UiT

THE ARCTIC
UNIVERSITY
OF NORWAY

Report: Smoking Variable Wrangling

2019

Faculty of Science and Technology

INF-8810 Special Curriculum – Complex Epidemiologic Study Designs

Jo Inge Arnes

Report: Smoking Variable Wrangling

lNF-8810, Complex Epidemiologic Study Designs

Jo Inge Arnes

January 24, 2019

Abstract

Smoking is the primary cause of lung cancer, and there is extensive research on smoking in relation to lung cancer because of this. In epigenome-wide epidemiologic studies that analyze blood to find biomarkers for lung cancer, it can be challenging to separate disease biomarkers and biological effects of smoking. To identify smoking-related effects, it is therefore interesting to analyze people without lung cancer. As a first step towards identifying such effects, we estimate smoking statuses by combining data from multiple surveys.

Acknowledgments

This report is part of the practical project for the individual special curriculum INF-8810 Complex Epidemiologic Study Designs at UiT, The Arctic University of Norway. The curriculum was created by professors Lars Ailo Bongo and Torkjel M. Sandanger. Professor Eiliv Lund also contributed by giving suggestions to the curriculum.

The practical part of this project was done in collaboration with Nikita Baiju. A special thanks to Dr. Therese Haugdahl Nøst for answering questions and giving advice during the project.

Contents

1 Introduction

It is well known that smoking causes lung cancer. It has also been shown that smoking causes alterations to gene expression and methylation levels [1][2]. However, it can be difficult to differentiate between a biomarker for cancer and for the exposure to smoking in itself. From a causal standpoint, a biomarker can be an intermediate in the natural history of the disease or an exposure marker. Whether it is one or the other is often not known, because of the current lack of knowledge about carcinogenic processes [3].

Our goal is to estimate smoking statuses for a control group consisting of healthy women from the NOWAC cohort. There are 1766 participants in the control group, for which we have blood samples analyzed for gene expression profiles. Each woman's smoking status is determined for the time of blood sample acquisition from 82 variables relevant for the task (Appendix C). The smoking statuses make it possible to later compare transcriptomic [4] data based on differences in smoking exposure. The estimated smoking status data can be used to detect changes in gene expressions, DNA methylation, and other values associated with tobacco smoking exposure.

The estimation of smoking status variables is mainly a question of wrangling data. Data wrangling is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics [5]. Challenges that complicate the data wrangling are the number of different, often overlapping smoking-related variables, and the many different questionnaires and questions involved. Each woman have answered between two and four questionnaires. Sometimes the answers are inconsistent, or the data punching erroneous. Issues with inconsistencies and data punching errors are generally common when using questionnaires, so this comes as no surprise.

We implemented a set of scripts for determining smoking status, passive smoking, intensity, duration, and time since cessation, in the R programming language [6]. In future work, we plan to combine the estimated smoking status values with gene expression data to find biomarkers for smoking in healthy people so that this can be adjusted for when researching disease biomarkers in people with cancer.

The project is done in conjunction with the Id-Lung project, which builds upon surveys and biological samples from the NOWAC cohort for its case-control studies. Since the project is part of a special curriculum, the report also includes essential concepts from epidemiology.

2 Ethical Considerations

The individuals who participate in epidemiologic studies must understand and consent to how their data will be used for research. The privacy of the participants must always be respected. The research is also unethical if it poses known risks of harm to the participants.

This project uses data from NOWAC. The Regional Committee for Medical Research Ethics has approved NOWAC, and all data are stored and handled according to the permission given by the Norwegian Data Inspectorate. All women have filled in informed consent.

We work with a de-identified [7] subset of data from NOWAC. The data only include variables relevant to smoking.

3 Background and Study Design Challenges

In this section, we present essential concepts and background information necessary to apprehend the project's context, challenges, and purpose.

3.1 Epidemiology and Study Designs

These sections describe topics in epidemiology that are relevant for this report, mainly based on the book Epidemiology by Leon Gordis [8].

Epidemiology is the study of the distribution and determinants of health-related states or events in specified populations and the application of this study to control health problems [9]. Before we conduct an epidemiologic study, we must first plan and design the study. There are many different study designs to choose from in epidemiology, each with characteristic strengths and weaknesses. A study design decides how individuals are selected and grouped, and outline how the study is carried out. The five types of study designs described in these sections are case-control, cohort, cross-sectional, hybrid, and systems epidemiologic.

The terms association, bias, confounding, and etiology are explained first, followed by descriptions of the study designs.

3.1.1 Bias

Bias comes in many forms and makes study results skewed and wrong because of systematic errors in the way we have designed, conducted, or analyzed a study. Reducing bias is an important goal in epidemiology.

The definition of bias is any systematic error in the design, conduct or analysis of a study that results in a mistaken estimate of an exposures effect on the risk of disease [9]. Two of the main categories are selection bias and information bias.

Selection bias are errors in how cases and controls, or exposed and non-exposed individuals are selected. If the selection causes the observation of associations, even if exposure and disease are not associated in reality, it is a selection bias. This should not be confused with problems in generalizability and external validity of a study due to how selections were made. Even if there is a problem with generalizability, the validity of the results can still be correct internally for the selected groups. With selection bias, on the other hand, the internal validity is also affected. Selection bias can be reduced as part of the study design.

Information bias can occur when some information regarding exposures and/or disease outcomes for participants is incorrect because it has been gathered inadequately. Misclassification bias is an information bias where, for example, some people with the disease in question are classified as healthy, or vice versa. In our project, we only included participants that were healthy. A misclassification would be if some of the participants, in reality, had cancer. Recall bias is a type of information bias caused by participants misremembering some answers to questions. In this project, handling recall bias is important. For example, the women had been asked how much they smoked between different ages. Precisely remembering the number of cigarettes smoked per day can be hard decades later. We therefore aim to reduce the bias by cross-checking answers from the same woman for comparable questions from different surveys.

3.1.2 Association, Etiology, and Confounding

In epidemiology, we seek to find risk factors, early symptoms, and causes of disease. We must differentiate between the risk factors and causes of the disease, and accompanying factors that are not causal. We should thus know the terms association, etiology, and confounding.

Etiology is the cause of disease, while association is that we observe factors that occur together. When analyzing associations and etiology, we should be conscious of confounding factors, where an association between two factors is not a cause-effect relationship. A factor can be associated with another without being caused by it, since there could be underlying circumstances on which the other factors depend.

3.1.3 Case-Control Studies

A case-control study compares persons with the disease (cases) with persons without the disease (controls). Cases and controls are first selected into separate groups. The researchers then examine the exposure histories for the groups. One way of gathering information about a specific exposure is to ask questions about past exposures. This is called a retrospective design. The gathered information can then be used to determine if an exposure increased the risk of developing the disease under study.

An advantage of case-control studies is that they are relatively inexpensive and require a small number of subjects. It is also possible to examine different exposures per study.

A challenge with case-control studies is that biases and problems with external validity easily can arise due to improper selection of controls. The selection of controls is regarded as one of the most difficult methodological problems in epidemiology. Another weakness of case-control studies is recall bias.

Case-control studies can be included as part of hybrid study designs, which are discussed in section 3.1.5.

3.1.4 Prospective Cohort Studies

A cohort study compares persons exposed to a suspected risk factor to non-exposed persons. Individuals who have not yet developed the disease are placed in different groups based on exposure. After some time, often years, a percent of the persons are expected to have developed the disease. This is called prospective design. The groups can then be compared to see if there is an association between the exposure and the disease.

A strength of cohort studies is that recall bias is much less prominent than in case-control studies, due to questions being asked pre-diagnostically to the disease. Also, participants can be followed up regularly, and data can be gathered for instance from blood samples before the disease shows any symptoms. It is also possible to study different diseases for the same exposure.

One significant weakness is that many participants are needed for studying rare diseases. It may also take many years before the disease develops in the groups.

NOWAC was designed as a prospective cohort study where the exposure in question was oral contraceptives, and the disease was breast cancer. It was also deliberately designed with hybrid designs in mind, and thus had extensive surveys, samples, and follow-ups. Hybrid designs are described in the next section.

3.1.5 Hybrid Study Designs

To tackle the challenges in cohort and case-control studies, a combination of the two study designs can be used. First, we can identify a population that will be followed up over time. This is the cohort. The participants in the cohort answer questionnaires, give blood samples, urine tests, and so on. Over a period of years, a small percentage of the participants will develop the disease. A case-control study can then be carried out by using these persons as cases, and a selected number of matching controls can be chosen from the cohort. There are two main categories of such hybrid designs: Nested case-control, and case-cohort. The difference is in the selection of controls. For nested case-control, controls are selected at the same time that a new case is discovered. For case-cohort, controls are picked out all at once at after a number of cases have been found.

There are at least four advantages of hybrid designs. First, recall bias is reduced because the questionnaires have been answered before the disease was diagnosed. Second, since samples are given years before disease, they can be used to identify pre-diagnostic risk factors. When, for instance, blood samples given after diagnosis are analyzed it can be hard to isolate early risk biomarkers from the rest. Third, it is less expensive than a traditional cohort study. Blood and other samples can be preserved at the beginning and during the study, but not used until needed in a much smaller, nested case-control. Then only a small number of specimens have to be analyzed in a laboratory. Fourth, since cases and controls are derived from the same original cohort, they are more likely to be comparable between them. As mentioned earlier, selecting controls is challenging with respect to selection bias or validity. A cohort-based case-control study reduces some of the risk for selection bias.

NOWAC is a prospective cohort study was designed to be usable for later hybrid designs. For example, NOWAC's design makes it possible for the Id-Lung project to design studies that select cases and controls from NOWAC.

3.1.6 Cross-Sectional Studies

Cross-sectional studies are used to investigate risk factors. They differ in that we simultaneously determine both exposure (or risk) and disease outcomes. A cross-sectional study is as a snapshot of the population at a particular time. We do not include when the diagnosis was set or how long it has lasted per individual, only that persons in the population have the disease or not at the given time. This is called prevalence. For this reason, cross-sectional studies are also known as prevalence studies. In this type of design, there are four groups:

- Exposed people with the disease
- Exposed without disease
- Non-exposed with the disease
- Non-exposed without the disease

This approach has several weaknesses. The temporal relationships are missing because there is no information about time. Consequently, it could be that the suspected risk factor is a result of the disease or for other reasons started after the onset of the disease. It could also be that many of the ill already have died, and therefore is not included in the cross-section. The risk factor may, in this case, be a survival factor instead, found in people who live longer with the disease. This is called survival bias.

