a simple command (or function-call) telling the system to create a workspace for the study. The system creates a data structure representing an empty workspace, which becomes available in the researcher's notebook. Included in the workspace is a default study design specification. The researcher specifies the study's overall design by adding groups and stratifications to the design specification. The system keeps a data structure representing this design within the study design specification. The researcher specifies the data sets that will be used, including the target versions. The system keeps this information in the workspace. The researcher then defines queries for the different groups and strata. The system keeps the queries but does not yet run them to fetch data. At this point, the researcher wants to inspect the data, which is an optional step. The system runs the queries on demand and makes the data available. After inspecting the data, the researcher defines how data will be analyzed by composing statistical methods and computations from standard or custom packages. These can be associated with specific groups or strata, and sequences of computations can be defined. The system keeps this in the workspace. The researcher instructs the system to execute the entire study, and the system executes the study by fetching necessary data and running computations as specified. It does this by delegating work to the storage and computational systems, such as data lakes and Apache Spark. It makes the resulting data available in the researcher's notebook environment. The researcher can then further explore and visualize the results with other tools.

The researcher can save the workspace at any point. Previously saved workspaces can be loaded and run. The researcher can modify individual parts of the workspace and execute the updated study.

Design study

To easily specify new study designs, we must provide a user interface (UI) to the system that is user-friendly and practical. Several options exist:

- A graphical UI for specifying study designs
- A human-readable text-based format for defining studies (XML, JSON, YAML)
- A software package integrated into a development environment commonly used in the researcher's field (R-studio)
- A domain-specific language (DSL) for defining study designs

Regardless of how we present the study design specification UI to the researcher, the specified designs must internally be represented in a machine interpretable manner that is useable later for the automatic execution of the study. The study design specifications describe what the researcher wishes to do, but not the details of how. The exact decision on how data retrieval and execution is performed is left to other parts of the system. This type of abstraction ensures that changes in implementation details, or even the replacement of whole subsystems, can be contained to the parts that retrieve data and execute the study without requiring changes to other parts. Equally important, the abstraction makes it possible automatically to optimize how the study is performed.

Data storage

Data is central in epidemiological research, but managing all the technical aspects of data is complicated and bears little relevance to the researcher's aims. For example, a considerable amount of time is spent on data wrangling due to impractical data structures or lack of consistent structures. Each project typically operates on smaller, custom data sets that have been extracted manually from the primary data sets. The data sets are stored in simple text-based formats on shared disks. The included fields and names are inconsistent across data sets. Sometimes the researchers will make personal copies of the data set file, with various changes that they have made. With the advent of multi-omics, the amount of data can potentially become very large, which will require a more professional approach to data management. The system should hide the technical details surrounding data and instead provide the researchers with simple, uniform data access.

Today, a variety of production-quality data storage solutions are available. It is crucial to investigate which type of solution best suits the system because there are significant differences between them. Examples of storage types are:

- Relational database management systems (RDBMS), including data warehouses: PostgreSQL, MS-SQL
- Key-value stores: Redis, Memcached
- Column stores or column formats: Cassandra, Parquet
- Graph databases: Neo4j, OrientDB
- Files in distributed file systems: Hadoop Distributed File System (HDFS), Tachyon
- A combination of the above, termed polyglot persistence (Sadalage and Fowler 2013)
- Data lakes (Miloslavskaya and Tolstoy 2016): Azure Data Lake, AWS Data Lake

A layer of abstraction should be created for easy and uniform access to the data, hiding the underlying data structures and storage systems. By abstracting the underlying storage mechanisms away from the rest of the system, it is easier to evolve or replace the storage solution as we discover opportunities for improvements. ADAM (Massie et al. 2013) is a set of formats, APIs, and processing stage implementations for genomic data. It has a layered design with a "narrow waist" in the middle, also termed an hourglass model (Beck 2019). The narrow-waist layer consists of a data schema, implemented with Apache Avro (The Apache Avro Project) that separates the details of the storage layers from the overlying layers. A similar approach may prove useful in our system.

Fill study design with data

After specifying a study design, the researcher must be able to query and retrieve the data for the study. First, one or more data sources are chosen. We should enable access to the data in a uniform manner and structure the data according to standard schemas. Next, the researcher defines queries that select and transform data for the study's different groups and strata, such as cases, controls, with spread, without spread. The queries are attached to the study design specification.

From the technical side, the queries should be attached to the study design but not immediately executed. The system should be allowed to run queries in the same context as the computations. This can prevent inefficient spilling of data to disk between the steps. It can also enable automatic query optimizations. There are several options for query languages, e.g., the query syntax could be SQL-like or fluent (Fowler 2005). LINQ (Torgersen 2007) or Resilient Distributed Datasets (RDD) (Zaharia et al. 2012) are examples that support deferred execution and both types of syntaxes.

The resulting data must have a structure recognizable by the computational and statistical methods in the next step of the workflow. Again, we need good abstractions.

Computations and statistical methods

The researcher should be able to choose from ready-made calculations and statistical methods and possibly define custom ones. Functions for common computations and statistical methods can be packaged in a reusable manner that is independent of a particular study. The statistical methods for curve groups (Lund E 2016) and classify strata (Holden 2015) are candidates for such packages. Novel statistical methods for systems epidemiology will likely be developed in the future. The system must support both ready-made packages, as well as custom packages. A statistician can implement functions, possibly in collaboration with scientific programmers, and epidemiologists can then apply the functions in various studies. A challenge is to define standards for functions and packaging that covers the needs of existing and future statistical methods.

The computations involved in omics analysis are often time-consuming and resource-heavy. Care should be taken to choose an underlying platform that performs well for the computations encountered in systems epidemiology. Apache Spark (Zaharia et al. 2010) is a unified analytics engine for large-scale data processing that could be used as an integral part of the system. Recent versions of Spark support R (The R Project for Statistical Computing), which is a programming language and environment for statistical computing often used in epidemiology.

Structure results for further exploration and visualization

After applying computations and statistical methods, it should be easy for the researcher to explore and visualize the data further. Because many general-purpose tools and software packages already exist that are excellent for data exploration and visualization, the results generated by the system should be usable within the context of such software packages and tools. We can achieve this by structuring data in a standard format so that the researcher can either use the result datasets directly or import them into their software tool of choice, such as an R environment.

Reproducibility

It has been claimed that there is a reproducibility crisis in science. *Nature* (Baker 2016) asked 1576 researchers questions about reproducibility. They found that 90% answered that there was either a slight or significant crisis. More than 70% had tried and failed to reproduce other scientists' experiments. More than half of the scientists had experienced that they were unable to reproduce their own exper-

iments. There are several reasons for the crisis – for example, selective reporting or low statistical significance. At other times it can be challenging to know how to repeat the experiment correctly. In the latter case, we can benefit from having a system that can automatically rerun previous experiments using the same steps and data.

The system's study design specifications, dataset selections, queries, and statistical methods can be saved together as a complete workflow. As long as the underlying data stay unchanged, the experiments can be reloaded and automatically repeated. The system must track changes to data and support data versioning. By specifying target data versions for the workflows, we can ensure that the experiment's data stays the same between runs.

CONCLUSION

We have described the complex NOWAC study, the many different types of data, and that the data can be combined in a large number of ways. The many combinations allow us to create many new system epidemiological study designs. We have also given a step-by-step example of a system epidemiological design.

The beauty of complex studies such as NOWAC is the opportunities for new studies that arise. However, opportunities can be lost because extensive studies are time-consuming and costly. By finding a quick way to create designs using existing data, we can perform initial explorations to investigate if a hypothesis is worth researching more extensively.

As a solution, we have proposed a computer systems architecture to support the swift design of system epidemiological studies and exploration of hypotheses.

ACKNOWLEDGEMENTS

We wish to thank Professor Eiliv Lund for providing invaluable input and feedback on the paper.

REFERENCES

Apache Spark [Internet]. Available from: https://spark.apache.org

Baglietto L, Ponzi E, Haycock P, Hodge A, Bianca Assumma M, Jung CH et al. DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. Int J Cancer. 2017 Jan 1; 140(1): 50–61. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/ijc.30431

Baiju N, Sandanger TM, Sætrom P, Nøst TH. Gene expression in whole-blood reflects smoking exposure among cancer-free women in the Norwegian Women and Cancer postgenome cohort. Submitted.

Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016 May 26; 533(7604): 452–454. Available from: https://www.nature.com/articles/533452a

Barnung RB, Nøst TH, Ulven SM, Skeie G, Olsen KS. Coffee Consumption and Whole-Blood Gene Expression in the Norwegian Women and Cancer Post-Genome Cohort. Nutrients. 2018 Aug 9; 10(8): 1047.

Beck M. On The Hourglass Model. Communications of the ACM. 2019 Jul; 62(7): 48–57. Available from: https://cacm.acm.org/magazines/2019/7/237714-on-the-hourglass-model/fulltext

Bingham S, Riboli E. Diet and cancer--the European Prospective Investigation into Cancer and Nutrition. Nat Rev Cancer. 2004 Mar; 4(3): 206–215. Available from: https://www.nature.com/articles/nrc1298

Breschi A, Gingeras TR, Guigo R. Comparative transcriptomics in human and mouse. Nat Rev Genet. 2017 Jul; 18(7): 425–440. Available from: https://www.nature.com/articles/nrg.2017.19

Castagne R, Kelly-Irving M, Campanella G, Guida F, Krogh V, Palli D, et al. Biological marks of early-life socioeconomic experience is detected in the adult inflammatory transcriptome. Sci Rep. 2016 Dec 9; 6: 38705. Available from: https://www.nature.com/articles/srep38705

Cockburn A. Writing effective use cases. Series: The Crystal Collection for software professionals. 1st ed. Addison-Wesley Professional; 2000. pp 304.

Dumeaux V, Børresen-Dale AL, Frantzen JO, Kumle M, Kristensen VN, Lund E. Gene expression analyses in breast cancer epidemiology: the Norwegian Women and Cancer postgenome cohort study. Breast Cancer Res. 2008 Feb; 10(1): 1–8. Available from: http://breast-cancer-research.com/content/10/1/R13

Dumeaux V, Fjukstad B, Fjosne HE, Frantzen JO, Holmen MM, Rodegerdts E, et al. Interactions between the tumor and the blood systemic response of breast cancer patients. PLoS Comput Biol. 2017 Mar 7; 13(9): e1005680. Available from: https://doi.org/10.1371/journal.pcbi.1005680

de Visser KE, Eichten A, Coussens LM. Paradoxical roles of the immune system during cancer development. Nat Rev Cancer. 2006 Jan; 6(1): 24–37. Available from: https://www.nature.com/articles/nrc1782

Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. Nat Commun. 2015 Dec 15; 6: 10192. Available from: https://www.nature.com/articles/ncomms10192

Fjukstad B. Toward Reproducible Analysis and Exploration of High-Throughput Biological Datasets [Doctoral thesis]. Tromsø: UiT, The Arctic University of Norway; 2019. 149 pp. Available from: https://munin.uit.no/handle/10037/14576

Fjukstad B, Shvetsov N, Nøst TH, Bøvelstad H, Halbach T, Holsbø E et al. Reproducible data management and analysis using R. bioRxiv. 644625, in press. Available from: https://www.biorxiv.org/content/10.1101/644625v1

Foulds L. The natural history of cancer. J Chronic Dis. 1958 Jul; 8(1): 2–37. Available from: https://www.sciencedirect.com/journal/journal-of-chronic-diseases/vol/8/issue/1

Fowler M. FluentInterface. At martinfowler.com [Internet]. Accessed 06.06.2019. Available from: https://martinfowler.com/bliki/FluentInterface.html

Garcia-Campos MA, Espinal-Enriquez J, Hernandez-Lemus E. Pathway Analysis: State of the Art. Front Physiol. 2015 Dec 17; 6: 383. Available from: https://www.frontiersin.org/articles/10.3389/fphys.2015.00383/full

Gram IT, Sandin S, Braaten T, Lund E, Weiderpass E. The hazards of death by smoking in middle-aged women. Eur J Epidemiol. 2013 Sep 29; 28(10), 799–806. Available from: https://link.springer.com/article/10.1007/s10654-013-9851-6

Grizzi F, Chiriva-Internati M. Cancer: looking for simplicity and finding complexity. Cancer Cell Int. 2006 Feb 15; 6(1): 4. Available from: https://cancerci.biomedcentral.com/articles/10.1186/1475-2867-6-4

Hasin Y, Seldin M, Lusis M. Multi-omics approaches to disease. Genome Biol. 2017 May 5; 18(1): 83. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1215-1

Holden L. Classify strata. Oslo: Norwegian Computing Center; SAMBA/11/15; 2015. pp 28. Available from: https://www.nr.no/directdownload/1426685952/classify_strata_holden2015.pdf

Imperial College London. Institute of Systems and Synthetic Biology [Internet]. Accessed: 06.06.2019. Available from: https://www.imperial.ac.uk/systems-biology/about-the-institute/

Kleppman M. Designing Data-Intensive Applications: the big ideas behind reliable, scalable, and maintainable systems. 1st ed. Sebastapol, CA: O'Reilly Media; 2017. pp 569. Available from: https://books.google.no/books?id=zFheDgAAQBAJ&lpg=PP1&lr&hl=no&pg=PP1#v=onepage&q&f=false

Lund E. Personal communication. Meeting at Institute for Informatics about BoCD. Tromsø, 2019.

Lund E, Dumeaux V. Systems epidemiology in cancer. Cancer Epidemiol Biomarkers Prev. 2008 Nov; 17(11): 2954–2957. Available from: https://cebp.aacrjournals.org/content/17/11/2954.long

Lund E, Dumeaux V, Braaten T, Hjartaker A, Engeset D, Skeie G et al. Cohort profile: The Norwegian Women and Cancer Study--NOWAC--Kvinner og kreft. Int J Epidemiol. 2008 Feb; 37(1): 36–41. Available from: https://academic.oup.com/ije/article/37/1/36/763947

Lund E, Holden L, Bøvelstad H, Plancade S, Mode N, Günther CC, et al. 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 Med Res Methodol. 2016 Mar 5; 16(1): 28. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0129-z

Lund E, Plancade S, Nuel G, Bøvelstad H, Thalabard JC A processual model for functional analyses of carcinogenesis in the prospective cohort design. Med Hypotheses. 2015 Oct; 85(4): 494–497. Available from: https://www.sciencedirect.com/science/article/pii/S0306987715002704?via%3Dihub

Massie M, Nothaft FA, Hartl C, Kozanitis C, Schumacher A, Joseph AD et al. ADAM: Genomics Formats and Processing Patterns for Cloud Scale Computing. Technical Report No. UCB/EECS-2013-207. Electrical Engineering and Computer Sciences, University of California at Berkeley; 2013. pp 22. Available from: https://www2.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-207.pdf

Mestas J, Hughes CC. Of mice and not men: differences between mouse and human immunology. J Immunol. 2004 Mar 1; 172(5): 2731–2738. Available from: https://www.jimmunol.org/content/172/5/2731.long

Miloslavskaya N, Tolstoy A. Big Data, Fast Data and Data Lake Concepts. Procedia Computer Science. 2016; 88: 300–305. Available from: https://reader.elsevier.com/reader/sd/pii/S1877050916316957

Kvinner og kreft, Blodprøve og biopsi [Internet]. Accessed: 28.07.2020. Available from: https://site.uit.no/kvinnerogkreft/blodprove-og-biopsi/

National Institute of Health, National Human Genome Research Institute. The Human Genome Project [Internet]. Accessed: 15.11.2019. Available from: https://www.genome.gov/human-genome-project

Norwegian Computing Central [Internet]. Available from https://www.nr.no/en

Notebook. Chapter 3.2 in Xie Y, Allaire JJ, Grolemund G (eds) R Markdown: The Definitive Guide [Internet]. Accessed: 15.11.2019. Available from: https://bookdown.org/yihui/rmarkdown/notebook.html

Sadalage PJ, Fowler M. NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. New Jersey: Pearson Education, Inc; 2013. pp 164.

The Apache Avro Project [Internet]. Available from: https://avro.apache.org

The EPIC Study [Internet]. Available from: https://epic.iarc.fr

The Jupyter Notebook [Internet]. Available from: https://jupyter.org

The Norwegian Women and Cancer Study, NOWAC [Internet]. Accessed: 06.06.2019. Available from: https://site.uit.no/nowac/methodological-description/timeline/

The R Project for Statistical Computing [Internet]. Available from: https://www.r-project.org

Torgersen M. Querying in C#: how language integrated query (LINQ) works. In: Proceeding OOPSLA´07 Companion to the 22nd ACM SIGPLAN conference on Object-oriented programming systems and applications companion, Montreal, Quebec, Canada, Oct 21–25, 2007. New York: ACM Press; 2007. pp 852–853.

UiT The Arctic University of Norway. Id-Lung [Internet]. Available from: https://en.uit.no/forskning/forskningsgrupper/gruppe?p_document_id=507532.

Vailati-Riboni M, Palombo V, Loor JJ. What Are Omics Sciences? In: Ametaj B (eds) Periparturient Diseases of Dairy Cows. Cham: Springer; 2017. pp. 1–7. Available from: https://link.springer.com/chapter/10.1007%2F978-3-319-43033-1_1#citeas

van der Wel KA, Östergren O, Lundberg O, Korhonen K, Martikainen P, Andersen AN, Urhoj SK. A gold mine, but still no Klondike: Nordic register data in health inequalities research. Scand J Public Health. 2019 Aug;47(6):618–630. Available from: https://journals.sagepub.com/doi/10.1177/1403494819858046

van Veldhoven K, Polidoro S, Baglietto L, Severi G, Sacerdote C, Panico S, et al. Epigenome-wide association study reveals decreased average methylation levels years before breast cancer diag-

nosis. Clin Epigenetics. 2015 Aug 4; 7: 67. Available from: https://clinicalepigeneticsjournal.biomedcentral.com/articles/10.1186/s13148-015-0104-2

Zaharia M, Chowdhury M, Das T, Dave A, Ma J, McCauley M et al. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In: NSDI´12 Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, San Jose, CA, USA, Apr 25–27, 2012. Berkeley: USENIX Association Berkeley; 2012(2–2). Available from: https://www.usenix.org/system/files/conference/nsdi12/nsdi12-final138.pdf

Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I. Spark: Cluster computing with working sets. In: HotCloud'10 Proceedings of the 2nd USENIX conference on Hot topics in cloud computing, Boston, MA, UAS, Jun 22–25, 2010. Berkeley: USENIX Association Berkeley; 2010(10–10): p. 95. Available from: https://www.usenix.org/legacy/events/hotcloud10/tech/full_papers/Zaharia.pdf

Lund, E. (Ed.) (2020). *Advancing Systems Epidemiology in Cancer. Exploring Trajectories of Gene Expression.*
Oslo: Scandinavian University Press.
DOI: https://doi.org/DOI: 10.18261/9788215041193-2020-03

3. Reproducible Data Management and Analysis Using R

Bjørn Fjukstad, Nikita Shvetsov, Therese H. Nøst, Hege Bøvelstad, Till Halbach, Einar Holsbø, Knut Hansen and Lars Ailo Bongo

Abstract Standardizing and documenting computational analyses is necessary to ensure reproducible results. We describe an R-based implementation of data management and preprocessing that is well integrated with the analysis tools typically used for statistical analysis of omics data. We have used these tools to organize data storage and documentation, and to standardize the analysis of gene expression data, in the Norwegian Women and Cancer study.

Keywords Gene expression data | data management | documentation | R-script. NOWAC

INTRODUCTION

Reproducibility is necessary to advance science and to leverage scientific results (Reality check on reproducibility 2016). This requires the implementation of best practices for data management and analysis. Such best practices are also necessary for large and complex projects in which data collection, analysis, and interpretation may span decades, and is therefore done in several iterations, by different people. We have observed this need in systems epidemiology (Lund and Dumeaux 2008). In addition, the need is recognized in the STROBE-ME (Gallo et al. 2011) initiative to strengthen the reporting of observational studies in molecular epidemiology.

There are many approaches, systems, and tools for data storage and processing that solve many of the technical challenges of ensuring reproducible analyses (Ivie and Thain 2018). To make it easy to find relevant data for re-analysis or re-interpretation, the data can be organized in file system structures, databases, or in other indexable storage systems. To keep track of different versions of files, we can use a versioned file system, or version control systems, such as git, that are widely

adopted in software engineering. To document the tools, parameters, and reference databases used in an analysis we can use frameworks such as CWL (Amstutz et al. 2016), Galaxy (Afgan et al. 2016), Snakemake (Köster and Rahmann 2012), Spark (Zaharia et al. 2016) or an in-house solution such as our Walrus system (Fjukstad et al. 2018). All of these frameworks provide an interface to set up an analysis pipeline, either as a text file or using a Graphical User Interface (GUI), and then execute it. To record provenance and keep track of the intermediate files, we can implement and run the analysis in, for example, Galaxy, Spark or Pachyderm (Pachyderm – Scalable, Reproducible Data Science 2019). However, there is a need to adapt these systems and tools to the needs of typical omics data analysis workflows.

In this paper we describe our lessons learned throughout 10 years of transcriptomics data analysis in the Norwegian Women and Cancer (NOWAC) study (Lund et al. 2008). We use these to propose an approach to maintain, preprocess, and facilitate statistical analyses in complex systems epidemiology datasets. The approach ensures reproducibility, and we believe that it is well adapted for omics data analyses. It enables us to achieve reproducible research through the four steps described above. First, we use R since it has many up-to-date and actively maintained packages for analyzing, plotting, and interpreting data (for instance, Bioconductor (Gentleman et al. 2004) and the Comprehensive R Archive Network (The Comprehensive R Archive Network 2019). Second, we have developed an R pipeline package with code and the datasets from the NOWAC study. We document all datasets thoroughly and use version control to track both datasets and code over time. Third, we have developed an interactive web application, the Pippeline, to perform the standardized preprocessing steps for gene expression datasets. Fourth, we export the data as a git repository and RStudio project file to encourage reproducible analyses. Fifth, we have developed our own best practices to report results and share analyses through reproducible analysis reports.

The article is organized as follows. After the description of the datasets at hand and the given context, we detail how omics data analysis was done previously, and what challenges this implied. We then discuss the requirements for our new approach and describe the solution in detail, together with an explanation of the corresponding methodology and best practices. We briefly discuss limitations of our work before concluding.

DATA ANALYSIS LESSONS LEARNED IN THE NORWEGIAN WOMEN AND CANCER STUDY

Our approach is based on the 10 years of transcriptomics data analysis in the NOWAC study. This is a prospective population-based cohort that tracks 34% (170 000) of all Norwegian women born between 1943 and 1957 (Lund et al. 2008). We started the data collection in 1991 with surveys that cover topics including: the use of oral contraceptives and hormonal replacement therapy, reproductive history, smoking, physical activity, breast cancer, and breast cancer in the family. We also periodically update the study with data from the Norwegian Cancer Registry, and the Cause of Death Registry. In addition to the questionnaire data, we collected blood samples from 50 000 women, as well as more than 300 biopsies. From the biological samples we generated the first microarray-based gene expression dataset in 2009, and later miRNA, DNA methylation, metabolomics, and RNA-seq datasets.

The data in the NOWAC cohort allows for a number of different study designs. While it is a prospective cohort study, we can also draw a case-control study from the cohort, or a cross-sectional study. We have published papers analyzing the questionnaire data (e.g. Busund et al. 2018, Gram et al. 2016), and many research papers that investigate the questionnaire data together with the gene expression datasets (e.g. Olsen et al. 2013, Dumeaux et al. 2010). We have also used the gene expression datasets to explore gene expression signals in blood and interactions between the tumor and the blood systemic response of breast cancer patients (Holden et al. 2017, Dumeaux et al. 2017). Some analyses have resulted in patents (Dumeaux and Lund 2014) and commercialization efforts. There are still, however, many unexplored areas in the NOWAC datasets.

In the NOWAC study we are a group of researchers, PhD students, post docs, technical staff, and administrative staff. The researchers are from statistics, medicine, epidemiology, and computer science. The administrative and the technical staff are responsible for managing the data—both data collection and data delivery to researchers. The interdisciplinary work and the complexity of the studies makes data management and analysis especially challenging.

Data management and analysis

Surveys are the traditional data collection method in epidemiology. Today, however, questionnaire responses are increasingly integrated with molecular data, but surveys are still important for designing a study that can answer particular research questions. In this section we describe how such omics data analysis was done in

NOWAC before we developed our approach. We believe many studies have been, or are still, analyzing epidemiological data using a similar practice, and that our approach and lessons learned presented here will be useful for these studies.

In the NOWAC study we have stored the raw survey and registry data in an in-house database. Researchers apply to get questionnaire data variables exported from the database by scientific staff. This was typically done through SAS scripts that did some preprocessing, e.g. selecting applicable variables or samples, before the data was sent to researchers as SAS data files. The downstream analysis was typically done in SAS. Researchers used e-mail to communicate and send data analysis scripts, so there was no central hub with all the scripts and data.

In addition to the questionnaire data, the NOWAC study also integrates with registries (cancer and death) that are updated regularly. The datasets received from the different registries are typically delivered as comma-separated values (CSV) files to our scientific staff, which are then processed into a standardized format. Since the NOWAC study is a prospective cohort, some women are expected to get cancer and move from the list of controls into the list of cases. This also requires updating their status in the analyses using gene expression data, and makes it necessary to keep track of the case-control changes.

In the NOWAC study, we have analyzed our biological samples in labs outside our research institution. The received raw instrument datasets are then stored on a local server and made available to researchers on demand. Because of the complexity of the biological datasets, many of these require extensive preprocessing before they are ready for analysis.

Issues in previous practice

Over nearly a decade of experiences from transcriptomics data analysis, we identified a set of issues with our previous practice that prevented us from fully ensuring reproducible data analysis:

1. It was difficult to keep track of the available datasets, how they were combined, and to determine how these had been processed. We had no standard data storage platform or structure, and there were limited reports for exported datasets used in different research projects.
2. There was no standard approach to preprocessing and initiating data analysis. This was because the different datasets were analyzed by different researchers at different points in time, and there were few practices for the sharing reusable code between projects.

3. It became difficult to reproduce the results reported in our published research manuscripts. This was due to the lack of standardized preprocessing, sharing of analysis tools in their various versions, and full documentation of the analysis process.

ENABLING REPRODUCIBLE DATA ANALYSES

To solve the above issues and enable easily reproducible research in the NOWAC study, we developed a system for managing and documenting the available datasets, a standardized data preprocessing and preparation system, and a set of best practices for data analysis and management. We first identified a set of requirements for a system to manage and document the different datasets:

1. It should provide users with a single interface to access the datasets, their respective documentation, and utility functions to access the raw and preprocessed data.
2. It should be capable of handling datasets in the order of a few gigabytes and simultaneously retain interactive computation time for the analyses.
3. It should provide version history for the data and analysis code and tools.
4. The system should provide reproducible data analysis reports for modified datasets.
5. It should be portable and reusable by other systems or applications.
6. The system should be able to handle access management, data protection, and privacy concerns such as anonymization.

To satisfy the above requirements we developed the NOWAC R package, a software package in the R programming language to provide access to all data, documentation, and utility functions.

We also identified a set of requirements for this data preprocessing and preparation system:

1. The data preprocessing and preparation system should provide users with an interactive point-and-click interface to generate analysis-ready datasets from the NOWAC study.
2. It should use the NOWAC R package to retrieve datasets.
3. It should provide users with a list of possible options for filtering, normalization, and other options required to preprocess a microarray dataset.
4. It should generate a reproducible report along with any exported dataset.
5. It should export the data in a format that encourages following best practices for reproducible research in further downstream analyses.

Finally, we developed a set of best practices for data analysis in our study. In the remainder of the section we detail how we built the NOWAC package, the Pippeline, and discuss best practices for data analysis.

The NOWAC R package: data management

The NOWAC R package is our solution for storing, documenting, and providing utility functions to parse and process the raw omics data in the NOWAC study (Figure 3.1). We use git to version-control both the analysis code and datasets and store the repository on a self-hosted git server. We bundle together all datasets in the NOWAC package. This includes questionnaire, registry, and gene expression datasets. Because these are small by modern standards (currently all datasets are less than 10 GB), we are able to distribute them with our R package. Some datasets require preprocessing and quality control steps such as the removal of observations marred by technical artefacts (we sometimes refer to this as outlier removal) before the analysts explore the datasets. For this, we store the raw datasets and the results of quality assessment. We store links to the raw datasets in their original file format and as R data files to simplify importing in R. In addition, we store the R code we used to generate the R objects. For clarity, we decorate the scripts with specially formatted comments that can be used with knitr (Dynamic Documents with R and knitr) to automatically generate data analysis reports. The reports highlight the transformation of the data from raw to processed and detail all information necessary to reproduce the entire processing, such as the specification of removed samples.

We have documented every raw dataset in the NOWAC R package. The documentation includes information such as data collection date, instrument types, the persons involved with data collection and analysis, and pre-processing methods. When users install the NOWAC R package, the documentation is used to generate interactive help pages which they can browse in R, either through the command line or through an integrated development environment (IDE) such as RStudio. We can also export this documentation to a range of different formats, and researchers can also view them in the RStudio interface. Figure 3.2 shows the user interface of RStudio where the user has opened the documentation page for one of the gene expression datasets.

We use a single repository for the R package and put each dataset into a git submodule (Figure 3.3). This allows us to separate access to the datasets from the documentation and analysis code for data security and privacy reasons. Everyone with access to the repository can view the documentation and analysis code, but only a few have

access to the data. Submodules allow us to keep the main repository size small, while still versioning the data. The NOWAC R package also provides various utility functions to process the raw datasets, and helper functions to retrieve questionnaire data.

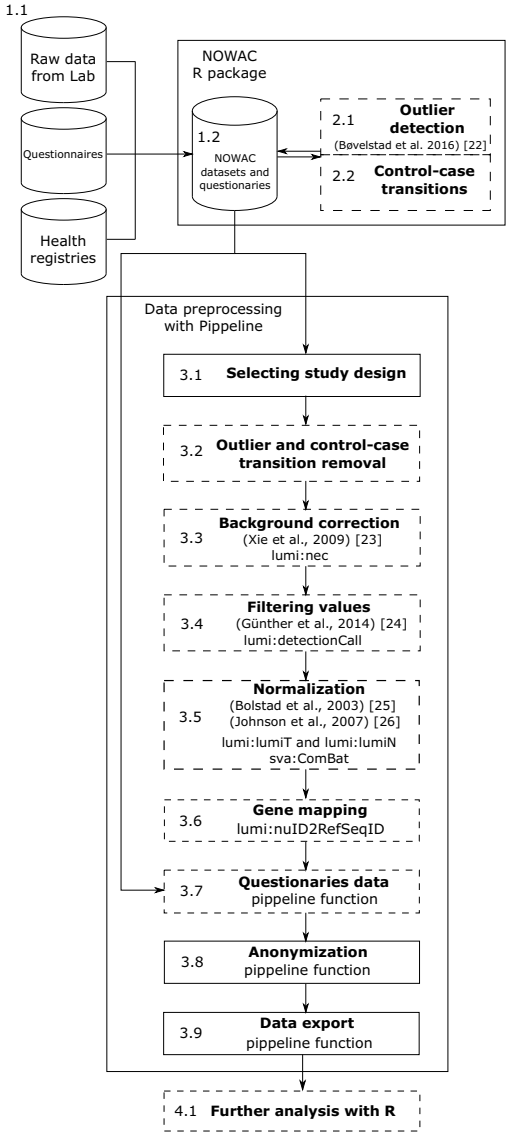


Figure 3.1. The standardized data processing pipeline for gene expression data preprocessing in the NOWAC study. Steps with a dashed border are optional, while steps with a solid border are mandatory. Further details are available in Bolstad et al. 2003, Johnson et al. 2007, Xie et al. 2009, Günther et al. 2014, Bøvelstad et al. 2017.

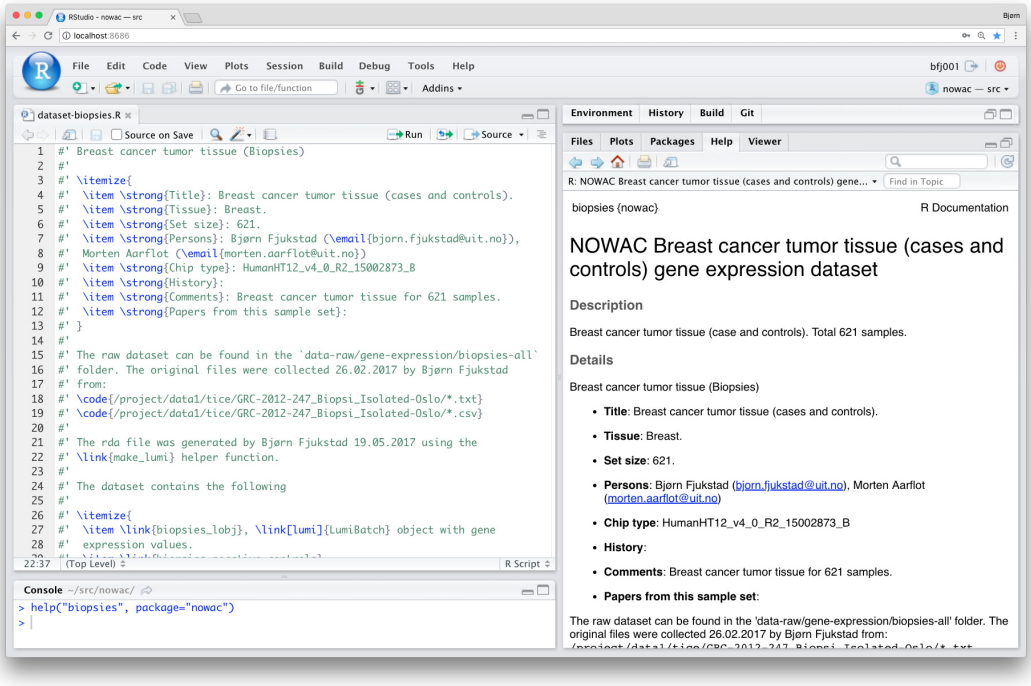


Figure 3.2. A screenshot of the user interface in R Studio, viewing the documentation help page for the "Biopsies" dataset in the NOWAC study. The right-hand panel shows the documentation generated by the code in the top left panel. The bottom left panel shows the R command that brought up the help page.

