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.

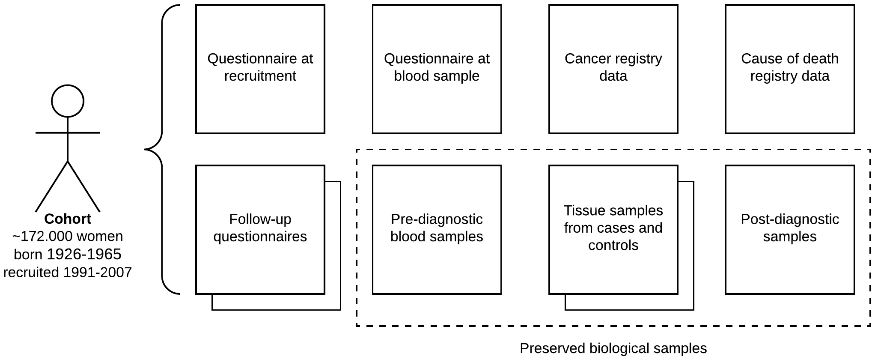
Cohort
~172.000 women
born 1926-1965
recruited 1991-2007
Questionnaire at recruitment
Questionnaire at blood sample
Cancer registry data
Cause of death registry data
Follow-up questionnaires
Pre-diagnostic blood samples
Tissue samples from cases and controls
Post-diagnostic samples
Preserved biological samples

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 different 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.

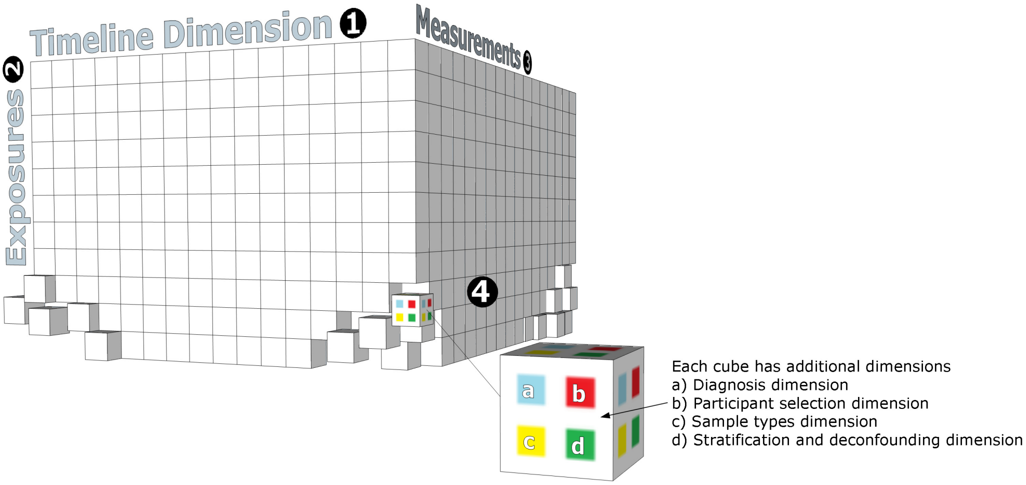
Timeline Dimension 1
Exposures 2
Measurements 3
4
a
b
c
d
Each cube has additional dimensions
a) Diagnosis dimension
b) Participant selection dimension
c) Sample types dimension
d) Stratification and deconfounding dimension

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.

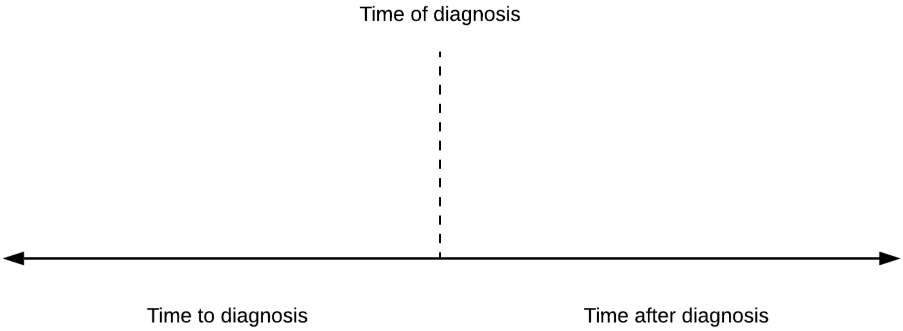

Figure 2.3. Time to diagnosis, time of diagnosis, and time after diagnosis.

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.

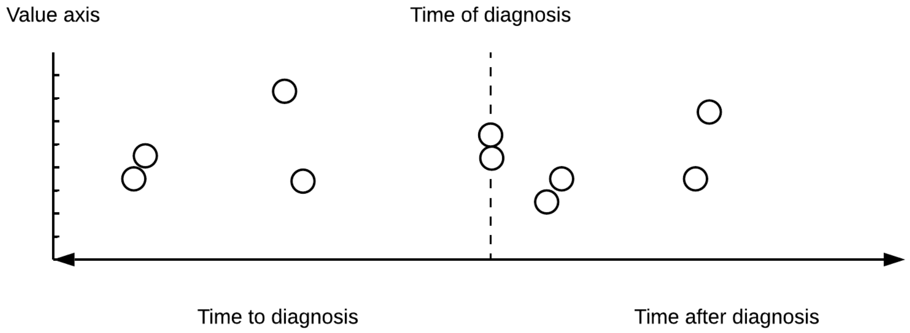

Figure 2.4. Sample data points positioned by distance from diagnosis. The value axis does not have to be linear; it can be logarithmic or other.

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.