In the project for this report, a cross-sectional study is conducted only for the groups without the disease.

3.1.7 Systems Epidemiology

Systems epidemiology [10] is a research discipline focused on analyzing gene expressions, functions, and pathways to improve the understanding of biological processes in humans in relation to diseases. It is described as the counterpart to systems biology, which has been defined as a discipline that seeks to determine how complex biological systems function by integrating experimentally derived information through mathematical and computing solutions [11]. Systems epidemiology researches the complex processes and causalities of diseases, such as cancers, by analyzing the transcriptome obtained from blood samples and other biological materials as part of prospective cohort designs. Systems epidemiologic designs make it possible to look at the observed gene expression profiles resulting from the complex real-life situation, with hundreds of different exposures interacting with genetic predisposition and the risk of cancer.

Lund et al. [3] describe a processual approach to systems epidemiology that can be used for prospective designs. Instead of focusing on risk estimation, which is typical in classical analyses, the suggestion is to look backward from the time of diagnosis and observe the processes that lead up to disease. The observed changes should thus not only be viewed as static biomarkers and risk factors but part of a multistage process leading to cancer.

3.2 NOWAC

The Norwegian Woman and Cancer Study (NOWAC) [12] is a cohort study consisting of survey information from 170 000 women, and a biobank with more than 60 000 blood samples. NOWAC was initially started to study the connection between oral contraceptives and breast cancer but is now also used as a base for several other projects, such as lung cancer research in the Id-Lung project. The surveys and preserved blood samples from NOWAC enable epigenome-wide association studies for understanding the interaction between genetic and environmental factors and understanding the etiology of cancers.

One of the motivations for the study was to overcome the bias-problems with earlier case-control studies of the association between oral contraceptives and breast cancer. When the prospective cohort study started, it was also designed to cover many other risk factors (exposures). This deliberate design choice was made to facilitate hybrid designs.

NOWAC has several times invited new women to participate. Between 1991–97 altogether 179 387 women were invited of whom 102 540 women (57%) aged 30–70 years returned a questionnaire to the Institute of Community Medicine, University of Tromsø, Norway. During 2003–06 another 130 577 women born 1943–57 were invited of whom 63 232 (48.4%) returned an eight pages questionnaire. [12]

From 2003, the scope of the study was expanded by including biological material for gene expression profiling. This 'post-genome NOWAC cohort' includes both normal and malignant peripheral blood and breast tissue. From 2006, the collection of breast cancer biopsies was started. In 2007, tissue samples from healthy women in matched control groups were collected. [12]

NOWAC is also part of the European Prospective Investigation into Cancer and Nutrition (EPIC) [13][14] collaboration.

3.2.1 Passive and Active Follow-Up

NOWAC have followed up the enrolled women both passively and actively.

Passive follow-up is based on linkage to the Cancer Registry of Norway, and the register of death certificates. [12]

For active follow-up, NOWAC collaborated for the tumor sampling with 11 major Norwegian hospitals and the Norwegian Breast Cancer Group (NBCG), which consists of clinicians at most Norwegian hospitals [12]. The biopsies were stored frozen. An additional tube of citrate blood was collected together with a questionnaire of two pages. The active follow-up has now ended.

3.2.2 Timeline

The participants have answered questionnaires up to four times: One baseline, two follow-ups, and one around the time of blood sample collection. The exact questions that they were given depended on what series they were in because the questionnaires were changed several times over the years.

Figure 1 shows the timeline for the NOWAC surveys and blood samples. Women were recruited during different time periods. The baseline questionnaires were given at the time of recruitment, which happened three times between 1991-97, and five times between 2003-07. After around 3-9 years they answered the first follow-up survey, which is shown as green boxes in the figure. The third survey was answered 6-8 years after the second. In the figure, red drops are shown on the right side of some of the boxes. This is when the blood sample was collected for the given group of women, which also had an included survey. The number ranges shown in the colored boxes are the questionnaire series used for the group.

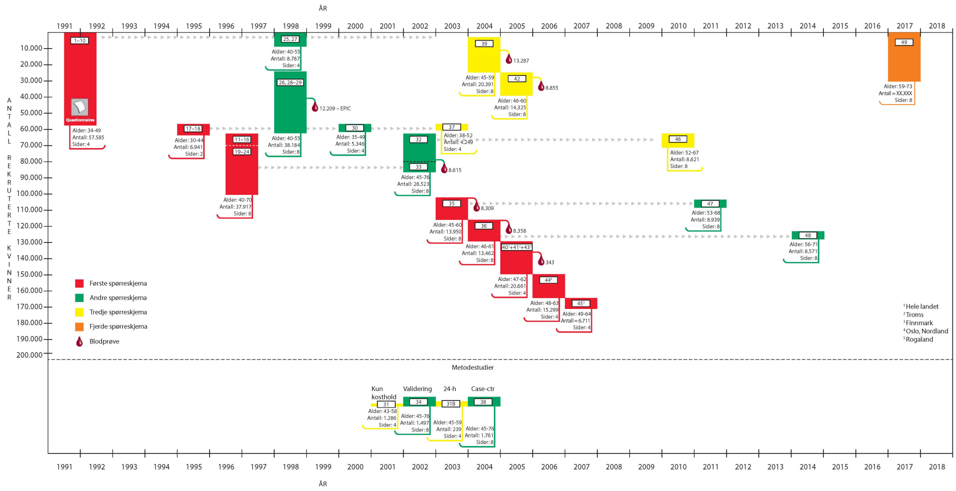


Figure 1: A larger version can be found in Appendix A. The figure shows the number of women recruited (red boxes), timing of second (green boxes) and third questionnaire mailings (yellow boxes) and collection of blood samples within the Norwegian Women and Cancer (NOWAC) study according to year of enrolment, age and length of questionnaires with number of blood samples in the EPIC and post-genome cohort biobanks.

3.3 Id-Lung

Our project is part of the Id-Lung [15] project that uses the NOWAC cohort, in combination with other cohorts. Id-Lung's main aim is to identify clinically relevant biomarkers of metastatic lung cancer through analysis of DNA methylation, gene expression and miRNA in blood. Id-Lung uses blood samples prior to clinical diagnosis obtained from the NOWAC cohort. The Id-Lung project team comprise a consortium consisting of six European institutions with expertise in systems epidemiology, bioinformatics/biostatistics, genomics, and cancer biology.

3.4 Data Wrangling Tools

This section presents and compares tools for wrangling data.

Before data can be analyzed, it is usually necessary to clean and transform the raw dataset into a more manageable format. Often, new values are calculated from the original data and added to the dataset. The process of adjusting the dataset to make it suitable for further analysis is called data wrangling. At present, data wrangling requires much manual work and is considered a time-consuming step in the data analysis workflow.

We chose to use the programming language R [6] for data wrangling. R is an open source programming language and environment popular among data scientists. It has powerful functionality for statistical analyses, including graphical visualizations. Even though R is somewhat specialized for data analytics and statistics, it is at the time of writing the 12th most popular programming language according to the TIOBE-index [16], which also includes general purpose languages.

Although it is common to use programming languages like R for data wrangling, there are disadvantages to the programmatic approach. The flexibility of a full-fledged scripting language also makes it more complicated to understand and use correctly. For an unseasoned programmer, it can be challenging to write and to reason about the code, and thus there is a high risk of making mistakes in the code. The programming environment usually gives little feedback to aid the programmer in verifying that the code behaves as expected.

An alternative is to use a specialized data wrangling tool that provides a more visual and interactive approach. Trifacta [17] is a company founded by researchers from Stanford and UC Berkeley. Their product, Trifacta Wrangler, promises faster and more intuitive data preparation. It lets users wrangle data through a graphical user interface that organizes and structures the data, and gives immediate visual feedback to the user. It also gives automatic suggestions on how to transform the data. By using the graphical interface, the user creates steps for transforming the data. Before adding a new transformation, Trifacta Wrangler previews the result, which makes it easy for the user to verify that the result is as expected. The user can create most of the steps via simple built-in functions. Trifacta Wrangler also includes a domain specific programming language that can be used to define customized steps when the built-in functions do not suffice. Before choosing a data wrangling tool, consideration should be made whether support for big datasets is required. Trifacta Wrangler is capable of wrangling big data through its Apache Spark [18], Hadoop [19], and cloud platform support.

Trifacta can perform data wrangling equivalent to the R-scripts for this project. However, for the dataset in this project, it is still a time consuming and complicated task. For example, it is complicated to calculate an average value from different columns in a row and to store the result in the same row. The main advantages are visual feedbacks, previews of transformations, automatic indications of problems and solution suggestions, and predefined functions for common tasks, all of which reduce the risk of making mistakes. Appendix B shows a screenshot of Trifacta Wrangler.

After completing the data wrangling, it can be interesting to explore the data further. Tableau [20] is a product for exploring, visualizing, and analyzing data via a graphical user interface. Tableau includes the data wrangling tool Tableau Prep but can also be used in combination with Trifacta and similar tools.

We do not give strong suggestions about which approach to choose, because all approaches have distinct advantages and disadvantages, but it is always useful to compare available options before deciding.

3.5 Testing and Evaluation

In this section, we introduce basic software testing concepts and how we evaluated our scripts.

It is human to err, and evaluation and testing is an integral part of the software development process. In the software industry, it is estimated that between 25% and 50% of the development time and cost is related to evaluation and testing [21].

There are many types of tests, which can be grouped into layers of abstraction. On the lowest abstraction level, the developer programs detailed tests for the automated verification of small functions and units of code. This is called unit testing. The developers who write the code should themselves write the unit tests. Unit tests are called white box or glass box because they require knowledge of the inner workings of the code. The purpose of unit tests and other tests at the lower abstraction levels is to verify the correctness of the code. However, unit tests can neither guarantee that the units work correctly together as a whole nor that the final program serves its intended purpose. It is, for example, possible to create a program that is technically flawless and at the same time failing to meet the needs of its users. Another example is a program containing logical errors that are not coding errors but errors that stem from incorrect reasoning or insufficient communication between humans in the planning phase. This is the reason for systematically testing at additional, higher levels of abstraction. At the higher levels, domain experts and people with other areas of expertise than the software developers will usually be involved. Black box testing is used, which means that the tests do not require knowledge of code details. The higher the abstraction level, the more manual testing is generally required.