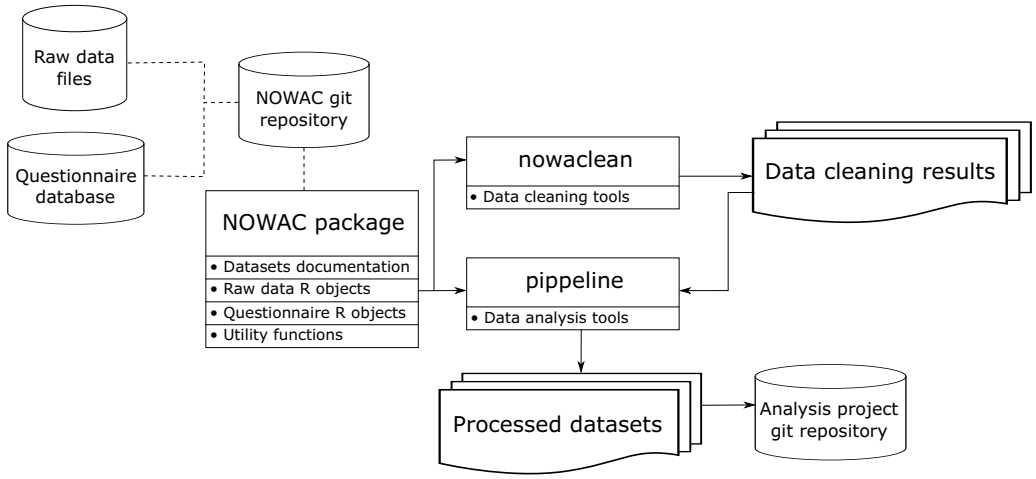


Figure 3.3. NOWAC R package and Pippeline deployment.

Pippeline: Interactive Preprocessing Web Application

The use of the biological data in the NOWAC study in a research project comprises four steps (Figure 3.1). First, as explained above, the raw datasets are added to the NOWAC R package and documented thoroughly by a data manager. Second, we perform manual quality assessment of the biological datasets. We add information about technical outliers to the NOWAC R package along with reports that describe why an observation is marked as an outlier. Third, the data manager generates an analysis-ready gene expression dataset for subsequent analysis using the interactive Pippeline tool as described below. Fourth, researchers further analyze the exported dataset with their tools of choice, following best practices for reproducible data analysis.

We have developed our preprocessing pipeline for gene expression data as a point-and-click web application called Pippeline. Pippeline generates an analysis-ready dataset by integrating biological datasets with questionnaire and registry data, all found in our NOWAC package. It allows selecting study design, removing already-discovered technical outliers, data normalization methods, filter values, and questionnaire fields. It presents the user with a list of possible processing options. We provide summary statistics for samples and probes about the changes made on each processing step in real time, so Pippeline users can see how each preprocessing step changes the number of samples and probes in the dataset (Figure 3.4). Pippeline exports the analysis-ready R data files, an R script that has all the choices and selections made during the preprocessing, and an R markdown which contains a human-readable report that can be used in the Methods section of a paper. The R script enables reproducing the output and intermediate data if needed.

Pippeline is implemented in R as a web application using the Shiny framework. It uses the NOWAC R package to retrieve all datasets. Intermediate data processing is implemented by creating temporary R files with the necessary functions for steps that are executed when the user interacts with the web application. Since our microarray datasets are small, the processing is very fast. Usually the Pippeline processing takes about 40 seconds. In the final step of Pippeline we create a git repository with the output files, clone the repository to a folder on the user's home directory, and create an RStudio project.

Study-specific data analysis and result interpretation using R

Study-specific analyses are done by researchers using their methods and tools of choice, for example in RStudio. To encourage best practices for reproducible research, we provide the following measures:

First, Pippeline exports the data as an RStudio project file with the data stored in a git repository. RStudio provides a graphical user interface for using git to version the code and data. This makes it easy to start using version control for any researcher.

Second, the NOWAC R package provides documentation about datasets. Missing information and corrections can be added either as suggested changes to the package or as issues to the package repository. This makes it easy to keep the documentation up to date.

Third, we provide a server with the necessary computational resources and software for the analysis, and we do not allow the data to be copied to another system. This makes it easy to keep all the data and code in one system, and to employ proper access management.

Finally, we encourage using best practice regarding software and data management in our research group, and we give tutorials and workshops to teach these practices.

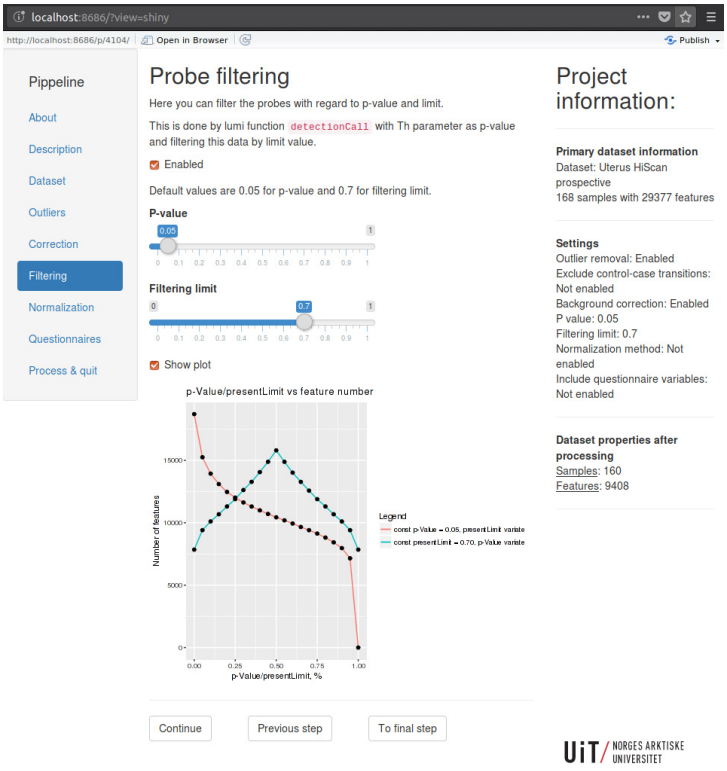


Figure 3.4. A screenshot of Pippeline's web interface. In the filtering step, users specify the p-value and filtering limit for excluding gene expression probes in the dataset.

Deployment of the NOWAC R package and Pippeline

We have deployed the NOWAC R package and Pippeline on two machines, and in addition we use our University's storage system for raw data storage, and a database server for the questionnaire data. The storage machine runs the git and git-lfs servers used by the NOWAC package, and by the individual research projects. Only a few selected users have access to this machine. Another computer is used by the NOWAC researchers for their study-specific analyses to run the Pippeline. This machine has an RStudio server that the user can access through the browser. The machine also has home directories for the research projects. Finally, the researchers have their own laptops and workstations, used solely to establish a connection to the servers. No data should be copied out of the servers.

Datasets stored in NOWAC R package and processed by Pippeline

To date we have used the NOWAC package and Pippeline for our 11 microarray datasets, but we are in the process of adding other data types also including microRNA, targeted RNA-seq, and methylation. The storage usage for the NOWAC package is 1.6 gigabytes including all R data objects. The total Pippeline output is 917 megabytes. The raw microarray (text) files are 8.7 gigabytes in size, but the corresponding R objects are more efficiently stored.

Best practices for reproducible epidemiological data analysis

From our experiences we have developed a set of best practices for data analysis. These apply to researchers, developers, and the technical staff managing the data in a research study:

First, document every step in the analysis. Analysis of modern datasets is a complex exercise with the potential to introduce errors at every step. Analysts often use different tools and systems that require a particular set of input parameters to produce results. Thoroughly document every step from raw data to the final tables that go into a manuscript.

In the NOWAC study, we write help pages and reports for all datasets, and the optional pre-processing steps.

Second, generate reports and papers using code. With tools such as R Markdown and knitr there are few reasons for decoupling analysis code from the presentation of the results through reports or scientific papers. Doing so ensures the correctness of reported results from the analyses, and greatly simplifies reproducing the results in a scientific paper.

In the NOWAC study we produce reports from R code to document preprocessing of delivered data. When a researcher requests access to a dataset, we export the dataset with R, and produce a report that contains information on what is in the dataset, possible filtering that have been made, and who exported the data.

Third, version-control everything. We track changes to source code and datasets with modern version control systems (VCS). Both code and data change over the course of a research project. With VCS it is possible to retrace changes and the person responsible for them. It is often necessary to roll back to previous versions of a dataset or analysis code, or to identify the researchers who worked on specific analyses. In the NOWAC study we encourage the use of git to version control both source code and data.

Fourth, collaborate and share code through source code management (SCM) systems. Traditional communication through e-mail makes it difficult to keep track of existing analyses and their design choices both for existing project members and new researchers. With SCM hosting systems such as GitHub, GitLab or Azure DevOps, the development of analysis code becomes more transparent to other collaborators, and encourages collaboration. It also simplifies the process of archiving development decisions such as choosing a normalization method.

In the NOWAC study we collaborate on data analysis through a self-hosted GitLab installation, the first single application for the entire DevOps lifecycle. We believe the ready-made git repository output from Pippeline encourages good software development practices and provides a good foundation for effective collaborative work.

Limitations

A potential drawback of using an R package that is version-controlled in git to manage, document, and analyze research datasets for researchers, is the requisite programming skills. While the topic of software engineering best practices may be absent in the current research training of many researchers, we believe such skills will become increasingly common in the scientific community.

One possible limitation of our NOWAC R package is its size. Microarray datasets are relatively small compared to sequencing data, so new datasets may require using extensions to git such as git-lfs, as we used in Walrus (Fjukstad et al. 2018). Since we are currently expanding the NOWAC package and creating interactive pipelines like the Pippeline workflow for RNA-seq, methylation, and microRNA datasets, this may become necessary.

CONCLUSIONS

We have proposed an approach to enabling reproducible analyses for epidemiological omics data analyses. Our solution consists of several software tools embedded in the proper methodology, as well as best practices, and solves a number of challenges previously encountered in omics studies. Among the advantages of our approach are the proper separation of datasets and tools, access management, anonymization, tracking of software versions and dataset changes, documentation of processing steps and corresponding parameters, as well as cross-platform support, and an easy-to-use graphical interface. It is also very fast. While we have applied our approach to a specific epidemiological research study for successful verification, we believe that it is generalizable to other biomedical analyses and even other scientific disciplines.

The NOWAC R package, without our data and data documentation, is available at: https://github.com/uit-hdl/nowaclite

Pippeline and a description of our microarray preprocessing pipeline are available at: https://github.com/uit-hdl/pippeline

A demo dataset from Bøvelstad et al. 2017 is available at: https://doi.org/10.18710/FGVLKS

ACKNOWLEDGEMENTS

The NOWAC study was supported by a grant from the European Research Council (ERC-AdG 232997 TICE).

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.

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.

REFERENCES

Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Cech M et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 2016 Jul 8; 44(W1): W3–W10. Available from: https://academic.oup.com/nar/article/44/W1/W3/2499339

Amstutz P, Crusoe MR, Tijanic C, Chilton J, Heuer M, Kartashov A. Common Workflow Language, v1.0. Posted: 08.07.2016. Available from: https://figshare.com/articles/Common_Workflow_Language_draft_3/3115156

Azure DevOps [Software; Internet]. Available from: https://azure.microsoft.com/en-us/services/devops/

Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003 Jan 22; 19(2): 185–193. Available from: https://academic.oup.com/bioinformatics/article/19/2/185/372664

Busund M, Bugge NS, Braaten T, Waaseth M, Rylander C, Lund E. Progestin-only and combined oral contraceptives and receptor-defined premenopausal breast cancer risk: The Norwegian Women and Cancer Study. Int J Cancer. 2018 Jun 1; 142(11): 2293–2302. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/ijc.31266

Bøvelstad HM, Holsbø E, Bongo LA, Lund L. A Standard Operating Procedure For Outlier Removal In Large-Sample Epidemiological Transcriptomics Datasets. bioRxiv 2017 May 31. Available from: https://www.biorxiv.org/content/10.1101/144519v1

Dynamic Documents with R and knitr. CRC Press [Internet]. Accessed: 27.02.2019. Available from: https://www.crcpress.com/Dynamic-Documents-with-R-and-knitr/Xie/p/book/9781498716963

Dumeaux V, Fjukstad B, Fjosne HE, Frantzen JO, Holmen MM, Rodegerdts E et al. Interactions between the tumor and the blood systemic response of breast cancer patients. PLoS Comput Biol. 2017 Sep 28; 13(9): e1005680. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005680

Dumeaux V, Lund E. Gene expression profile in diagnostics [Patent]. World Intellectual Property Organization 2014 May 30: WO2014081313 AI. Available from: https://patentimages.storage.googleapis.com/f3/58/0e/82a897ca036515/WO2014081313A1.pdf

Dumeaux V, Olsen KS, Nuel G, Paulssen RH, Børresen-Dale AL, Lund E. Deciphering normal blood gene expression variation--The NOWAC postgenome study. PLoS Genet. 2010 Mar; 6(3): e1000873. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000873

Fjukstad B, Dumeaux V, Hallett M, Bongo L. Reproducible Data Analysis Pipelines for Precision Medicine. bioRxiv. 2018 Jun 25. Available from: https://www.biorxiv.org/content/10.1101/354811v1.full

Gallo V, Egger M, McCormack V, Farmer PB, Ioannidis JP, Kirsch-Volders M et al. STrengthening the Reporting of OBservational studies in Epidemiology-Molecular Epidemiology (STROBE-ME): an extension of the STROBE statement. Eur J Epidemiol. 2011 Oct; 26(10): 797–810. Available from: https://link.springer.com/article/10.1007%2Fs10654-011-9622-1

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004; 5(10): R80. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2004-5-10-r80

GitHub [Development platform; Internet]. Available from: https://github.com

GitLab [Application; Internet]. Accessed: 13.05.2019. Available from: https://about.gitlab.com

Gram IT, Little MA, Lund E, Braaten T. The fraction of breast cancer attributable to smoking: The Norwegian women and cancer study 1991–2012. Br J Cancer. 2016 Aug 23; 115(5): 616–623. Available from: https://www.nature.com/articles/bjc2016154

Günther CC, Holden M, Holden L. Preprocessing of gene-expression data related to breast cancer diagnosis. Oslo: Norwegian Computing Center; 2014. pp 41. Available from: https://www.nr.no/files/samba/smbi/note2015SAMBA3514preprocessing.pdf

Holden M, Holden L, Olsen KS, Lund E. 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. 2017; 7: 11–28. Accessed: 27.02. 2019. Available from: https://www.dovepress.com/local-in-time-statistics-for-detecting-weak-gene-expression-signals-in-peer-reviewed-article-AGG

Ivie P, Thain D. Reproducibility in Scientific Computing. ACM Comput Surv 2018 Jul; 51(3): 63. Available from: http://ccl.cse.nd.edu/research/papers/repro-survey.pdf

Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007 Jan; 8(1): 118–127. Available from: https://academic.oup.com/biostatistics/article/8/1/118/252073

Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*. 2012 Oct 1; 28(19): 2520–2522. Available from: https://academic.oup.com/bioinformatics/article/28/19/2520/290322

Lund E, Dumeaux V. Systems epidemiology in cancer. Cancer Epidemiol Biomarkers Prev. 2008 Nov; 17(11): 2954–2957. Available from: https://cebp.aacrjournals.org/content/17/11/2954.long

Lund E, Dumeaux V, Braaten T, Hjartaker A, Engeset D, Skeie G et al. Cohort profile: The Norwegian Women and Cancer Study--NOWAC--Kvinner og kreft. Int J Epidemiol. 2008 Feb; 37(1): 36–41. Available from: https://academic.oup.com/ije/article/37/1/36/763947

Olsen KS, Fenton C, Frøyland L, Waaseth M, Paulssen RH, Lund E. Plasma fatty acid ratios affect blood gene expression profiles--a cross-sectional study of the Norwegian Women and Cancer Post-Genome Cohort. PloS One 2013 Jun 26; 8(6): e67270. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0067270

Pachyderm [Data Science Platform; Internet]. Accessed: 13.05.2019. Available from: https://www.pachyderm.io/

Reality check on reproducibility [Editorial]. Nature. 2016 May 26; 533(7604): 437. Available from: https://www.nature.com/articles/533437a

R Markdown [Software; Internet]. Accessed 13.05.2019. Available from: https://rmarkdown.rstudio.com/

Shiny [Software; Internet]. Accessed: 13.05.2019. Available from: https://shiny.rstudio.com/

The Comprehensive R Archive Network [Software; Internet]. Accessed: 13.05.2019. Available from: https://cran.r-project.org/

Xie Y, Wang X, Story M. Statistical methods of background correction for Illumina BeadArray data. Bioinformatics. 2009 Mar 15; 25(6): 751–757. Available from: https://academic.oup.com/bioinformatics/article/25/6/751/251033

Zaharia M, Xin RS, Wendell P, Das T, Armbrust T, Dave A et al. Apache Spark: A Unified Engine for Big Data Processing. Commun ACM. 2016 Oct; 59(11): 56–65. Available from: https://m-cacm.acm.org/magazines/2016/11/209116-apache-spark/abstract

Lund, E. (Ed.) (2020). *Advancing Systems Epidemiology in Cancer. Exploring Trajectories of Gene Expression.*
Oslo: Scandinavian University Press.
DOI: https://doi.org/10.18261/9788215041193-2020-04

4. Practical and Ethical Issues in Establishing a Collection of Normal Breast Tissue Biopsies—Part of the NOWAC Post-Genome Cohort

Sanda Krum-Hansen and Karina Standahl Olsen

Abstract For tissue-based studies of breast cancer, getting access to truly normal, well-annotated tissue can be a challenge. To address that need, we collected 368 breast tissue biopsies and buffered blood samples from healthy postmenopausal women. Volunteers were part of the Norwegian Women and Cancer (NOWAC) Post-genome cohort, recruited through the national mammography screening program. The NOWAC normal breast tissue biobank for gene expression analysis will provide a correct basis for comparison in case-control studies.

Keywords normal breast tissue | biobank | breast cancer

BACKGROUND

Epidemiology and risk factors of breast cancer

Breast cancer is the most frequent type of cancer among females worldwide. The latest GLOBOCAN report estimated approximately 2.1 million newly diagnosed breast cancers in 2018 (Bray et al. 2018).The incidence of breast cancer varies significantly around the world, but is increasing in most countries (Bray et al. 2018). The high incidence in developed countries has to some extent been counterbalanced by a reduction in mortality. Early diagnosis due to mammographic screening, improved treatment, secondary prophylaxis and follow-up have improved the outcome for breast cancer patients. The 5-year survival rate in Norway is 90.4%—yet breast cancer is the leading course of cancer-related deaths among females

(Cancer Registry of Norway 2017). The increasing incidence and improved survival rate results in high prevalence of the disease. Since the treatment is associated with severe side effects over a long period, the burden of the disease is large.

The current body of evidence suggests that genetic structure and internal and external risk factors, as well as their interactions, combine to constitute the causes of breast cancer. Two major risk factors are gender and age. Other causal factors relate to the levels of endogenous hormones determined by age at the first menstruation, age at menopause, age at first birth, and number of births, as well as use of oral contraceptives and hormone therapy (HT) (Kaminska et al. 2015). Lifestyle factors regarded as risk factors include lack of physical activity, obesity, alcohol consumption, smoking, night shift work, exposure to radiation, and possibly diet (Sun et al. 2017). Hereditary breast cancer accounts for 5–10% of cases (Apostolou and Fostira 2013), making non-hereditary risk factors the major drivers of incidences of breast cancer.

Breast cancer characteristics

Breast cancer is a heterogeneous disease both etiologically and genetically. It consists of several sub-types with different molecular profiles, and biological and clinical behavior. Different sub-groups are associated with different risk profiles and present a big challenge for clinical management. In clinical practice, an array of methods is used to determine which sub-type the patient has: tumor-node-metastasis (TNM) staging, histological sub-typing, tumor grade, tumor invasion in lymphatic and vascular tissue, axillary lymph node status, immune-histochemical staging providing estrogen and progesterone receptor status, presence of human epidermal growth factor receptor 2 (HER2) receptor, and Ki67 marker. These factors describe the tumor biology regarding hormone sensitivity and tumor aggressiveness, guide decision-making for treatment, and predict the prognosis.

Today there are efficient surgical and medical treatments available, but we are unable to determine specifically which type of treatment the individual patient needs, often implying overtreatment. There is a need for better prognostic and predictive markers to individualize the treatment in order to provide the best treatment for patients with high-risk profiles, and to avoid overtreatment of patients with a low risk profile.

Normal breast tissue histology and development

The human breast is an apocrine gland designed to produce milk, and breast tissue is heterogeneous and complex in composition. The breast consists of three main

components: the skin, containing areola and nipple, the subcutaneous adipose tissue (white fat tissue), and the glandular tissue (functional tissue of the breast) including both parenchyma and stroma. The parenchyma is divided into 15–25 lobes, each made up of 20–40 lobules. The structure is based on a branching duct system that leads from the collecting ducts to the terminal duct-lobular units (TLUs). The TLUs are the functional unit of the breast tissue and sites of milk production. The terminal collecting ducts drain the milk from TDLUs into 4–18 lactiferous ducts, which drain to the nipple. The inter- and perilobular connective tissue surrounding the TLUs and lobules contain fibrovascular tissue and white adipose tissue. Fibrous stroma provides the background architecture for the glandular tissue, as well as nutrition and protection. The proportion of adipose and fibrous tissues varies from one woman to another and changes in the same person over time.

Breast tissue development occurs in defined stages: embryonic, pre-pubertal, pubertal, pregnancy, lactation and involution. The tissue only reaches its final level of development during the last stages of pregnancy, and if pregnancy does not occur, it is never reached. During menopause, the glandular tissue is progressively atrophied. The lobules decrease in size and number, mainly through progressive involution of the milk-producing acini. Fibrous tissue is also replaced by adipocytes. However, the extensive use of hormonal replacement therapy has considerably changed the appearance of this postmenopausal breast tissue.

Biobanking of normal breast tissue for research

Tissue-based studies of breast carcinogenesis utilize breast cancer tissue and different types of non-cancerous breast tissue, sometimes called normal breast tissue, as control for comparison. Most commonly used non-cancerous breast tissue is derived from reduction mammoplasty either from breast cancer patients, of unaffected breast for symmetry in breast cancer patients, or from healthy women operated for cosmetic purposes. Other sources of non-malignant breast tissue used in research include tissue from prophylactic mastectomy, neighboring breast tissue from women with benign breast lesions, excess tissues with benign histological appearance collected from surgical procedures, or unaffected ipsilateral or contralateral breast tissue from patients with breast cancer.

Usually there is a medical reason to surgically remove tissue—for example in prophylactic mastectomy for high risk of breast cancer due to gene mutations, or removal of benign lesions due to pathological features. Therefore, this type of tissue is not suitable for use as "normal" tissue. Breast tissue collected by reduction

mammoplasty, selected on the basis of convenience, may be the best representative of normal tissue. It is plentiful and removed for cosmetic reasons, not because of clinical abnormalities or high-risk profiles. However, none of these tissues have been found suitable as a substitute for truly normal breast tissue in studies of breast cancer carcinogenesis (Ambaye et al. 2009, Graham et al 2010, Degnim et al. 2012, Tadler et al. 2014, Acevedo et al. 2019).

Today there are several breast cancer tissue biobanks around Europe, North and South America, Asia and Australia, but to our knowledge the only biobank that collects truly normal breast tissue is the Susan G Komen for the Cure Tissue Bank (KTB) at Indiana University Simon Comprehensive Cancer Center in the USA (Sherman et al. 2012). There, tissue has been collected from volunteers of all ethnicities aged 18 and upward. Several articles have been published using this material. Radovic et al. 2014 concluded that breast tissue from healthy volunteers acts as a superior normal breast tissue control. The same source of tissue has been used in Pardo et al. 2014, where the author analyzed the transcriptome of normal, healthy, pre-menopausal breast tissue using next-generation sequencing.

In order to move breast cancer research forward, there is a need for well-annotated collections of breast tissue from healthy women (Thompson et al. 2008, Eccles et al. 2013). Adequate control tissue will help shed light on pre-clinical molecular events, and provide the correct basis for comparison in case-control studies. The overall goal of this study was to establish a biobank of normal breast tissue biopsies. The biobank was established for the purpose of describing baseline gene expression patterns in normal breast tissue of postmenopausal women. We will also explore the variation of gene expression in normal breast tissue following exposure to known breast cancer risk factors (smoking, alcohol consumption, HT use, obesity and parity), and finally, we will use the normal breast tissue in future case-control studies.

METHODS

The normal breast tissue biopsy study, part of the NOWAC Postgenome cohort

This study is part of the Norwegian Women and Cancer (NOWAC) Postgenome cohort. NOWAC is a national, prospective study started in 1991, where breast cancer is the most important endpoint (Lund et al. 2008). The study included 150 000 women born 1943–1957, who to date have answered between one and three questionnaires. During the period 2003–2006 we built a unique biobank by collecting

blood samples, buffered to protect the mRNA gene expression profile, from 50 000 NOWAC participants. These samples constitute the major part of the NOWAC Postgenome cohort. Furthermore, starting in 2006 and in collaboration with 11 Norwegian hospitals, we collected buffered blood samples and tissue samples from 400 women with breast cancer tumors at the time of diagnosis. These women were also participants in NOWAC, they were born between 1943–1957, and were diagnosed with breast cancer during the period 2006–2011. Until that time, there was no suitable tissue material available that expressed the normal pattern of variation in gene expression in the relevant age group. To address that need, during the period 2010–2012 we collected breast tissue and buffered blood samples from 368 healthy women. Volunteers for this part of the study were recruited from the NOWAC cohort through the national mammography screening program, which they were participating in at the time.

Recruitment of study participants

Recruitment to the study and the tissue collection took place at the Breast Diagnostic Center at the University Hospital of Northern Norway (UNN), Tromsø, Norway. Inclusion criteria were as follows: enrolled in the NOWAC cohort, born between 1943 and 1957, and consent given. The radiographer (not affiliated with the NOWAC study) asked women, when presenting at the mammography screening unit, if they would consider participating in this study. If answering positively, the candidate would meet after the screening procedure for written and oral information and to get answers to any questions they may have had. The women who agreed to participate were asked to sign a written, informed consent form. All participants completed a two-page questionnaire regarding menopausal status, weight and height, exposure to smoking and alcohol consumption, use of HT and other medication. Exclusion criteria included previous history of breast cancer, positive mammogram, other relevant malignant diseases, and use of anticoagulation therapy with Coumadin (Marevan), Heparin, Persantine, or Plavix. Use of acetylsalicylic acid was not an exclusion criterion.

Procedures for tissue and blood sampling

Core biopsies of normal breast tissue were obtained immediately after mammography, from the gland tissue of the upper lateral quadrant of the left breast. The tissue biopsy was taken with the women in declined position on the examination bed. The skin was disinfected with chlorhexidine solution in alcohol prior to incision.

Intradermal local anesthesia was applied using 2 ccl of 1% Lidocaine. A 3 mm skin incision was performed with a scalpel. With ultrasound guidance, a cylinder biopsy was taken with a needle size 14 gauge in a biopsy pistol, by an experienced radiologist. Compression bandage was placed at the biopsy site, which was to be kept in place until the next day. No further activity restriction was advised. During the study, no systematic follow-up has been undertaken. The biopsy was immediately placed in RNAlater for RNA stabilization (Qiagen, Hilden, Germany), and kept at room temperature for <24 hours until storage in a freezer at –70˚C.

Two vials of blood were taken by standard venipuncture (phlebotomy) with hypodermic butterfly needle on a closed system to the vacuum test tubes. One of the blood samples was taken using the PAXgene Blood RNA collection system (Preanalytix/Qiagen, Hombrechtikon, Switzerland), which contains a buffer for stabilizing the mRNA gene expression profile during long-term storage. The other blood sample was mixed with standard citrate solution. Blood samples were kept at –70˚C until further use. The blood sampling was performed before the tissue sampling.

RESULTS

We collected 368 biopsies of normal breast tissue from postmenopausal women. The rate of inclusion of all women invited to participate was 64%. A linkage to the Norwegian Cancer Registry 3 years after the sampling period ended resulted in five biopsies being excluded due to breast cancer diagnosis within 3 years after the biopsy was taken, and one due to a prior lymphoma diagnosis with unknown treatment. We used 16 biopsies for testing of different laboratory methods. A total of 311 biopsies were included for further analysis, which matched the number of cancer biopsies in our biobank collected for a comparative study.

All participants were advised to contact a physician in case of any suspicion of adverse reaction or complication such as hematoma, infection, or pain. No case of allergic reaction to the local anesthesia was registered. One participant directly reported a hematoma at the biopsy site. She was examined by a surgeon, who found a 3 cm hematoma, but no treatment or follow-up was considered necessary.

Characteristics of women included in this study

Characteristics of the 311 women included in the final study sample is summarized in Table 4.1. All participants were post-menopausal, and the average age was 60 years. The population, as a whole, were slightly overweight after WHO standard, with average BMI 26,2. Most of the women had given birth (have completed

full term pregnancy), and the average number of children was 1,9. The highest number of children was 8. A majority of the women (79%) had consumed alcohol during the week before sampling, and 21 % had been smoking during the week prior to biopsy sampling. Very few participants (8,4 %) used HT for menopausal symptoms. The majority of participants (70 %) used different types of medication in the week prior to blood sampling, either alone or in combination. The most frequent types were blood pressure medication, anti-cholesterol drugs, and synthetic thyroid hormone, followed by ASA (aspirin) and NSAIDs.

Table 4.1. Characteristics of the study population (n=311)

Characteristics	Mean/Frequency	Missing
Age, mean (SD)	60,1 (3,9)	0
BMI, mean (SD)	26,2 (4,5)	4
Parity (n, %)		0
Yes	256 (82,3)	
No	55 (17,7)	
N children (mean, SD)	1,9 (1,2)	0
Smoking (n, %)		0
Yes	66 (21,2)	
No	245 (78,8)	
HT use (n, %)		1
Yes	26 (8,4)	
No	284 (91,6)	
Alcohol (n, %)		6
Yes	241 (79)	
No	64 (21)	
Medication use (n, %)		
Any medication	216 (70,8)	6
Blood pressure alone or in comb. with antiarrhythmic	56 (18,4)	
Anti-cholesterol	36 (11,8	
Levaxin (synthetic thyroid medications)	30 (9,8)	
Asthma/allergy	23 (7,5)	
NSAIDs alone or in combination with Paracetamol	22 (7,2)	
Albyl (acetylsalicylic acid)	19 (6,2)	
Other	30 (9,8)	

Abbreviations: BMI, body mass index; HT, hormone therapy; NSAID, non-steroidal anti-inflammatory drugs; SD, standard deviation.

DISCUSSION

Above we have described the process of establishing a biobank of normal breast tissue biopsies from 311 postmenopausal women. In the following we discuss practical aspects of establishing the biobank, as well as ethical considerations, and highlight some factors that enabled the successful establishment of the NOWAC normal breast tissue biobank.

Where to find volunteers and how to recruit them?

The process of recruiting healthy volunteers for an invasive procedure may, if not planned properly, render the final study sample heavily affected by selection bias, subsequently reducing the generalizability of any findings. To reduce selection bias, our starting point was the nationally representative NOWAC study, as well as the national mammography screening program. The screening program invites all Norwegian women aged 50–69 years to mammography every other year, free of charge. Hence, an important success factor for this study was the use of the local screening facility, which enabled us to contact all eligible women in the region.

Prior to our work, the same facility had completed two small surveys (unpublished) to start the process of assessing the feasibility of collecting tissue biopsies from healthy women. The first was conducted to register discomfort and possible complications associated with the biopsy procedure and was based on interviews with 100 women who had undergone this procedure. The women were asked about pain, bleeding, hematoma, and infections. The result was consistent with the impression from the clinical work that biopsy taking is virtually painless and there is a very low risk for complications associated with the procedure. The second survey aimed to determine whether it would be possible to collect breast tissue biopsies from healthy women. We asked 81 women who participated in the mammography screening program if, hypothetically, they would be willing to have a breast biopsy taken to be used for research purposes. After receiving written and oral information, 12% answered no, 14% needed more information, and 74% answered yes. These results gave important cues on feasibility.

Collaboration with clinicians

The local mammography screening facility handles about 40 invitations every day. The NOWAC study has been collaborating with the facility since March 2002, when approximately 2 000 blood samples were collected for a different NOWAC project. The facility also played an active role in recruiting partners for a cancer

biopsy study at eleven of the country's hospitals. This close and long-standing collaboration is another important success factor for the present project. The screening facility already had valuable experience in contributing to research during their clinical everyday setting. Though the environment was familiar with research, it was necessary to make a detailed plan and spend time to figure out the most feasible way to complete all the steps with the clinical personnel involved. This included having the same person involved every day, who was familiar with the hospital environment and the department's work, as well as being involved in the research project.

The biopsy procedure involved is virtually painless, with a very low complication rate, and was performed by an experienced radiologist within the well-established framework of the screening facility, minimizing the risk of unforeseen incidents. All women were given information on actions to be taken in the case of complications. Since the procedures took place in the hospital setting, any complication or injury would be reported as a patient injury according to established national guidelines. Women were encouraged to contact the screening facility if a suspicion of a complication should arise after leaving the department. Complications requiring immediate treatment outside opening hours would be attended by the staff in the emergency room. These actions were largely comparable to actions to be taken in case of complications after any breast tissue biopsy procedure, and put no extra burden on the clinical staff.

Ethical aspects

In accordance with legal requirements for research on human biological material and personal data (The Health Research Act, Chapters 3-7), the Regional Committee for Medical and Health Research Ethics of Northern Norway (REC North) approved the protocol for the present study, and the Data Protection Authority granted a license for the use of health-related data. However, the project was planned some years ago, before the European Union issued the new General Data Protection Regulation (GDPR) in 2018. In Norway, GDPR was implemented at the national level through a new Personal Data Act, also in 2018. The risk of misuse of personal information, or the risk of loss of control of the personal information, is present in the current project, but this risk is by no means greater here than in comparable projects. These aforementioned risks are the focus of GDPR, and after its implementation, data-handling procedures have also been improved for the NOWAC project.