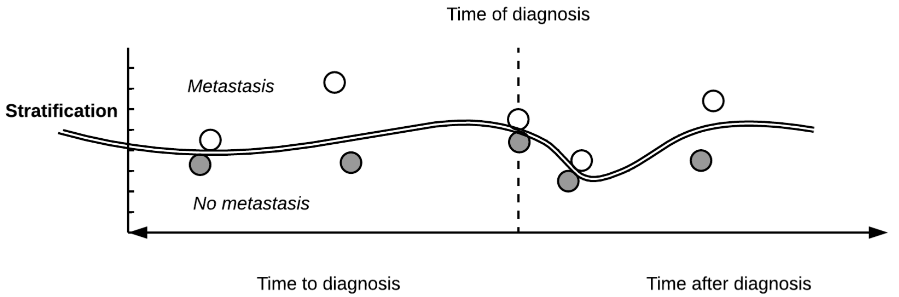

Figure 2.5. Stratification of data points. In this example, the white-filled circles represent women with metastasis, and the grey-filled circles represent women without.

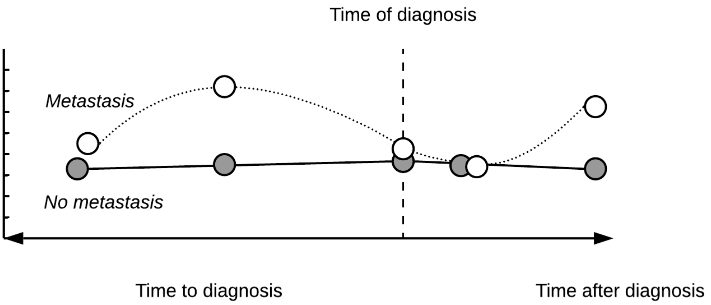

Figure 2.6. An illustration of estimated curves or trajectories for each stratum. The curves for the two strata are different.

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.

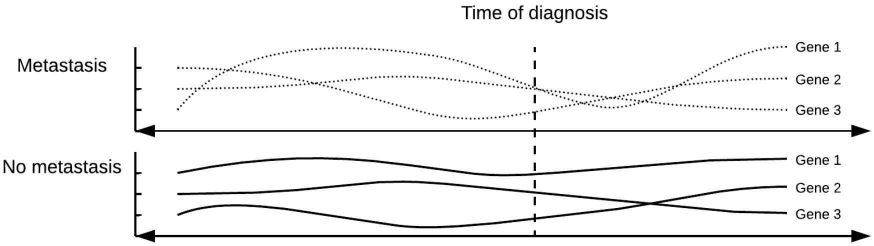

Figure 2.7. For each biological sample, we can measure the levels of many different expressed genes. For each, we can imagine a separate curve per strata. In the illustration, only the expression levels for "Gene 1" differ between the two strata. Note that we are not restricted to gene expressions. Other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

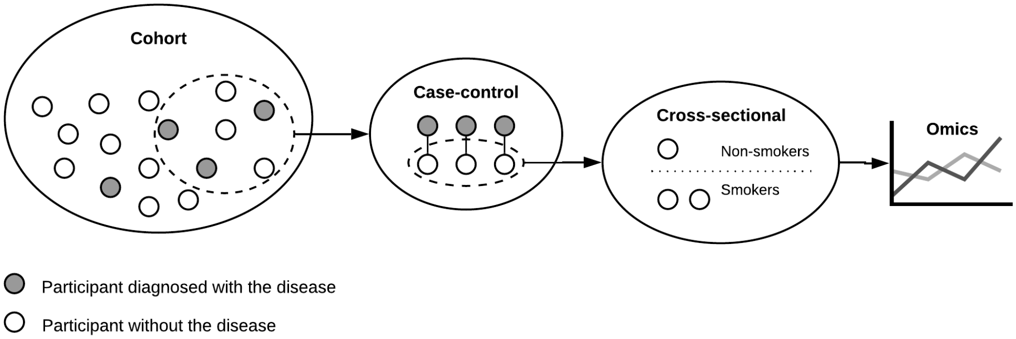
Cohort
Case-control
Cross-sectional
Non-smokers
Smokers
Omics
Participant diagnosed with the disease
Participant without the disease

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.

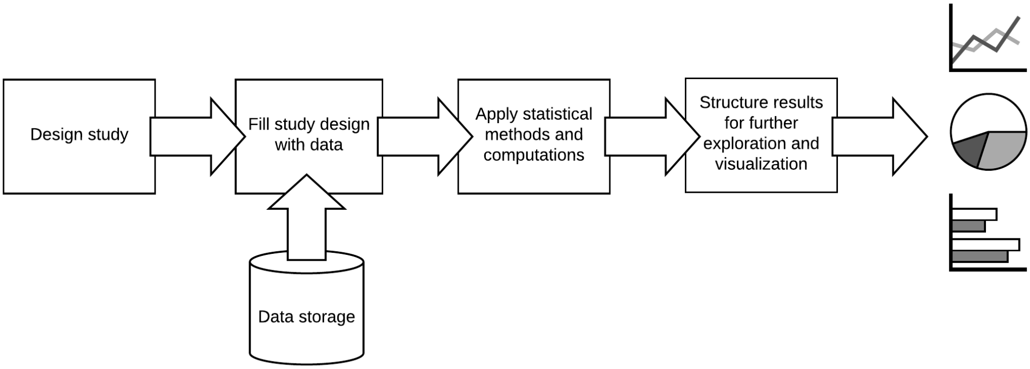
Design study
Fill study design with data
Apply statistical methods and computations
Structure results for further exploration and visualization
Data storage

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 experiments. 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 diagnosis. 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