We evaluated and tested our R-scripts in two ways:

1. We wrote test functions to test our functions. This is an example of unit testing.
2. A domain expert ran the scripts against another, smaller dataset. The results were compared to already known results that existed for this other dataset.

4 Project Design and Implementation

This project aims to categorize healthy women into groups based on smoking exposure, which in turn can be used for identifying transcriptomic changes associated with smoking. Lung cancer studies can then adjust for these changes to better identify lung cancer biomarkers.

The women in this project are a control group of 1766 women selected from NOWAC for a case-control study by Id-Lung. We use a de-identified dataset extracted from the NOWAC questionnaires with 82 data variables per woman (Appendix C). Of these, 77 are smoking-related.

This project has a cross-sectional study design. Cross-sectional study designs are prone to survival bias and misinterpretations caused by missing temporal relationships. Since we only use controls, which are healthy women, the typical types of bias related to cross-sectional study designs do not apply.

We have developed scripts using the programming language R to create five additional variables for smoking status, passive smoking status, smoking intensity, years of smoking, and time since cessation at the time of blood collection.

4.1 Combination of Multiple Questionnaire Surveys

The smoking status will be used to adjust for smoking in transcriptomic analyses. We therefore calculate smoking status at the time of blood sample collection. The women answered a questionnaire at the time of blood sampling, but to establish a more complete picture of the smoking statuses, we use data from the other questionnaires. The participants have answered one to three other surveys with several years in between. The number of years between surveys vary greatly and depends on the recruitment time and participant series (Figure 1). The blood sample was given either after the first, second, or third survey. We are not interested in follow-up surveys given after the blood sampling. Consequently, we must determine the last survey the woman answered before giving blood. We ignore follow-up surveys answered by the woman after the time of blood sample acquisition.

To alleviate recall bias, we combine the data in all questionnaires before the blood sample. There are many overlapping questions in the different surveys, and often the participants answered slightly differently for the corresponding question in two or more surveys. By comparing the answers or calculating average values, we can make more accurate estimates. For example, by comparing a woman's answers about smoking habits in different questionnaires, we can discover that she previously was a smoker but no longer is. Average smoking intensities can be calculated from smoking intensity variables in different questionnaires.

Missing data is also an issue, but the information is often in relevant answers in a different questionnaire.

4.2 Multiple Variables with Similar Information

For each follow-up stage in NOWAC, the participants have answered a new questionnaire. For the different participant series, different versions of questionnaires have been used for the same follow-up stage. Consequently, different women have been asked slightly different sets of questions depending on when they were recruited. The result is that many variable values are missing per woman. Often the reason for a missing value is that the woman has not been asked the question assigned to that particular variable, but instead a similar question assigned to another variable. Therefore, the script must check multiple variables per participant to find the value of interest in our calculations.

4.3 Smoking Status

Smoking status describes the person's smoking habits. We define four possible values for smoking status:

- Current
- Former
- Never
- Undefined (*NA*)

The implementation of smoking status is included in Appendix H.

Determining the value for a current smoker is straightforward. Only the values from the blood sample questionnaire are used. If the person answered that she did smoke, then she is classified as a current smoker. If the answer is missing, we check values for two questions telling how many cigarettes she smoked today and yesterday. If she had smoked both days, then she is classified as a current smoker. We decided to require both to be greater than zero because we did not want to conclude too strongly. If a woman had not answered the question for current smoking and also had replied that she only smoked one of the days, it is possible that she was an occasional smoker that rarely smoked.

To identify the never-smokers, from the women not already classified as current smokers, we use variables for questionnaires answered earlier than the blood sample date. Women who consistently have answered that they do not smoke and never have are good candidates for being classified as never-smokers. Additionally, we use values for the historical average smoking intensities. We estimate the average smoking intensity for a woman based on her answers about the number of cigarettes per day at different ages. By using this information, we avoid misclassification in cases when data are missing for the other variables. If the estimated average smoking intensity is consistent with never having smoked, then the woman is finally categorized as a never-smoker.

We identify former smokers from women not classified as current or never smokers. Those who answered yes to ever having smoked, in any of the questionnaires, are classified as former smokers. To find former smokers, we must additionally find women that were current smokers at one point during the study, and then had quit at a later follow-up. If the blood sample questionnaire stated that someone did not smoke, but we found answers in surveys prior to the blood sample saying that the person was a smoker or had smoked at some time in their life, then this person was a former smoker.

4.4 Passive Smoking

In addition to smoking status, we want information about passive smoking for the individuals, but we quickly realized that mixing data about smoking and passive smoking into one variable would be troublesome. A decision was thus made to keep them separate. It is easy to combine the two variables if needed at a later stage. The implementation of passive smoking status is included in Appendix I.

To illustrate the problems in combining the two values, assume that a former smoker worked at a place where she was exposed to smoking. Should this person be given the status as a former smoker or current passive smoker? Or, what if a woman was exposed to passive smoking in childhood. Should she be counted as a former smoker, even though she otherwise would count as a never-smoker? We could create smoking statuses that captured all combinations, but the number of different types of statuses would quickly become impractical.

Exposure to passive smoking can come from many sources. Our dataset contains variables concerning passive smoking gathered from a variety of surveys for the different series. Not all women had been asked the same set of questions and the questions differ among the series. Generally, the data for passive smoking can be categorized into questions about childhood, adulthood home environment, and workplace. For passive smoking, we define five categories:

- Exposed in childhood
- Exposed in adulthood
- Exposed both in childhood and adulthood
- Never exposed
- Undefined (*NA*)

We calculate these statuses by using variables from different questionnaires. As with smoking status, we first determine which of the questionnaires (baseline, follow-up one, or follow-up two) is closest to the blood sample collection, as there is no data about passive smoking in the blood sample questionnaire. We use all questionnaires prior to a woman's blood sample in our calculations.

Then we find all participants that have answered at least once that they were exposed to passive smoking during childhood.

In the next step, we select all who were non-exposed to smoking in childhood. If they have somewhere answered that they were exposed at home or at work, then we classify them as adulthood passive smoker.

For this calculation, as with most of the others, several variables must be checked for a variety of values. For example, the values for variables ROKBOR, yROKBOR, and yROKARB are checked if the closest questionnaire was the second or third. The variable ROKBOR belongs to the first survey. It tells if the participant was exposed to passive smoking at home. The variable yROKBOR is the corresponding variable found in the second survey. The second survey also includes information about passive smoking in the workplace in the variable yROKARB.

To illustrate how many variables typically are encoded in the dataset, the possible values for yROKBOR are listed here:

- 0 = Yes
- 1 = No
- 2 = Living with an ex-smoker
- 3 = Living with a cigarette smoker
- 4 = Living with a pipe smoker
- 9 = Living with a party smoker

- 98 = Uncertain

To find the women that were both exposed to passive smoking as a child and as an adult, we find all persons already classified as childhood passive smokers, and then, in addition, apply the same criteria as used for adulthood passive smoking. This combination overwrites the status of persons already classified as childhood passive smokers if they were also adulthood passive smokers so that the status is set to a value indicating both.

Finally, the non-exposed group are the women that have not already been classified, and that never had answered yes to exposure in childhood, at home as an adult, or at their workplace.

The remaining persons are not assigned a passive smoking status but are left with the empty value NA.

4.5 Exposure Levels

We believe that the physiological changes associated with smoking may depend on the dosage and the time since exposure, just as other chemical substances typically do. It is therefore useful to estimate how much, when, and for how long a person smoked. We use these values to calculate a smoking index; a number that is an indication of the exposure level. For this purpose, we calculate three variables:

- Average smoking intensity (cigarettes per day)
- Number of years as a smoker (duration)
- Number of years since cessation

4.5.1 Smoking Duration

The smoking duration tells how many years a woman has smoked. If the participant was a current smoker, then we subtract the age at blood sample from the start age of smoking to find the duration. For former smokers, the number of years since cessation must be found in addition to the age at blood sample and start age of smoking. The implementation is included in Appendix K.

4.5.2 Age at Blood Sample

The woman's age at blood sample is needed to estimate how long she has been a smoker or years since quit. The script code is included in Appendix J.

To find the woman's age at blood sample, we subtract the year that the blood sample was collected from the birth year. The blood sample date is stored as an integer with the day as the most one or two significant digits, then the month as the two next digits, and then the year as the least two significant digits. To extract the two year-digits, we simply use a modulo of 100 to get the remainder. This is converted to a four digit year by either adding 1900 or 2000 depending on the size of the two last digits to determine the correct millennium. The birth year is given as two digits, and since none of the participants were born in the current millennium, we just add 1900. The two four-digit years are subtracted to find the age at blood sample.

4.5.3 Start and Cessation Age

To calculate start and cessation age, several variables from the dataset are used. Some of variables are straightforward with simple values, others are complex sets of age intervals. The implementation is included in Appendix L and Appendix E.

To find the start age we first inspect the simple variables. Such variables exist in the baseline and the two follow-up questionnaires. We start with the questionnaire before blood sample collection. In total, there are four possible simple variables for start age: SIGALDER, YSIGALDER, YSIGALDERB, ZSIGALDER.

To reduce recall bias, we calculate an average the variables concerning the time before blood sample collection.

In some cases, the only variable with any value is from the last questionnaire, even though the blood sample had been taken earlier than this. If we lack all other simple variables, then we use this variable (ZSIGALDER).

If we still have not found a start age, we continue further to a more complicated approach involving age intervals for smoking.

The baseline and first follow-up questionnaires have sets of variables for smoking intensity between different age intervals. Three sets of variables for the baseline questionnaire can contain this kind of data, depending on which series the participant belong. One set of variables has been registered for the second questionnaire. The second questionnaire is only used if the blood sample was taken at a later point. The variables are shown in table 1.