The need for close regulation of biomedical research dates back to atrocities during the Second World War, which led to the emphasis on human rights in the

Nuremberg Code of 1947. A main point in the Code stated that participation in research must be voluntary. Furthermore, the World Medical Association's Declaration of Helsinki (1964) focused on obligations of the researchers and the research institutions, and stressed the concept of informed consent (Fisher 2006). That the consent must be voluntary or free means that the individual included in the research shall not decide his/her position through a process characterized by coercion or pressure. Likewise, situations that do not include direct coercion can mean an unacceptable weakening of the consent that was given. Our participants were already part of the NOWAC study when they were invited for the biopsy study. Potentially, this could contribute to a feeling of pressure to participate in the biopsy study. We, the researchers, regarded this project as a continuation of the ongoing NOWAC study, and this backdrop may have put an indirect pressure on the women at the point of invitation. Still, the option to decline participation was always clearly communicated, both orally and in writing, hence we conclude that the principle of voluntary participation was never challenged.

The principle of informed consent entails that the individual being subjected to research must be aware of the study's methodology/procedures, purposes, and the type of results expected. The information given to participants must include a description of any expected inconvenience, discomfort, or risk that may be inflicted. This principle may be regarded as particularly important when performing an invasive procedure on healthy volunteers who would not otherwise undergo such procedures. Further, as the material collected in our study will be used for genomic profiling (mRNA gene expression analysis and potentially DNA profiling), care must be taken to ensure that participants understand the information that was given. The participants may have different experiences and assumptions when they internalize and interpret the information. We did not undertake any evaluation of the participants' understanding of the scientific content of the project, but each woman spoke personally to our radiologist, with ample opportunity to ask questions. Legislation on this topic focuses only on groups of people that may be noncompetent to give consent (e.g. persons under the age of 18, or for medical reasons). Hence, some questions may be ethically interesting, but will not have any practical consequences for our project. As examples, one might ask if it would be ethically acceptable to include participants if we discovered that they had not understood the information correctly. In addition, what about individuals who did not want to read the information that was given, but nonetheless wished to participate in the project?

One of our pre-study surveys assessed the healthy women's willingness to donate a breast tissue biopsy. The majority (74%) were willing to donate, and many women expressed a high degree of motivation to continue contributing to research

on breast cancer. Contributing biological sample material to research may be viewed in different ways. The biopsy may be a viewed as a gift or a donation, with no expectation of receiving anything in return. It may also be viewed as a transaction. It that case, the regional ethical committee would act as the real estate agent, looking out for the donors rights, and the consent form may be regarded as the contract between the two parties in the transaction. Viewed as a transaction, there is an expectation of receiving something in return, in this case somewhat distant "payments" such as knowledge of breast cancer, and better treatment. Another option for how to view the act of contributing a biopsy would be as an act of reciprocity. Modern-day medicine is an empirical science which has been built on the knowledge generated from the general population and from patients. Patients today expect to receive the latest treatments that are developed on the basis of this knowledge, and as such, they are morally obliged to contribute to that same knowledge base. In this normative ethics setting, the consent may be viewed as an expression of gratitude toward previous sample donors, of acknowledgment of the moral obligation to contribute, of the will to contribute, and of trust in that the donated material will be used as intended.

We do not have information on each woman's motivation to contribute to the study, but some external factors may also be at play. The city of Tromsø is small, with only 72 000 inhabitants. The city's one university is young and was founded in 1968 during a period of strong growth for the city, and, naturally, its foundation contributed to this growth. Today, the university is one of the city's two largest workplaces, along with the university hospital. These aspects contribute to the fact that the university is a strong part of the city's identity and the inhabitants are well known for contributing to research (Jacobsen et al. 2012). Hence, the feeling of reciprocity, grounded in normative ethics, and supported by favorable local conditions, may be important aspects for the high participation rates in the present study.

There is an ongoing debate on whether researchers should be obliged to return information on health-related aspects to research participants (Klingstrom et al. 2018). However, the present study and its analytical methodology is purely explorative in nature. No clinical relevance of potential findings based on our chosen analytical methods has been established (low clinical validity), and any findings would be non-actionable (i.e. the participant or clinicians could not take action to improve the risk or progression of a potential disease) (Klingstrom et al. 2018). Based on the limited clinical relevance of any findings in this project, any results were unlikely to affect the patient's need for further information, or for their consent. Hence, in this project giving feedback to participants was not considered as relevant, and this was stated in the information given to participants.

Strength and weaknesses

Firstly, the women were recruited from the mammography screening program, not referred from a physician due to symptoms or suspicion of breast pathology. Their biopsies are therefore representative of truly normal breast tissue, and the women have the same risk of developing breast cancer as any other women in the same age group. Since all women were NOWAC participants, extensive information on exposures in the past can be retrieved from questionnaires answered prior to the initiation of the biopsy study. Further strengths of the study include the high inclusion rate (64%) and the high number (368) of biopsies sampled via a standard procedure, which ensures low technical variability. The blood samples were taken at the same time as the biopsies, enabling a valid comparison of gene expression profiles in two different tissues.

One weakness of the study pertains to the risk of selection bias. Our participants were recruited at the mammography screening facility in Tromsø, hence, at one single location. As a consequence, there is the possibility of geographical differences compared to the average Norwegian population regarding the gene expression in relation to different types of exposures. It should be mentioned that the blood and tissue samples were collected by random and continuous invitation during the whole 2-year period, so we expect minimal influence of seasonal biorhythms.

Due to heterogeneity of breast tissue, one single biopsy is not representative of the entire breast. Studies have shown intra-individual variability in composition of breast biopsies, and its impact on gene expression (Chollet-Hinton et al. 2018). This fact has important implications for studies based on normal breast tissue, including our own study. Since our inclusion rate was high and the complication rate turned out to be almost nil, we could have chosen to sample several biopsies from different areas of the same breast via the same skin incision. This can be considered for future trials, taking the varying biopsy composition into account. On the same note, our biopsies are whole tissue biopsies containing multiple cell types which may confound gene expression results. The biopsies were not histologically controlled/evaluated, so we do not have information on the ratio between different cell types. The biopsies were taken from the upper lateral area of the breast, known for a higher density of glandular tissue, in order to reduce the amount of adipocytes and increase mRNA output amounts. However, the biopsies were collected from postmenopausal women. The quantity of glandular tissue decreases with age, and our biopsies likely contain a higher proportion of fat and less glandular tissue compared to samples taken from younger women.

CONCLUSION

The work presented shows that establishing a collection of normal breast tissue samples is feasible and doable. Enabling factors for the present study included largely unbiased access to eligible participants, and close collaboration with clinicians during all steps of the sampling procedures. Furthermore, the source population of the present study has a high degree of health literacy and willingness to participate in research, which contributes to a high participation rate. The NOWAC normal breast tissue biobank for gene expression analysis will provide much-needed information on pre-clinical molecular events and a correct basis for comparison in case-control studies.

REFERENCES

Acevedo F, Armengol VD, Deng Z, Tang R, Coopey SB, Braun D et al. Pathologic findings in reduction mammoplasty specimens: a surrogate for the population prevalence of breast cancer and high-risk lesions. Breast Cancer Res Treat. 2019 Jan; 173(1): 201–207. Available from: https://link.springer.com/article/10.1007%2Fs10549-018-4962-0

Ambaye AB, MacLennan SE, Goodwin AJ, Suppan T, Naud S, Weaver DL. Carcinoma and atypical hyperplasia in reduction mammaplasty: increased sampling leads to increased detection. A prospective study. Plast Reconstr Surg. 2009 Nov; 124(5): 1386–1392. Available from: https://insights.ovid.com/pubmed?pmid=20009822

Apostolou P, Fostira F. Hereditary breast cancer: the era of new susceptibility genes. Biomed Res Int. 2013; 2013: 747318. Available from: https://www.hindawi.com/journals/bmri/2013/747318/

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018 Nov; 68(6): 394–424. Available from: https://onlinelibrary.wiley.com/doi/full/10.3322/caac.21492

Cancer Registry of Norway. Cancer in Norway 2017 – Cancer incidence, mortality, survival and prevalence in Norway. Oslo: Cancer Registry of Norway, 2018

Chollet-Hinton L, Puvanesarajah S, Sandhu R, Kirk EL, Midkiff BR, Ghosh K et al. Stroma modifies relationships between risk factor exposure and age-related epithelial involution in benign breast. Mod Pathol. 2018 Jul; 31(7): 1085–1096. Available from: https://www.nature.com/articles/s41379-018-0033-7

Degnim AC, Visscher DW, Hoskin TL, Frost MH, Vierkant RA, Vachon CM et al. Histologic Findings in Normal Breast Tissues: Comparison to Reduction Mammaplasty and Benign Breast Disease Tissues. Breast Cancer Res Treat. 2012 May; 133(1): 169–177. Available from: https://link.springer.com/article/10.1007%2Fs10549-011-1746-1

Eccles SA, Aboagye EO, Ali S, Anderson AS, Armes J, Berditchevski F et al. Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer. Breast Cancer Res. 2013 Oct 1; 15(5): R92. Available from: https://breast-cancer-research.biomedcentral.com/articles/10.1186/bcr3493

Fischer BA 4th. A summary of important documents in the field of research ethics. Schizophr Bull. 2006 Jan; 32(1): 69–80. Available from: https://academic.oup.com/schizophreniabulletin/article/32/1/69/2888604

Graham K, de las Morenas A, Tripathi A, King C, Kavanah M, Mendez J et al. Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. Br J Cancer. 2010 Apr 13; 102(8): 1284–1293. Available from: https://www.nature.com/articles/6605576

Jacobsen BK, Eggen AE, Mathiesen EB, Wilsgaard T, Njolstad I. Cohort profile: the Tromso Study. Int J Epidemiol. 2012 Aug; 41(4): 961–967. Available from: https://academic.oup.com/ije/article/41/4/961/683871

Kamińska M, Ciszewski T, Łopacka-Szatan K, Miotła P, Starosławska E. Breast cancer risk factors. Prz Menopauzalny. 2015 Sep; 14(3): 196–202. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4612558/

Klingstrom T, Bongcam-Rudloff E, Reichel J. Legal & ethical compliance when sharing biospecimen. Brief Funct Genomics. 2018 Jan 1; 17(1): 1–7. Available from: https://academic.oup.com/bfg/article/17/1/1/3782585

Lund E, Dumeaux V, Braaten T, Hjartåker A, Engeset D, Skeie G et al. Cohort profile: The Norwegian Women and Cancer Study – NOWAC – Kvinner og kreft. Int J Epidemiol. 2008 Feb; 37(1): 36–41. Available from: https://academic.oup.com/ije/article/37/1/36/763947

Pardo I, Lillemoe HA, Blosser RJ, Choi M, Sauder CAM, Doxey DK et al. Next-generation transcriptome sequencing of the premenopausal breast epithelium using specimens from a normal human breast tissue bank. Breast Cancer Res. 2014 Mar 17; 16(2): R26. Available from: https://breast-cancer-research.biomedcentral.com/articles/10.1186/bcr3627

Radovich M, Clare SE, Atale R, Pardo I, Hancock BA, Solzak JP et al. Characterizing the heterogeneity of triple-negative breast cancers using microdissected normal ductal epithelium and RNA-sequencing. Breast Cancer Res Treat. 2014 Jan; 143(1): 57–68. Available from: https://link.springer.com/article/10.1007%2Fs10549-013-2780-y

Sherman ME, Figueroa JD, Henry JE, Clare SE, Rufenbarger C, Storniolo AM. The Susan G. Komen for the Cure Tissue Bank at the IU Simon Cancer Center: a unique resource for defining the "molecular histology" of the breast. Cancer Prev Res (Phila). 2012 Apr; 5(4): 528–535. Available from: https://cancerpreventionresearch.aacrjournals.org/content/5/4/528.long

Sun YS, Zhao Z, Yang ZN, Xu F, Lu HJ, Zhu ZY et al. Risk Factors and Preventions of Breast Cancer. Int J Biol Sci. 2017 Nov 1; 13(11): 1387–1397. Available from: https://www.ijbs.com/v13p1387.htm

Tadler M, Vlastos G, Pelte MF, Tille JC, Bouchardy C, Usel M et al. Breast lesions in reduction mammaplasty specimens: a histopathological pattern in 534 patients. Br J Cancer. 2014 Feb 4; 110(3): 788–791. Available from: https://www.nature.com/articles/bjc2013708

Thompson A, Brennan K, Cox A, Gee J, Harcourt D, Harris A et al. Evaluation of the current knowledge limitations in breast cancer research: a gap analysis. Breast Cancer Res. 2008; 10(2): R26. Available from: https://breast-cancer-research.biomedcentral.com/articles/10.1186/bcr1983

Lund, E. (Ed.) (2020). *Advancing Systems Epidemiology in Cancer. Exploring Trajectories of Gene Expression.*
Oslo: Scandinavian University Press.
DOI: https://doi.org/10.18261/9788215041193-2020-05

5. Woes of The Practicing Omics Researcher

Einar Holsbø and Kajsa Møllersen

Abstract Omics researchers routinely use hypothesis tests. These tests can lead to highly inefficient use of omics data. Through a familiar example, we show the need for exploratory approaches and show how common statistical tools such as p-values and confidence intervals can be used for exploratory omics research. We discuss the often-misunderstood hypothesis test and emphasize its lesser known flexibility. This work is an effort to improve the use of statistical tools in omics by non-statisticians.

Keywords applied statistics | null-hypothesis significance testing | exploratory analyses | omics

TO BE TRAPPED BETWEEN TOOLS AND MATERIALS

Red herrings and missed opportunities

In the world of omics, where so little is known about the biological mechanisms, and the potential clinical impact of new discoveries is beyond what we can imagine, there are pitfalls everywhere. The researcher can easily get stuck in the middle between the dataset and the statistical tools for data analysis. The researcher uses statistical tools to aid them in describing and understanding the data, and ultimately the underlying biology that led to the observations in the data at hand.

The statistical toolbox for omics research is enormous and ever-expanding. New tools are invented, and old ones are refined, according to how data generation changes with technological development and how new computational approaches arise.

Still, the omics researcher reaches for the good old familiar toolset, some out of sheer habit, others for the very good reason that results are better when using a simple but familiar tool than some really great tool that you don't know how to operate.

The human brain is wired to find patterns. It is, on the other hand, highly unreliable in the face of randomness, demonstrated famously in e.g. Tversky and Kahneman 1971. We can easily find patterns where there are none. But as researchers we hope to avoid falling into the ditch of red herrings, and we hope that trusting what others publish does not lead us into this ditch. Our statistical tools can help us weigh patterns against randomness.

But neither do we wish to end up in the other ditch, that of missed opportunity, where important biological findings are not communicated because of the uncertainty that surrounds them. Statistical tools cannot be used to answer biological questions, but are excellent at quantifying the uncertainty in the connections between data, assumptions, and underlying biological mechanisms.

Common tools: hypothesis tests, p-values, and confidence intervals

The hypothesis test is arguably the statistical workhorse of our time. This tool is designed to give a yes/no answer to a certain claim and is used in studies with a simple and focused research question. The most common null hypothesis is that the mean level of the observations is 0, the difference between the mean levels of two groups is 0, or the mean difference between paired observations in two groups is 0. Researchers too seldom devote their time to stating a different null hypothesis, e.g. the mean level of the observations is smaller than X, where X then represents some threshold for an interesting finding, or biological or clinical significance. Even more rarely, the researcher states a hypothesis about something else than the mean, e.g., the variance or the effect size. In the following, we use the most common null hypotheses as examples not to lose our readers in unfamiliar mathematical details. However, we emphasize that hypothesis testing as a statistical tool is not about the mean and 0; the choice of test statistic and null hypothesis is a choice to be made by the researcher.

The hypothesis test requires the calculation of a p-value to be compared against a threshold for "statistical significance," α. This α is usually governed by conventions particular to a certain research area.

The α has a direct impact on the scientific literature. Figure 5.1 shows the distribution of p-values below .1 published in a selection of top medical journals. Note the spikes at round numbers. There is reason to suspect that many have ended up in the ditch of missed opportunity due to p-values that were just above the α and hence never communicated their results at all.

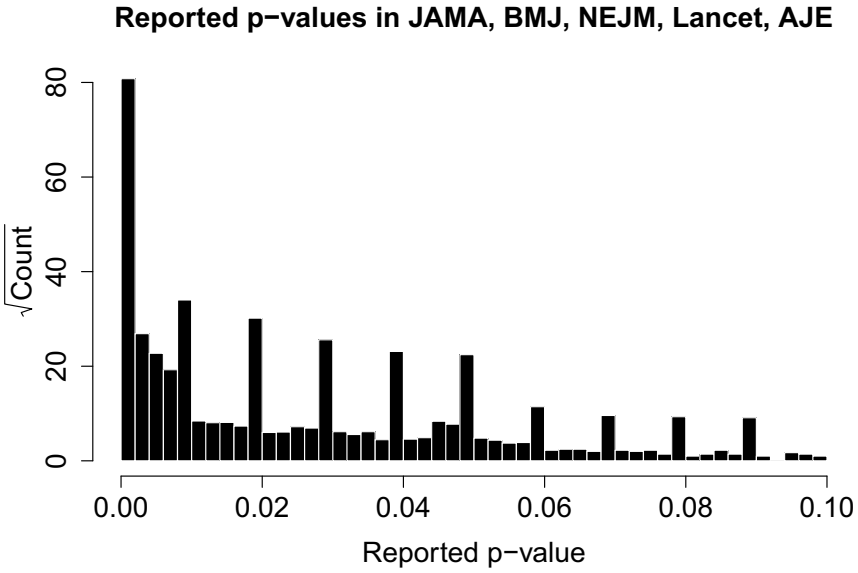


Figure 5.1. Distribution of p-values below .1 reported in various top medical journals. Figure generated from data in Jager and Leek 2014.

For those findings that are published, the p-value is presented to the reader, who can make up their own opinion about whether the uncertainty is acceptable for whatever further action they might want to take. The confidence interval and the p-value are closely related: If the 95% confidence interval excludes the specific amount stated for the hypothesis test (the 0 or the X), the null hypothesis will be rejected at a 5% level.

For some reason, the confidence interval is often not reported, despite being available—literally—at the push of a button. Whereas the p-value says something about the uncertainty, the confidence interval complements that information perfectly by providing the likely range of the mean (or variance or effect size) supported by the data.

SCIENTIFIC RAW MATERIALS: DATASET SIZES

Figure 5.2 shows typical sample sizes in transcriptomic studies based on humans over the years since 2001. The middle 90% of this distribution today lies between 3 and 84 observations. These quantiles have barely moved since the beginning. In other words, the majority of transcriptomic datasets has been of modest size for

the past 15 years, and this does not seem to be changing soon. The larger data sets are likely to be from consortia that pool many smaller data sets; the largest one comprises over 13 000 observations.

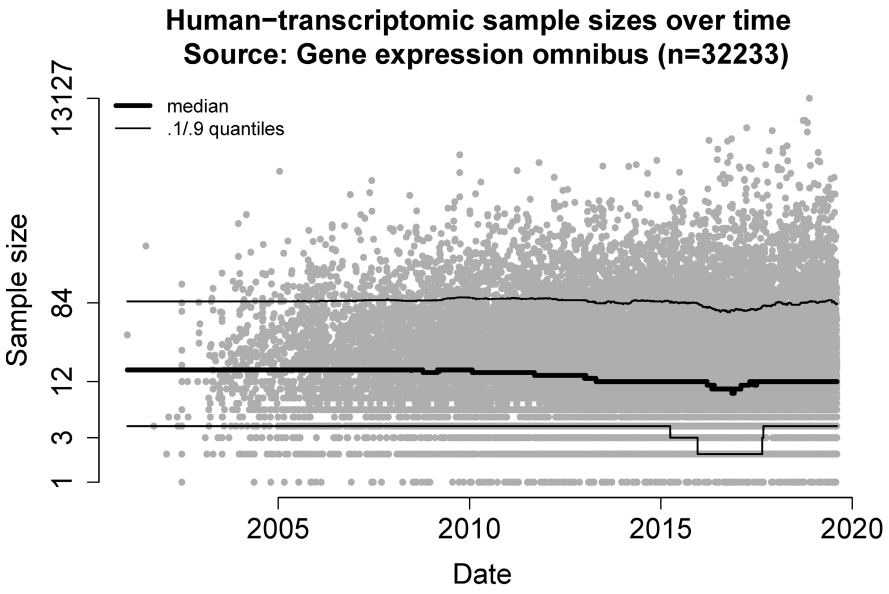


Figure 5.2. Sizes of human-derived transcriptomics data published in the Gene Expression Omnibus between January 2001 and August 2018. The plotted sizes are logarithmic, but the axis annotation is in the real sizes.

THE LIMITS OF OUR TOOLS

Since most people seem to be working with fewer than 100 observations, it is a valuable exercise to explore what is feasible to achieve with a standard type of analysis. As an example, consider data that comprise a single gene expression level on 100 subjects; assuming convenient, nearly normal data, conventional power calculations tell us that the smallest gene expression level we can reliably detect with 95% confidence ("detection" here interpreted as a confidence interval excluding zero) is, roughly, $\frac{4\sigma}{\sqrt{100}} = .4\sigma$. Here, σ is the standard deviation with which the data came to be (for technical details, see the appendix).

Usually we are interested in comparing groups, such as smokers and non-smokers, or cancer cases and controls without a cancer diagnosis. If we have our 100

subjects partitioned into two groups of 50, assuming equal variances, the smallest difference in gene expression we can detect increases to .8σ.

Often, we are also interested in some stratification of the target population. Perhaps we know that the cancer cases naturally split into two subgroups that we expect to have different gene expression levels—say, metastatic and non-metastatic cancers. We split the cancer cases into two subgroups of 25. We estimate the gene expression level in each subgroup, and can expect to detect a gene expression level no smaller than 1.1σ. If we wish to compare the two subgroups against one another, **their** true difference can be no smaller than 1.6σ. To clarify: if the gene expression level in one subgroup is 1.6σ, it has to be nothing at all or at least 3.2σ in the other.

REALITY CHECK

Observations from the real world

A particular microarray dataset from the Norwegian Women and Cancer (NOWAC) study (Lund et al. 2008) contains gene expression measurements from blood samples. Upon providing a blood sample, the women who participate in this study fill out a questionnaire. One of the questions on this form is whether the participant has recently been smoking. It is plausible that this affects blood gene expression in the short term, see Huan et al. 2016.

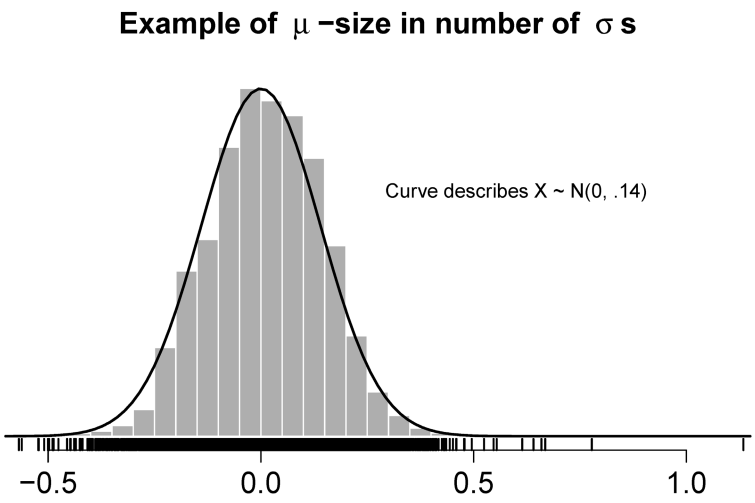


Figure 5.3. Mean observed gene expression level as number of estimated standard deviations in a Norwegian Women and Cancer dataset. The ticks along the x axis show all genewise means estimated in 88 case-control pairs.

If we compare $\log_2$ expression levels of current smokers and nonsmokers in 6664 genes, we get Figure 5.3, which shows the distribution of differences in number of σs for all genes in these data. A difference of $.8\sigma$ is rare. Only 1 in 200 differences are greater than $.4\sigma$. However, keep in mind that the blood transcriptome is notoriously variable, and it is natural to expect no difference in most genes.

Testimation bias

With 100 observations and sample mean on the scale of Figure 5.3, there is clearly some friction between what we would like our tools to be able to do and what we can realistically hope for. Our tools are underpowered to deal with such small differences in means: the ditch of missed opportunity opens up before us.

Missing an opportunity, the making of a "Type 2 error" in the technical language, is one kind of error we might make with an underpowered test. Another interesting and worrisome type of error we might make is one of magnitude.

The magnitude error or "Type M error" (Gelman and Tuerlinckx 2000) is the expected absolute ratio of the estimated mean (the sample mean) to the true mean (the population mean). This is a measure of exaggeration. A Type M error of 2 implies that the estimated means are on average two times the size of the true means. This is half-jokingly called "testimation bias" as it results from looking at the estimate only after it passes a statistical test.

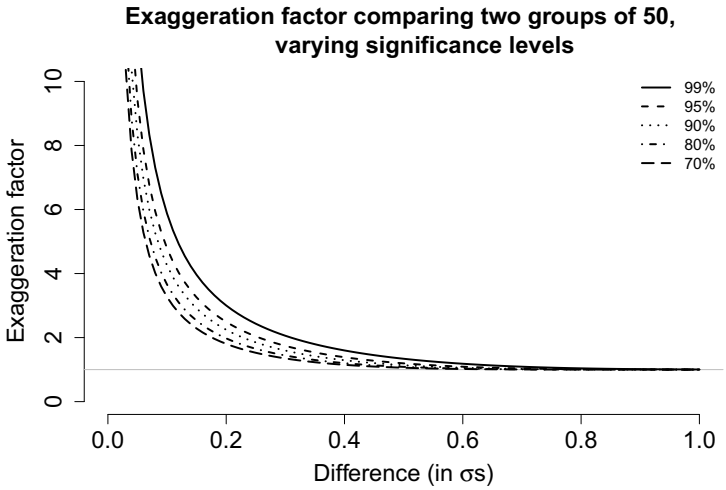


Figure 5.4. Magnitude (or exaggeration ratio) errors comparing two groups of 50 for a range of true differences and a selection of confidence levels. The horizontal grey line denotes no exaggeration.

Figure 5.4 shows the M error for sample means between 0 and 1σ. Once we start losing statistical power (recall the smallest-detectable mean of .8σ) the M error shoots up quite quickly, and overestimating the mean gene expression level by a factor of 2 is quite realistic.

What can we expect?

Assume that Figure 5.3 is representative of gene expression in the whole-blood transcriptome (it need not be, but they are real, plausible data): 90% of the sample means are smaller than .2σ. But we do not expect 90% of all genes to be active in any given process, far from it. For now, assume we are looking for the top 10%, those sample means larger than .2σ, or about 600 genes in these particular data. If we assume $\sigma = 1$, this .2σ corresponds to a fold change of about 1.15.

Table 5.1. Overestimation and power for the largest 10% of mean gene expressions in Figure 5.3

Confidence	99%	95%	90%	80%	70%
Magnitude error	2.3	2.0	1.8	1.6	1.5
Power	12%	28%	39%	53%	63%

Under the assumptions above, Table 5.1 shows the expected power and M error for different confidence levels. We can expect population means to be overestimated by a factor of 1.5–2. Depending on what types of error we want to make (which ditch we prefer to fall into), it might make sense to work at unconventionally low confidence levels: at 70% confidence we're starting to see tolerable power.

Familiar tools—new applications

Omics studies containing several thousand features (as the 6664 genes in the above example) are not simple and focused studies that are cut out for hypothesis testing. Even if they were simple and focused, the data we generate comprise too much noise and too few observations for hypothesis testing to be useful. Whereas it is certainly possible to draw statistically valid conclusions by applying adjustment for multiple testing (e.g., Bonferroni or FDR), the minimum detectable difference increases, and hence the power to detect small differences drops; see Figure 5.5 and Figure 5.6.

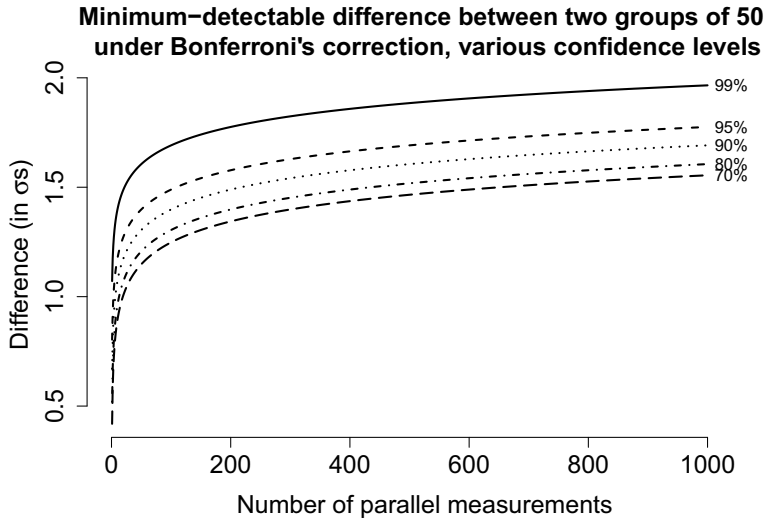


Figure 5.5. Smallest detectable difference between two groups of 50 as a function of the number of hypotheses.

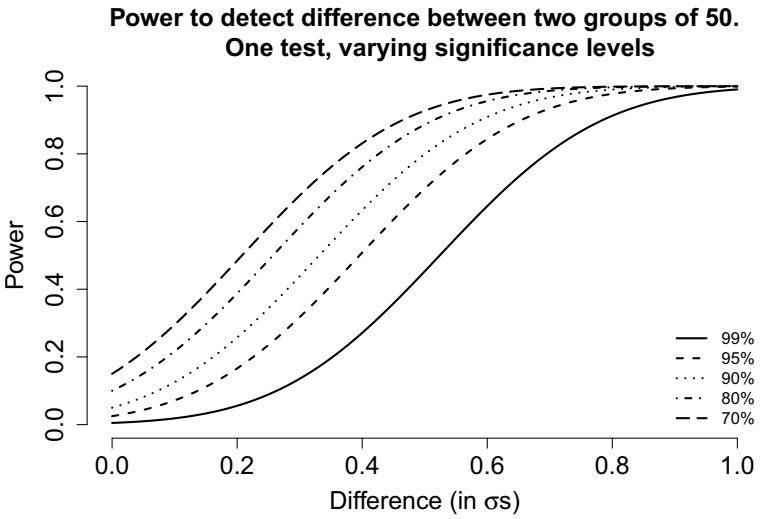


Figure 5.6. The power to detect a given difference between two groups of 50 with a single test at various confidence levels.

Figure 5.7 show the statistical power to detect differences when doing various numbers of parallel tests, corrected by a Bonferroni adjustment. With 1000 tests we struggle to detect differences smaller than 1σ at the 95% level.

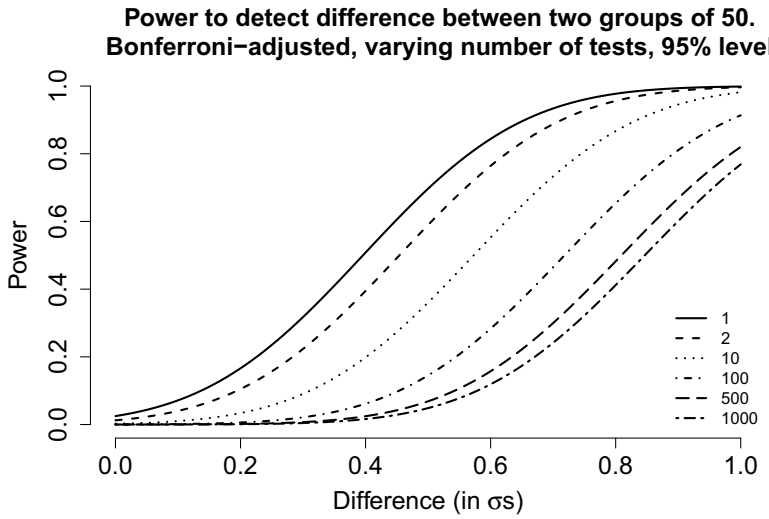


Figure 5.7. The power to detect a given difference between two groups of 50 at 95% confidence level, when we perform a certain number of parallel tests and adjust by the Bonferroni method.

Studying thousands of features is rarely about answering yes/no questions, but rather about seeing how the land lies so that more focused studies can be directed towards interesting questions. The design is clearly exploratory. The main question for the exploratory omics researcher is not "Are my findings statistically significant?" but "Which findings are biologically interesting?".

Exploratory research is not an easy way out of weak statistical findings. It requires the same rigor as anything that labels itself as scientific. "Which findings are biologically interesting?" is a question that needs an answer before the data are analyzed, in the same way that a hypothesis needs to be stated before the data are analyzed. The answer requires knowledge about the specific field of study. It is not a statistical matter; it is biological, medical, or otherwise. The answer needs to be as clear as any hypothesis, e.g. "effect size larger than X" or "difference larger than Y."

The exploratory analysis does not stop with identifying interesting findings. There is still a need to quantify the uncertainty of the relation between the findings in the sample and the underlying population from which the sample is drawn. Fortunately, we already possess the tools to do so: confidence intervals and p-values!

EXPLORATORY OMICS ANALYSES

Exploratory p-values

Results from scientific studies are never reported as a yes/no answer to a hypothesis test: they are always accompanied by p-values and discussion on the possible impact of the findings. One such impact is change in medical procedures, an area where omics research has made contributions (Vieira and Schmitt 2018). Erroneously changing a cancer treatment procedure is much more costly than choosing to carry on as usual when, in fact, a change would have been better. This is also true in many other aspects of real life and science. A key aspect in hypothesis testing is an agreement on how strong the statistical evidence must be to conclude that the results from the study support e.g. a change in cancer treatment procedure. The statement of the null and the alternative hypotheses must be tailored to answer a specific question of clinical, biological, societal or scientific interest.

Most omics studies are not designed to support a decision regarding medical procedures—they are conducted with the sole aim of gaining more knowledge, and their results are published to communicate that knowledge.

Fisher 1955 describes the alternative to the yes/no hypothesis testing framework as follows:

> The worker's real attitude in such a case [i.e., where p-values are large] might be, according to the circumstances:
>
> The possible deviation from truth of my working hypothesis, to examine which the test is appropriate, seems not to be of sufficient magnitude to warrant any immediate modification.
>
> Or it might be:
>
> The deviation is in the direction expected for certain influences which seemed to me not improbable, and to this extent my suspicion has been confirmed; but the body of data available so far is not **by itself** [emphasis added] sufficient to demonstrate their reality.

He goes on:

> What we look forward to in science is further data, probably of a somewhat different kind, which may confirm or elaborate the conclusions we have drawn; but perhaps of the same kind, which may then be added to what we have already, to form an enlarged basis for induction.

As an example of exploratory analysis, consider again the 6664 genes:

- Decide what is an interesting finding, e.g. mean gene expression above *X*, from the biological perspective.
- Identify the subgroup among your 6664 genes that have a mean gene expression larger than *X*.
- Calculate the confidence interval of the mean gene expression for each of the genes in that group.
- Calculate the p-value (under the assumption that the mean value is smaller than *X*) for each gene.
- Present mean value, confidence interval and p-value to the reader.

Ultimately, the results will form the basis for a decision, e.g. to conduct a new study with 50 genes on a similar population. At this point it might make sense to start thinking of sharp hypothesis tests.

A bulwark against overexcitement

The overestimation of sample means mentioned earlier is an unfortunate reality for the omics researcher no matter whether they choose interesting genes based on a p-value, or a threshold for the sample mean or sample effect size. Clearly, we cannot report on the biological implications for all genes: some selection has to happen. Once you select some candidates *because* they stand out against the rest, you can expect "regression to the mean" to apply (Senn 2011). That is, it is very likely that the population mean of the top candidates on average are smaller than they appear in the data.

In fact, even if the population mean is 0, large sample means are likely if variation is large, which it will be if the sample size is small, if there are large natural variations in the data, or if the data is noisy. This is illustrated in Figure 5.8. Under the narrow distribution an observed mean larger than $\bar{x} = 3$ is unlikely, but under the broad distribution it should not provoke too much surprise. Both are, however, centered on 0: no effect.

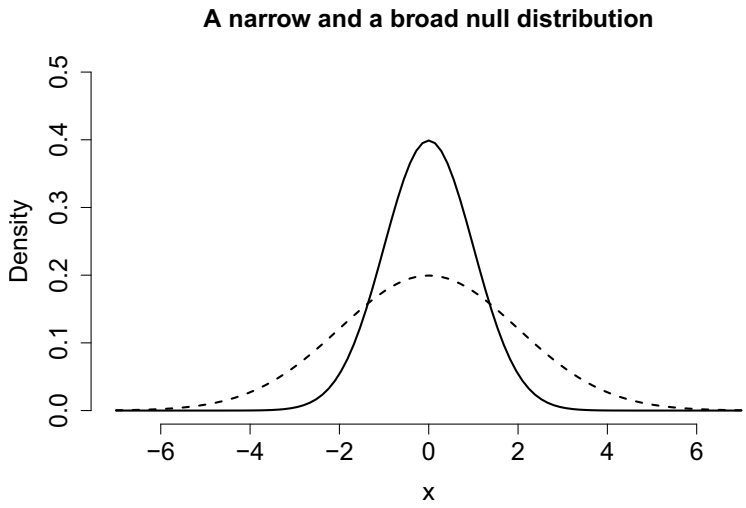


Figure 5.8. Two distributions centered at 0: one broad, one narrow.

Given Figure 5.3, should we believe that the sample mean of size 2σ reflects a population mean of the same size? Probably not. There are ways to guard against massive overestimates due to noise. Those who prefer to make Bayesian arguments ameliorate the problem by assigning less credence to extreme results, essentially requiring stronger evidence for large effects than for small effects. To do this we summarize our assumptions about realistic effects and their uncertainty with a probability distribution, the *prior distribution*, and augment this with a model for the data through Bayes' rule, a basic result from probability theory.

What results is the *posterior distribution*: a probability distribution for a variable given prior assumptions and data. This distribution lies somewhere between the prior assumptions and the data likelihood, and yields both our estimate and its uncertainty, e.g. by its mean and the middle 95% of the distribution (called a credible interval to mark the philosophical distinction from a confidence interval).

The estimate of the mean likewise lies somewhere between the observed effect and the prior assumptions, the exact location depending on the strength of the data.

Such an analysis is not entirely different from the usual mean estimate and confidence interval approach, in fact i) the classical confidence interval and mean estimate result from assigning equal prior probability to all observations in a Bayesian argument; and ii) the two approaches are identical when there is ample data.

But we do not have ample data, so the prior assumptions matter. For our purpose of comparing groups, the value in making a stronger prior assumption lies in

the fact that as long as we place most of our credibility around 0, the Bayesian approach will be more conservative than the frequency approach. It both pushes estimates closer to 0 (known as "regularization" or "shrinkage") and yields broader uncertainty intervals (Gelman and Tuerlinckx 2000).

An explorative Bayesian analysis might proceed as follows:

- Decide what is an interesting finding, e.g., in terms of mean value, etc., from the biological perspective.
- Decide what is a reasonable assumption about the variable a priori. Using Figure 3.6 we might decide that a normal centered on 0 with a standard deviation of perhaps .15 is reasonable (for this example found eyeballing the fit). The normal is also a reasonable data model since we are comparing means.
- Calculate the posterior distribution of the mean given data.
- Use the posterior mean as population mean estimate, use it to identify interesting genes.
- Use the middle (e.g.) 95% of the posterior distribution as credible interval.
- Report prior assumptions, data likelihood, and posterior summaries.

For technical reasons, the Bayesian argument usually requires more compute-intensive procedures. This probably deters more people than it should. For common analyses, such as the comparison of two groups, or basic regression models, most modern statistical software provides comparatively simple interfaces, such as the brms package for R. It is a larger challenge that the approach requires different thinking from what is usually taught, so a researcher must find the use of regularizing priors valuable enough that they are willing to invest time in understanding the framework.

DISCUSSION

The critique against hypothesis testing (or more specifically the null hypothesis significance testing, NHST) has been going on for decades. Besides the early correspondence-like articles of Fisher and Neyman and Pearson, it is rarely clear *who* the critique is aimed at. Surely, it makes no sense to criticize the hypothesis test for merely existing, as in the much-cited and much-discussed paper of Cohen 1994:

> What's wrong with NHST? Well, among many other things, it does not tell us what we want to know.

This is like criticizing a saw for not being a hammer. To be fair, Cohen finishes the same sentence with

> …and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!

It is not clear who "we" are in this scenario. If you are truly after the effect size, or predicting the outcome of individuals, or estimates of posterior probabilities, it is a waste of time to reach for the NHST tool, and it is not clear who (if anyone) is forcing "us" to do so.

Trafimow and Rice 2009 also criticize the hypothesis test for not answering the question to which they want an answer: the probability of the null hypothesis being true. Interestingly, none of the alternatives to NHST that they list, including their own, gives an answer. This is not strange. The Truth of the null lies outside of statistics: we represent the null by an idealized mathematical construct that can never be strictly speaking True. It has never been more or less than a model. But the model can be (and is) very useful for measuring evidence and uncertainty. Trafimow's very important contribution to the NHST discussion is his statement as an editor (Trafimow 2014), where he welcomes alternatives to NHST, and so-called null findings.

Most omics studies are interested in the characteristics of a population. Findings regarding 100 specific breast cancer patients are interesting (to others than the patients themselves) only if they also say something about the population of breast cancer patients. This requires quantification of the findings' uncertainty. There are many ways to quantify this uncertainty—many tools in the statistical toolbox to choose from. Many tools require certain training in statistics and programming. All tools require correct use in order to avoid making invalid conclusions. In addition, for publishing purposes, the editors, reviewers and readers must be given the opportunity to understand the tools that were used.

The data we generate in omics do not for now support a confirmatory testing procedure. This is OK. Exploratory research has value in itself; we cannot know where to go in this young field unless we open our eyes and look at some data. The use of statistical tools is an important part of both exploratory and confirmatory analyses, but the biologist's expertise is just as important. There is no way for a statistician without biological training to say what an interesting effect is, just as there is no way for a biologist without statistical training to say what a valid data analysis is.

However, not all biologists have access to a statistician, and not all statisticians have access to a biologist. We have made some brief suggestions for approaches to keep in the toolbox for those who feel caught in the conflict between tools and raw

materials. We have done so because we suspect that many who work with omics data do not have a clear idea of what their raw materials can support, and that many feel as though the NHST tool is the only way to lend legitimacy to a scientific argument.

P-values and confidence intervals are tools that the omics researcher already has in their toolbox. They can legitimately and fruitfully be used in exploratory analysis. Bayesian tools are not necessarily better, worse, or more complicated, but they are different and require a different mindset. They might therefore be more difficult to communicate to reviewers and readers in the omics world.

TECHNICAL DETAILS

Below we provide some technical details. Everywhere we have assumed that we know σ, which of course we never do; adjusting for this will widen confidence intervals and make smaller effects harder to discern from noise. We have also mostly assumed a normal model, which is probably too light in the tails compared with real-world scenarios. Adjusting for this will widen confidence intervals, etc.

The calculation of smallest detectable means

We base the calculations of means and differences of means on the rudimentary facts of the normal distribution. A rule-of-thumb says that 95% of the mass of a normal distribution lies within two standard deviations of the mean (the truth is closer to 1.96). Since any particular realization from a normal distribution can easily fall (i.e. about 95% of the time) within two standard deviations of the mean μ, and since we put a confidence interval $\pm 2\sigma$ around this realization, it follows that μ must lie four standard deviations away from 0 for the interval to reliably exclude 0.

The rest follows from the standard error of the mean: $\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$. Assuming two independent groups of 50 with equal variances, the difference in their means has a standard error of $\sqrt{2\sigma_{\hat{\mu}}^2} = \sqrt{\frac{2\sigma^2}{50}} = \frac{\sigma}{\sqrt{25}}$. So finally, we require that $\mu > \frac{4}{\sqrt{25}}\sigma = .8\sigma$.

In the Bonferroni multiple testing scenario, we have simply extended this calculation using the implied number of standard errors for different confidence levels from the appropriate quantiles of a t_{50} distribution ($\frac{\alpha/2}{M}$).

Power calculations

The *power* is the probability of detecting what we set out to detect, e.g. mean value larger than X, if the true mean value is in fact larger than X. The power depends on the size of the dataset, the variance, the true value of the mean, and assumptions regarding the hypothesis statements, e.g. a Gaussian or a Student's t distribution. In the examples of this manuscript, the important take-home message is the approximate magnitude of the power or the mean value.

The power calculations have been conducted using the R function power.t.test.

MAGNITUDE ERRORS

These errors are straightforward derivations assuming a hierarchical model where $\mu \sim N(0, \tau)$ and $y \sim N(\mu, \sigma)$; see Gelman and Tuerlinckx 2000. We use the R package retrodesign to compute them.

Real-world expectations of error

For the real-world expected errors in Table 5.1, we perform a Monte Carlo integration over the empirical distribution of sample means, displayed in Figure 5.3. We isolate the sample means larger than $.2\sigma$ and sample from a kernel density estimate of their distribution to calculate the different quantities on display.

P-values and confidence intervals for variances and effect sizes

When doing hypothesis testing, and calculating p-values and confidence intervals regarding the mean, the calculations are based on the assumption that the sample mean has a specific Gaussian distribution. When doing hypothesis testing regarding variances or effect sizes, we need an assumption regarding the sample variance or effect size. For the sample variance, it can be shown that $\frac{(n-1)s^2}{\sigma^2}$ has a χ^2_{n-1} distribution, for a Gaussian distributed population, and from that p-values and confidence intervals are calculated the same way as for the mean.

The term effect size refers to the mean or difference in means relative to the variance. Its confidence interval is slightly more complicated, but certainly within reach. Kelley 2007 offers an introduction to the confidence intervals and an explanation of the much-used Cohen's d effect size, including its bias, and a method and an R package for the confidence interval based on the non-central t-distribution.

Bayesian data analysis

Gelman et al. 2014 is probably the best-known textbook on Bayesian data analysis. It is quite technical. McElreath 2018 focuses much more on the practicalities of doing such analyses and is quite light on mathematical theory. Spiegelhalter, Abrams and Myles 2004 provide a wealth of healthcare applications.

REFERENCES

Cohen J. The earth is round (p < .05). American Psychologist. 1994; 49(12): 997–1003. Available from: https://doi.org/10.1037/0003-066X.49.12.997

Fisher R. Statistical Methods and Scientific Induction. Journal of the Royal Statistical Society: Series B (Methodological) 1955; 17(1): 69–78. Available from: https://doi.org/10.1111/j.2517-6161.1955.tb00180.x

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. Series: Chapman & Hall/CRC Texts in Statistical Science. 3rd ed. Boca Ration: CRC Press, Taylor & Francis Group; 2013. pp 675.

Gelman A, Tuerlinckx F. Type S Error Rates for Classical and Bayesian Single and Multiple Comparison Procedures. Computational Statistics. 2000; 15 (3): 373–390. Available from: https://link.springer.com/article/10.1007/s001800000040

Huan T, Joehanes R, Schurmann C, Schramm K, Pilling LC, Peters MJ et al. A whole-blood transcriptome meta-analysis identifies gene expression signatures of cigarette smoking. Hum Mol Genet. 2016 Nov 1; 25(21): 4611–4623. Available from: https://doi.org/10.1093/hmg/ddw288.

Jager LR, Leek JT. An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics. 2014 Jan; 15(1): 1–12. Available from: https://doi.org/10.1093/biostatistics/kxt007

Kelley K. Confidence Intervals for Standardized Effect Sizes: Theory, Application, and Implementation. Journal of Statistical Software. 2007 May; 20(8): 1–24. Available from: https://EconPapers.repec.org/RePEc:jss:jstsof:v:020:i08

Lund E, Dumeaux V, Braaten T, Hjartåker A, Engeset D, Skeie G et al. Cohort profile: The Norwegian Women and Cancer Study--NOWAC--Kvinner og kreft. Int J Epidemiol. 2008 Feb; 37(1): 36–41. Available from: https://doi.org/10.1093/ije/dym137

McElreath R. Statistical Rethinking: A Bayesian Course with Examples in R and Stan. Series: Chapman & Hall/CRC Texts in Statistical Science. 2nd ed. Boca Ration: CRC Press, Taylor & Francis Group; 2018. pp 469.

Senn S. Francis Galton and regression to the mean. Significance. 2011 Aug 25; 8(3): 124–126. Available from: https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2011.00509.x

Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Series: Statistics in Practice, Vol 13. Chichester: John Wiley & Sons, Ltd; 2004. pp 408.

Trafimow D. Editorial. Basic Appl Soc Psych. 2014 Feb 10; 36(1): 1–2. Available from: https://doi.org/10.1080/01973533.2014.865505

Trafimow D, Rice S. A Test of the Null Hypothesis Significance Testing Procedure Correlation Argument. J Gen Psychol 2009 Jul; 136(3): 261–270. Available from: https://doi.org/10.3200/GENP.136.3.261-270

Tversky A, Kahneman D. Belief in the Law of Small Numbers. Psychol Bull 1971; 76(2): 105–110. Available from: https://psycnet.apa.org/fulltext/1972-01934-001.pdf

Vieira AF, Schmitt F. An Update on Breast Cancer Multigene Prognostic Tests – Emergent Clinical Biomarkers. Front Med (Lausanne) 2018 Sep 4; 5: 248. Available from: https://doi.org/10.3389/fmed.2018.00248

Lund, E. (Ed.) (2020). *Advancing Systems Epidemiology in Cancer. Exploring Trajectories of Gene Expression.*
Oslo: Scandinavian University Press.
DOI: https://doi.org/10.18261/9788215041193-2020-06

6. Statistics of Sparsely Sampled Curves

Marit Holden and Lars Holden

Abstract We develop new statistical methods for analyzing sparsely sampled curves that vary in time. The typical dataset is differences in log gene expressions from case-control pairs for a large number of genes sampled relative to time of diagnosis. We focus on weak signals in the gene expression in many genes instead of strong signals in a few genes. The methods are based on moving windows in time, hypothesis testing, dimension reductions and randomization of the time to observation.

Keywords Sparsely samples curves | case control differences | time to diagnosis | hypothesis testing | randomization

INTRODUCTION

In this chapter, we describe statistical methods for analyzing the data described in the previous paragraphs. The methods may, however, be applied more generally. We only need observations on irregular points in time from one or several strata from many different response variables, and we test whether the response variables are stationary and whether there are differences between the strata. The approaches are based on using a moving window in time and hypothesis testing.

The lack of repeated samples, in addition to irregularly sampled data and the quite small number of case-control pairs (around 400–500 or less), is a challenge, especially in analyses stratified on important clinical parameters. The aim of some of the analyses is therefore to show that there are changes over time or between strata, but without identifying single differentially expressed genes or gene sets. Normally, when identifying such genes or gene sets, we test multiple hypotheses and consequently need to adjust for this, for example by controlling the false discovery rate (FDR). For some of our datasets, we would then find no or very few genes. We therefore need approaches with fewer tests that include many genes in

order to avoid the problem of low power, noisy data, and multiple testing. Such approaches can be viewed as an effective method for dimension reduction in studies of functional genomics.

In such approaches, we need to decide how many genes to include when testing different kinds of hypotheses. Assume the genes have been ranked so that those that are varying most significantly have the highest rank. In (Holden 2015a, Holden 2015b), we showed that if there is a difference in average value of $X_{g,c}$ between the strata for some of the genes, but we do not know which genes, and the difference is normally distributed, then the statistical tests are strongest for a small rank. We concluded that if the distribution has heavier tails than the normal distribution, we should focus on the few genes with the strongest signal. On the other hand, if the distribution has a less heavy tail, for example a constant difference in the average value, then the statistical test is strongest for a larger rank, often larger than (closer to) the number of genes with a difference in average value between the strata.

This chapter discusses the analysis of functions $f_{a,g}(t)$ for different strata a,b, ... and many genes g where the function varies with time t. These functions may be denoted as trajectories. The challenge in our case is that the functions are sparsely sampled, the different strata are sampled at different points in time, and there is considerable noise in the sampling. We wish to identify the time dependency in the data and the differences between the strata. We expect to find small differences in many genes instead of larger differences in one or a few genes. This chapter summarizes some of the applied work.

In order to study how gene expression profiles vary with time in the years before or after a cancer diagnosis, several datasets showing gene expression in blood have been collected. Each such dataset consists of cases diagnosed with cancer and healthy controls. Each case and control belong to one case-control pair. The case and control of a case-control pair each gave one blood sample, and they are matched by time of blood sampling and year of birth. In the statistical analyses, we use the differences of the log2 gene expression levels, $X_{g,c}$, for each case-control pair c and gene g. Here, $c = 1, \ldots, M$, where M is the number of case-control pairs, and $g = 1, \ldots N_g$, where N_g is the number of genes. As the time intervals between blood sampling and cancer diagnosis vary from case to case, the case-control pairs provide information on the sparsely sampled curves describing gene expression over time some years before or after diagnosis.

Let t_c be the time interval between blood sampling and cancer diagnosis for case-control pair c, where $t_1 \le t_2 \le \ldots \le t_M$. We assume the log2 gene expressions $X_{g,c}$ follow a smooth function in time $f_{s(c),g}(t_c) = E\{X_{g,c}\}$, where $s(c)$ is the stratum

of case c. We estimate the function $f_{s(c),g}(t)$ by taking an average of the observations $X_{g,c}$ from stratum $s(c)$ in an interval in time. The variance of $X_{g,c}$ is estimated from the variance of the observations in the same interval.

METHODS FOR ANALYZING COMPLEX CURVES

Moving window in time

A central part of the statistical methodology is to examine how gene expression varies with time. By dividing the entire time period into shorter time periods and computing different kind of statistics for each such time period, we simplify the problem to examining how these statistics vary with time. We use overlapping time periods by using a moving window in time. The statistics computed for a time period are independent of time. As the distribution of the gene expressions may vary with time, the lengths of the time periods should be chosen such that we obtain as short time periods as possible. However, to obtain as good estimates as possible for each time period, there should be as many case-control pairs as possible within each time period. There is a trade-off between these two wishes. To conclude, we define $M - L + 1$ time periods $[t_1, t_L]$, $[t_2, t_{L+1}]$, ... , $[t_{M-L+1}, t_M]$ where L is chosen such that we obtain short time periods with many case-control pairs. Typically, we let $L \approx M/4$.

Compared to an approach where the time periods are not overlapping, an advantage with the moving window approach is that we are better able to identify the points in time relative to diagnosis where changes in gene expression occur.

Randomization for estimating null distributions and p-values

In most hypothesis tests, we compute p-values by estimating the null distribution for the statistic of the hypothesis test by randomizing the data, i.e. interchanging covariates (time to diagnosis, case/control, etc.) between the patients. In the randomization we preserve critical properties of the genes (level of expression, complex correlation between genes, etc.) and randomize only what is connected to the changes in time, stratum or case/control status. This randomization defines the null distribution for the test statistic that is used when finding the p-value.

We can randomize the data either by randomizing the case and control in each case-control pair, by randomizing the case-control pairs between the periods, or by randomizing between the two strata within the time period. Note that all three randomization strategies maintain the correlation structure between the genes for

each case-control pair. Also note that each randomization of the data leads to a different ordering of the genes if the genes are ordered according to the statistic of the hypothesis test.

The p-value of the test is set to $\frac{K+1}{N_S+1}$, where N_S is the total number of randomizations and K is the number of randomizations out of N_S with a more extreme statistic than the statistic for the real data. When we test one hypothesis for each gene, we take multiple testing into account by using the Benjamin-Hochberg procedure for controlling the false discovery rate, FDR (Reiner et al. 2003).

Extracting information from a time period

When extracting information for a time period, we use the gene expression data in that time period and ignore information about time when we compute different kinds of statistics, identify differentially expressed genes etc. We have used the following statistics for a time period:

Ordered standard deviation, mean and weight for a time period

Let $m_{p,g}$ be the sample mean and $s_{p,g}$ be the sample standard deviations for the differences in $\log_2$ gene expressions for gene g in time period p. The variable $m_{p,g}$ is an approximation to $f_{s(c),g}(t_c) = E\{X_{g,c}\}$ and $s_{p,g}$ an approximation to the standard deviation of X_{gc} in the interval. Let $m_{p,g,a}(m_{p,g,b})$ be the sample mean and $s_{p,g,a}$ ($s_{p,g,b}$) be the sample standard deviations for the differences in $\log_2$ gene expression for gene g in time period p for stratum a (b). We define the statistics $s_{p,(g)}$, $m_{p,(g)}$, and $w_{p,(g)}$ from these sample means and standard deviations as follows:

- $s_{p,(g)} = s_{p,g'}$, where $s_{p,g'}$ has rank g when the $s_{p,g}$'s for period p are sorted in increasing order. Rank 1 corresponds to the smallest of the $s_{p,g}$'s for period p.
- $m_{p,(g)} = |m_{p,g'}|$, where $|m_{p,g'}|$ has rank g when the $|m_{p,g}|$'s for period p are sorted in decreasing order. Rank 1 corresponds to the largest of the $|m_{p,g}|$'s for period p.
- Let $w_{p,g} = \frac{m_{p,g,a} - m_{p,g,b}}{\sqrt{s^2_{p,g,1} + s^2_{p,g,0}}}$ be the weight for gene g in time period p, i.e. a measure of the difference between the two strata. $w_{p,(g)} = |w_{p,g}|$ where $|w_{p,g'}|$ has rank g when the $|w_{p,g}|$'s for period p are sorted in decreasing order. Rank 1 corresponds to the largest of the $|w_{p,g}|$'s for period p.

In some approaches, we use the sample mean $m_{p,g}$ directly, without ranking.

The number of differentially expressed genes for a time period

We identify the significantly differentially expressed genes in the time period using the Bioconductor R-package limma, linear models for microarrays (Ritchie et al. 2015), where the response is $X_{g,c}$, i.e. the difference in log2 gene expression between the case and the control of a case-control pair.

In the next section, we describe how these statistics can be used for examining how gene expression varies over time, between strata, and between cases and controls

Using information extracted for time periods

Finding signal in the data

The objective is to be able to identify small changes that vary slowly in time and/or between strata, by using a large number of genes in each hypothesis test and predictor.

For examining whether there are differences between cases and controls, between strata or in time, we test following hypotheses using the three statistics' standard deviation, mean and weight defined in the previous section:

H0-case-ctrl: The expectation of $X_{g,c}$ is zero. This means that there is no difference between the expectations of the $\log_2$ gene expression values for the cases and controls. It implies that $f_{s(c),g}(t_c) = 0$. If the null hypothesis is false, the expectation will be different from zero for some periods and genes. We test the hypothesis by using the statistic $m_{p,(g)}$.

H0-time: The distribution of $X_{g,c}$ is not associated with the time to diagnosis. This means that the expectation and standard deviation of $X_{g,c}$ are the same in all time periods. It implies that $f_{s(c),g}(t_c)$ is constant in time. 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 $m_{p,(g)}$.

H0-node: The expectation of $X_{g,c}$ is not associated with stratum. This means that the expectations for the two strata are equal for all genes g and time to diagnosis t. It implies that $f_{s(c),g}(t_c)$ is the same for all strata. 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 $w_{p,(g)}$.

The null distribution of each 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.

We will reject the H0-case-ctrl hypothesis if the hypothesis $m_{p,(g)}>0$ is rejected for at least one time period p and rank g, where g belongs to a subset of the N_g ranks. In practice, we have chosen to let the subset of ranks consist of ranks between approximately 1% and 25% of the number of genes, so that the subset contains both relatively low and high ranks. This means that H0-case-ctrl is rejected based on a very large number of hypotheses, that are also highly positively correlated, and we therefore needed to adjust for multiple testing. The approaches for rejecting the H0-time and H0-node hypotheses are similar. Besides rejecting the three null hypotheses, the hypothesis tests for the statistics for each time period and rank can be used for illustrating how the p-values are associated with the time to diagnosis.

Changes in the number of differentially expressed genes over time

In this approach, we examine changes in gene expression over time by examining how the number of differentially expressed genes between cases and controls varies with time. The time curve in this case consists of the number of differentially expressed genes in each time period. Such time curves give an indication of when there is a large difference between cases and controls before or after diagnosis. For comparing different strata, we can compare the time curves of the strata.

When testing whether the number of differentially expressed genes are different for two strata, we use n_s, the number of genes that are differentially expressed between cases and controls in at least one time period, as test statistic for stratum s.

When comparing stratum a and stratum b, we cannot directly compare n_a and n_b if M_a, the number of case-control pairs in stratum a, is much larger than M_b, the number of case-control pairs in stratum b. To use the same number of case-controls pairs when comparing the strata, we test the following hypothesis.

H0-strata1: The number of differentially expressed genes between cases and controls is different for stratum a and b. Assume $M_b < M_a$. We want to estimate the null distribution for n_a when the sample size of stratum a is M_b, and then compare this distribution to n_b. The null distribution is found using simulation by repeatedly sample M_b case-control pairs from stratum a, and then compute the number of differentially expressed genes for each sampled dataset. The p-value of the hypothesis test is computed by comparing the samples of the null distribution to n_b.

Identifying significant genes based on area between curves

In this section we also estimate the function $f_{s(c),g}(t_c) = E\{X_{g,c}\}$ by taking the mean of observations of $X_{g,c}$ in an interval of time. However, here we analyze the properties of the function in the entire time period at the same time instead of for each interval in time separately. Then we are able to analyze all the data from a gene instead of focusing on properties in an interval. This may be used to identify genes that are significantly different between strata and predict strata from the gene observations.

We estimate the area between two curves $f_{a,g}(t)$ and $f_{b,g}(t)$ by the test statistics $V_g = |f_{a,g} - f_{b,g}| = \int |f_{a,g}(t) - f_{b,g}(t)| \mathrm{dt}$ where the two strata are denoted a and b, respectively. This is equal to the weighted sum of the absolute value of the differences in average gene expression between the two strata in each time interval, where the weight depends on the length of the time interval. The area is large if there is a large difference between the curves and we neglect whether this is due to different average value or one is increasing and the other is decreasing in time.

We test the following hypothesis:

H0-strata2: The functions $f_{a,g}(t)$ and $f_{b,g}(t)$ are equal. For each gene g, we compare the observed V_g with the same variable from a simulated distribution where we resample the variables $X_{g,c}$ for all the genes simultaneously by randomizing the stratum $s(c)$ for the case-control pairs. We maintain the observations for each gene and the number of observations from each stratum. We have made N_g simultaneous tests and need to use the methods for adjusting for multiple testing.

Prediction of stratum based on local statistics

The weights $w_{p,(g)}$ can be estimated for each rank g from data in period p for a training dataset. The stratum of the case of a new case-control pair, i.e. a case-control pair that does not belong to the training set, can then be predicted based on the score

$$z = \sum_{g=1}^{n} \delta_{p,(g)} w_{p,(g)} x_{(g)} ,$$

where $x_{(g)}$ is the difference in $\log_2$ gene expressions of the new case-control pair and $\delta_{p,(g)}$ is 1 if the weight $w_{p,g'}$ is positive, and –1 otherwise, where $|w_{p,g'}| = w_{p,(g)}$. The n genes with highest absolute value of the weights are used for computing the score, where n is a number less than or equal to the number of genes, N_g. Large values of z indicate that the new case belongs to stratum a. If $z > c$, for some arbi-

trary threshold c we conclude that the new case belongs to stratum a, otherwise we conclude that the new case belongs to stratum b. We may set $c = 0$ if it is not more important to avoid false classification in one stratum relative to the other and if

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

where $m_{p,(g),a}$ and $m_{p,(g),b}$ are the sample means that are used when computing $w_{p,(g)}$. Increasing (decreasing) c results in fewer false positives (negatives) at the cost of more false negatives (positives).

The available datasets are too small to be divided into a training and validation set. When predicting the stratum of the cases in a dataset, we should therefore use a leave-one-out or k-fold cross validation approach. When using the leave-one-out approach, we predict the stratum of case j using weights $w_{p,(g)}$ that have been estimated using the dataset where case-control pair j has been excluded. The k-fold cross validation approach is similar, except that we divide the dataset into k folds and predict the stratum of the cases in fold f using weights $w_{p,(g)}$ that have been estimated using the dataset where the case-control pairs in fold f have been excluded.

EXAMPLES OF USE OF THE METHODS

In this section, we give examples from papers where the methods described above have been used.

Finding signal in the data

In a previous methodological study, time was categorized in three non-overlapping periods (Lund et al. 2016); see Chapter 8. The aim in that study was to show that there is signal in the data, but without showing where in time the changes in gene expression occurred or which genes were involved. The main idea used in the paper is that genes can be grouped into curve groups, each curve group corresponding to genes with a similar development over time (Figure 6.1). Based on these curve groups, we tested a set of hypotheses that determined whether there is development in gene expression levels over time, and whether this development varies among different strata. For a breast cancer dataset in the Norwegian Women and Cancer (NOWAC) postgenome cohort, the curve group analysis revealed that development of gene expression levels varied in the last years before breast cancer diagnosis, and that this development differed by lymph node status and participa-

tion in the Norwegian Breast Cancer Screening Program. The effect of the participation may be due to different treatment for the participating women representing the majority of the population.

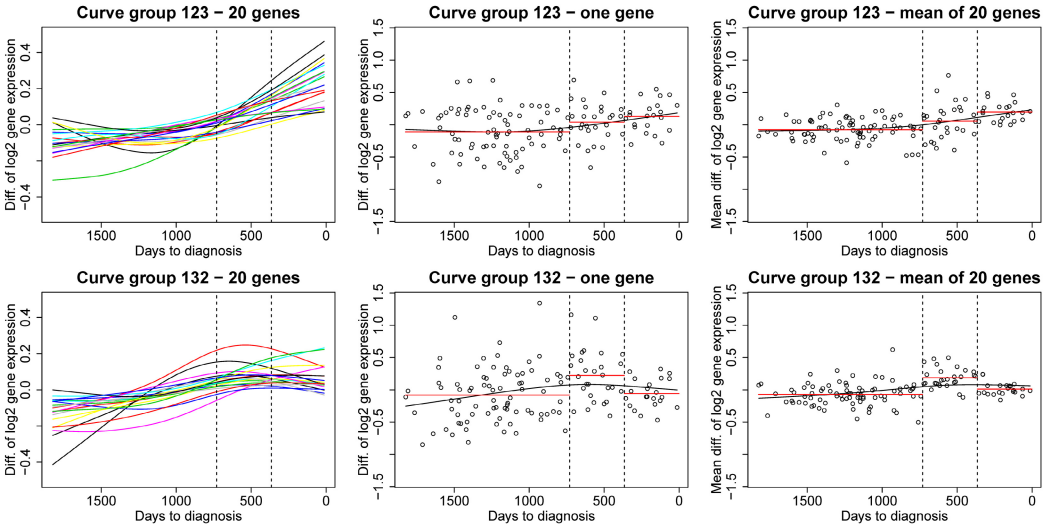


Figure 6.1. Examples of curve groups according to time to diagnosis. Example of two different curve groups: curve group '123' (upper panel, gene expression values increasing with time) and curve group '132' (lower panel, highest gene expression value in the middle time period). In the left panels curves with the gene expression differences for 20 genes from the given curve group are plotted. For illustrational purposes, the curves have been estimated from the data using splines. In the middle panels the data for one of the 20 genes are shown with the corresponding spline-estimated curve. The points represent the differences in gene expression for each case-control pair. The mean value in each or the three time periods is shown in red. The right panels are similar to the middle panels except that the data points that are plotted are the mean values computed over the 20 genes in the left panel.

Partly to be able to better identify the points in time relative to cancer diagnosis where changes in gene expression occur, we developed methods based on a moving window in time. This approach includes time in a more continuous manner than the approach based on curve groups. In Holden et al. 2017 we used moving windows and randomization for the same dataset as we used in Lund et al. 2016, Chapter 8. The null hypotheses of no differences between cases and controls, no time-dependent changes, and no differences between different strata were all rejected. The main conclusion of the analyses was that there are time-dependent changes of the blood transcriptome up to eight years before breast cancer diagnosis.

Changes in the number of differentially expressed genes over time

In Chapter 9 we examined the changes in gene expression after diagnosis for a breast cancer dataset in the Norwegian Women and Cancer (NOWAC) postgenome cohort. We stratified stage in invasive or metastatic breast cancer, and vital status in dead or alive at the end of follow-up, and observed a significant increase of differentially expressed genes among women with metastatic disease who later died both compared to invasive cases that survived ($p = 0.001$) and to metastatic cases that survived ($p = 0.024$). To illustrate the difference between strata, we made heatmaps for the most differentially expressed genes over time; see Figure 6.2 below.