Variable (Age Interval)			
ROK1 (10-14)	ROYKANT1014 (10-14)	ROYKANT1 (10-19)	yROYKANT1 (10-19)
ROK2 (15-19)	ROYKANT1519 (15-19)	ROYKANT2 (20-29)	yROYKANT2 (20-29)
ROK3 (20-24)	ROYKANT2029 (20-29)	ROYKANT3 (30-39)	yROYKANT3 (30-39)
ROK4 (25-29)	ROYKANT3039 (30-39)	ROYKANT4 (40-49)	yROYKANT4 (40-49)
ROK5 (30-34)	ROYKANT4049 (40-49)	ROYKANT5 (50-59)	yROYKANT5 (50-59)
ROK6 (35-39)	ROYKANT50MM (50-more)	ROYKANT6 (60-69)	yROYKANT6 (60-69)
ROK7 (40-44)			
ROK8 (45-49)			

Table 1: The table shows the different variables for smoking intensity at different age intervals. For each column, the variable name is to the left and the age interval in parenthesis to the right of the variable name.

To find the start and quit age, we inspect these variables, from youngest to oldest age, to find the first non-empty value, which is the interval for the start age. Then we continue until we find an empty value or the end of the interval. This is a candidate for the quit age interval, for former smokers.

The approach to finding the quit/cessation age is much the same as for start age. The time of cessation is only calculated for former smokers.

4.5.4 Smoking Intensity

To calculate the average intensity of smoking, we use both simple values and intensities for age intervals. The implementation is included in Appendix G.

Two simple variables available are ZROKSIST5 and ZROKSIST8, which belong to the second follow-up questionnaire. They give the values for smoking intensities for either the last 5 or 8 years. In all the variables, including those for age intervals, the number of cigarettes is not given directly but encoded as a number or option representing an interval from-to a number of cigarettes per day. We have created a function to convert between an encoded number, and select the average of that interval. The conversions to the number of cigarettes are given in table 2.

Value	Intensity
0	1-300 per year
1	1-4 per day
2	5-9 per day
3	10-14 per day
4	15-19 per day
5	20-24 per day
6	25 or more per day
98	Uncertain

Table 2: The table shows the mapping between the numeric values for the age intervals and the number of cigarettes per day.

The value 98 has a special meaning (uncertain) and we must make sure to ignore this value in our calculations. Note that most intervals are 10 years of length, but some are 5 years. We have to take this into account when we calculate smoking intensity averages. This is solved by combining two 5-year intervals to one 10-year interval by taking the average and creating a new variable for the combined interval.

5 Results and Discussion

The goal of the data wrangling scripts is to create five new variables for each woman in the dataset. We first present the resulting variable values and then discuss information bias, registration errors, ambiguous variables, zero or non-available values, and missing values.

The variables and possible values created by the scripts are:

- SmokingStatus (Smoking Habits)
  - Current
  - Former
  - Never
  - NA
- PassiveSmoking (Exposure to Passive Smoking)
  - Passive_childhood
  - Passive_adulthood
  - Passive_both
  - Not_PassiveSmoker
  - NA
- SmokingDuration (Number of Years as a Smoker)
  - Decimal
  - NA
- Intensity (Average Cigarettes per Day)
  - Decimal
  - NA
- TSC (Time Since Quit in Years)
  - Integer
  - NA

5.1 Results

Table 3 shows the distribution of the women in the four smoking status classes. 25% were current smokers, 38% were former smokers, and 36% were never-smokers. 1% (22 women) were unclassified. Most of these 22 were likely never-smokers or former smokers with low intensity, but we could not conclude because of missing data. Data could be missing because the women had not answered some of the required questions or due to errors introduced during data registration.

Smoking Status	No. of Women
Current	433
Former	673
Never	638
NA	22
Sum	**1766**

Table 3: The table shows the number of women matching each smoking status. NA means that the status was non-conclusive.

Table 4 shows how many women who were classified as having the different *passive* smoking statuses. About 12% had not been exposed to passive smoking, 10% could not be classified with certainty, and a total of 78% had been exposed to passive smoking at some point. Most of the participants, 73%, had been exposed in childhood. About 25% had been exposed as an adult, either at home or at work, and about 20% had been exposed both as a child and in adult life.

Passive Smoking	No. of Women
Both Childhood and Adulthood	358
Only Childhood	923
Only Adulthood	91
Non-exposed	222
NA	172
Sum	**1766**

Table 4: The table shows the number of women matching each passive smoking smoking status. NA means that the status was non-conclusive.

Table 5 shows the number of cigarettes per day the women had smoked on average throughout all their smoking years. For 600 women, the smoking intensity was undefined (NA). Of these, 561 were never-smokers. The value NA is used for never-smokers instead of zero because the value 0 has a special meaning. The value 0 means less than 300 cigarettes per year, not zero cigarettes. For the remaining 39 women, the script could not calculate smoking intensity due to insufficient data. This can be seen from Table 6 for smoking status and intensity. Most notably, 77 of the never-smokers had an intensity of 0, which could mean that they have smoked some cigarettes in their life. The exact intensity is difficult to know for these, since the value of zero could mean anything from one to three hundred cigarettes per year. It could also be a registration error, meaning that the real intensity was actually zero cigarettes. Since they have answered that they were never-smokers, we believe that they probably did not smoke in the upper range of this interval per year.

Average Number of Cigarettes Per Day	No. of Women
0*	151
2.5	329
7	328
12	241
17	90
22	21
27	6
NA	600
Sum	**1766**

* 0 = Less than 300 per year

Table 5: The table shows the number of women who matched the estimated average number of cigarettes per day.

Smoking Status	**0***	**2.5**	**7**	**12**	**17**	**22**	**27**	**NA**	**Sum**
Current	8	55	148	146	54	11	2	9	433
Former	66	270	180	95	36	10	4	12	673
Never	77	0	0	0	0	0	0	561	638
NA	0	4	0	0	0	0	0	18	22
Sum	151	329	328	241	90	21	6	600	1766

* 0 = Less than 300 cigarettes per year

Table 6: The contingency table shows the number of women for the smoking intensities and statuses.

Figure 2 and 3 shows the cumulative frequencies for smoking duration for current and former smokers. Less than 5.4% of the current smokers had smoked less than 30 years. Smoking durations for former smokers were more evenly distributed, and generally lower than for current smokers. In this group, 71.4% had smoked less than 30 years.

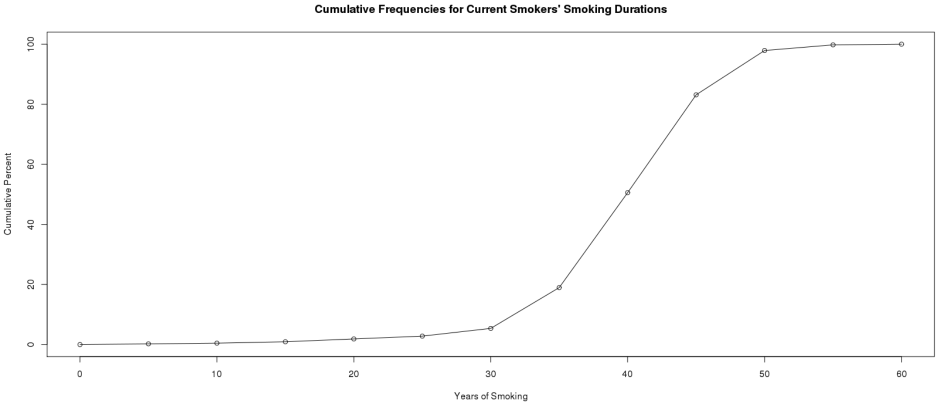


Figure 2: Cumulative frequency graph for current smokers' number of years as smokers

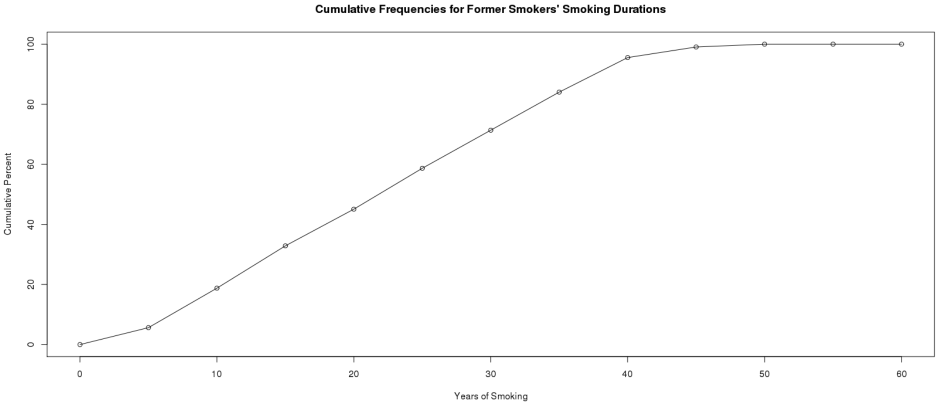


Figure 3: Cumulative frequency graph for former smokers' number of years as smokers

As seen in figure 4 and 5, smoking intensities were also different for current and former smokers. More of the former smokers smoked less per day than the current smokers when they were smokers. It may be important to note that time since exposure is not the only difference between former and current smokers, but that dosage and duration differ as well.

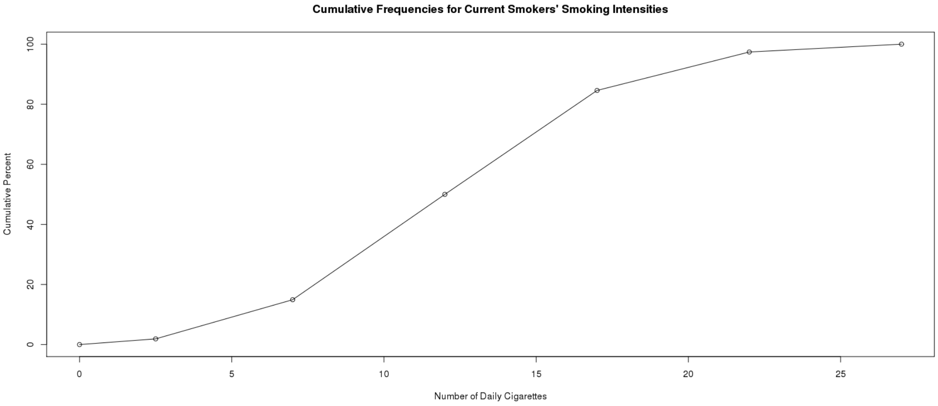


Figure 4: Cumulative frequency graph for current smokers' average number of cigarettes per day

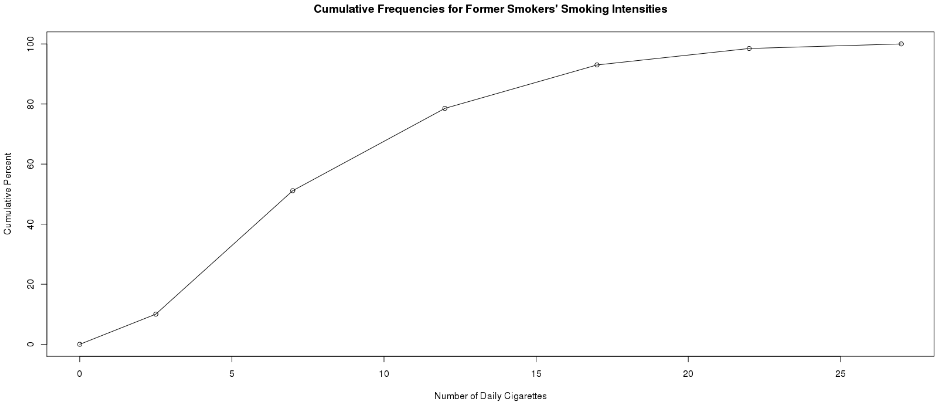


Figure 5: Cumulative frequency graph for former smokers' average number of cigarettes per day

5.2 Information Bias

In this subsection, we discuss how information bias may affect our results. Recall bias is a type of information bias that is likely to affect all answers about past events. Questions about passive smoking in childhood, the first cigarette, age at onset of smoking, the number of cigarettes per day in a period of years, and time of cessation can all be challenging to remember exactly in retrospect. Even questions about recent events can be wrong, such as the answers about passive smoking exposure at the workplace, or how much a person smoked yesterday. Our general strategy was to combine data from several questionnaires and calculate averages when possible.

5.3 Registration Errors

Registration errors can be a challenge when working with questionnaire data. In the earlier years of the NOWAC study, answers from the questionnaires were manually registered (punched) into the data system. Later, answers were scanned automatically. In both cases, errors can occur. One example is that we discovered an individual who was 3 years old when she started smoking. We thought that this seemed unlikely and cross-checked with a corresponding answer from another survey. There the start age was given as 23. This age is much more probable, and the age of 3 was probably a punching error where one digit was missing. We assume that calculating averages reduces the problem, and that these types of outliers would be accounted for in more robust statistical computations in later stages.

5.4 Ambiguous Variables

The meaning of a variable value can be unclear or inconsistent. During our work, we discovered that the interpretation of the data in some cases could be ambiguous. Answers from different series of questionnaires had been stored in the same variable, or in different variables that we treated as having the same meaning. Unfortunately, the questions that the women had been asked in the different series sometimes had slightly different meanings. One example is that the questions ”when did you start smoking” and ”when did you smoke your first cigarette” are slightly different. A person could have taken their first cigarette at some early age, but not started smoking regularly for many years later. In our calculations, we used average values based on several variables.

5.5 Zero or Not Available

There is a differences between the value 0 and the value *NA* in the dataset. In time intervals regarding smoking between ages X and Y, zero could mean anything from 1 to 300 cigarettes per year. We thus had to include the intervals with 0 as a period of smoking. The value *NA* was interpreted as no smoking and was not counted in the calculations.

5.6 Missing Values

In this subsection, we discuss how we handled that values required for the estimation a result were missing. When none of the needed values for calculating our variables were available, we did not try to impute them. Instead, we left those variables having *NA*-values for the given person. We believe that it would be worse to impute wrong results based on incomplete data, than to leave these as empty. Empty values can easily be filtered at a later stage, while wrong values can pollute future data analyses that are based on our results.

6 The Road Ahead

In this section, we outline possible future work and the road ahead.

In our results, we have generated variable values for intensity, duration, and time since cessation. These three variables can be used in a formula for calculating a comprehensive smoking index (CSI) for lung cancer [22]. The mathematical representation of smoking history is an important tool in the analysis of epidemiological and clinical data. CSI is a single aggregate measure of smoking exposure that incorporates intensity, duration, and time since cessation. This index, which may be incorporated in any regression model, also depends on a half-life and a lag parameter that have to be fixed a priori or estimated by maximizing the fit.

The CSI-values, the smoking statuses, and the information about passive smoking exposures are useful for systems epidemiologic analyses related to smoking exposure. The aim is then to assess if there are differentially expressed genes related to different levels of measures of smoking habits in blood from disease-free women. After changes in gene-expressions have been identified, the results can potentially be applied in analyses of women with lung cancer to differentiate between gene-expressions changes caused by smoking exposure and lung cancer.

7 Conclusion

In this report, we have estimated smoking status, passive smoking, smoking intensity, years of smoking, and years since cessation for a control group of 1766 women from the NOWAC cohort. We have described the motivation, background, design, implementation, results, and the road ahead for the project. To estimate the variable values, we developed scripts in the R programming language. The results from the scripts can be used in future work in conjunction with lung cancer research.

References

[1] Laura Baglietto, Erica Ponzi, Philip Haycock, Allison Hodge, Manuela Bianca Assumma, Chol-Hee Jung, Jessica Chung, Francesca Fasanelli, Florence Guida, Gianluca Campanella, et al. Dna methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. *International journal of cancer*, 140(1):50–61, 2017.

[2] Francesca Fasanelli, Laura Baglietto, Erica Ponzi, Florence Guida, Gianluca Campanella, Mattias Johansson, Kjell Grankvist, Mikael Johansson, Manuela Bianca Assumma, Alessio Naccarati, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nature communications*, 6:10192, 2015.

[3] Eiliv Lund, Sandra Plancade, Gregory Nuel, Hege Bøvelstad, and Jean-Christophe Thalabard. A processual model for functional analyses of carcinogenesis in the prospective cohort design. *Medical hypotheses*, 85 (4):494–497, 2015.

[4] Sciencedirect about transcriptome. `https://www.sciencedirect.com/topics/neuroscience/transcriptome`. Accessed: 2018-14-12.

[5] Data wrangling. `https://en.wikipedia.org/wiki/Data_wrangling`. Accessed: 2018-12-12.

[6] R programming language. `https://www.r-project.org/`. Accessed: 2018-12-12.

[7] About de-identified data. `https://www.hopkinsmedicine.org/institutional_review_board/hipaa_research/de_identified_data.html`. Accessed: 2018-12-12.

[8] Leon Gordis. *Epidemiology.* Saunders, 5th edition, 2014.

[9] Miquel Porta. *A Dictionary of Epidemiology.* Oxford University Press, 5th edition, 2008.

[10] Eiliv Lund and Vanessa Dumeaux. Systems epidemiology in cancer. *Cancer Epidemiology and Prevention Biomarkers*, 17(11):2954–2957, 2008.

[11] The syngenta innovation centre on systems biology, imperial college london. `https://www.imperial.ac.uk/syngenta-uic`. Accessed: 2019-10-01.

[12] Eiliv Lund, Vanessa Dumeaux, Tonje Braaten, Anette Hjartåker, Dagrun Engeset, Guri Skeie, and Merethe Kumle. Cohort profile: the norwegian women and cancer study—nowac—kvinner og kreft. *International journal of epidemiology*, 37(1):36–41, 2007.

[13] Sheila Bingham and Elio Riboli. Diet and cancer—the european prospective investigation into cancer and nutrition. *Nature Reviews Cancer*, 4(3):206, 2004.

[14] N Slimani, R Kaaks, P Ferrari, C Casagrande, F Clavel-Chapelon, G Lotze, A Kroke, D Trichopoulos, A Trichopoulou, C Lauria, et al. European prospective investigation into cancer and nutrition (epic) calibration study: rationale, design and population characteristics. *Public health nutrition*, 5(6b):1125–1145, 2002.

[15] Id-Lung research group. `https://en.uit.no/forskning/forskningsgrupper/gruppe?p_document_id=507532`. Accessed: 2018-12-12.

[16] TIOBE index. `https://www.tiobe.com/tiobe-index/`. Accessed: 2019-14-01.

[17] Trifacta. `https://www.trifacta.com/`. Accessed: 2019-14-01.

[18] Apache spark. `https://spark.apache.org/`. Accessed: 2019-14-01.

[19] Apache hadoop. `https://hadoop.apache.org/`. Accessed: 2019-14-01.

[20] Tableau. `https://www.tableau.com/`. Accessed: 2019-14-01.

[21] Tim Koomen and Martin Pol. *Test Process Improvement: A step-by-step guide to structured testing*. Addison-Wesley, 1999.

[22] Karen Leffondré, Michal Abrahamowicz, Yongling Xiao, and Jack Siemiatycki. Modelling smoking history using a comprehensive smoking index: application to lung cancer. *Statistics in medicine*, 25(24):4132–4146, 2006.