We also observed a second transient increase in blood gene expression a few years after diagnosis in metastatic cases, hypothetically representing a capitulation of the immune system.

Identifying significant genes based on area between curves

The analysis of complex curves was extended in order to identify genes that are significantly different between two strata The method tests differences in a non-parametric time development relative to time of diagnosis of the gene expressions from different strata using the area between the curves in a long time period.

The method was tested on case-control differences in log2 gene expressions in a post diagnostic time period separating between the women who survived and the women who died of breast cancer. The method clearly showed non-linear changes, with rapid transient mostly increasing fold changes, in cases who later died. Survivors had no changes. For cases that died, this transient increase was followed by a regression towards the gene expression profiles of survivors. For 9786 genes, the integrated area from 18 months to 8 years was highly significant ($p < 0.00001$) among women who died. There were indications of a stronger relationship in metastatic cases alone.

Figure 6.3 below shows the $f_{s(c),g}(t_c)$ curves for the 100 most significant genes in the period after diagnosis for the women that survived and died of breast cancer.

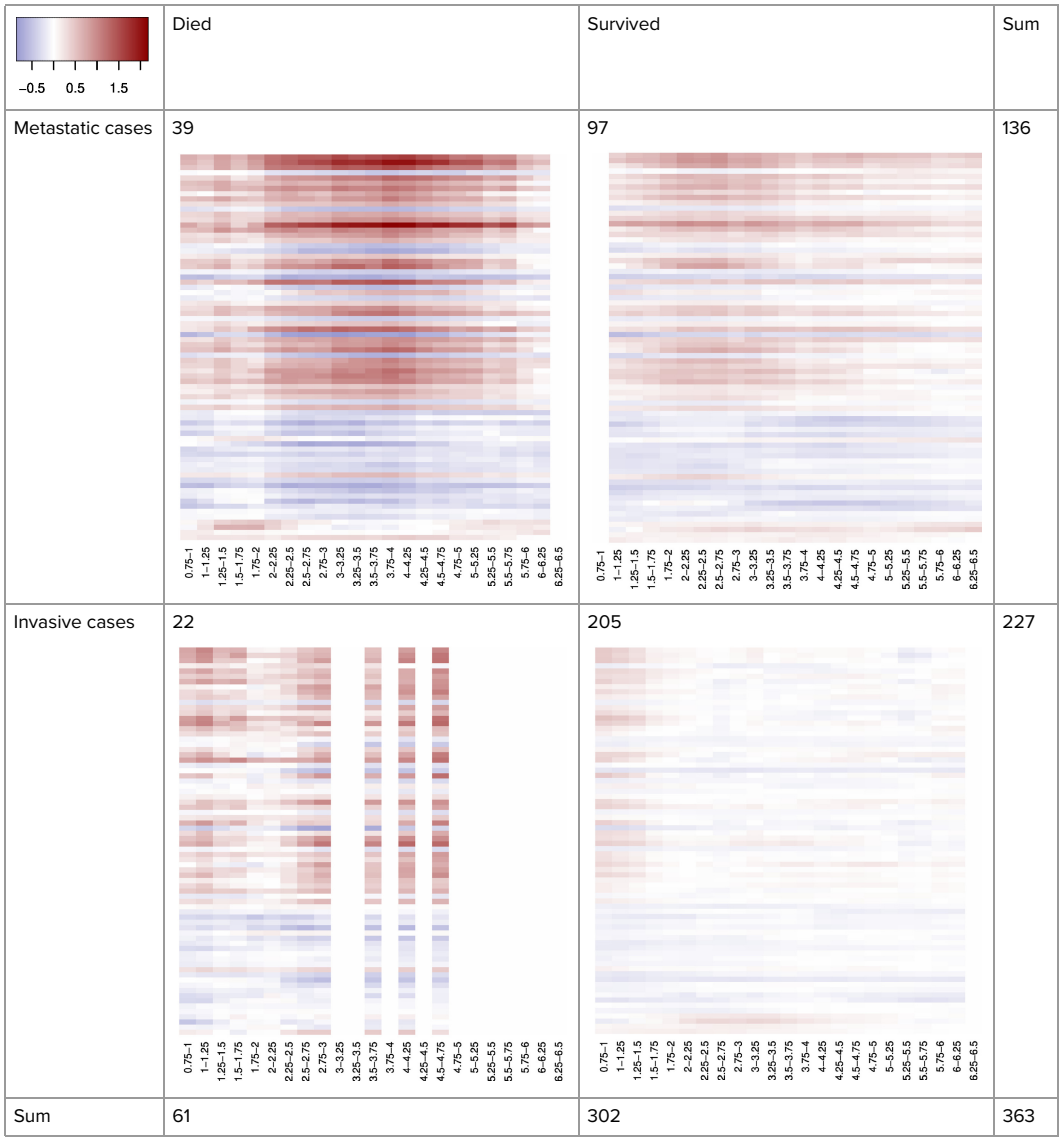


Figure 6.2. Heat maps with 74 selected genes in each stratum; invasive versus metastatic and alive or dead at end of follow-up. The heat maps show log fold change for each gene (y-axis) for each quarter of the years after diagnosis (x-axis).

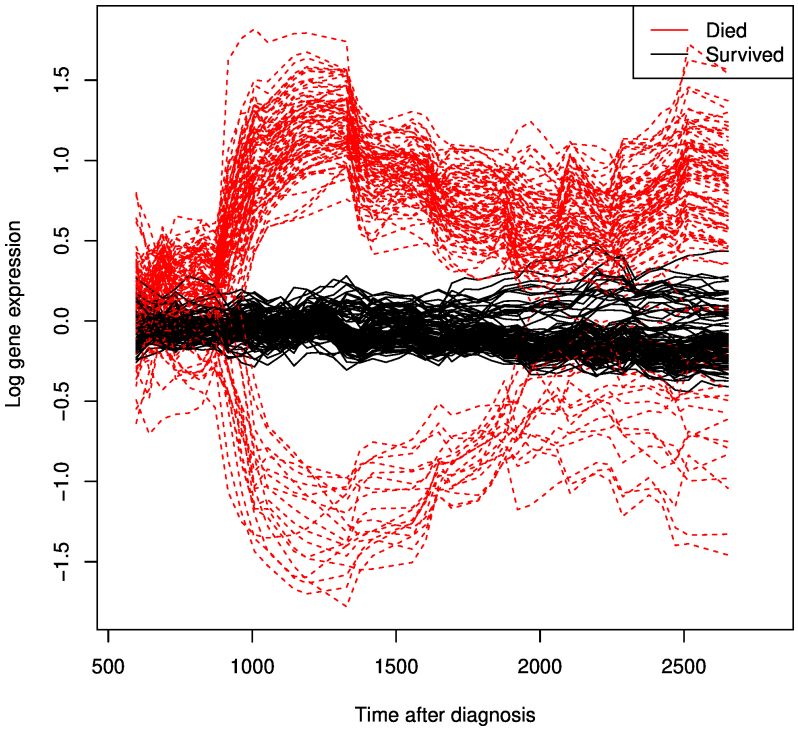


Figure 6.3. The 100 most significant genes in the period after diagnosis of breast cancer. The data is scaled such that variance in the data is 1 for all the genes. Hence, the vertical axis does not give information about the fold change.

This systems epidemiology approach provides a proof of concept for the use of gene expression as an individualized biomarker of prognosis related to death or not. Since we had no prior knowledge of the shape of differences in gene expressions as a function of time relative to diagnosis, we needed a non-parametric model that identified all possible changes in trajectories. The aim of the study was 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.

Prediction of stratum based on local statistics

In Holden et al. 2017, we described *Prediction of stratum based on local statistics*, to illustrate how the predictive power of the test varies with time. In Figure 6.4 below, we observe that for screening detected cancers the probability of correct prediction of metastasis status was best in year 1 before diagnosis compared to year 3 and 4 before diagnosis for clinically detected cancers.

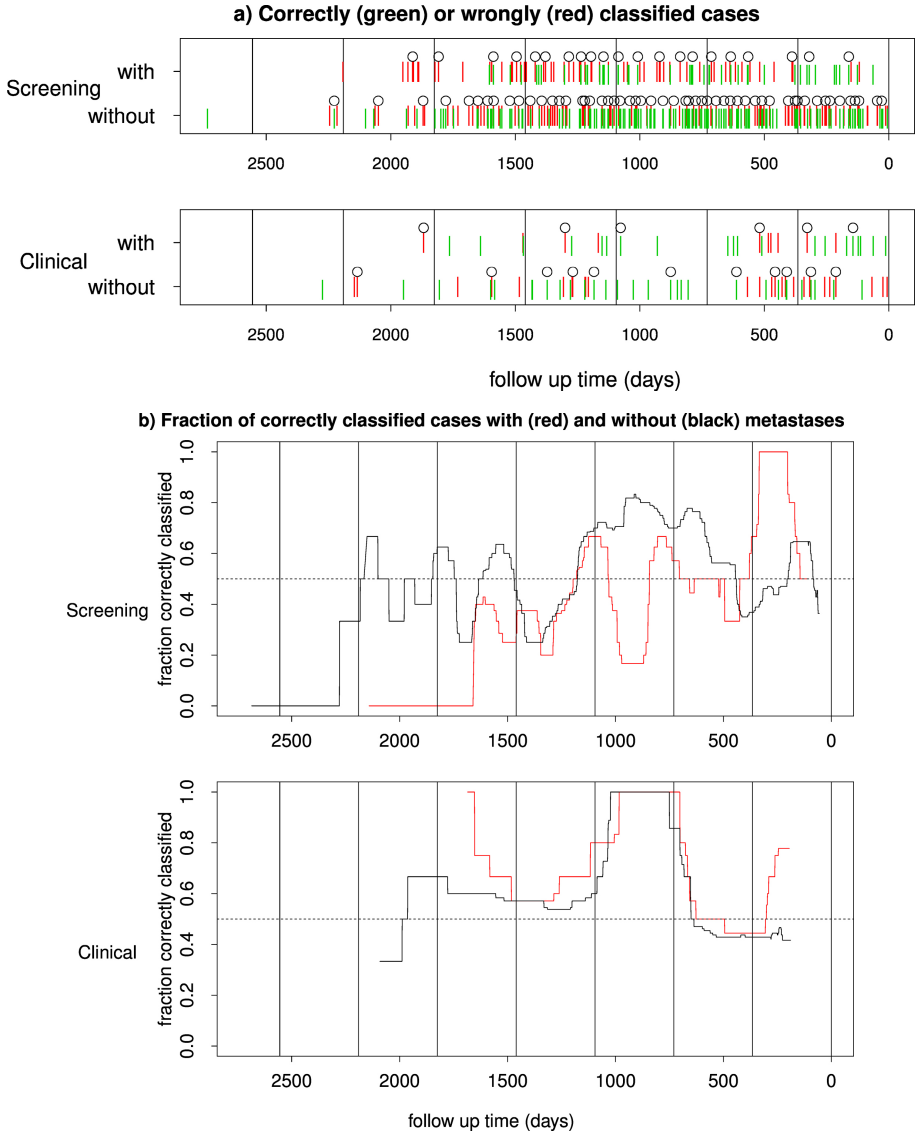


Figure 6.4. Prediction results. a) Correctly (green) or wrongly (red) classified cases plotted against time to diagnosis for the screening (upper panel) and the clinical group (lower panel). A circle is plotted above every fifth case. Long vertical lines are plotted to indicate the years. On the y-axis "with" means cases with metastases and "without" means cases without metastases. b) Fraction of correctly classified cases with (red) and without (black) metastases 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 (clinical) or 100 days (screening). The resulting curve is then smoothed using a median filter with a window size of one year (clinical) or 100 days (screening).

Identifying significant genes with only one time period

The mortality of breast cancer is strongly associated with parity, i.e. the number of children to whom a woman has given birth (Lund 1990). In Lund 2018, we used one of the breast cancer datasets in the Norwegian Women and Cancer (NOWAC) postgenome cohort mentioned in the previous sections to examine whether there is an association between gene expression and parity. In that study we used only one time period, which means that we ignore information about time and as a consequence will mainly be able to identify genes with expressions that do not vary with time in the years before/after diagnosis. As there is a large body of evidence demonstrating the long-lasting protective effect of each full-term pregnancy (FTP) on the development of breast cancer (BC) later in life, this is reasonable.

We used the Bioconductor R-package Limma, linear models for microarrays (Ritchie et al. 2015), to identify the genes that were influenced by parity. In the linear model, the responses were $X_{g,c}$, the differences in the log2 gene expression for each case-control pair, while we included the parity of the control and the parity of the case as covariates. In the analyses, we merged parities 1–3 and 4–6 so that the parity data consisted of three different values: 0, 1–3, and 4–6. The merging was done in order to reduce the effect of the highest parities. We identified gene sets that were influenced by parity using Limma in the same way as we did for individual genes, by using enrichment scores for gene sets instead of differences in the log2 gene expressions as responses in the linear model. The enrichment scores for gene sets were obtained from the $X_{g,c}$s using the Bioconductor R-package gene set variation analysis, GSVA (Hänzelmann et al. 2013).

We found that 756 genes showed linear trends in cancer-free controls, false discovery rate (FDR) 5%, but this was not the case for any of the genes in the breast cancer cases. Gene Set Enrichment Analysis, GSEA of immunologic gene sets, C7 collection in Molecular Signatures Database, MSigDB (GSEA MSigDB, Subramanian et al. 2005) revealed 215 significantly enriched human gene sets (FDR 5%). These marked differences in gene expression and enrichment profiles of immunologic gene sets between breast cancer cases and healthy controls suggest an important protective effect of the immune system on breast cancer risk.

REFERENCES

GSEA MSigDB, Gene Set Enrichment Analysis (GSEA). Molecular Signatures Database (MSigDB). Internet. UC San Diego and BROAD Institute; Available from: http://software.broadinstitute.org/gsea/index.jsp

Holden L. Classify strata. Oslo: Norwegian Computing Center; SAMBA/11/15; 2015b. pp 28. Available from: https://www.nr.no/directdownload/1426685952/classify_strata_holden2015.pdf

Holden L. Time development of gene expression; Oslo: Norwegian Computing Center; SAMBA/35/15; 2015a. pp. 13. Available from: https://www.nr.no/files/samba/smbi/note2015SAMBA3515timeDevelopmentGenes.pdf

Holden M, Holden L, Olsen KS, Lund E. 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. 2017; 7: 11–28. Available from: https://www.dovepress.com/local-in-time-statistics-for-detecting-weak-gene-expression-signals-in-peer-reviewed-article-AGG

Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics. 2013 Jan 16; 14: 7. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-7

Lund E. Childbearing in marriage and mortality from breast cancer in Norway. Int J Epidemiol. 1990; 19(3): 527–531. Available from: https://academic.oup.com/ije/article-abstract/19/3/527/760486?redirectedFrom=fulltext

Lund E, Holden L, Bøvelstad H, Plancade S, Mode N, Günther C-C et al. A new statistical method for curve group analysis of longitudinal gene expression data illustrated for breast cancer in the NOWAC Post-genome Cohort as a proof of principle. BMC Med Res Methodol. 2016 Mar 5; 16: 28. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0129-z

Lund E, Nakamura A, Snapkov I, Thalabard JC, Olsen KS, Holden L, et al. Each pregnancy linearly changes immune gene expression in the blood of healthy women compared with breast cancer patients. Clin Epidemiol. 2018 Aug 6; 10: 931–940. Available from: https://www.dovepress.com/each-pregnancy-linearly-changes-immune-gene-expression-in-the-blood-of-peer-reviewed-article-CLEP

Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics. 2003 Feb 12; 19(3): 368–375. Available from: https://academic.oup.com/bioinformatics/article/19/3/368/258230

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20; 43(7): e47. Available from: https://academic.oup.com/nar/article/43/7/e47/2414268

Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005 Oct 25; 102(43): 15545–15550. Available from: https://www.pnas.org/content/102/43/15545.long

Lund, E. (Ed.) (2020). *Advancing Systems Epidemiology in Cancer. Exploring Trajectories of Gene Expression.* Oslo: Scandinavian University Press.
DOI: https://doi.org/10.18261/9788215041193-2020-07

7. Seven Levels of Analogical Fallacies—From Mice to Women

or from reductionist experiments in mice to functional transcriptomics in humans

Eiliv Lund

Abstract Interpretations of findings in transcriptomic analyses as part of systems epidemiology are usually based on analogies from mostly reductionist experiments on mice. Such transfer of knowledge from one scientific discipline to another depends on the validity of comparisons. The potential fallacies of analogical thinking cover all aspects of the differences between mice and humans, genetically and in lifestyle. We need better classification of the experimental information in standard databases.

Keywords Analogical fallacy | reductionist experiment | observational data | immunology | lifestyle

The interpretations of results from gene expression analyses in humans are heavily dependent on reductionist experiments in animal or in human cell lines. The most commonly used animal is the mouse (Breschi et al. 2017). However, to go from mice experiments to human evidence is a long and winding path signposted with analogies. The traditional definitions of analogy is a comparison between "things that have similar features, often used to help explain a principle or idea" (Cambridge Dictionary) "or one thing or another, typically for the purpose of explanation or clarification" (Oxford Dictionaries).

Systems epidemiology (Lund and Dumeaux 2008), the rapidly growing interest and potential for analyses of functional genomics in human studies, confronts

researchers with interpretations of statistical associations based on biological knowledge. The explanations of the epidemiological findings will depend mostly on information from animals, particularly mice. This confrontation could deepen the canyon between the scientific disciplines of basic biology and epidemiology, especially concerning the comparability or validity of transferring information from one biological species to another. We will discuss the use of information from reductionist experiments in the interpretations of functional genomics—from mice to men, as the editorial of *Nature Medicine* put it (Editorial 2013); or, in our case, from mice to women. The transfer is mostly based on analogies with little specific knowledge about the nature of these biological and methodological differences in different species. In epidemiology, analogy was originally one of the criteria of causality (Bradford Hill 1965), but over time lost its meaning due to the potential for fallacious thinking, and it became difficult to assume that results from a study on one disease could be generalized to other research areas. Today, epidemiologists working with analyses of functional genomics import knowledge of function from databases of information on basic biology. It is a concern that the information collected for systems biology cannot necessarily be used for interpretations of the biology of statistical associations.

The purpose is to describe some of the many analogical fallacies that should be considered, from mice reductionist experiments to human lifestyle, and in addition to discuss the upcoming issue of the validity of mice experiments in relation to the immune system.

SEVEN LEVELS OF ANALOGICAL FALLACIES

The wide variety of analogical fallacies, from reductionist experiments to observational studies of humans, can be classified into at least seven levels. Normally, these levels are not discussed in depth in epidemiological papers. Here we will focus on the use of information gathered from mice experiments and its relevance for the interpretation of functional genomic findings in epidemiology. The seven levels are shown in Box 7.1, where the epidemiological use of information is compared to the reductionist situation.

Box 7.1

One: genomic differences
In most mice, experiments are inbred mostly from the same strain or genetically modified. In women there are large differences between individuals of the genome at the level of single nucleotide polymorphisms. The reductionist experiments can consequently say nothing about the effect of genetic variation in the mice model. For humans, enormous efforts have been made over the last two decades to understand the importance of this genetic variation on cancer risk.

Two: lifestyle
Mice are normally treated in the lab with standard diets. It is observed that changing the diet also changes the effect of experiments related to diet. Among women, diets are very different, from vegans to people on low-carb, high-fat diets. Many diets still depend on local conditions and culture. This diversity of diets influences the gene expression or functional genomics, and must be taken into account in the analyses.

Three: sterile labs
Most labs are built for being non-pathogenic. The mice are living under a controlled infection environment. The women attract continuously through life new infections from bacteria and viruses. This has an important consequence for studies of the immune system. In blood of laboratory mice, the immune system shows no effects of a long infectious life, in fact the immune system more like human newborn. In women the immune system accumulates experiences over years.

Four: physiology, hormones, and immune system
There are many differences in physiology and hormone levels in mice and women. Most important for female cancers are probably differences related to fertility and lactation. Mice have no regular cycles while women have a regular complicated hormonal regulation of the menstruation giving her opportunity for a pregnancy every month in a period of more than 30 years. In addition, mice have an immune system more directed towards the old innate system versus humans with a more developed adaptive. During pregnancy, the fetus acts like a pseudo semi-allocraft.

Five: differences in tissue sampling
Mice will often be killed at time for sampling of biological material. Samples can therefore be from liver, bone marrow or spleen. In humans blood sampling is the most easy and common sampling procedure. Some human studies use blood from umbilicus at time of birth.

Six: gender or sex
Most mice experiments have used male mice. This had some consequences for studies of hormone related products. The gender is not always given in the articles or different to find. In human gender is a standard factor for stratification. The use of specific gender in the experiments are often not discussed. In epidemiology almost no studies use analyses of male and female together.

Seven: nature of tumors
In mice most tumors are induced. They are not malignant. The incidence of cancer in mice naturally is lower than in human with no outcome studies in cancer. While epidemiology has hundreds of thousands of participants a lab cannot hold more than a few mice over their lifetime. The construction of tumors is therefore either chemically or through genetic manipulation. There is more heterogeneity in breast cancer in women than in mice.

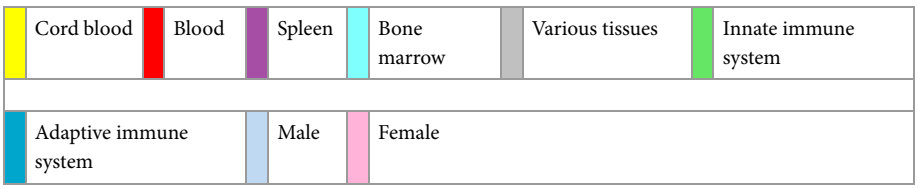
Example: Gene set enrichment analysis

The following is an example of the differences of the parameters between mice and human information taken from an analysis of gene expression in relation to parity among breast cancer patients and controls; for its design, see Lund et al. 2018.

Table 7.1. Top 10 gene sets for humans and mice with information on experimental design (Lund et al. 2018)

gene-SetID	score-Diff	p-value	FDR q-value	source	immune system	parity	cells	sex
Human								
GSE3982	0.023	3.4E-05	0.027	Cord blood	I	NA	Macrophages	NA
GSE2770	-0.026	5.9E-05	0.027	Cord blood	A	NA	CD4+ T cells	NA
GSE16385	0.047	7.0E-05	0.027	Blood	I	NA	Monocytes	NA
GSE1460	0.034	1.0E-04	0.027	Cord blood / Blood	A	NA	CD4+ T cells	NA
GSE13411	-0.035	1.6E-04	0.027	Spleen	A	NA	B cells	M / F
GSE2770	0.027	1.7E-04	0.027	Cord blood	A	NA	CD4+ T cells	NA
GSE29618	-0.029	1.9E-04	0.027	Blood	I	NA	DC	NA
GSE17974	0.028	1.9E-04	0.027	Cord blood	A	NA	CD4+ T cells	NA
GSE2770	-0.028	1.9E-04	0.027	Cord blood	A	NA	CD4+ T cells	NA
GSE29615	0.038	2.2E-04	0.027	Blood	I / A	NA	PBMCs	NA
Mouse								
GSE17721	0.018	1.1E-05	0.027	Bone marrow	I	NA	DC	F
GSE14769	0.032	1.2E-05	0.027	Bone marrow	I	NA	Macrophages	NA
GSE3691	0.037	3.3E-05	0.027	Various tissues	I	NA	DC	F
GSE37301	0.035	3.6E-05	0.027	Bone marrow	I / A	NA	HSCs, CLPCs	NA
GSE32034	0.028	3.7E-05	0.027	Various tissues	I	NA	Monocytes	M
GSE17721	0.026	4.5E-05	0.027	Bone marrow	I	NA	DC	F
GSE21063	0.036	4.5E-05	0.027	Spleen	A	NA	B cells	NA
GSE11924	-0.029	6.5E-05	0.027	Spleen	A	NA	CD4+ T cells	NA
GSE28237	0.035	7.7E-05	0.027	Spleen	A	NA	B cells	NA
GSE13547	0.034	7.8E-05	0.027	Spleen	A	NA	B cells	NA

I – innate immune system, A – adaptive immune system, DC – dendritic cells, PBMCs – peripheral blood mononuclear cells, HSCs – hematopoietic stem cells, CLPCs – common lymphoid progenitor cells M – male, NA – not applicable, F – female



A total of 588 gene sets were identified from the C7 collection in Molecular Signatures Database (GSEA MSigDB) that were significantly enriched when the parity of the controls varied (FDR 5%). Experimentally produced gene sets are submitted to MSigDB from researchers using both *in vivo* and *in vitro* material, as well as both human cells or tissues and animal models. Of our 588 enriched gene sets, 215 were derived from human data and 373 were derived mostly from mouse data. Detailed information on the top 10 gene sets from each species showed a discrepancy in the tissues analyzed: blood and cord blood in humans versus bone marrow and spleen in mice (Table 7.1). These tissue sources are closely related to the immune system; in humans, cord blood is related to the innate immune system and blood is mainly related to the adaptive immune system, while for mice bone marrow is associated with the innate immune system and spleen with the adaptive immune system. We had information on gender for only one of the top 10 human gene sets, and four of the mouse gene sets, even after contacting the main authors of the associated publications. The five gene sets from human cord blood mostly represented the adaptive immune system, but immune signatures from cord blood have not been found to be representative of adult immune response, instead resembling a newborn response (Beura et al. 2016). None of the MSigDB gene sets from either humans or mice included information on the number of full-term pregnancies. Female laboratory mice are generally nulliparous for breeding reasons.

In Gene Set Enrichment Analysis, GSEA we focused our analysis on the C7 collection in Molecular Signatures Database, MSigDB (GSEA MSigDB), consisting of gene sets related to the immune system. Interestingly, the greatest portion of these gene sets was obtained from animal studies. It is nearly impossible to make an unambiguous conclusion from these results due to the complexity of the gene sets' data and lack of essential information in the gene sets' descriptions (e.g. we were unable to ascertain the gender of blood donors for most human experiments even after contacting the authors of publications). We observed a clear interspecific difference between components of the immune system to which the gene sets are related (in mice, most of the gene sets are of innate immunity origin, while in humans the immunity origin is adaptive) and between the sources of the cells for the experiments. While the latter can be explained by technical and ethical considerations, the former raises yet another concern on the validity of results obtained in animal models.

DISCUSSION

We have proposed at least seven levels of analogical fallacy; some could think of more. In addition, an intense debate is ongoing related to the scientific validity of reductionist experiments (Gould et al. 2015). Surprisingly, many experiments cannot be repeated due to lack of information, and many that have been repeated show different results. Together, this has raised the question of the use of experimentally derived information to find new drugs and for the testing of human patient populations. In spite of attempts to improve the scientific standard, not much has been achieved (Enserink 2017).

Comparative transcriptomics in humans and mice have shown a number of limitations (for an overview, see Breschi et al. 2017). These include incomplete transcriptomic characterization, difficulties in identifying orthologue phenotypes and cells. Emerging technologies could improve our understanding of the conditions under which the mouse is an acceptable model of human physiology. Current limitations of mouse models are well known in the research into carcinogenesis. Due to differences in duplication time, lifespan and cancer susceptibility, the success rate of translation from animal models to human clinical trials has been less than 10% (Editorial 2012).

While the similarities between the human and mouse genome are acceptable for mechanistic research, the murine genome being 12% smaller. Around 90% of each genome can be portioned into conserved syntenic regions, but only 40% of the nucleotides. The remaining 60% might be due to changes during evolution. For protein-coding genes, around 20-30% are either one-to-many or many-to-many orthologous relationships. Since these changes could reflect human disease phenotypes, care should be taken with interpretations. For the long-coding RNAs the situation is even more diverse, and even more so for small-coding RNAs (miRNA) where only a small fraction has a defined orthology.

Comparison of different mouse models of breast cancer showed that many of the characteristics gave tumors with distinctive and homogenous expression patterns within each strain, but none of the models had all of the expression profiles of a specific human subtype (Herschkowitz et al. 2007). In fact, in mice the development of spontaneous mammary tumors is linked to the infection of mice with either exogenous or endogenous viruses (Russo 2015).

Consequently, similarities in gene expression between species do not necessarily reflect the functional gene expression signals in humans. The immune system could be the most challenging field for comparative transcriptomics. As an example, the transcriptional responses to trauma such as burns or accidents have highly similar genomic responses in humans, while the responses in corresponding

mouse models correlate poorly with human conditions. No such patterns can be identified in mice (Seok et al. 2013). Even if new RNA-seq technology is introduced that can give more detailed information, the mouse and human samples are not comparable since the former is taken from animal tissue while the latter mostly comes from peripheral blood. Another important aspect of comparative transcriptomics is the growing concern that laboratory mice do not reflect relevant aspects of the human immune system (Beura et al. 2016, Maizels and Nussey 2013). Laboratory mice are inbred and genetically similar, with genetic manipulations. In addition, they live in abnormally hygienic conditions in sterile laboratories. It has been shown that laboratory mice have an immune system more like that of newborns, not adult humans. They lack effector-differentiated and mucosally distributed memory T-cells. Altering the life conditions of the mice changes the cellular composition of the innate and adaptive immune system to be more like adult humans. Restoring physiological microbial exposures in the laboratory could improve the relevance of mice models for studies of immunology in free-living humans (Beura et al. 2016). These results can also be discussed from an evolutionary setting (Maizels and Nussey 2013) in which the immune system has evolved over years. The genuine importance of the environment limits the generalization of results obtained in good, controlled laboratory settings compared to the challenges of life in the wilderness.

In addition, there are many more methodological and statistical problems in the experimental research on cancer (Holman et al. 2016). Small sample sizes, loss of animals randomly or as outliers, and the statistical testing of many hypotheses can all increase the probability of false positive results.

Very specifically, gender is often forgotten in environmental toxicology, and often in other studies too (Liang et al. 2018). This will result in noise in the analyses if there are different transcriptomic effects of chemicals or other lifestyle factors.

There is also very sparse comparative data about transcriptional changes associated with the carcinogenic process over time.

The translation of results from basic biology reductionist experiments to human drugs or tests have been high on the research agenda for many years. In one comment in 2012, the question was how to raise standards for preclinical cancer research in order to increase the reliability of preclinical cancer studies (Begley and Ellis 2012). Six years later, an editorial in *BMJ* had the heading "We need better animal research, better reported" (Goodlee 2018). Another comment in *Nature* bore the title "Consider drug efficacy before first-in-human trials" (Kimmelman and Federico 2017). Some specific problems are related to the use of translational mouse models in oncology drug development (Gould et al. 2015). The misleading

mouse studies have wasted medical resources, but what is worse, led to unnecessary clinical trials with thousands of patients (Perrin 2014). The sloppy reporting on animal studies continues in 2017 (Enserink 2017).

CONCLUSION

Epidemiologists should be very careful in using animal model (non-human) experimentally derived information in order to interpret findings of functional genomics as part of systems epidemiology. The databases should more clearly define the methodology of the design in reductionist experiments.

REFERENCES

Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Comment. Nature. 2012 Mar 28; 483(7391): 531–533. Available from: https://www.nature.com/articles/483531a

Beura LK, Hamilton SE, Bi K, Schenkel JM, Odumade OA, Casey KA. Normalizing the environment recapitulates adult human immune traits in laboratory mice. Nature. 2016 Apr 28; 532(7600): 512–516. Available from: https://www.nature.com/articles/nature17655

Bradford Hill A. The Environment and Disease: Association or Causation? Proc R Soc Med 1965 May; 58: 295–300. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1898525/

Breschi A, Gingeras TR, Guigó R. Comparative transcriptomics in human and mouse. Nat Rev Genet. 2017 Jul; 18(7): 425–440. Available from: https://www.nature.com/articles/nrg.2017.19

Cambridge Dictionary [Internet]. Accessed: 15.11.2019. Available from: https://dictionary.cambridge.org/dictionary/english/analogy

Must try harder [Editorial]. Nature 2012 Mar 28; 483(7391): 509. Available from: https://www.nature.com/articles/483509a

Of men, not mice [Editorial]. Nat Med 2013 Apr 4; 19(4): 379. Available from: https://www.nature.com/articles/nm.3163

Enserink M. Sloppy reporting on animal studies proves hard to change. Science. 2017 Sep 29; 357(6358): 1337–1338. Available from: https://science.sciencemag.org/content/357/6358/1337.long

Goodlee F. We need better animal research, better reported. Editor´s Choice. BMJ. 2018 Jan 11; 360: k124. Available from: https://www.bmj.com/content/360/bmj.k124

Gould SE, Junttila MR, de Sauvage FJ. Translational value of mouse models in oncology drug development. Nat Med 2015 May; 21(5): 431–439. Available from: https://www.nature.com/articles/nm.3853

GSEA MSigDB, Gene Set Enrichment Analysis (GSEA). Molecular Signatures Database (MSigDB). Internet. UC San Diego and BROAD Institute; Available from: http://software.broadinstitute.org/gsea/index.jsp

Herschkowitz JI, Simin K, Weigman VJ, Mikaelian I, Usary J, Hu Z et al. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. Genome Biol 2007; 8(5): R76. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2007-8-5-r76

Holman C, Piper SK, Grittner U, Diamantaras AA, Kimmelman J, Siegerink B et al. Where have all the rodents gone? The effect of attrition in experimental research on cancer and stroke. PLoS Biol. 2016 Jan 4; 14(1): e1002331. Available from: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002331

Kimmelman J, Federico C. Consider drug efficacy before first-in-human trials. Comment. Nature. 2017 Feb 2; 542(7639): 25–27. Available from: https://www.nature.com/articles/542025a

Liang X, Feswick A, Simmons D, Martyniuk CJ. Environmental toxicology and omics: A question of sex. J Proteomics. 2018 Feb 10; 172: 152–164. Available from: https://www.sciencedirect.com/science/article/pii/S1874391917303263?via%3Dihub

Lund E, Dumeaux V. Systems epidemiology in cancer. Cancer Epidemiol Biomarkers Prev. 2008 Nov; 17(11): 2954–2957. Available from: https://cebp.aacrjournals.org/content/17/11/2954.long

Lund E, Nakamura A, Snapkov I, Thalabard JC, Olsen KS, Holden L, et al. Each pregnancy linearly changes immune gene expression in the blood of healthy women compared with breast cancer patients. Clin Epidemiol. 2018 Aug 6; 10: 931–940. Available from: https://www.dovepress.com/each-pregnancy-linearly-changes-immune-gene-expression-in-the-blood-of-peer-reviewed-article-CLEP

Maizels RM, Nussey DH. Into the wild: digging at immunology's evolutionary roots. Nat Immunol. 2013 Sep; 14(9): 879–883. Available from: https://www.nature.com/articles/ni.2643

Oxford Dictionaries [Internet]. Accessed: 15.11.2019. Available from: https://en.oxforddictionaries.com/definition/analogy

Perrin S. Preclinical research: Make mouse studies work. Comment. Nature. 2014 Mar 26; 507(7493): 423–425. Available from: https://www.nature.com/articles/507423a#article-comments

Russo J. Significance of rat mammary tumors for human risk assessment. Toxicol Pathol. 2015 Feb; 43(2): 145–170. Available from: https://journals.sagepub.com/doi/full/10.1177/0192623314532036?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed

Seok J, Warren HS, Cuenca AG, Mindrinos MN, Baker HV, Xu W et al. Genomic responses in mouse models poorly mimic human inflammatory disease. Proc Natl Acad Sci USA. 2013 Feb 26; 110(9): 3507–3512. Available from: https://www.pnas.org/content/110/9/3507.long

Lund, E. (Ed.) (2020). *Advancing Systems Epidemiology in Cancer. Exploring Trajectories of Gene Expression.*
Oslo: Scandinavian University Press.
DOI: https://doi.org/10.18261/9788215041193-2020-08

8. 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

Eiliv Lund, Lars Holden, Hege Bøvelstad, Sandra Plancade, Nicolle Mode, Clara-Cecilie Günther, Gregory Nuel, Jean-Christophe Thalabard and Marit Holden

Abstract The understanding of changes in temporal processes related to human carcinogenesis is limited. Here we compile trajectories of differential expression of genes, based on measurements from many case-control pairs. We propose a new statistical method that does not assume any parametric shape for the gene trajectories. This new statistical approach had good properties in terms of statistical power and type 1 error under minimal assumptions. It was able to discriminate between groups of genes with non-linear similar patterns before diagnosis.