Appendices

A NOWAC Timeline

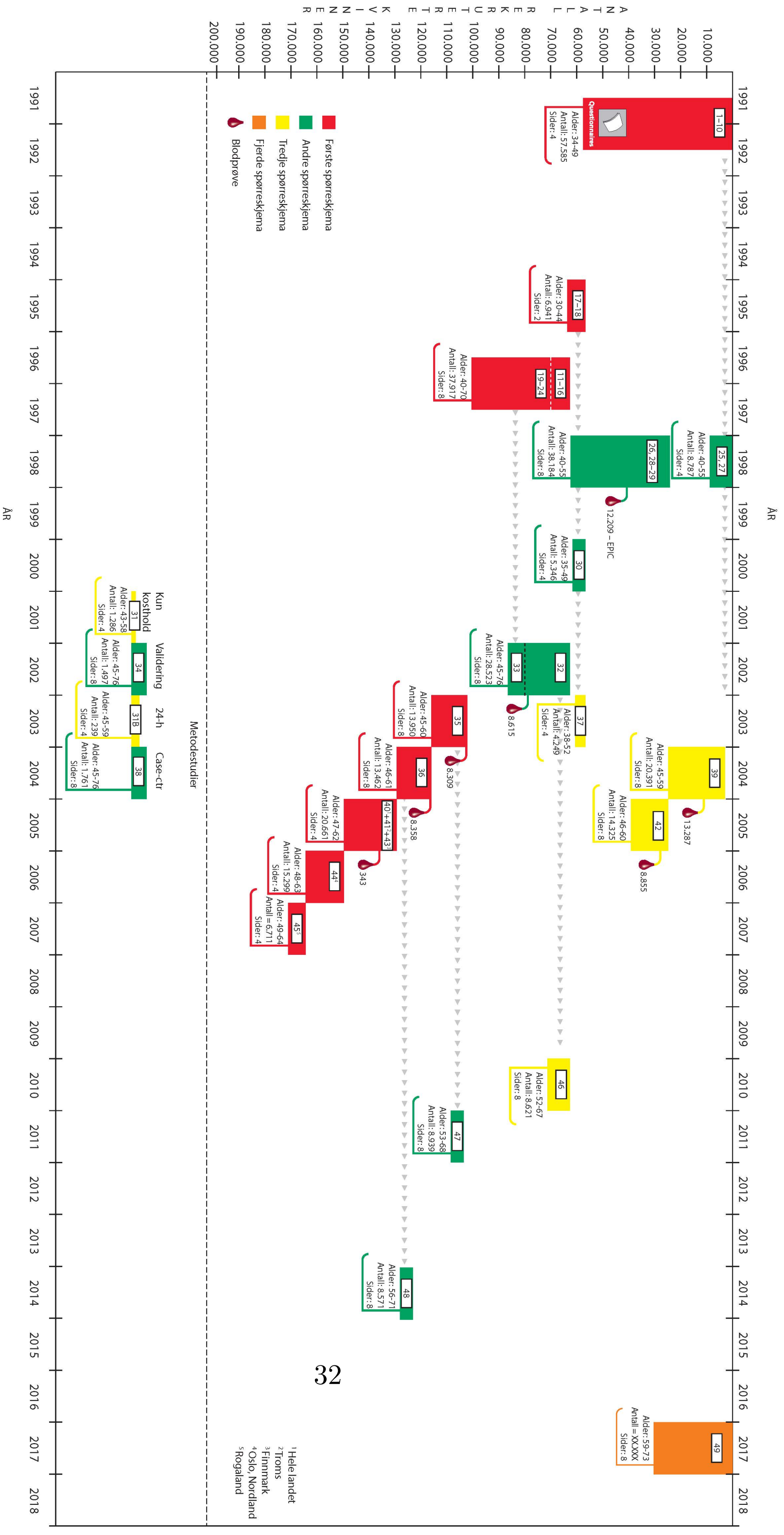
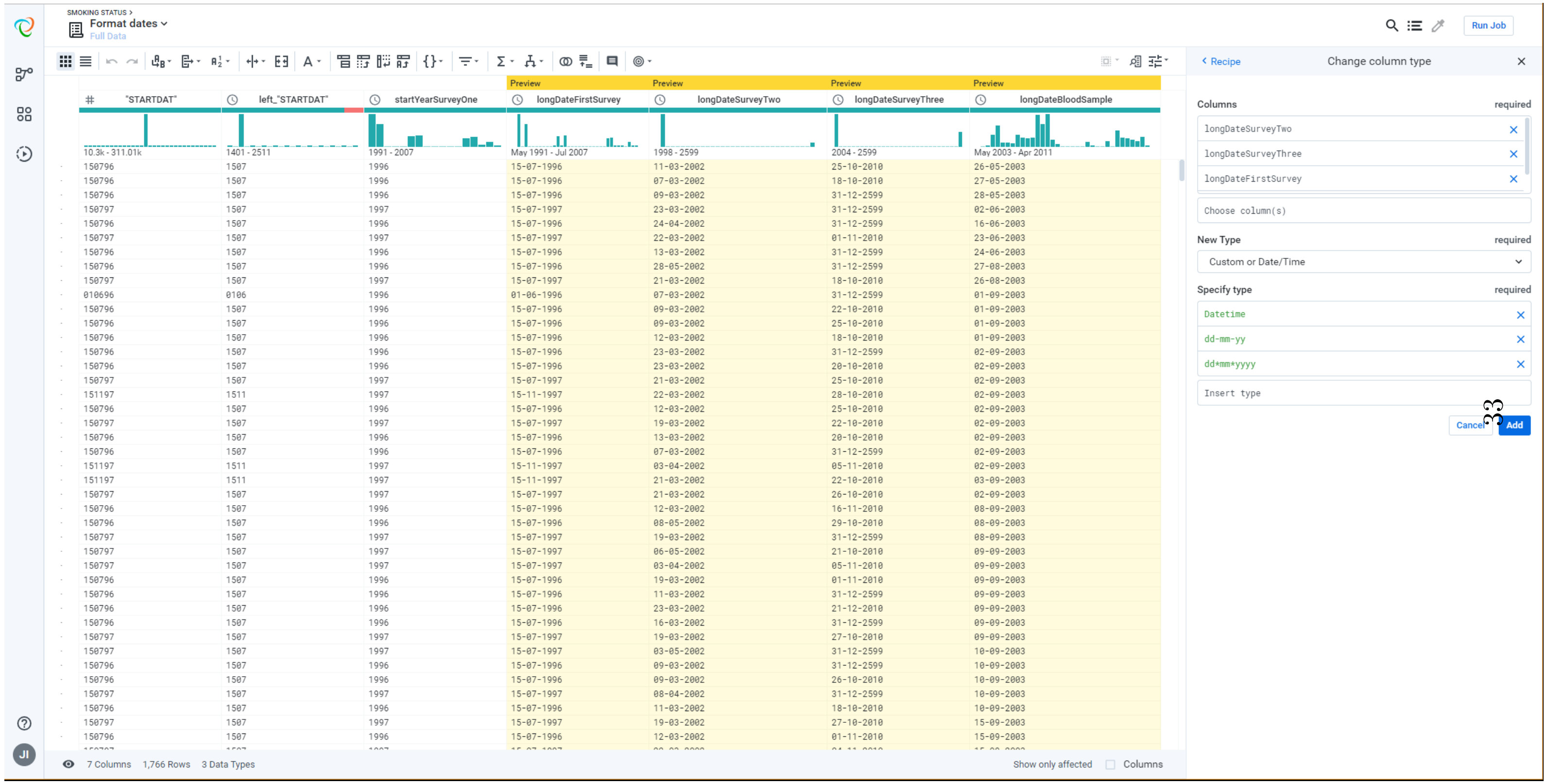


Figure 6: Number of women recruited (red boxes), timing of second (green boxes) and third questionnaire mailings (yellow boxes) and collection of blood samples within the Norwegian Women and Cancer (NOWAC) study according to year of enrolment, age and length of questionnaires with number of blood samples in the EPIC and post-genome cohort biobanks

B Trifacta Screenshot



			Preview	Preview	Preview	Preview
# "STARTDAT"	left_"STARTDAT"	startYearSurveyOne	longDateFirstSurvey	longDateSurveyTwo	longDateSurveyThree	longDateBloodSample
10.3k - 311.01k	1401 - 2511	1991 - 2007	May 1991 - Jul 2007	1998 - 2599	2004 - 2599	May 2003 - Apr 2011
150796	1507	1996	15-07-1996	11-03-2002	25-10-2010	26-05-2003
150796	1507	1996	15-07-1996	07-03-2002	18-10-2010	27-05-2003
150796	1507	1996	15-07-1996	09-03-2002	31-12-2599	28-05-2003
150797	1507	1997	15-07-1997	23-03-2002	31-12-2599	02-06-2003
150796	1507	1996	15-07-1996	24-04-2002	31-12-2599	16-06-2003
150797	1507	1997	15-07-1997	22-03-2002	01-11-2010	23-06-2003
150796	1507	1996	15-07-1996	13-03-2002	31-12-2599	24-06-2003
150796	1507	1996	15-07-1996	28-05-2002	31-12-2599	27-08-2003
150797	1507	1997	15-07-1997	21-03-2002	18-10-2010	26-08-2003
010696	0106	1996	01-06-1996	07-03-2002	31-12-2599	01-09-2003
150796	1507	1996	15-07-1996	09-03-2002	22-10-2010	01-09-2003
150796	1507	1996	15-07-1996	09-03-2002	25-10-2010	01-09-2003
150796	1507	1996	15-07-1996	12-03-2002	18-10-2010	01-09-2003
150796	1507	1996	15-07-1996	23-03-2002	31-12-2599	02-09-2003
150796	1507	1996	15-07-1996	23-03-2002	20-10-2010	02-09-2003
150797	1507	1997	15-07-1997	21-03-2002	25-10-2010	02-09-2003
151197	1511	1997	15-11-1997	22-03-2002	28-10-2010	02-09-2003
150796	1507	1996	15-07-1996	12-03-2002	25-10-2010	02-09-2003
150797	1507	1997	15-07-1997	19-03-2002	22-10-2010	02-09-2003
150796	1507	1996	15-07-1996	13-03-2002	20-10-2010	02-09-2003
150796	1507	1996	15-07-1996	07-03-2002	31-12-2599	02-09-2003
151197	1511	1997	15-11-1997	03-04-2002	05-11-2010	02-09-2003
151197	1511	1997	15-11-1997	21-03-2002	22-10-2010	03-09-2003
150797	1507	1997	15-07-1997	21-03-2002	26-10-2010	02-09-2003
150796	1507	1996	15-07-1996	12-03-2002	16-11-2010	08-09-2003
150796	1507	1996	15-07-1996	08-05-2002	29-10-2010	08-09-2003
150797	1507	1997	15-07-1997	19-03-2002	31-12-2599	08-09-2003
150797	1507	1997	15-07-1997	06-05-2002	21-10-2010	09-09-2003
150797	1507	1997	15-07-1997	03-04-2002	05-11-2010	09-09-2003
150796	1507	1996	15-07-1996	19-03-2002	01-11-2010	09-09-2003
150796	1507	1996	15-07-1996	11-03-2002	31-12-2599	09-09-2003
150796	1507	1996	15-07-1996	23-03-2002	21-12-2010	09-09-2003
150796	1507	1996	15-07-1996	16-03-2002	31-12-2599	09-09-2003
150797	1507	1997	15-07-1997	19-03-2002	27-10-2010	09-09-2003
150797	1507	1997	15-07-1997	03-05-2002	31-12-2599	10-09-2003
150796	1507	1996	15-07-1996	09-03-2002	31-12-2599	10-09-2003
150796	1507	1996	15-07-1996	09-03-2002	26-10-2010	10-09-2003
150797	1507	1997	15-07-1997	08-04-2002	31-12-2599	10-09-2003
150796	1507	1996	15-07-1996	11-03-2002	18-10-2010	10-09-2003
150797	1507	1997	15-07-1997	19-03-2002	27-10-2010	15-09-2003
150796	1507	1996	15-07-1996	12-03-2002	01-11-2010	15-09-2003