Keywords transcriptomics | gene expression | breast cancer | carcinogenesis | curve group statistics

Previously published in *BMC Medical Research Methodology,* 2016; 16, 28.

BACKGROUND

The assumption of systems epidemiology (Lund & Dumeaux, 2008) is that functional aspects of the human carcinogenic process can be detected in the blood as gene expression patterns before cancer diagnosis, either as active signals or as passive information. A recent editorial in *Nature Medicine* ("Of men, not mice", 2013) stressed that if we are to understand the carcinogenic process, research needs to shift from mouse models to a "human model". However, the peculiarities and timescale of cancer development in humans essentially force us to rely on observational studies. The prospective design is clearly the best design if one wants to incorporate the time aspect of carcinogenesis and changing exposures. However, practical considerations frequently force us to use a nested case-control design within the cohort, which keeps part of the advantage of the previous design. Analyses of somatic mutations in cancer genome studies have revealed the huge diversity of mutational processes that occurs during carcinogenesis (Alexandrov et al., 2013). One explanation for this observation could be that multiple mutational processes operate differently within biological processes depending on subtypes of cancer, thus giving a jumbled composite signature. In order to avoid jumbled composite signatures, functional analyses in observational studies must be stratified by important clinical information such as lymph node status and exposures to potential carcinogens.

One approach for prospective functional genomic studies is to compile trajectories based on measurements from many case-control pairs in order to study the carcinogenic process (Lund & Plancade, 2015). The trajectory of a gene is defined as the curve showing the changes in gene expression levels in the blood as a function of time to cancer diagnosis, and consists of a nested case-control design of the differences in gene expression levels between cases and controls.

Our overall aim was to develop statistical methods for exploring the changes in gene expression in years before diagnosis as part of a processual approach (Lund & Plancade, 2015), not to estimate risk.

There is no prior knowledge about the form of the trajectory of gene expression for any of the thousands of genes. This lack of a priori information normally demands an agnostic approach (Spitz & Bondy, 2010), i.e. considering all genes as equal and adjusting for multiple testing using a false discovery rate (Reiner, Yekutieli, & Benjamini, 2003). Here, however, we present a new statistical method to study trajectories. We applied this new method in a prospective analysis of women with breast cancer in the Norwegian Women and Cancer (NOWAC) postgenome cohort (Dumeaux et al., 2008). The trajectories were analyzed within strata of different biological stages in carcinogenesis of breast cancer within the screening or

outside as clinical cancer, but without identifying single genes or conducting pathway analyses.

METHODS

The new statistical approach is described below. As a "proof of concept" we carried out an analysis in a nested case-control design in the Norwegian Women and Cancer postgenome cohort. For each incidence of breast cancer identified through linkage to the Norwegian Cancer Registry, a control was drawn from blood samples collected at the same time and year of birth. This ensured the same storage time and no effect of age between cases and controls. The pairs of cases and controls were kept together throughout all laboratory procedures in order to reduce batch effects. For more details, see later under Epidemiological design and study population.

Statistical methods

The new statistical method for curve group analyses is based on a set of hypothesis testing that we developed in order to detect changes in gene expression levels over time, and whether these changes, if they exist, differ among strata. This method is able to identify small changes that vary slowly over time and/or among strata, by using a large number of genes in each analysis. In order to define test statistics that measure the development of differential gene expression levels over time and differences among strata, we have introduced the concept of curve groups, where each curve group consists of genes that have a similar development over time, i.e. similar differential trajectories. These methods are described in detail below:

Let $X_{g,p}$ be the $\log_2$-expression difference for gene g and the matched case-control pair p. Each case-control pair belongs to a stratum s and a time period t. We wanted to test whether $X_{g,p}$ is independent of the time period, and whether there is no difference among the strata, i.e. $X_{g,p}$ is independent of stratum. Figure 8.1 gives an overview of the different tests and the variables used in these tests, the strata used in the analyses, and the table and figures where the results are shown.

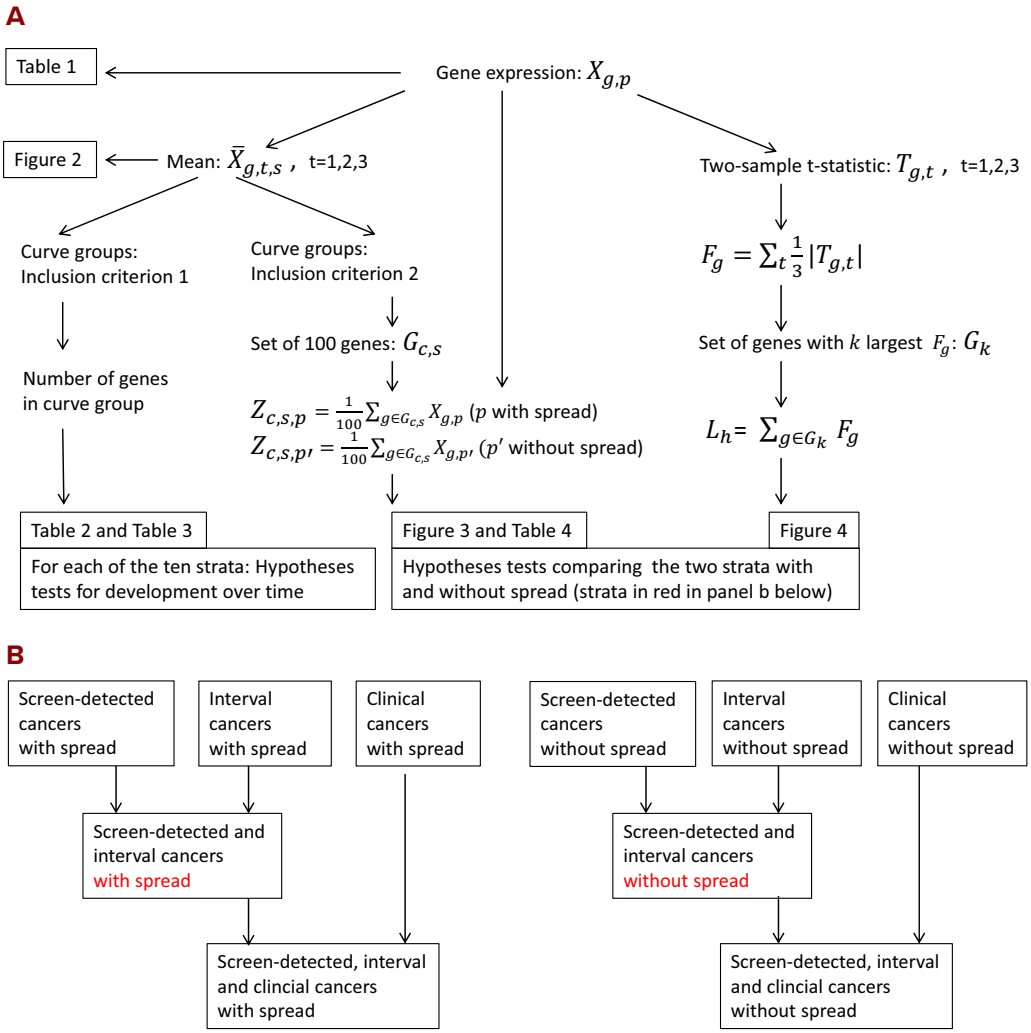


Figure 8.1. Overview of hypothesis tests, variables, strata, tables and figures. A. Illustration of the relationship between the data $X_{g,p}$, the different hypothesis tests, the variables used in these tests, and which tables and figures that show the results from the tests. B. Overview of the different strata.

In the illustrative application, analyses were either conducted within strata of lymph node status at breast cancer diagnosis (positive or 'with spread', and negative or 'without spread') or with respect to breast cancer screening visits (detection categories); cancers diagnosed during screening visits were considered 'screen-detected cancers'; cancers diagnosed within 2 years of last screening visit were

considered 'interval cancers'; and cancers diagnosed clinically in women that did not attend screening or had not attended screening for more than 2 years were considered 'clinical cancers' (Table 8.1).

Table 8.1. Number of case-control pairs in each stratum and time period with gene expression data $X_{g,p}$

Strata		Year before diagnosis (time period)		
Detection category	Lymph node status	5-3 (3)	2 (2)	1 (1)
Screen-detected cancers[a]	With spread	41	11	6
	Without spread	118	42	43
Interval cancers[b]	With spread	28	9	6
	Without spread	30	15	10
Clinical cancers[c]	With spread	11	8	10
	Without spread	28	12	13

[a]Diagnosed at a screening visit.
[b]Diagnosed within 2 years of a screening visit.
[c]Diagnosed clinically and did not attend the screening program or diagnosed clinically more than 2 years after a screening visit.

Hypothesis test for development over time in each stratum

For each stratum, we tested whether $X_{g,p}$ is independent of the time period in a global test since we are interested in weak signals from many genes, not signals that may only be identified in a single gene. To define a test statistic that measures development over time, we used curve groups. The follow-up time was divided into three time periods $t = 1,2,3$ where $t = 1$ is 0–1 year before cancer diagnosis, $t=2$ is 1–2 years before cancer diagnosis, and $t = 3$ is 3–5 years before cancer diagnosis.

- For a given stratum s, a gene g can belong to zero or one of six curve groups based on the average (mean) of the data over all case-control pairs in the stratum in each of the three time periods. These averages were denoted $X_{g,3,s}$, $X_{g,2,s}$ and $X_{g,1,s}$, respectively, and the curve groups are defined based on the ordering of these three averages. In order to search for curves with changes over time, we defined six potential curve groups that changed from time period to time period, called '123', '132', '213', '231', '312', and '321', respectively. The three numbers that denote each curve group represent the level of the average gene expres-

sion of time period 3 (left number), the level of the average gene expression of time period 2 (middle number) and the level of the average gene expression of time period 1 (right number). For example, if $X_{g,3,s} < X_{g,2,s} < X_{g,1,s}$, and these three averages are not too similar (to be defined later), gene g belong to curve group '123', indicating an increasing gene expression level over time when approaching the time of diagnosis, with gene expression level 1 in time period 3, gene expression level 2 in time period 2 and gene expression level 3 in time period 1 (closest to the time of diagnosis). If the three averages are too similar, gene g does not belong to any curve group. See Figure 8.2 for an illustration of the concept of curve groups.

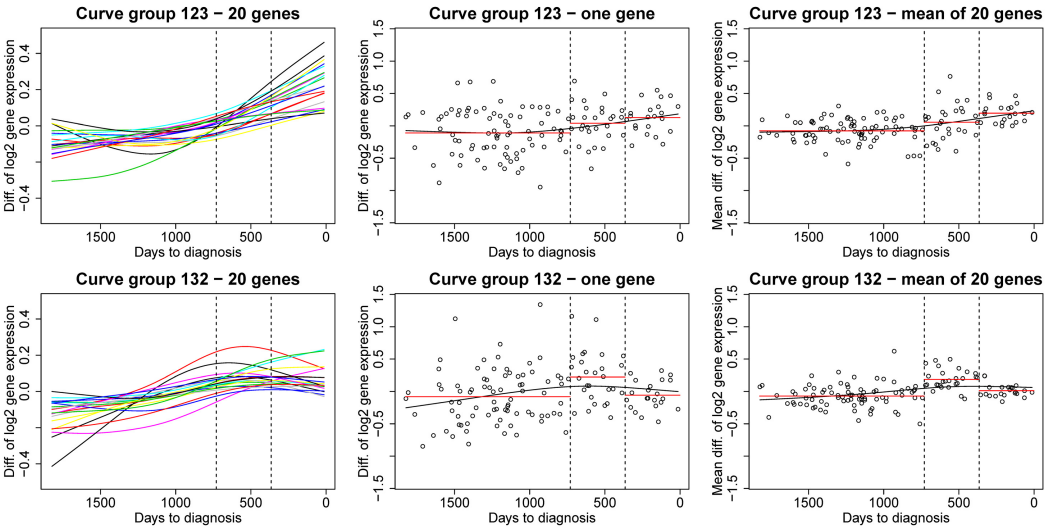


Figure 8.2. Examples of curve groups according to time to diagnosis. Example of two different curve groups: curve group '123' (*upper panel*, gene expression values increasing with time) and curve group '132' (*lower panel*, highest gene expression value in the middle time period). In the left panels curves with the gene expression differences $X_{g,p}$ for 20 genes from the given curve group are plotted. For illustrational purposes, the curves have been estimated from the data using splines. In the middle panels the data $X_{g,p}$ for one of the 20 genes are shown with the corresponding spline-estimated curve. The points represent the differences in gene expression $X_{g,p}$ for each case-control pair. The mean value in each time period, $X_{g,3,s}$, $X_{g,2,s}$ and $X_{g,1,s}$, is shown in red. The right panels are similar to the middle panels except that the data points that are plotted are the mean values computed over the 20 genes in the left panel.

- Each curve group included only genes with a significant change in expression level over time. This was done by testing whether the smallest and largest values

of $X_{g,3,s}$, $X_{g,2,s}$ and $X_{g,1,s}$ were different using a two-sample t-test (assuming unequal variances). Let $p_{g,c}$ be the p-value of this test. Depending on the statistical question at hand, we defined two alternative criteria for concluding that a gene g belongs to the curve group c:

- Inclusion criterion 1: Gene g belongs to curve group c if $p_{g,c}$ is below a predefined limit α.
- Inclusion criterion 2: Gene g belongs to curve group c if gene g is among the M genes with lowest $p_{g,c}$. See more in the next section.

To test for the development of gene expression levels over time, for each stratum we counted the number of genes that belong to the curve group using inclusion criterion 1. We then performed seven hypothesis tests: one global test and one for each of the six curve groups in each stratum. In the global test the test statistic is the total number of genes that belong to any one of the six curve groups, while in the test for individual curve groups the test statistic is the number of genes that belong to the curve group in question. If the conclusion of the hypothesis test was that there were more genes in the curve groups than what was expected by chance, we concluded that there was a significant development over time for some of these genes.

Hypothesis test for comparing two strata

Let us consider in our illustrative example two strata, such as "with spread" and "without spread" at the time of diagnosis. We wanted to test whether there were differences in gene expression levels between these two strata, using information from several genes. For each curve group c, stratum s and case-control pair p, we defined a curve group variable $Z_{c,s,p}$ as follows: we selected the genes that belonged to curve group c for stratum s using inclusion criterion 2 with M=100. Let $G_{c,s}$ denote this set of genes. The curve group variable $Z_{c,s,p}$ for case-control pair p was then computed as the average value of the data $X_{g,p}$ over the genes in $G_{c,s}$:

$$Z_{c,s,p} = \frac{1}{100} \sum_{g \in G_{c,s}} X_{g,p} \, .$$

We could then test whether the variables $Z_{c,s,p}$ were different between the two strata for case-control pairs p either for all time periods combined or for each time period separately. Note that the genes were selected based on data from stratum s, but the variable may have been calculated for case-control pairs p in any stratum. For example, assume that we wanted to test if there was a difference in gene expression level between case-control pairs in the stratum with spread versus the stratum

without spread for curve group '123'. Assume that the set of 100 genes $G_{123,\text{with spread}}$ was selected using criterion 2 in the stratum with spread. We would then have calculated $Z_{123,\text{with spread},p}$ for all case-control pairs p in the stratum with spread and $Z_{123,\text{with spread},p'}$ for p' in the stratum without spread, and tested if the difference was larger than expected by chance. Note that testing the strata with spread versus without spread may also be performed with the set of genes $G_{123,\text{without spread}}$ selected from the without spread stratum or from any of the other defined strata.

An alternative statistic for comparing two strata

The test described above focuses on genes that belong to the same curve group. We also constructed a hypothesis test to compare the difference in development over time between two strata that did not depend on curve groups. This test statistic was constructed by first computing the two-sample t-statistic $T_{g,t}$ and comparing the difference in gene expression levels between the two strata for each gene g and time period t. We defined $F_g = \Sigma_t w_t |T_{g,t}|$ as the weighted sum of the absolute values of the t-statistics for gene g with weight w_t. Furthermore, the test statistic was defined as $L_k = \Sigma_{g \epsilon Gk} F_g$, where G_k is the set of genes with the k largest F_g values, i.e. L_k is the sum of the k largest F_g values. We observe that L_k is a weighted sum of t-statistics. We used equal weights $w_t = 1/3$ for each time period. Alternatively, the weights could be selected either as proportional to the number of case-control pairs in each time period or with larger values for the case-control pairs in a time period closer to the time of diagnosis. We then performed a global test including all three time periods, and separate tests for each time period, in which only data corresponding to each time period were included. This test performed very well on several simulated datasets with a different development over time or different gene expression levels for some genes for two strata. For details see Holden (2015).

Computing p-values—permutation tests

We computed p-values in all the tests described above by estimating the null distribution for the statistic of the hypothesis test by randomizing the data. In the randomization, we preserve critical properties of the genes (level of expression, complex correlation between genes, etc.) and randomize only what's connected to the evolution over time and stratum. This randomization defines the null-distribution for the test statistic that is used when finding the p-value. In hypothesis tests for development over time in a single stratum, the null model was estimated by ran-

domizing case-control pairs for that stratum between time periods, while in the hypothesis tests comparing two strata, the null model was estimated by randomizing case-control pairs between the two strata for each time period. Note that these randomization algorithms maintained the correlation structure between the genes for each case-control pair. Also note that the curve groups were redefined before a sample of the null model was computed from a randomized dataset. The p-value of the test was set to $(K + 1)/(N + 1)$, where N is the total number of randomizations and K is the number of randomizations out of N with a more extreme statistic than the statistic for the real data (Phipson & Smyth, 2010). In the results presented we used $N = 1000$.

Illustrative example: epidemiological design and study population

The NOWAC study is a nation-wide population-based cancer study that was initiated in 1991 (Lund et al., 2008), and the postgenome cohort has been described previously in detail (Dumeaux et al., 2008). Briefly, random samples of women were drawn from the Central Person Register by Statistics Norway based on their unique national birth number. Selected women were sent an invitation that included information on blood sample collection and an 8-page questionnaire, on which their national birth number was replaced by a serial number. The linkage file for the national birth number and the serial number was kept at Statistics Norway. The questionnaires were returned to the Department of Community Medicine, University of Tromsø. Non-responders were mailed one or two reminders. Of all invited women, 97.2 % agreed to give a blood sample. These women were sent a blood sampling kit including another 2-page questionnaire and one PAXgene tube (PreAnalytiX GmbH, Hembrechtikon, Switzerland) with a buffer or stabilization agent for mRNA in order to improve the quality of gene expression for genome-wide microarray analyses. These kits were mailed in batches of 500, with one reminder sent after 4–6 weeks. Blood was primarily drawn at family doctors' offices with the doctors then sending the samples as biological material overnight to Tromsø, where they were immediately frozen. Between 2003 and 2006, 48 692 blood samples were included in the NOWAC postgenome biobank, and these women make up the NOWAC postgenome cohort.

A nested case-control design was chosen in order to reduce batch effects in the laboratory and also for the high cost of each analysis. For each case of breast cancer, a control from the same batch of 500 women in the postgenome cohort was assigned, matched by time of blood sampling and year of birth, to be analyzed together with the case.

The controls are used to establish the average (mean) gene expression level in individuals without cancer and to allow exposure-adjusted analyses to be performed. The expression level of a gene not involved in the carcinogenetic process will exhibit variability dependent on day-to-day changes in exposures such as environment and nutrition, resulting in random fluctuations of the difference in gene expression between case and matched control around a population-average constant over time, whereas the difference in expression level of genes related to different stages of the carcinogenetic process may vary over time in a non-random way, thus exhibiting some non-random trend. The changes in genes related to the carcinogenic process could be complicated by other effects of exposures to the carcinogens (Lund & Plancade, 2012).

Follow-up and registry information

Cases of invasive breast cancer diagnosed in the NOWAC postgenome cohort through the end of 2009 were identified through linkage to the Cancer Registry of Norway. Altogether, 637 cases of invasive breast cancer were reported. After removing outliers and ineligible cases including women with distant metastases, the study consisted of 441 case-control pairs. Information on lymph node status at breast cancer diagnosis was based on the pTNM information included in the Cancer Registry of Norway. Detection categories were also obtained from the Cancer Registry of Norway, which updates this data regularly through linkage to the screening database kept by the National Breast Cancer Screening Program (Hofvind, Geller, Vacek, Thoresen, & Skaane, 2007).

Ethical issues

The NOWAC study was approved by the Norwegian Data Inspectorate and the Regional Ethical Committee of North Norway (REK). The linkages of the NOWAC database to national registries such as the Cancer Registry of Norway and registries on death and emigration was approved by the Directorate of Health. The women were informed about these linkages. Furthermore, the collection and storing of human biological material was approved by the REK in accordance with the Norwegian Biobank Act. Women were informed in the letter of introduction that the blood samples would be used for gene expression analyses.

Laboratory procedures

Microarray data

The extraction and microarray services were provided by the Genomics Core Facility, Norwegian University of Science and Technology, Trondheim, Norway. To control for technical variability such as different batches of reagents and kits, day-to-day variations, microarray production batches, and effects related to different laboratory operators, each case-control pair was kept together throughout all extraction, amplification, and hybridization procedures. RNA extraction was performed using the PAXgene Blood miRNA Isolation kit according to the manufacturer's instructions. RNA quality and purity were assessed using the NanoDrop ND 8000 spectrophotometer (ThermoFisher Scientific; Wilmington, Delaware, USA) and Agilent bioanalyzer (Agilent Technologies, Palo Alto, CA, USA), respectively. RNA amplification was performed on 96 plates using 300 ng of total RNA and the Illumina TotalPrep-96 RNA Amplification Kit (Ambio, Inc., Austin, Texas, USA). The amplification procedure consisted of reverse transcription with a T7 promotor and ArrayScript, followed by a second-strand synthesis. In vitro transcription with T7 RNA polymerase using a biotin-NTP mix produced biotinylated cRNA copies of each mRNA in the sample. All case-control pairs were run on either the IlluminaHumanAWG-6 version three expression bead or the HumanHT-12 version 4. Outliers were excluded after visual examination of dendrograms, principal component analysis plots and density plots. Individuals that were considered borderline outliers were excluded if their laboratory quality measures where below given thresholds (RIN value <7, 260/280 ratio <2, 260/230 ratio <1.7, and $50 < \text{RNA} < 500$).

Preprocessing of microarray data

The dataset was preprocessed as previously described (Günther, Holden, & Holden, 2014). The dataset, which consisted of 441 case-control pairs and 30 046 probes, was background-corrected using negative control probes and normalized on the original scale using quantile normalization. Data from the two Illumina chips (HumanWG-6 v3 and HumanHT-12 v4) were combined on identical nucleotide universal identifiers (Du, Kibbe, & Lin, 2007). We retained probes present in at least 1 % of the individuals, i.e., in at least nine of the 882 individuals. If a gene was represented with more than one probe, only one was selected, resulting in a dataset with 11 431 probes. The probes were translated to genes using the IlluminaHumanAll.db database (Carlson, n.d.). Finally, the $\log_2$-differences of the gene expression levels for each case-control pair were computed and used in the statistical analyses. Additional adjustments for possible batch effects were unnecessary as the case-control pairs were kept together throughout the laboratory processes.

RESULTS

Hypothesis tests for development over time in each stratum

A time trend was considered to be present if there were more genes in the curve groups than expected by chance. The number of case-control pairs stratified according to lymph node status and detection category is shown in Table 8.1. First, we stratified all case-control pairs by lymph node status (Tables 8.2 and 8.3). The results were not significant, indicating no changes in gene expression levels over time. We then stratified all screening and interval cancers by lymph node status, which rendered a highly significant global test ($p=0.01$), and more p-values less than 0.05 than expected by chance (Tables 8.2 and 8.3). Finally, we stratified by all detection categories and lymph node status. This analysis showed that the effect was mainly restricted to interval cancers with spread (global test; $p=0.02$). In these tests the inclusion criterion 1 had value $\alpha=0.01$. The results depend on the α-value, but the results were not very sensitive to the choice of α-value (data not shown). Tables 8.4 and 8.5 show the observed number and the expected number of genes in each curve group analysis in Tables 8.2 and 8.3. Here it is important to note that the number of genes in each curve group is not too small (Tables 8.4 and 8.5). If this had been the case, it would have indicated that the chosen α-value - value was too small, weakening the power of the test.

Table 8.2. P-values obtained when testing whether there are more genes in the curve groups than what is expected by chance in different strata

p-value				
Curve group	Screen-detected, interval, and clinical cancers with spread	Screen-detected, interval, and clinical cancers without spread	Screen-detected and interval cancers with spread	Screen-detected and interval cancers without spread
Global	0.78	0.27	0.01	0.20
123	0.61	0.23	0.02	0.39
132	0.49	0.13	0.008	0.11
312	0.88	0.18	0.13	0.11
321	0.41	0.74	0.02	0.66
231	0.74	0.68	0.50	0.57
213	0.58	0.17	0.48	0.13

p-value						
Curve group	Screen-detected cancers with spread	Screen-detected cancers without spread	Interval cancers with spread	Interval cancers without spread	Clinical cancers with spread	Clinical cancers without spread
Global	0.36	0.43	0.02	0.46	0.40	0.81
123	0.10	0.33	0.21	0.89	0.06	0.34
132	0.38	0.19	0.009	0.32	0.51	0.63
312	0.83	0.30	0.07	0.21	0.98	0.81
321	0.18	0.90	0.05	0.40	0.22	0.66
231	0.33	0.63	0.21	0.83	0.94	0.93
213	0.70	0.27	0.29	0.16	0.90	0.59

Inclusion criterion 1 was used with $\alpha = 0.01$. P-values below 0.05 are highlighted in yellow.

Table 8.3. P-values for curve group variables $Z_{c,s,p}$ in the strata "screen-detected and interval cancers with spread" versus "screen-detected and interval cancers without spread"

p-value						
	Genes selected based on stratum s_1 = "Screen-detected and or interval cancers with spread" $Z_{c,s1,p}$			Genes selected based on stratum s_2 = "Screen-detected and interval cancers without spread" $Z_{c,s2,p}$		
Time period *t*	3	2	1	3	2	1
N1	69	20	12	69	20	12
N2	148	57	53	148	57	53
Curve group *c*						
123	0.22	0.59	0.02	0.53	0.11	0.08
132	0.90	0.005	0.004	0.71	0.11	0.009
312	0.80	0.27	0.15	0.04	0.009	0.001
321	0.12	0.98	0.24	0.35	0.72	0.15
231	0.26	0.45	0.78	0.34	0.38	0.23
213	0.53	0.45	0.65	0.36	0.04	0.08

P-values below 0.05 are highlighted in yellow. 'N1' is the number of case-control pairs in the stratum "Screening or interval with spread" in the time period t, while 'N2' is the number of case-control pairs in the stratum "Screening or interval without spread" in the time period *t*.

Table 8.4. Observed number of genes in each curve group with expected number of genes in parenthesis

Observed number of genes (expected number of genes)				
Curve group	Screen-detected, interval, and clinical cancers with spread	Screen-detected, interval, and clinical cancers without spread	Screen-detected and interval cancers with spread	Screen-detected and interval cancers without spread
Global	305 (513)	609 (535)	1360 (482)	708 (547)
123	47 (76)	97 (82)	259 (70)	69 (86)
132	69 (100)	171 (103)	518 (99)	205 (107)
312	37 (102)	145 (105)	171 (105)	203 (108)
321	66 (82)	40 (82)	314 (77)	46 (82)
231	38 (77)	44 (81)	48 (66)	51 (82)
213	48 (76)	112 (82)	50 (65)	134 (83)

Table 8.5. Observed number of genes and expected number in each curve group in the strata "screen-detected and interval cancers with spread" versus "screen-detected and interval cancers without spread"

Observed number of gene (expected number of genes)						
Curve group	Screen-detected cancers with spread	Screen-detected cancers without spread	Interval cancers with spread	Interval cancers without spread	Clinical cancers with spread	Clinical cancers without spread
Global	475 (464)	490 (547)	1233 (485)	471 (525)	448 (491)	302 (502)
123	139 (75)	78 (85)	101 (81)	33 (90)	233 (84)	83 (83)
132	81 (91)	141 (106)	515 (92)	96 (97)	52 (84)	54 (90)
312	43 (96)	107 (109)	237 (89)	123 (96)	18 (82)	40 (92)
321	115 (82)	29 (82)	213 (81)	71 (83)	101 (83)	45 (77)
231	63 (63)	46 (82)	92 (70)	31 (78)	21 (77)	27 (77)
213	34 (58)	89 (83)	75 (73)	117 (81)	23 (81)	53 (83)

Cases with a p-value below 0.05 is highlighted in yellow.

Hypothesis tests for comparing two strata

Based on the results from each stratum, we restricted our analysis to comparing gene expression levels in the strata "screening or interval with spread" and "screening or interval without spread" using the curve group variable $Z_{c,s,p}$ described in the methods section. P-values were obtained by testing whether the curve group variables $Z_{c,s,p}$

were different in the two strata; many were below 0.05 and some were smaller than 0.01 (Table 8.5). In Figure 8.3, we illustrated how to use the gene expression data to separate these two strata by showing the curve group variable $Z_{c,s,p}$ for each case-control pair p in the different strata. The plot shows that the difference between the two strata changes over time for the two most significant $Z_{c,s,p}$ variables. The differences between the strata with spread and without spread were larger in the year before diagnosis compared to earlier years, but even these differences were comparatively small.

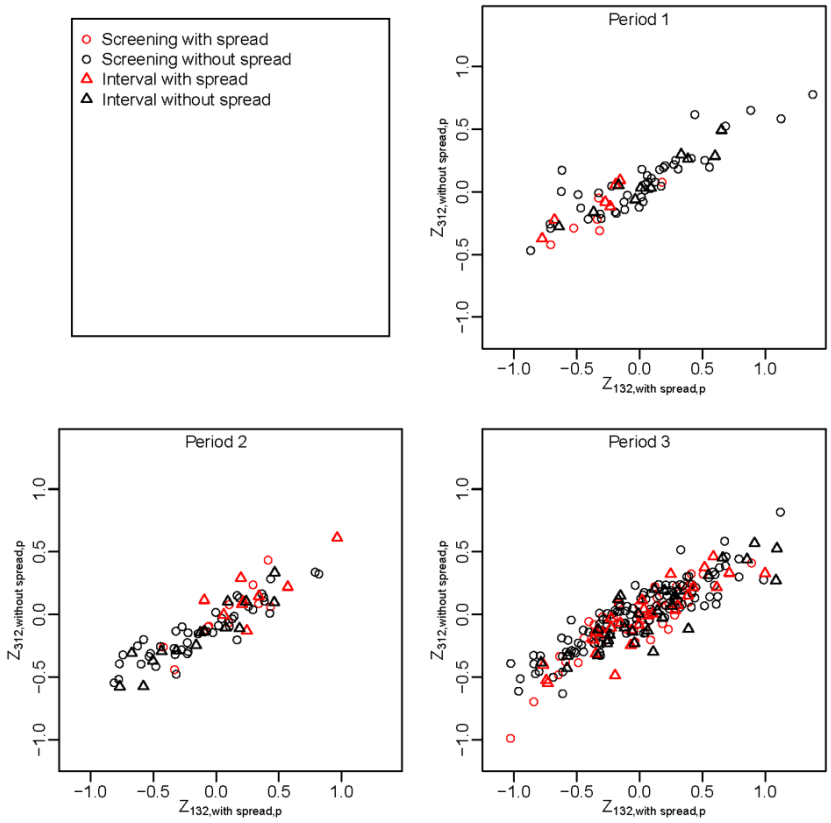


Figure 8.3. Distribution of case-control pairs for two curve groups stratified on spread in each time period. Plot of two of the most significant curve group variables, $Z_{132,\text{with spread},p}$ and $Z_{312,\text{without spread},p}$, for the three time periods. These variables are the sum of gene expression differences $X_{g,p}$ for genes selected from curve group 132 (high values in middle period) based on data with spread and curve group 312 (low values in middle period) based on data without spread. The data with spread (without spread) are first used to select two sets of genes, one set for each of the two curve-group variables. We may calculate both $Z_{132,\text{with spread},p}$ and $Z_{312,\text{without spread},p}$ for all case-control pairs from all strata. Note that the difference between the two strata varies between the periods.

In the methods section we introduced the statistic L_k, a weighted sum of t-statistics, as an alternative to the curve group variables $Z_{c,s,p}$ for comparing the gene expression levels of two strata. In Figure 8.4 we plot the p-value in a hypothesis test with L_k as test statistic against the number of genes k. The plot shows that the gene expression levels are different in the two strata. L_k is the sum of the k-largest weighted sums of t-statistics. Note (in Figure 8.4) that when we add more and more terms in the sum, the observation becomes more significant. When we used 50 genes, the p-value was about 0.05, and the p-value decreased to below 0.02 when we used the 1000 most significant genes. This indicates that the difference between the strata is present in a large number of genes, but so weak that the strongest result was only obtained when including a large number of genes. Also, time period 1, i.e. 0-1 year before diagnosis, contributed the most to the low p-values, which is in accordance with the results shown in Figure 8.3 and Table 8.5.

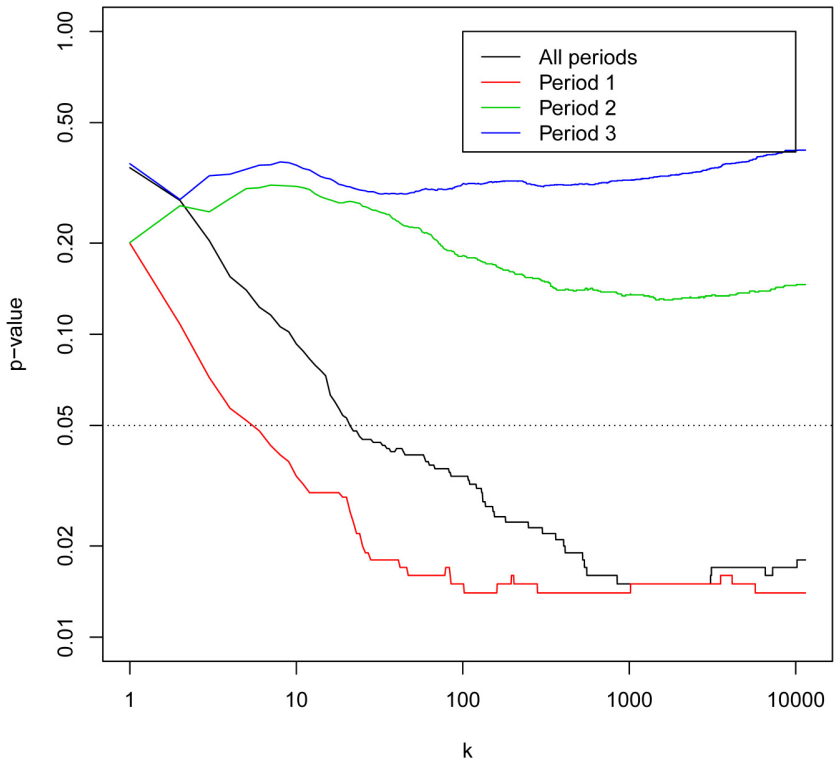


Figure 8.4. The relationship of p-values to number of genes in the test statistic L_k.
The p-value in a hypothesis test with test statistic L_k, a weighted sum of t-statistics, plotted against the number of genes k used in the calculation of L_k. The two strata that are compared in the t-statistics that are used for computing L_k are "Screening or interval with spread" and "Screening or interval without spread".

DISCUSSION

This methodological analysis has shown that it is possible to significantly discriminate the time trend of gene expression patterns observed before breast cancer diagnosis. The findings are based on an original approach for the statistical analysis of time-dependent curves of gene expression levels in the NOWAC postgenome cohort. These methods could also be used for other aspects of functional genomics such as methylation.

From a statistical point of view, since the publication of the seminal work by Cox (1972), the Cox proportional hazard model and its extension have been largely used by epidemiologists to analyze cohort studies with time-dependent covariates. This model has also been adapted to case-control designs (Aalen, Borgan, & Gjessing, 2008) and some extensions have been proposed for covariates measured with noise (Hu, Tsiatis, & Davidian, 1998) and time-dependent coefficients (O'Quigley, 2008). More recently, the adjunction of numerous covariates such as gene expression data has added some challenging statistical issues (Benner, Zucknick, Hielscher, Ittrich, & Mansmann, 2010). While the characteristics and the basic assumptions of the Cox model have been adapted to the dimensionality and the very specific paired design of the NOWAC postgenome cohort, the Cox model cannot be fully adapted to the estimation of changes in gene expression curves or to the biological interpretations of gene pathways.

The curve group approach can be viewed as an effective method for dimension reduction in studies of functional genomics. The grouping of the curves is not dependent on the individual testing of the curves for the more than 10 000 expressed genes; thus it mostly eliminates the false discovery rate of multiple testing. The strength of the curve group approach can be seen in the statistical power that was achieved even in strata with a low number of cases, such as the six cases with spread in two strata. We stratified the data based on the detection category and lymph node status. The Norwegian Breast Cancer Screening Program uses mammographic screening and started in 1996, with coverage of the entire population starting in 2005 (Hofvind et al., 2007). It has been estimated that the introduction of population-based breast cancer screening in Norway gave a mean sojourn time for invasive cancer of 4.0 years in women aged 50–59 years and 6.6 years for those 60–69 years (Weedon-Fekjaer, Lindquist, Vatten, Aalen, & Trettli, 2008). Analyses of breast carcinogenesis as a time-dependent process should therefore take into consideration that cases diagnosed within the screening program are diagnosed at an earlier phase of carcinogenesis and thus are not directly comparable to clinically detected cases. Lymph node status has been the most important prognostic factor for breast cancer survival for 100 years (Todd, Shoag, & Cadman,

1983; Cancer Registry of Norway, 1975). At time of diagnosis, we had a censored distribution of tumors where detection category determined the time of diagnosis, irrespective of the underlying carcinogenic process.

The prospective analyses of gene expression levels in the years preceding breast cancer diagnosis as assessed by the log-fold change between cases and controls showed significant differences in the curve groups after stratification by lymph node status and detection category. The analyses showed the ability to discriminate between different stages of the carcinogenic process. A previous analysis of a case-control study within NOWAC showed that differences in gene expression mainly reflect immune responses, but also genes related to cell control (Dumeaux et al., 2015). The analyses of trajectories could aid in understanding the time-dependent interaction between the immune response and carcinogenesis. Our findings should be further interpreted in relation to the biology of both single genes and gene pathways.

An agnostic search for time trends depends on a sensitive statistical approach. We have presented two novel statistical methods that demonstrated that the gene expression levels varied over time in the last years prior to breast cancer diagnosis, and that the development over time differed by lymph node status among women who attended the National Breast Cancer Screening Program in Norway (i.e., those with screen-detected or interval cancers). One of the methods focused on identifying genes with specific changes over time within a given lymph node status. The other method focused on differences in gene expression levels between lymph node statuses in the different time periods. Both methods focused on different aspects of functional time dependency of gene expression levels relative to time of breast cancer diagnosis, and both methods gave significant results when many genes were used. As gene expression data are very noisy, our methods used information from several genes simultaneously to increase the power of the hypothesis tests.

A potential weakness of the curve group approach is the increasing number of curve groups as observation time periods increases. When there are four time periods, 24 curve groups will be needed, and even more will be needed for five time periods.

Studies of gene expression levels in peripheral blood are challenging and have many difficulties and pitfalls. Most biobanks suffer from ubiquitous degradation by RNase, which reduces the quality of mRNA for whole genome analyses. Only samples that contain a specific buffer or are directly frozen in liquid nitrogen can be used for whole genome analyses. The signals related to carcinogenesis in the blood are expected to be much weaker than those in tumor tissue and can be con-

founded by signals from exposures to carcinogens or other lifestyle factors. The problem of noise due to the complicated study of carcinogenesis, the need for an adequate epidemiological design including exposure information and blood sampling, complicated technology, and the development of robust statistics, could make the approach unsuccessful. The prospective design of our study made it difficult to increase the statistical power, so our results should be interpreted with care.

To the best of our knowledge, the NOWAC postgenome cohort is the largest population-based prospective cancer study designed for transcriptomics due to the presence of buffered RNA. All parts of the analyses were done within the framework of the NOWAC study. In the NOWAC postgenome cohort, a single laboratory processed all samples using the same technology, thus reducing analytical bias and batch effects. The cohort design reduced selection bias. A weakness of a prospective study could be possible changes in case-control status as controls became cases over time, thus reducing the differences in gene expression levels within a case-control pair. We removed all case-control pairs in which controls were diagnosed with breast cancer or any other cancer within 2 years of blood sampling. The matching was done only for storage time and year of birth. Matching on other variables will eliminate the inclusion of these lifestyle factors in the analyses. If matched on e.g. smoking, we could not estimate the effect of smoking or any interactions with other risk factors. Unfortunately, there was no repeated sampling of blood, and no additional questionnaires were completed. Repeated measurements would secure better analyses, making it possible to use intra-individual comparisons over time.

CONCLUSIONS

The proposed statistical methods are sensitive for finding curve groups of genes, even for strata with few case-control pairs. This made it possible to describe and test non-linear relationships. Our findings could be viewed as a proof of concept of systems epidemiology, indicating the potential to include gene expression for functional analysis in prospective studies of cancer.

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.

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.

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.

AUTHORS' CONTRIBUTIONS

EL is PI of the NOWAC Study and initiated the methodological collaboration; JCT, HMB, SP, GN, and NM participated in the methodological developments; LH, MH, and C-CG developed the statistical methods. All authors have read and approved the final manuscript.

REFERENCES

Aalen, O.O., Borgan, Ø., & Gjessing, H.K. (2008). *Survival and event history analysis. A process point of view*. New York: Springer.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., ... Stratton, M.R. (2013). Signatures of mutational processes in human cancer. *Nature*, 500(7463), 415–421. doi: https://doi.org/10.1038/nature12477

Benner, A., Zucknick, M., Hielscher, T., Ittrich, C., & Mansmann, U. (2010). High-dimensional cox models: The choice of penalty as part of the model building process. *Biometrical Journal*, 52(1), 50–69. doi: https://doi.org/10.1002/bimj.200900064

Cancer Registry of Norway. (1975). *Survival of cancer patients: cases diagnosed in Norway 1953–1967*. Oslo, Norway: The Norwegian Cancer Society.

Carlson, M. (n.d). *lumiHumanAll.db: Illumina Human Illumina expression annotation data (chip lumiHumanAll)*. R Package version 1.22.0.[Database]

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

Du, P., Kibbe, W., & Lin, S. (2007). nuID: a universal naming scheme of oligonucleotides for Illumina, affymetrix, and other microarrays. *Biology Direct*, 2, 16. doi: https://doi.org/10.1186/1745-6150-2-16

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

Dumeaux, V., Ursini-Siegel, J., Flatberg, A., Fjosne, H.E., Frantzen, J.O., Holmen, M.M., ... Lund, E. (2015). Peripheral blood cells inform on the presence of breast cancer: a population-based case-control study. *International Journal of Cancer*, 136(3), 656–667. doi: https://doi.org/10.1002/ijc.29030

Günther, C., Holden, M., & Holden, L. (2014). Preprocessing of gene-expression data related to breast cancer diagnosis. *NR Note*, SAMBA/35/14, 7–41. Retrieved from https://www.nr.no/files/samba/smbi/note2015SAMBA3514preprocessing.pdf

Hofvind, S., Geller, B., Vacek, P.M., Thoresen, S., & Skaane, P. (2007). Using the European guidelines to evaluate the Norwegian breast cancer screening program. *European Journal of Epidemiology*, 22(7), 447–455. doi: https://doi.org/10.1007/s10654-007-9137-y

Holden, L. (2015). Classify strata. *NR Note*, SAMBA/11/15, 7–28. Retrieved from https://www.nr.no/files/samba/smbi/note2015SAMBA1115classifyStrata.pdf

Hu, P., Tsiatis, A.A., & Davidian, M. (1998). Estimating the parameters in the cox model when covariate variables are measured with error. *Biometrics*, 54(4), 1407–1419.

Lund, E., & Dumeaux, V. (2008). Systems epidemiology in cancer. *Cancer Epidemiology Biomarkers & Prevention*, 17(11), 2954–2957. doi: https://doi.org/10.1158/1055-9965

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., Plancade, S., Nuel, G., Bøvelstad, H., & Thalabard, J.C. (2015). A processual model for functional analyses of carcinogenesis in the prospective cohort design. *Medical Hypothese,* 85(4), 494–497. doi: https://doi.org/10.1016/j.mehy.2015.07.006

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

O'Quigley, J. (2008). *Proportional Hazards Regression*. New York: Springer.

Phipson, B. & Smyth, G.K. (2010). Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9, Article 39. doi: https://doi.org/10.2202/1544-6115.1585

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

Todd, M., Shoag, M., & Cadman, E. (1983). Survival of women with metastatic breast cancer at Yale from 1920 to 1980. *Journal of Clinical Oncology*, 1(6), 406–408.

Weedon-Fekjær, H., Lindqvist, B.H., Vatten, L.J., Aalen, O.O., & Tretli. S. (2008). Estimating mean sojourn time and screening sensitivity using questionnaire data on time since previous screening. *Journal of Medical Screening*, 15(2), 83–90. doi: https://doi.org/10.1258/jms.2008.007071

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 $s(c)$.

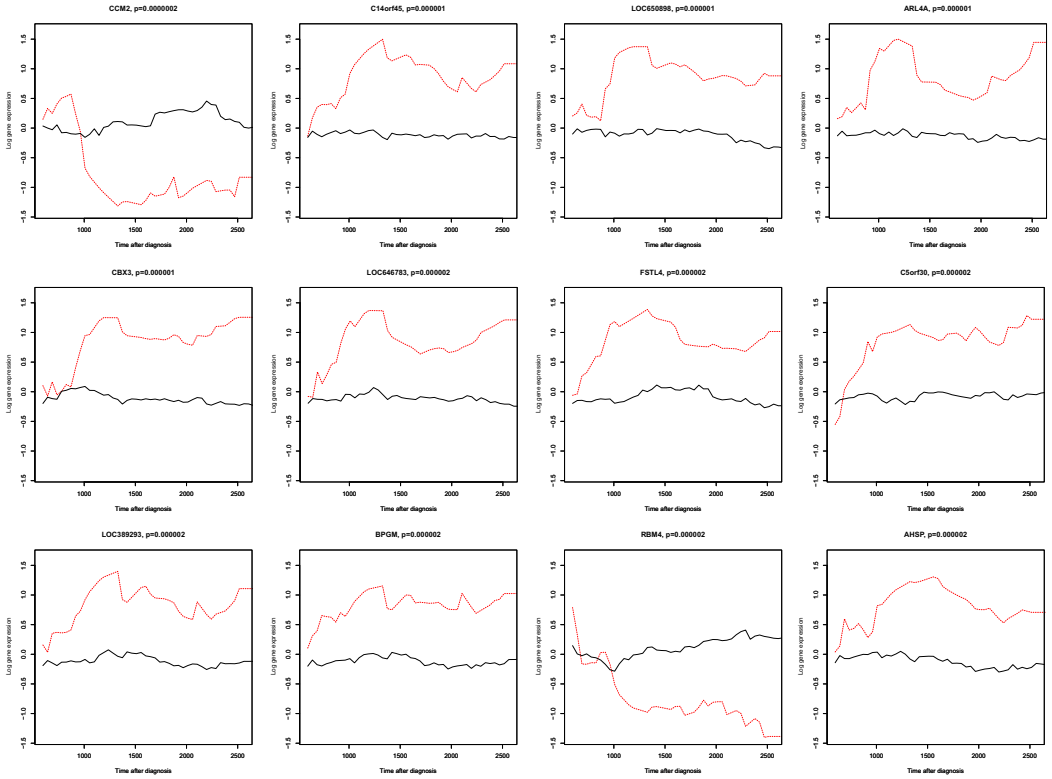


Figure 9.1. Log2 gene expression from the 12 case-control curves with the smallest p-value of the 8400 genes. The 12 p-values less than 0.00001. The figure uses normalized data as is used in the test statistics. The black continuous and the red dashed curves are the log2 gene expression from the case-controls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

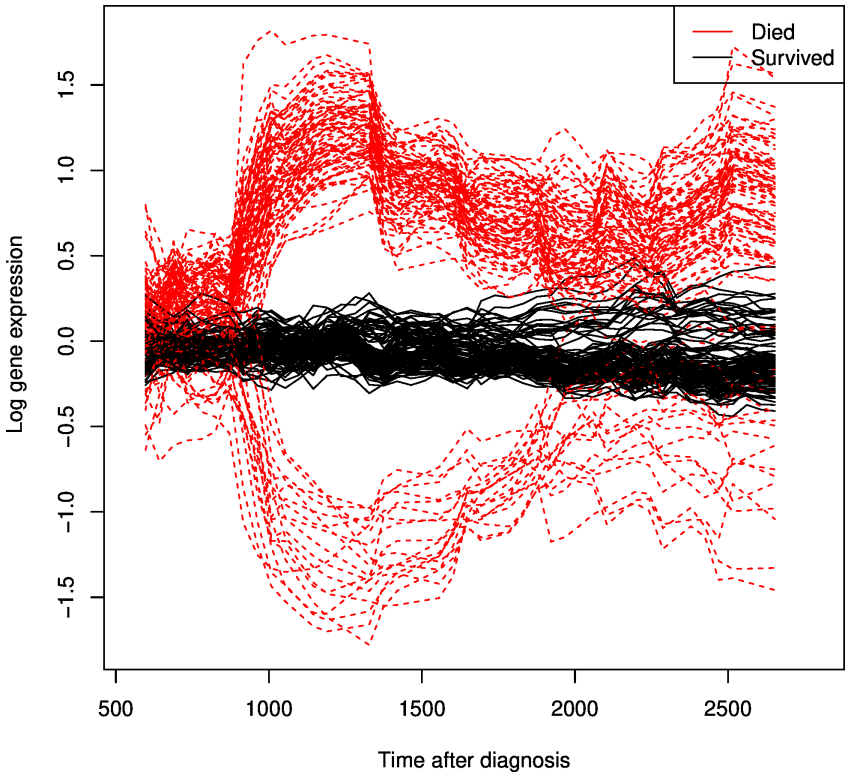


Figure 9.2. Log2 gene expression from the case-controls curves for the 12 genes with smallest p-value of the 8400 genes. The black continuous and the red dashed curves are the log2 gene expression from the case-controls who survived and died, respectively.

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.

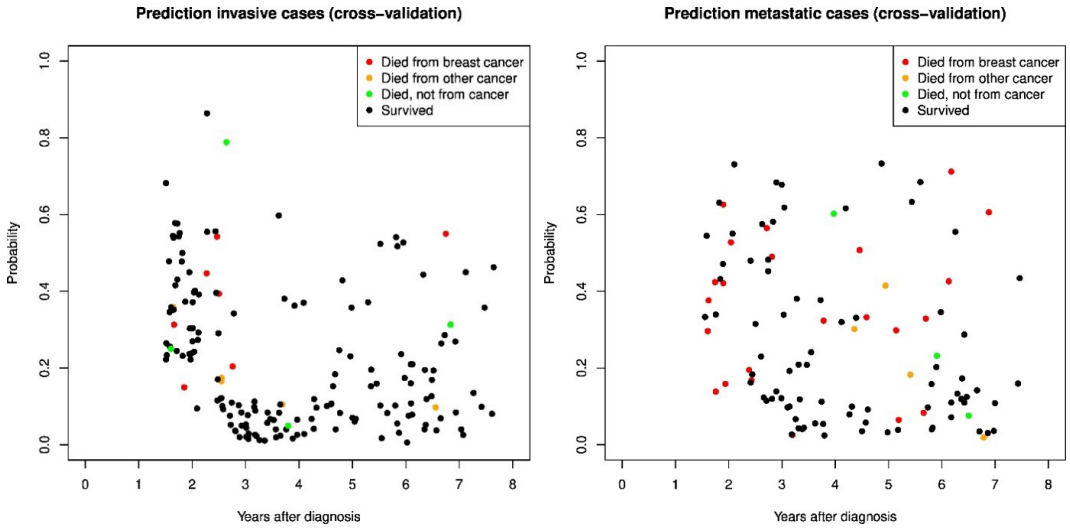


Figure 9.3. The prediction of dying from cancer for the cases who died of breast cancer, died of other types of cancers and died, but not from cancer, and cases who survived for each of the 278 case-control pairs. The figure to the left shows predictions for cases with invasive cancer, while the figure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

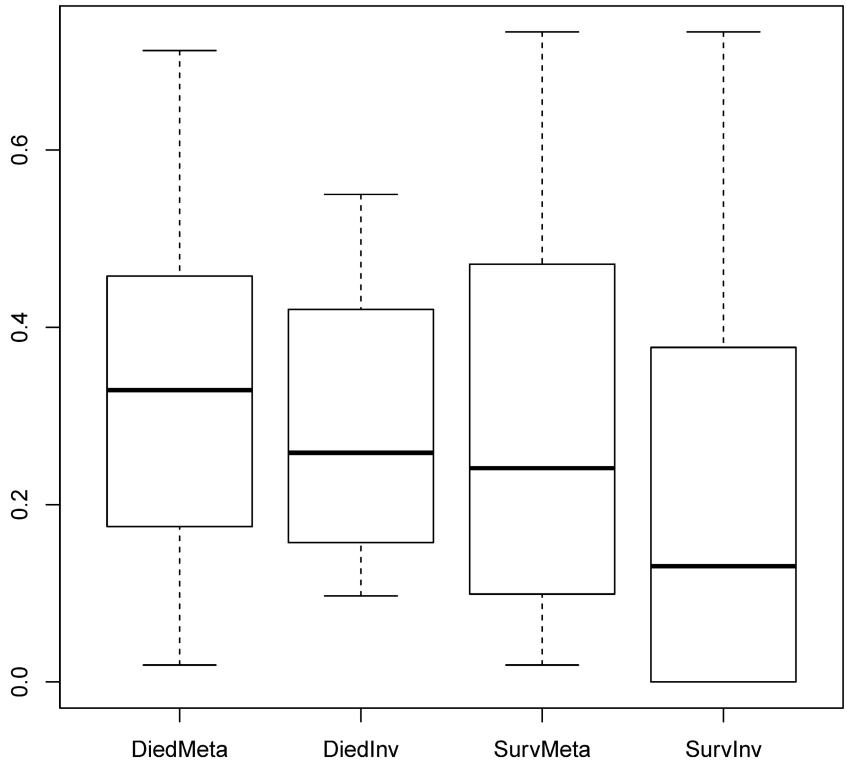


Figure 9.4. Boxplot of prediction of death from cross-validation of cases after 18 months from diagnosis. Horizontal lines describe 0.25, 0.5 and 0.75 quantiles. The number of cases and mean in the four categories are metastatic cancer where case died, no: 27, mean 0.33, invasive cancer where case died,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

Lund, E. (Ed.) (2020). *Advancing Systems Epidemiology in Cancer. Exploring Trajectories of Gene Expression.* Oslo: Scandinavian University Press.
DOI: https://doi.org/10.18261/9788215041193-2020-10

10. Hypotheses of Carcinogenesis—The Atavistic Theory

Lill-Tove Rasmussen Busund

Abstract A deep comprehension of what cancer is as a biological phenomenon is lacking. Several theories have been proposed and many of them do not necessarily contradict each other. One of the theories is the intriguing hypothesis that a cancer cell may be triggered by mutations, but is basically a self-activated throwback to an ancestral cell phenotype running its ancient core functionality by preserving its vital functions, such as survival and uncontrolled proliferation.

Keywords ancestral cell phenotype | core functionality | atavism

After decades of extensive research, our knowledge of the carcinogenic process has grown. Cancer is currently widely regarded as random oncogenic mutations accumulating in cells, leading to an evolutionary process of emerging hallmarks (Hanahan and Weinberg 2011) that are reminiscent of unicellular organisms. However, a deeper comprehension of what cancer is as a biological phenomenon is still lacking. Several theories have been proposed and many of them do not necessarily contradict each other.

One of the theories is the intriguing hypothesis that a cancer cell is a throwback to an ancestral cell phenotype. That essential idea was proposed in 1914 by Theodor Boveri, who characterized the malignant tumor cell as a previously normal and "altruistic" tissue cell changed into an "egoistical" mode with loss of functions. The latter cell had lost normal reactivity to the rest of the body by releasing its multiplication from restraint and tending toward primitive, unicellular properties (Boveri 2008). In recent decades, essential elements of Boveri's idea have been resurrected in an atavistic hypothesis of cancer which regard cancer as an ancient and systematic program of emergency survival procedures that preserves the two most vital core functions—survival and proliferation—in response to a damaging environment. Since cell survival and proliferation is deeply integrated in normal cell

physiology, upregulated genes coding for these normal traits are more likely to escape rather than alarm the immune system's surveillance, which is one of the hallmarks of carcinogenesis. The hypothesis further suggests that carcinogenesis may be triggered by mutations, but its basic cause is a self-activation of a very old, programmed and deeply embedded toolkit of emergency survival programs (Davies and Lineweaver 2011). Evidence supporting this hypothesis has long remained non-observational, until recently.

More than one hundred years ago, William Coley noticed that some cancer patients showed spontaneous remission following severe infection (Hoption Cann et al. 2003). The traditional explanation was that infection boosted the reactions of the immune system, which then destroyed the cancer by chance. The atavistic hypothesis, on the other hand, proposes that at least part of the reason for Coley's results is that cancer tumors are more vulnerable to infection than the rest of the body because, via their throwback to the ancestral phenotype, they have decoupled from the adaptive immune system. In other words, the tumor has regressed to an immunocompromised state and is thereby left unprotected against infection.

The role of immune system cells in carcinogenesis has been studied for decades, and the particular importance of the adaptive immune system has been unraveled. The basic idea behind immunotherapy, which has received increasing attention as a new way to combat cancer, is to boost the body's immune system—both the innate and the adaptive—through a diversity of sophisticated strategies in order to improve the immune response against cancer (Dempke et al. 2017). The early results are encouraging; combinations of immunotherapy, novel targeted therapies, and conventional chemotherapy might be especially promising. Nevertheless, additional basic, translational and clinical studies with long follow-up time are crucial to unraveling immunotherapy as a breakthrough in cancer treatment.

Another discovery made by Otto Warburg a century ago showed that cancer cells often switch to fermentation, especially when the oxygen tension is low and the glucose concentration is high (Otto 2016). Fermentation processes create less energy but more biomass compared to normal human cells, which use oxygen to generate energy. The atavistic hypothesis seems to suggest that the ancestral cancer cells have proliferative advantages in low oxygen surroundings, since they are reversed to the evolutionary time more than one billion years ago when the first multicellular organisms evolved in a far less oxygen-rich atmosphere. Based on the Warburg effect, hyperbaric oxygen combined with a low-glucose have been studied *in vitro* and in patient trials. The mechanisms are complex and combinations of graded oxygen-glucose concentrations with novel therapeutics need to be studied in further detail.

The majority of cancers follow a predictable pattern of clinical development: a tumor grows in an organ, and if it is not cured, some of the cancer cells leave the primary tumor, spread via lymph or blood and invade remote organs, where they create metastases. Metastatic cancers are often beyond being cured and the vast majority of patients who die from cancer die from their metastases, not from their primary tumors. Cancer development goes through several distinctive functional hallmarks, including survival of the neoplastic cells, uncontrolled proliferation, increased motility, evasion of the immune system, and establishment of its own blood supply (Hanahan and Weinberg 2011). All these traits improve the survival and sustainability of the cancer cells over a relatively short period of time. The somatic mutation theory has challenges in explaining how random mutations accumulating in cells over time can confer so many improved functions in a single tumor. It also seems paradoxical that increasingly damaged genomes are able to code for proteins, resulting in gained functions in such a systematic and predictable behavior, acquiring the various hallmarks. Further, the non-neoplastic cells in a tumor—i.e. the non-mutated cells of the microenvironment—have shown tremendous impact on the overall survival of cancer patients. The predictable way that cancer progresses through its various stages of malignancy, both clinical and pathophysiological, indicates that cancer is not a case of randomly, damaged cells but a primitive cellular defense mechanism consisting of a systematic, programmed strategy as a response to environmental challenges.

Paul Davies, director of the Beyond Center for Fundamental Concepts in Science at Arizona State University, USA, describes the atavistic theory of cancer as

> a default state in which a cell under threat runs on its ancient core functionality, thereby preserving its vital functions, of which survival and uncontrolled proliferation is the most ancient, most vital and best preserved (Davies and Lineweaver 2011).

They have brought novelty to the atavistic theory through phylostratigraphic studies of the ages of genes by comparing how gene sequences diverge across many species, thereby enabling them to trace the evolutionary origin of genes involved in carcinogenesis (Bussey et al. 2017). In the general phylogenetic tree, the most widespread properties of the different organisms are usually the oldest, and can often be traced back to a common ancestor in the distant past. The atavistic theory hypothesizes that the oncogenes are clustered around the age of onset of multicellular organisms. The earliest traces of unicellular life can be dated back about 3.5 billion years and the onset of multicellularity first emerged about 1.5 billion years

ago. Through using gene databases, they found that evolutionary roots of cancer can be traced back to the early transitional phase from unicellularity to multicellularity, about 600 million years ago, before complex metazoans emerged. An estimation of the evolutionary ages of the genes in the human genome showed that genes younger than about 500 million years were more likely to be mutated in cancer, while genes older than a billion years tended to have fewer mutations than average. This is in accordance with the atavistic theory's prediction that older genes are likely to be less mutated than younger genes in cancer, since the oldest genes are expected to be responsible for the ancient core functions of the aggressive cancer cell. A comprehensive study has recently characterized cancer driver genes and mutations from several thousand tumor exomes (Bailey et al. 2018); another study shows that older genes are expressed at higher levels when cancer progresses to a more aggressive and advanced stage (Trigos 2019).

By investigating the functionality of the oncogenes, they have shown that genes older than 950 million years are strongly enriched for two core functions: control of the cell cycle, and repair of DNA double-strand breaks (Cisneros et al. 2017). The evolutionary history of these genes revealed that cancer genes implicated in DNA repair match up with mutated genes in stressed bacteria employed for a critical survival function. These ancient and essentially identical genes, discovered in the DNA of bacteria and cancer, are known to be associated with poor patient prognosis. Elevated mutation rates in neoplastic cells are among the main reasons why chemotherapy falters when neoplasms evolve drug-resistant variants.

Other phylostratigraphic studies suggests a link between cancer genes and the emergence of multicellular life. By analyzing the expression profile of xenograft tumors at different stages and various tumor samples, Chen et al. demonstrated an evolving convergence from multicellular state towards unicellular state in cancer expression profile and functional status (Chen et al. 2018) Although together these evidences demonstrated a general trend toward atavism in carcinogenesis, there are still large elements of the atavistic hypothesis that remain unanswered. In particular, the answer as to whether cellular atavism from multicellularity to unicellularity is the cause or the result of carcinogenesis remains elusive.

GENE EXPRESSION INFORMATION FROM THE NOWAC STUDY

Mutations induce disturbed gene regulatory networks at a very early stage in the carcinogenic process, leading to changes in the flow of signals between genes and between cells. The Norwegian Women and Cancer Study, NOWAC (Lund et al.

2008) is trying to identify early carcinogenic signatures of these gene network changes in blood. We aim to identify distinct gene signature hallmarks of cancer in white blood cells that may precede the clinically noticeable changes in tumor cells and tissues, thus providing an early warning of upcoming cancer.

In Chapter 8 we identified a gene profile in white blood cells in those women who died from their cancer within a few years. Some of these signals are upregulated, pre-vertebrate genes maintaining core functions such as survival, maintenance of cytoskeleton, proliferation, cilia orientation, nuclear membrane proteins, DNA damage repair, and oxygen utilization (Table 9.1). The intriguing finding—that the signals from the white blood cells disappeared after a short time—might be a result of decoupling and thereby a default surveillance of the aggressive, ancestral phenotype of the cancer cells by the phylogenetically more novel adaptive immune system. At the time of diagnosis, another study showed no strong correlation between genes expressed in cancer tissue compared to white blood cells, except for some highly immunogenic breast cancer subgroups (Dumeaux et al. 2017).

These findings are essentially in accordance with the atavistic hypothesis showing that white blood cells' responses to cells under threat run their core functionalities, preserving its most vital functions. These data are exclusively from GeneCards®. They should be further analyzed in comprehensive, inventory databases and applied by gene classification tools organizing genes around their function.

Table 10.1. Name of the ten most significantly upregulated genes in breast cancer during follow-up with their function taken from table 9.2 (Chapter 9). Function from the human gene database GeneCards®

Gene	Function
Significant up-regulated genes	*Mainly basic cell functions like cytoskeleton, cilia orientation, DNA damage, oxygen utilization*
CCM2	Cerebral Cavernous Malformations 2 is required for normal cytoskeletal structure, cell-cell interactions, and lumen formation in endothelial cells.
C14orf45	May act by mediating a maturation step that stabilizes and aligns cilia orientation.
ARL4A	ADP Ribosylation Factor Like GTPase 4A is related to *Mesodermal Commitment Pathway*. Gene Ontology annotations related to this gene include GTP binding and GTPase activity. An important paralog of this gene is *ARL4C*.
LOC650898	Involved in inducing the expression of cellular antiviral genes, including the interferon-β gene, in response to Pattern Recognition Receptors which modulate the strength and duration of the innate immune responses.

Gene	Function
CBX3	This protein is recruited to sites of DNA damage and double-strand breaks. This protein can bind lamin B receptor, an integral membrane protein found in the inner nuclear membrane.
C5orf30	May play a role in cilia membrane localization via its interaction with UNC119B and protein transport into photoreceptor cells.
FSTL4	Follistatin Like 4 shows Gene Ontology annotations related to calcium ion binding. Prognostic marker in pancreatic cancer. An important paralog of this gene is *FSTL5*.
BPGM	The protein 2,3-diphosphoglycerate (2,3-DPG) is a small molecule found at high concentrations in red blood cells, where it binds to and decreases the oxygen affinity of hemoglobin. Deficiency of this enzyme increases the affinity of cells for oxygen.

REFERENCES

Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018 Apr 5; 173(2): 371–385. Available from: https://www.cell.com/cell/fulltext/S0092-8674(18)30237-X?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS009286741830237X%3Fshowall%3Dtrue

Boveri T. Concerning the Origin of Malignant Tumours by Theodor Boveri. Translated and annotated by Henry Harris. J Cell Sci. 2008 Jan; 121 Suppl 1: 1–84. Available from: https://jcs.biologists.org/content/121/Supplement_1/1

Bussey KJ, Cisneros LH, Lineweaver CH, Davies PCW. Ancestral gene regulatory networks drive cancer. Proc Natl Acad Sci U S A. 2017 Jun 13; 114(24): 6160–6162. Available from: https://www.pnas.org/content/114/24/6160.long

Chen W, Li Y, Wang Z. Evolution of oncogenic signatures of mutation hotspots in tyrosine kinases supports the atavistic hypothesis of cancer. Sci Rep. 2018 May 29; 8(1): 8256. Available from: https://www.nature.com/articles/s41598-018-26653-5

Cisneros L, Bussey KJ, Orr AJ, Miočević M, Lineweaver CH, Davies P. Ancient genes establish stress-induced mutation as a hallmark of cancer. PLoS One. 2017 Apr 25; 12(4): e0176258. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0176258

Davies PC, Lineweaver CH. Cancer tumors as Metazoa 1.0: tapping genes of ancient ancestors. Phys Biol. 2011 Feb; 8(1): 015001. Available from: https://iopscience.iop.org/article/10.1088/1478-3975/8/1/015001

Dempke WCM, Fenchel K, Uciechowski P, Dale SP. Second- and third-generation drugs for immuno-oncology treatment-The more the better? Eur J Cancer. 2017 Mar; 74: 55–72. Available from: https://www.ejcancer.com/article/S0959-8049(17)30024-2/fulltext

Dumeaux V, Fjukstad B, Fjosne HE, Frantzen JO, Holmen MM, Rodegerdts E et al. Interactions between the tumor and the blood systemic response of breast cancer patients. PLoS Comput Biol. 2017 Sep 28; 13(9): e1005680. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005680

Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011 Mar 4; 144(5): 646–674. Available from: https://www.sciencedirect.com/science/article/pii/S0092867411001279?via%3Dihub

Hoption Cann SA, van Netten JP, van Netten C. Dr William Coley and tumour regression: a place in history or in the future. Postgrad Med J. 2003 Dec; 79(938): 672–680. Available from: https://pmj.bmj.com/content/79/938/672.long

Lund E, Dumeaux V, Braaten T, Hjartaker A, Engeset D, Skeie G et al. Cohort profile: The Norwegian Women and Cancer Study--NOWAC--Kvinner og kreft. Int J Epidemiol. 2008 Feb; 37(1): 36–41. Available from: https://academic.oup.com/ije/article/37/1/36/763947

Lund E, Holden M, Olsen KS, Thalabard JC, Busund LT, Holden L. Global gene expression levels in blood following a breast cancer diagnosis – clinical follow-up in the NOWAC postgenome cohort. Submitted 1.

Lund E, Holden M, Thalabard JC, Busund LT, Snapkov I, Holden L. Post-diagnostic single gene expression trajectories as signals of death – breast cancer as a proof of concept – the NOWAC postgenome cohort. Submitted 2.

Otto AM. Warburg effect(s) – a biographical sketch of Otto Warburg and his impacts on tumor metabolism. Cancer Metab. 2016 Mar 8; 4: 5. https://cancerandmetabolism.biomedcentral.com/articles/10.1186/s40170-016-0145-9

Trigos AS, Pearson RB, Papenfuss AT, Goode DL. Somatic mutations in early metazoan genes disrupt regulatory links between unicellular and multicellular genes in cancer. eLife. 2019 Feb 26; 8: e40947. Available from: https://elifesciences.org/articles/40947

Lund, E. (Ed.) (2020). *Advancing Systems Epidemiology in Cancer. Exploring Trajectories of Gene Expression.* Oslo: Scandinavian University Press.
DOI: https://doi.org/10.18261/9788215041193-2020-11

11. The Immuno-Carcinogen Theory of Cancer—The Lifelong Dynamic Interface Between the Immune System and the Carcinogen-Driven Carcinogenesis

Eiliv Lund

Abstract Over the decades, many theories or models of carcinogenesis have been proposed. Based on the systems epidemiology research on gene expression from immune cells in peripheral blood, the concept of the dynamic interface between the immune system and the carcinogen driven carcinogenesis is put forward. This combines traditional exposure research in cancer epidemiology with upcoming knowledge of the immunological response to cancer, from clones of cancer cells to clones of immune cells.

Keywords Immune system | risk factor | carcinogenesis | tumor tissue | blood | clones

This novel theory of carcinogenesis introduces the concept of *a lifelong dynamic interface* between the immune system and carcinogen-driven carcinogenesis. Through the analyses of trajectories of gene expression in peripheral blood from immune cells, the introduction of time-dependent changes in functional genomics has documented the responses of the immune system to the carcinogenic process. This dynamic interface could be looked upon as a *balance or war* between clones of immune cells and tumor cells. In a systems epidemiology design, the two forces can be weighed against each other: the traditional carcinogen-driven model against the immune defense system. The metaphor of war follows from lay state-

ments about cancer: the "war on cancer", "la lutte" (in French) or "sin livs kamp" (Norwegian). In popular speeches, researchers call TD8+ cells "killer cells".

The aim is to discuss the observational background for the theory, its relation to other models, the need for scientific collaborations between different disciplines, and finally, new challenges.

SYSTEMS EPIDEMIOLOGY STUDIES OF GENE EXPRESSION FROM PERIPHERAL IMMUNE CELLS IN BLOOD

We have demonstrated the potential for studies of trajectories in Chapters 8 and 9, The trajectories before (Lund et al. 2016, see Chapter 8, Holden et al. 2017) and after diagnosis (Chapter 9) demonstrate the time-dependent difference in gene expression from immune cells *dependent on stage*. Important for the interpretation of the interface is the finding of *lack of correlation* between gene expression in blood and in tumor tissue in the same individual at time of diagnosis; see Dumeaux et al. 2017. Blood is not a surrogate for tissue studies. The findings tell us that there is a dynamic interface between the immune system and the effect of carcinogens on the tissue cells changing over time. While invasive cancer shows few changes, in metastatic cancer relatively rapid changes are seen around and before diagnosis. More dramatic are the changes in gene expression after diagnosis in metastatic cases, with a second, transient increase. This effect can be found even more clearly in cases where the patient later dies (Lund submitted PLOS).

In other studies we found that hundreds of genes change their gene expression in blood as a consequence of increasing parity, the major protective factor for breast cancer; the more pregnancies, the more experiences of the immune system of the semi-allograft or fetus (Lund et al. 2018a). Here, the fetus is protected through a redirection of response away from the adaptive system towards the innate, a balance that is restored just before, at and after birth. The proposal that later the immune system will consider the cancer as a pseudo-semi-allograft, and with more experience immune cells or clones of cells, the better the success rate of elimination (Lund et al. 2018b).

HISTORICAL THEORIES

Over the last century, many theories of causes of cancer have been proposed by basic researchers and epidemiologists. In a review of carcinogenic models (Vineis

et al. 2010), five different models with their statistical methods were proposed: the mutational, genome instability, non-genotoxic, Darwinian, and tissue organization. Over the years, epidemiologists tried to use incidence rates to estimate the necessary number of stages for a cancer to develop, starting with the Armitage-Doll assumption of at least five stages (Armitage and Doll 1954). The number of stages was later reduced to two, the two-hit model proposed by Knudson (Knudson 2001). Today, most epidemiologists argue that cancer is caused by environmental carcinogens like smoking and radiation. In the paper by Peto and Doll in the early 1980s, "bad luck" was not necessary to explain the cancer epidemic (Doll and Peto 1981). Epidemiologists primarily use the relative risk between a carcinogen and cancer to discuss causality and prevention, but with no information on time dependency in the semi-parametric proportional hazard models. In basic research, the Hansemann-Boveri aneuploidy theory (Holland and Cleveland 2009) was based on observations of asymmetric mitoses in skin cancer. Warburg's theory was based on observations that cancers metabolize glucose via glycolysis (Hsu and Sabatini 2008). Today, basic researchers propose multistep carcinogenesis, such as the "bad luck" hypothesis explaining cancer as intrinsically random, and, therefore, unavailable, mutagenic events that dominate tumorigenesis (Tomasetti and Vogelstein 2015). This theory is unsupported by individual data and has been rejected by epidemiologists (Perduca et al. 2019).

Clinical researchers have mostly relied on basic research findings for new therapies. Now there is increasing concern about analogies from mice to human, from constructed diseases to human conditions (for further detail see Chapter 7).

THE IMMUNE SYSTEM AND CANCER—A LONG HISTORY

Accounts of the effects of the immune system on cancer patients have been recorded for centuries, such as the spontaneous regression of cancer, mostly in relation to serious infection (Hoption Cann et al. 2002). One of the first treatments of cancer was introduced in 1850 by French doctors, and they succeeded in treating two patients (Kaplon and Dieu-Nosjean 2018). Before and after the First World War, researchers performed systematic experiments on humans by injecting various bacteria, viruses and toxins (Kucerova and Cervinkova 2016). The high mortality due to the virulent disease used in the injections killed many patients, although many were cured too. After the Second World War, these accounts were dismissed and the spontaneous elimination of tumors in patients was considered impossible. Due to vaccinations and the more effective treatment of infectious diseases, it is possible that such cases were no longer seen. From basic research, many

studies pointed towards the concept of immune evasion (Hanahan and Weinberg 2011, Wang et al. 2017, Steven and Seliger 2018). This concept is partly compatible with the novel theory, but lack information on the exposures that are the driving forces. It is hard to imagine that the carcinogenesis of hormone-related breast cancer should be similar to smoking-induced lung cancer or HPV-induced cervical cancer. Thus, due to the different methods, observational confirmation has been lacking. New therapies called immunotherapies are based on several mechanisms in the tumor and the use of immune cells to kill the tumor cells. The new immunotherapies (De la Cruz and Czarniecki 2018) have changed the direction of cancer treatment towards immunology. Today, "hot" and "cold" cancers refer to the number of immune cells in the matrix around the cancer and the consequences for prognosis.

Another important aspect in understanding the role of the immune system in cancer and the novel theory is the increasing understanding of the role infections play in cancer development as causes or co-factors. In an overview the PAF (population-attributable factor) risk was calculated to be around 16% worldwide, but with large geographic differences (de Martel et al. 2012).

Infection is the main cause of HPV in cervical (Cohen et al. 2019) and pharyngeal cancer (Chen et al. 2019), some lymphomas (Molyneux et al. 2012), hepatitis C (Mina et al. 2015), and partly act as co-factors for helicobacter pylori in stomach cancer (Pereira-Marques et al. 2019). Research has been ongoing to link cancer of colon to the biota (Collins et al. 2011). The importance of chronic inflammation as an additional driver in many cancer sites has become a major research area (Qu et al. 2018). Other aspects of immunology and cancer promise new immunotherapies (De la Cruz and Czarniecki 2018) such as monoclonal antibodies and checkpoint inhibitors. Both cancer vaccines and oncolytic immunotherapy have the potential to improve survival (Guo et al. 2019). Epidemiologists should include inflammatory biomarkers in cancer research in order to obtain some indications of the reaction of the immune system towards the carcinogenic process (Brenner et al. 2014).

THE NEED FOR COLLABORATION ACROSS SCIENTIFIC DISCIPLINES

The different carcinogenic paradigms still exist side by side due to the lack of interaction between scientists of the differing traditions. The main reason has been the lack of models that might incorporate information from all three research disciplines in cancer (basal, clinical, and epidemiological research). Basic research uses

reductionist experiments with little option for experiments with different exposures over extended time, as in humans. In addition, the mice model is usually based on animal experiments in non-pathogenic laboratories, leaving the adult mice with an immune system comparable to a newborn human. We postulated that the experiences of the immune system could be important for the power of protection against transformed cancer cells. Clinicians almost only have studies of the cancer patient, with no possibility of looking back on lifestyle.

Obviously, new designs and technologies are necessary for an understanding of the dynamic interface between the immune system and the effect of carcinogens on tissue cells around the body. Through the systems epidemiology concept we have built a new biobank giving us the opportunity to follow up on gene expression in the blood from before diagnosis, at diagnosis, and after diagnosis. At the same time, we can collect either fresh tumor or normal tissue cells, or collect biopsies from the paraffin-embedded samples used for diagnostics of cancer. From the basic traditional prospective design, we can introduce different lifestyle factors through questionnaire information or the use of biomarkers. It is easy to add genomic information such as single nucleotide polymorphisms (SNPs).

The crucial difference between the dynamic interface theory and previous proposals is the willingness to view the different scientific disciplines as equally important. A novel theory must be able to synthesize previous ones.

The strength of the proposed theory is the combination of information from all three cancer research disciplines with the core concepts of dynamics over time and an interface in which tumor cells encounter the immune cells struggling for life. Here, immunology as both a basic research discipline and well-accepted disease-related research meets new possibilities in epidemiology that also include basic cell studies.

CHALLENGES OF THE DYNAMIC INTERFERENCE THEORY

This novel theory has some important implications or new hypotheses:

- Any substance inhibiting the immune system will work as a carcinogen. Carcinogens are the drivers of change from normal to cancer tissue cells. The immune system acts on the interface of the tissue to attack transformed cells. This illustrates the balance or war on cancer, but any substance inhibiting the immune system could act as driver of carcinogenesis.
- What is the effect of previous experiences of the immune system? The accumulated experiences of the immune system could be important for later resistance towards cancer development.

- An interesting hypothesis could be to search for the memory cells in the immune system, for the immune cells' victories, or the carcinogens' lost battles. Should we expect to find successful clones of immune cells, and if yes, how many different clones over a long life?

EPILOGUE

The metaphor of clone wars was chosen without knowledge of the clone war in *Star Wars*. Still, the metaphor is a nice one: The good guys create clones of 12 000 soldiers to defend their empire, but the bad ones among them implant a chip into the heads of the soldiers. When the command "Order 66" is given, the soldiers kill the good guys, the officers or the leaders. Fortunately, however, one good guy survives.

REFERENCES

Armitage P, Doll R. Br J Cancer. 1954 Mar; 8(1): 1–12. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2007940/

Brenner DR, Scherer D, Muir K, Schildkraut J, Boffetta P, Spitz MR et al. A review of the application of inflammatory biomarkers in epidemiologic cancer research. Cancer Epidemiol Biomarkers Prev 2014 Sep; 23(9): 1729–1751. Available from: https://cebp.aacrjournals.org/content/23/9/1729.long

Chen YP, Chan ATC, Le QT, Blanchard P, Sun Y, Ma J. Nasopharyngeal carcinoma. Lancet. 2019 Jul 6; 394(10192): 64–80. Available from: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(19)30956-0/fulltext

Cohen PA, Jhingran A, Oaknin A, Denny L. Cervical cancer. Lancet. 2019 Jan 12; 393(10167): 169–182. Available from: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)32470-X/fulltext

Collins D, Hogan AM, Winter DC. Microbial and viral pathogens in colorectal cancer. Lancet Oncol. 2011 May; 12(5): 504–512. Available from: https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(10)70186-8/fulltext

De la Cruz LM, Czarniecki BJ. Immunotherapy for breast cancer is finally at the doorstep: Immunotherapy in breast cancer. Ann Surg Oncol. 2018 Oct; 25(10): 2852–2857. Available from: https://link.springer.com/article/10.1245%2Fs10434-018-6620-5

de Martel C, Ferlay J, Franceschi S, Vignat J, Bray F, Forman D et al. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. Lancet Oncol. 2012 Jun; 13(6): 607–615. Available from: https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(12)70137-7/fulltext

Doll R, Peto R. The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. J Natl Cancer Inst. 1981 Jun; 66(6): 1191–1308.

Dumeaux V, Fjukstad B, Fjosne HE, Frantzen JO, Holmen MM, Rodegerdts E et al. Interactions between the tumor and the blood systemic response of breast cancer patients. PLoS Comput Biol. 2017 Sep 28; 13(9): e1005680. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005680

Guo ZS, Lu B, Guo Z, Giehl E, Feist M, Dai E et al. Vaccinia virus-mediated cancer immunotherapy; cancer vaccines and oncolytics. J Immunother Cancer. 2019 Jan 9; 7(1): 6. Available from: https://jitc.biomedcentral.com/articles/10.1186/s40425-018-0495-7

Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011 Mar 4; 144(5): 646–674. Available from: https://www.sciencedirect.com/science/article/pii/S0092867411001279?via%3Dihub

Holden M, Holden L, Olsen KS, Lund E. 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. 2017; 7: 11–28. Available from: https://www.dovepress.com/local-in-time-statistics-for-detecting-weak-gene-expression-signals-in-peer-reviewed-article-AGG

Holland AJ, Cleveland DW. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. Nat Rev Mol Cell Biol. 2009 Jul; 10(7): 478–487. Available rom: https://www.nature.com/articles/nrm2718

Hoption Cann SA, van Netten JP, van Netten C, Glover DW. Spontaneous regression: a hidden treasure buried in time. Med Hypotheses. 2002 Feb; 58(2): 115–119. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0306987701914690?via%3Dihub

Hsu PP, Sabatini DM. Cancer cell metabolism: Warburg and beyond. Cell 2008 Sep 5; 134(5): 703–707. Available from: https://www.cell.com/cell/fulltext/S0092-8674(08)01066-0?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867408010660%3Fshowall%3Dtrue

Kaplon H, Dieu-Nosjean MC. Quelle avenir pour les lymphocytes B infiltrant les tumeures solides. MedSci (Paris). 2018 Jan; 34(1): 72–78. Available from: https://www.medecinesciences.org/articles/medsci/pdf/2018/01/medsci20183401p72.pdf

Knudson AG. Two genetic hits (more or less) to cancer. Nat Rev Cancer. 2001 Nov; 1(2): 157–162. Available from: https://www.nature.com/articles/35101031

Kucerova P, Cervinkova M. Spontaneous regression of tumour and the role of microbila infection – possibilities for cancer treatment. Anticancer Drugs. 2016 Apr; 27(4): 269–277. Available from: https://insights.ovid.com/article/00001813-201604000-00001

Lund E, Holden L, Bøvestad H, Plancade S, Mode N, Günther CC et al. 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 Med Res Methodol. 2016 Mar 5; 16: 28. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0129-z

Lund E, Nakamura A, Snapkov I, Thalabard JC, Olsen KS, Holden L, et al. Each pregnancy linearly changes immune gene expression in the blood of healthy women compared with breast cancer patients. Clin Epidemiol. 2018a Aug 6; 10: 931–940. Available from: https://www.dovepress.com/each-pregnancy-linearly-changes-immune-gene-expression-in-the-blood-of-peer-reviewed-article-CLEP

Lund E, Rasmussen Busund LT, Thalabard JC. Rethinking the carcinogenesis of breast cancer: The theory of breast cancer as a child deficiency disease or a pseudo semi-allograft. Med Hypotheses. 2018b Nov; 120: 76–80. Available from: https://www.sciencedirect.com/science/article/pii/S0306987718305310?via%3Dihub

Mina MM, Luciani F, Cameron B, Bull RA, Beard MR, Lloyd AR. Resistance to hepatitis C virus: potential genetic and immunological determinants. Lancet Infect Dis. 2015 Apr; 15(4): 451–460. Available from: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(14)70965-X/fulltext

Molyneux EM, Rochford R, Griffin B, Newton R, Jackson G, Menon G et al. Burkitt´s Lymphoma. Lancet. 2012 Mar 31; 379(9822): 1234–1244. Available from: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(11)61177-X/fulltext

Perduca V, Alexandrov LB, Kelly-Irving M, Delpierre C, Omichessan H, Little MP, et al. Stem cell replication, somatic mutations and role of randomness in the development of cancer. Eur J Epidemiol. 2019 May; 34(5): 439–445. Available from: https://link.springer.com/article/10.1007%2Fs10654-018-0477-6

Pereira-Marques J, Ferreira RM, Pinto-Ribeiro I, Figueiredo C. Helicobacter pylori infection, the Gastric Microbiome and Gastric Cancer. Adv Exp Med Biol. 2019; 1149: 195–210. Available from: https://link.springer.com/chapter/10.1007%2F5584_2019_366

Qu X, Tang Y, Hua S. Immunological approaches towards cancer and inflammation: a cross talk. Front Immunol. 2018 Mar 20; 9: 563. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2018.00563/full

Steven A, Seliger B. The role of immune escape and immune cell infiltration in breast cancer. Breast Care (Basel). 2018 Mar; 13(1): 16–21. Available from: https://www.karger.com/Article/FullText/486585

Tomasetti C, Vogelstein B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science. 2015 Jan 2; 347(6217): 78–81. Available from: https://science.sciencemag.org/content/347/6217/78.long

Vineis P, Schatzkin A, Potter JD. Models of carcinogenesis: an overview. Carcinogenesis. 2010 Oct; 31(10): 1703–1709. Available from: https://academic.oup.com/carcin/article/31/10/1703/2476893

Wang M, Zhang C, Song Y, Wang Z, Wang Y, Luo F et al. Mechanism of immune evasion in breast cancer. OncoTargets Ther. 2017 Mar 14; 10: 1561–1573. Available from: https://www.dovepress.com/mechanism-of-immune-evasion-in-breast-cancer-peer-reviewed-article-OTT

Lund, E. (Ed.) (2020). *Advancing Systems Epidemiology in Cancer. Exploring Trajectories of Gene Expression.*
Oslo: Scandinavian University Press.
DOI: https://doi.org/10.18261/9788215041193-2020-12

Supplement

Workshop to Facilitate Cancer Systems Epidemiology Research

https://epi.grants.cancer.gov/events/systems-epidemiology/?cid=eb_govdel_en_egrp_newsletter_jan19

OVERVIEW

The availability of high-throughput -omic technologies, novel devices for exposure assessment, and electronic medical records have the potential to facilitate a more comprehensive study of risk factors contributing to development of and outcomes from cancer.

Despite individual successes at identifying genetic, biological, and environmental risk factors for cancer, much of the etiology remains unexplained. This may be due in part to the limited focus of many studies on a single or small set of risk factors or data types (i.e. measures such as DNA sequence, methylation data, variables from questionnaires). Moreover, many studies fail to consider the complexities and interrelations among multiple risk factors and associated outcomes. For example, each individual risk factor, such as a single dietary component or genetic polymorphism, occurs in a broader biological (e.g. pathways) or societal (e.g. individual in social network) context which could modulate the effect of individual risk factors on disease. Further, many risk factors for disease can be highly correlated with possible interactive, synergistic, or attenuating effects. Importantly, risk factors can change over time.

A more comprehensive, systems modeling-based type of approach, which accounts for multiple dimensions, integration of diverse data types, and changes over time, is needed to better understand contributors to disease and treatment outcomes and provide clues for improved intervention.

PURPOSE

The objective of this workshop was to facilitate interdisciplinary discussion about the application of systems modeling approaches for population-based cancer epidemiology research. By bringing together scientists from various fields that use systems modeling, the workshop aimed to:

- Identify ideas and strategies to improve understanding of systems modeling among population scientists and epidemiology amongst modelers
- Share lessons learned in the application of systems approaches from other fields (e.g. cancer biology)
- Identify potential high-impact use cases for systems modeling in population science
- Increase understanding of potential barriers to and facilitators of taking a system modeling approach in population science (including dataset availabilities, data and methods needs)
- Establish new collaborative interdisciplinary relationships between statisticians, mathematicians, computer scientists, bioinformaticians, epidemiologists, and clinicians

SYSTEMS EPIDEMIOLOGY KEY REFERENCES

Reviews/Commentaries

Bennet BJ et al. (2015) Nutrition and the science of disease prevention: a systems approach to support metabolic health. *Ann N Y Acad Sci*; 1352:1–12.

Burke T et al. (2017) Rethinking Environmental Protection: Meeting the Challenges of a Changing World. *Environmental Health Perspectives*; 125(3): A43–A49.

Cornelis MC and Hu FB (2013) Systems Epidemiology: A New Direction in Nutrition and Metabolic Disease Research. *Curr Nutr Rep*; 2(4).

Cronbach LJ and Meehl PE (1955) Construct Validity in Psychological Tests. *Psychol Bull*; 52(4):281–302.

Dammann O et al. (2014) Systems Epidemiology: What's in a Name? *Online Journal of Public Health Informatics*; 6(3): e198.

Diez Roux AV (2011) Complex Systems Thinking and Current Impasses In Health Disparities Research. *Am J Public Health*; 101(9):1627–1634.

Hammond RA (2009) Complex Systems Modeling for Obesity Research. *Preventing Chronic Disease*; 6(3):A97.

Joffe M et al. (2012) Causal Diagrams in Systems Epidemiology. *Emerg Themes Epidemiol*; 9(1):1.

Krauth SJ et al. (2019) A Call for Systems Epidemiology to Tackle the Complexity of Schistosomiasis, Its Control, and Its Elimination. *Trop Med Infect Dis*; 4(1): 21.

Lee BY et al. (2017) A Systems Approach to Obesity. *Nutr Rev*; 75(suppl 1):94–106.
Lund E and Dumeaux V (2008) Systems Epidemiology in Cancer. *Cancer Epidemiol Biomarkers Prev*; 17(11).
Ritchie MD et al. (2015) Methods of Integrating Data to Uncover Genotype-Phenotype Interactions. *Nat Rev Genet*; 16(2):85–97.
Rothman KJ and Greenland S. (2005) Causation and Causal Inference in Epidemiology. *American Journal of Public Health*; 95.S1 (2005): S144–S150.
Rutter H et al. (2017) The Need for a Complex Systems Model of Evidence for Public Health. *The Lancet*; 390(10112):2602–2604.
Warnecke et al. (2008) Approaching Health Disparities from a Population Perspective: The National Institutes of Health Centers for Population Health and Health Disparities. *Am J Public Health*; 98(9): 1608–1615.

Research Studies

Curtis C et al. (2012) The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups. *Nature*, 486(7403): 346–352.
Finucane HK et al. (2018) Heritability Enrichment of Specifically Expressed Genes Identifies Disease-Relevant Tissues and Cell Types. *Nat Genet*; 50(4): 621–629.
Hiatt RA et al. (2014) A Multilevel Model of Postmenopausal Breast Cancer Incidence. *Cancer Epidemiol Biomarkers Prev*; 23(10): 2078–2092.
Jones AP et al. (2006) Understanding Diabetes Population Dynamics Through Simulation Modeling and Experimentation. *Am J Public Health*; 96(3):488–494.
Kaligotla C et al. (2018) Modeling an Information-Based Community Health Intervention on the South Side of Chicago. In: 2018 Winter Simulation Conference (WSC) 2018 Dec 9 (pp. 2600–2611). IEEE.
Kettunen J et al. (2016) Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun*; 7:11122.
Price ND et al. (2017) A Wellness Study of 108 Individuals Using Personal, Dense, Dynamic Data Clouds. *Nat Biotechnol*; 35(8): 747–756.
Shaman J and Karspeck A (2012) Forecasting Seasonal Outbreaks of Influenza. *Proc Natl Acad Sci USA*; 109(50): 20425–20430.

DISSEMINATION & IMPLEMENTATION PANEL

Aarons G et al. (2014) Aligning leadership across systems and organizations to develop a strategic climate for evidence-based practice implementation. *Annu Rev Public Health*; 34:255–274.
Lindau ST et al. (2016) CommunityRx: A Population Health Improvement Innovation that Connects Clinics to Communities. *Health Affairs*, 35(11):2020–2029.
Mandelblatt et al. (2016) Collaborative Modeling of the Benefits and Harms Associated With Different U.S. Breast Cancer Screening Strategies. *Ann Intern Med*. 2016;164:215–225.
Supplement on Breast Cancer Simulation Modeling in CISNET. *Medical Decision Making*; 39(2).

Lund, E. (Ed.) (2020). *Advancing Systems Epidemiology in Cancer. Exploring Trajectories of Gene Expression.* Oslo: Scandinavian University Press.
DOI: https://doi.org/10.18261/9788215041193-2020-13

First Author Nota

Arnes, Jo Inge (1973). Began his professional career in systems development two decades ago. He is currently a PhD student at UiT The Arctic University of Norway, Departement of Computer Science. His experience includes systems used in healthcare.

Busund, Lill Tove Rasmussen (1959). Professor, UiT The Arctic University of Norway and University Hospital in Northern Norway, Tromsø, Norway. Scientific area: clinical diagnostic pathology, translational cancer research with focus on prognostic and predictive markers in lung, prostate, breast cancers and sarcomas.

- Moi, L., Braaten, T., Al-Shibli, K., Lund, E., & Busund, L.T.R. (2019). Differential expression of the miR-17-92 cluster and miR-17 family in breast cancer according to tumor type; results from the Norwegian Women and Cancer (NOWAC) study. *Journal of Translational Medicine*, 17, 1, 334.
- Richardsen, E., Andersen, S., Melbø-Jørgensen, C., Rakaee, M., Ness, N., Al-Saad, S., ... Busund, L.T. (2019). MicroRNA 141 is associated to outcome and aggressive tumor characteristics in prostate cancer. *Scientific Reports*, 9, 1, 386.

Fjukstad, Bjørn (1990). System developer at DIPS, previously PhD student at the Department of Computer Science, UiT The Arctic University of Norway.

- Fjukstad, B., Dumeaux, V., Hallett, M., & Bongo, L.A. Reproducible Data Analysis Pipelines for Precision Medicine. (2019). In *Proceedings of the 27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)* (pp. 299–306). Pavia, Italy: IEEE Computer Society.
- Fjukstad, B., Dumeaux, V., Standahl Olsen, K., Hallett, M., Lund, E., Bongo, L.A. (2017). Building Applications For Interactive Data Exploration In Systems Biology. In *ACM-BCB´17: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 556–561). New York, United States: Association for Computing Machinery.

Holden, Lars (1959). Managing director, Norwegian Computing Center. Scientific area: Applied mathematics also including genetics, Stochastic models, Spatial statistics, Markov Chain Monte Carlo (MCMC).

- Holden, L. (2019). Mixing of MCMC algorithms. *Journal of Statistical Computation and Simulation*, 89, 12.

- Ferkingstad, E., Holden, L., & Sandve, G.K. (2015). Monte Carlo Null Models for Genomic Data. *Statistical Science*, 30, 1, 59–71.
- Sandve, G.K., Gundersen, S., Rydbeck, H., Glad, I.K., Holden, L., Holden, M., ... Hovig, E. (2010). The Genomic HyperBrowser: inferential genomics at the sequence level. *Genome Biology*, 11, 12, R121.

Holden, Marit (1962). Chief research scientist, Norwegian Computing Center. Scientific area: Bioinformatics, Functional genomics, Deep learning, Image analysis, Pattern recognition, Markov Chain Monte Carlo (MCMC).

- 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.
- Eikvil, L., & Holden, M. (2014). Evaluation of Binary Descriptors for Fast and Fully Automatic Identification. In *Proceedings of 22nd International Conference on Pattern Recognition* (pp. 154-159). Stockholm, Sweden: IEEE Computer Society.

Holsbø, Einar (1985). Postdoctor, UiT The Arctic University of Norway, Tromsø, Norway. Scientific area: Population surveys in north Norway.

- Holsbø, E. (2018). *Small data: practical modeling issues in human-model -omic data.* (PhD thesis). Department of Computer Science, Faculty of Science and Technology, UiT The Arctic University of Norway, Tromsø, Norway.

Krum-Hansen, Sanda (1968). **Medical doctor, PhD student,** Stavanger university hospital, UiT The Arctic University of Norway, Scientific area: breast cancer, clinic, molecular epidemiology, functional genomics.

- Olsen, K.S., Holsbø, E., Rognmo, K., Krum-Hansen, S., & Lund, E. (2015). Stress related to a suspicious mammogram – potential transcriptomic effects. *Norsk Epidemiologi*, 25, Suppl. 1, 51.
- Ritte, R., Lukanova, A., Tjønneland, A., Olsen, A., Overvad, K., Mesrine, S., … Kaaks, R. (2012). Height, age at menarche and risk of hormone receptor-positive and -negative breast cancer: A cohort study. *International Journal of Cancer*, 132, 11, 2619–2629.

Lund, Eiliv (1947). Professor emeritus at the UiT Arctic University of Norway, and a part-time researcher at the Norwegian Cancer Registry. Scientific area: cancer

epidemiology with main focus on breast cancer, and since 2001 studies on functional genomics.

- Lund, E. (1990). Number of children and death from hormone-dependent cancers. *International Journal of Cancer,* 46, 6, 998–1000.
- Lund, E., Nakamura, A., Snapkov, I., Thalabard, J.C., Olsen, K.S., Holden, L., Holden, M. (2018). Each pregnancy linearly changes immune gene expression in the blood of healthy women compared with breast cancer patients. *Clinical Epidemiology*, 10, 931–940.

Affiliations

Eiliv Lund: Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway; elu000@post.uit.no

Jo Inge Arnes: Department of Computer Science, UiT The Arctic University of Norway, Tromsø, Norway; jo.i.arnes@uit.no

Lars Ailo Bongo: Department of Computer Science, UiT The Arctic University of Norway, Tromsø, Norway; labongo@gmail.com

Bjørn Fjukstad: Department of Computer Science, UiT The Arctic University of Norway, Tromsø, Norway

Nikita Shvetsov: Department of Computer Science, UiT The Arctic University of Norway, Tromsø, Norway

Therese H. Nøst: Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway

Hege Bøvelstad: Norwegian Institute of Public Health, Oslo, Norway

Till Halbach: Department of Applied Research in Information Technology, Norwegian Computing Center, Oslo, Norway

Einar Holsbø: Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway; einar.j.holsbo@uit.no

Knut Hansen: Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway

Sanda Krum-Hansen: Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway; sanda.krum-hansen@uit.no

Karina Standahl Olsen: Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway

Marit Holden: Department of Statistical Analysis, Machine Learning and Image Analysis, Norwegian Computing Center, Oslo, Norway

Lars Holden: Norwegian Computing Center, Oslo, Norway

Kajsa Møllersen: Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway; kajsa.mollersen@uit.no

Lill Tove Rasmussen Busund: Institute for Medical Biology, UiT The Arctic University of Norway, Tromsø, Norway; lill.tove.rasmussen.busund@unn.no

Jean-Christophe Thalabard: MAP5, Universite Rene Descartes, Paris, France

Sandra Plancade; INRA, Paris, France