Change column type

Columns required
- longDateSurveyTwo
- longDateSurveyThree
- longDateFirstSurvey
- Choose column(s)

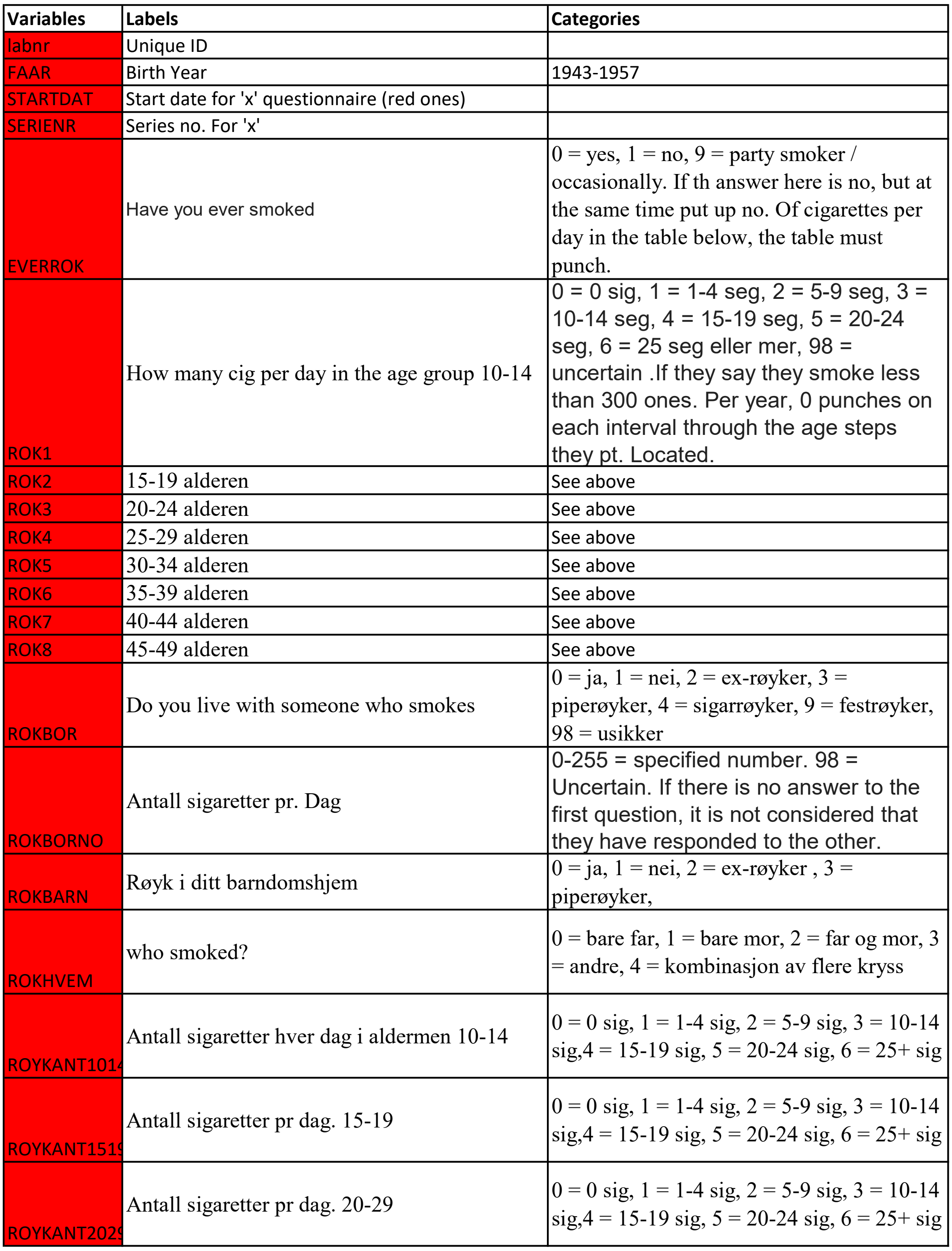
New Type required
- Custom or Date/Time

Specify type required
- Datetime
- dd-mm-yy
- dd*mm*yyyy
- Insert type

Cancel Add

Figure 7: Screenshot of Trifacta Wrangler. This is a cloud version accessed via a browser. The yellow columns are previews of the changes made by the step defined on the right side of the screen.

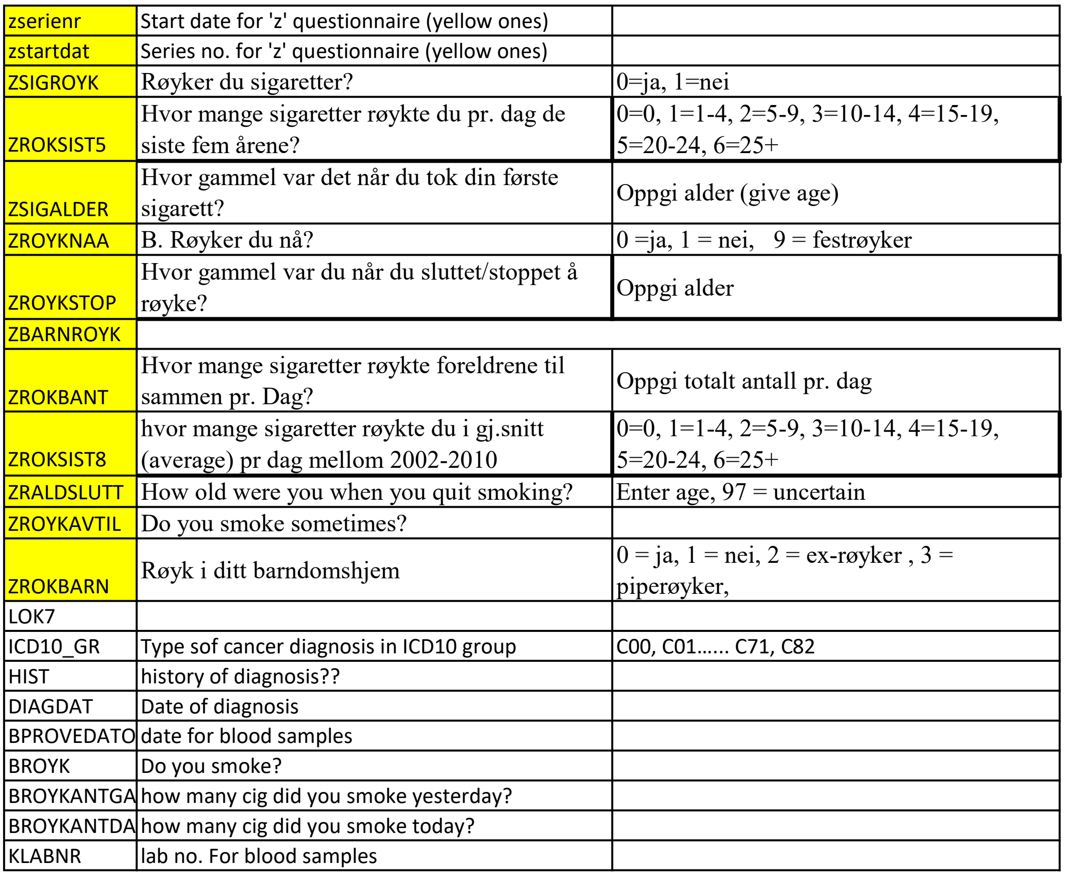
C Smoking Variables



Variables	Labels	Categories
labnr	Unique ID	
FAAR	Birth Year	1943-1957
STARTDAT	Start date for 'x' questionnaire (red ones)	
SERIENR	Series no. For 'x'	
EVERROK	Have you ever smoked	0 = yes, 1 = no, 9 = party smoker / occasionally. If th answer here is no, but at the same time put up no. Of cigarettes per day in the table below, the table must punch.
ROK1	How many cig per day in the age group 10-14	0 = 0 sig, 1 = 1-4 seg, 2 = 5-9 seg, 3 = 10-14 seg, 4 = 15-19 seg, 5 = 20-24 seg, 6 = 25 seg eller mer, 98 = uncertain .If they say they smoke less than 300 ones. Per year, 0 punches on each interval through the age steps they pt. Located.
ROK2	15-19 alderen	See above
ROK3	20-24 alderen	See above
ROK4	25-29 alderen	See above
ROK5	30-34 alderen	See above
ROK6	35-39 alderen	See above
ROK7	40-44 alderen	See above
ROK8	45-49 alderen	See above
ROKBOR	Do you live with someone who smokes	0 = ja, 1 = nei, 2 = ex-røyker, 3 = piperøyker, 4 = sigarrøyker, 9 = festrøyker, 98 = usikker
ROKBORNO	Antall sigaretter pr. Dag	0-255 = specified number. 98 = Uncertain. If there is no answer to the first question, it is not considered that they have responded to the other.
ROKBARN	Røyk i ditt barndomshjem	0 = ja, 1 = nei, 2 = ex-røyker , 3 = piperøyker,
ROKHVEM	who smoked?	0 = bare far, 1 = bare mor, 2 = far og mor, 3 = andre, 4 = kombinasjon av flere kryss
ROYKANT1014	Antall sigaretter hver dag i aldermen 10-14	0 = 0 sig, 1 = 1-4 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig
ROYKANT1519	Antall sigaretter pr dag. 15-19	0 = 0 sig, 1 = 1-4 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig
ROYKANT2029	Antall sigaretter pr dag. 20-29	0 = 0 sig, 1 = 1-4 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig

ROYKANT3039	Antall sigaretter pr dag. 30-39	0 = 0 sig, 1 = 1-4 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig
ROYKANT4049	Antall sigaretter pr dag. 40-49	0 = 0 sig, 1 = 1-4 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig
ROYKANT50M	Antall sigaretter pr dag. More than or equal to 50	0 = 0 sig, 1 = 1-4 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig
ROYKNAA	B. Røyker du nå?	0 =ja, 1 = nei, 9 = festrøyker
ROKBANT	Hvor mange sigaretter røykte foreldrene til sammen pr. Dag?	Oppgi totalt antall pr. dag
ROYKNAAB	??	0 , 1
ROYKANT1	10-19 alderen: s95/96	0 = 0 sig, 1 = 1-5 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig
ROYKANT2	20-29 alderen: s95/96	0 = 0 sig, 1 = 1-5 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig
ROYKANT3	30-39 alderen: s95/96	0 = 0 sig, 1 = 1-5 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig
ROYKANT4	40-49 alderen: s95/96	0 = 0 sig, 1 = 1-5 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig
ROYKANT5	50-59 alderen: s95/96	0 = 0 sig, 1 = 1-5 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig
ROYKANT6	60-69 alderen: s95/96	0 = 0 sig, 1 = 1-5 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig
SIGROYK	Røyker du sigaretter?	0=ja, 1=nei
SIGALDER	Hvor gammel var det når du tok din første sigarett?	Oppgi alder (give age)
ystartdat	Start date for 'y' questionnaire (green ones)	
yserienr	Series no. for 'y' questionnaire (green ones)	
yEVERROK	Have you ever smoked	0 = yes, 1 = no, 9 = party smoker / occasionally. If th answer here is no, but at the same time put up no. Of cigarettes per day in the table below, the table must punch.
yROYKAR1		0-6
yROYKAR2		0-6
yROYKNAA	B. Røyker du nå?	0 =ja, 1 = nei, 9 = festrøyker

yROKBOR	Do you live with someone who smokes	0 = ja, 1 = nei, 2 = ex-røyker, 3 = piperøyker, 4 = sigarrøyker, 9 = festrøyker, 98 = usikker
yROKBORNO	Antall sigaretter pr. Dag	0-255 = specified number. 98 = Uncertain. If there is no answer to the first question, it is not considered that they have responded to the other.
yROKANT1	10-19 alderen:	0 = 0 sig, 1 = 1-5 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig
yROKANT2	20-29 alderen:	0 = 0 sig, 1 = 1-5 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig
yROKANT3	30-39 alderen:	0 = 0 sig, 1 = 1-5 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig
yROKANT4	40-49 alderen:	0 = 0 sig, 1 = 1-5 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig
yROKANT5	50-59 alderen:	0 = 0 sig, 1 = 1-5 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig
yROKANT6	60-69 alderen:	0 = 0 sig, 1 = 1-5 sig, 2 = 5-9 sig, 3 = 10-14 sig,4 = 15-19 sig, 5 = 20-24 sig, 6 = 25+ sig
ySIGROYK	Røyker du sigaretter?	0=ja, 1=nei
yROYKSTOP	How old were you when you quit smoking?	Enter age, 97 = uncertain
yROYKARB	How long (hrs) / per day are you exposed to smoking at work??	
yROKBARN	Røyk i ditt barndomshjem	0 = ja, 1 = nei, 2 = ex-røyker , 3 = piperøyker,
yROKHVEM	who smoked?	0 = bare far, 1 = bare mor, 2 = far og mor, 3 = andre, 4 = kombinasjon av flere kryss
yROYHJEM	How long (hrs) / per day are you exposed to smoking at home??	
YSIGALDER	Hvor gammel var det når du tok din første sigarett?	Oppgi alder (give age)
YRALDSLUTT	How old were you when you quit smoking?	Enter age, 97 = uncertain
yroykavogtil	Do you smoke sometimes?	
YROKBANT	Hvor mange sigaretter røykte foreldrene til sammen pr. Dag?	Oppgi totalt antall pr. dag
YSIGALDERB	Hvor gammel var det når du tok din første sigarett?	
YROKIKKENO	ikke nå (not smoking now)	



zserienr	Start date for 'z' questionnaire (yellow ones)	
zstartdat	Series no. for 'z' questionnaire (yellow ones)	
ZSIGROYK	Røyker du sigaretter?	0=ja, 1=nei
ZROKSIST5	Hvor mange sigaretter røykte du pr. dag de siste fem årene?	0=0, 1=1-4, 2=5-9, 3=10-14, 4=15-19, 5=20-24, 6=25+
ZSIGALDER	Hvor gammel var det når du tok din første sigarett?	Oppgi alder (give age)
ZROYKNAA	B. Røyker du nå?	0 =ja, 1 = nei, 9 = festrøyker
ZROYKSTOP	Hvor gammel var du når du sluttet/stoppet å røyke?	Oppgi alder
ZBARNROYK		
ZROKBANT	Hvor mange sigaretter røykte foreldrene til sammen pr. Dag?	Oppgi totalt antall pr. dag
ZROKSIST8	hvor mange sigaretter røykte du i gj.snitt (average) pr dag mellom 2002-2010	0=0, 1=1-4, 2=5-9, 3=10-14, 4=15-19, 5=20-24, 6=25+
ZRALDSLUTT	How old were you when you quit smoking?	Enter age, 97 = uncertain
ZROYKAVTIL	Do you smoke sometimes?	
ZROKBARN	Røyk i ditt barndomshjem	0 = ja, 1 = nei, 2 = ex-røyker , 3 = piperøyker,
LOK7		
ICD10_GR	Type sof cancer diagnosis in ICD10 group	C00, C01…… C71, C82
HIST	history of diagnosis??	
DIAGDAT	Date of diagnosis	
BPROVEDATO	date for blood samples	
BROYK	Do you smoke?	
BROYKANTGA	how many cig did you smoke yesterday?	
BROYKANTDA	how many cig did you smoke today?	
KLABNR	lab no. For blood samples	

D Main Script

```
# main_smoking_status.R
# Main script for smoking status calculations

rm(list=ls())
load("PATH_TO_DATA/Controldataset_GEdata_Smokingvars_171018.RData")

if (!require(dplyr)) {
  install.packages(dplyr)
  library(dplyr)
}

if (!require(gdata)) {
  install.packages(gdata)
  library(gdata)
}

source("utility_functions.R")
source("find_closest_quest.R")
source("smoking_intensity.R")
source("smoking_status_nfc.R")
source("passive_smoking_status.R")
source("age_at_blood_sample.R")
source("duration_of_smoking.R")
source("time_since_cessation.R")

Qscontrols <- findClosestQuest(Qscontrols)
Qscontrols <- smokingIntensity(Qscontrols)
Qscontrols <- smokingStatusNfc(Qscontrols)
Qscontrols <- passiveSmokingStatus(Qscontrols)
Qscontrols <- ageAtBloodSample(Qscontrols)
Qscontrols <- smokingDuration(Qscontrols)
Qscontrols <- timeSinceCessation(Qscontrols)
```

E Utility Functions

```
# utility_functions.R
# Utility functions

# If the input is a character string, then trim it and convert to number, or
# else return input as it is
convertStringToNumber <- function (numberStr) {
  if (is.character(numberStr)) {
    return(as.numeric(trim(numberStr)))
  }

  return(numberStr)
}

# For an integer, use the last two numbers as the year and make it four digits
extractYearFromIntegerDate <- function (intDate) {
  # Modulo (%%) operator to extract the last two digits
  year <- intDate %% 100

  # NOWAC participants not born before 1943
  if (year >= 43) {
    year <- year + 1900
  } else {
    year <- year + 2000
  }

  return(year)
}

# Find average of values given a vector of column names. Columns containing NA
    will be skipped.
# Note: 98 has a special meaning for smoking intensity options. Be aware that we
    are ignoring
# these values.
calculateAverageOption <- function (obs, colNames) {
  if (length(colNames) < 1)
    return(NA)

  dividend <- 0
  divisor <- 0

  for (colName in colNames) {
    option <- obs[colName]

    if (!is.na(option) & !is.null(option)) {
      option <- convertStringToNumber(option)

      if (option != 98) {
        divisor <- divisor + 1
        dividend <- dividend + option
      }
    }
  }

  if (divisor != 0) {
    return(dividend / divisor)
  } else {
    return(NA)
  }
}

# Estimate start age from intervals
estimateStartAndStopAgeFromIntervals <- function (obs) {
  estimatedStart <- NA
  estimatedEnd <- NA

  if (obs["ClosestQuest"] != "x") {
    # yROYKANT1-yROYKANT6
    colNames <- sprintf("yROYKANT%d", seq(1:6))
    i <- 0
    for (colName in colNames) {
      colValue <- convertStringToNumber(obs[colName])

      if (!is.na(colValue)) {
        if (is.na(estimatedStart)) {
          # First interval after NA is estimated Start age
          # The average age start at 12 and increase by 5 for each interval
          estimatedStart <- 14.5 + i
        }
      } else if (!is.na(estimatedStart)) {
        # The loop continues here, and the previous interval before new NA is
            estimated stop age.
      estimatedStop <- 14.5 + (i - 10)

      # Jump out of loop, we have found the value for start and stop age
      break
    }

    i <- i + 10
  }
}

if (!is.na(estimatedStart) & !is.na(estimatedEnd)) {
  return(c(estimatedStart, estimatedEnd))
}

# ROK1-ROK8
colNames <- sprintf("ROK%d", seq(1:8))
i <- 0
for (colName in colNames) {
  colValue <- convertStringToNumber(obs[colName])

  if (!is.na(colValue)) {
    if (is.na(estimatedStart)) {
      # First interval after NA is estimated Start age
      # The average age start at 12 and increase by 5 for each interval
      estimatedStart <- 12 + i
    }
  } else if (!is.na(estimatedStart)) {
    estimatedStop <- 12 + (i - 5) # The loop continues here, and the previous
         interval before new NA is estimated Stop age
    break # Jump out of loop, we have found the value for start and stop age
  }

  i <- i + 5
}

if (!is.na(estimatedStart)) {
  return(c(estimatedStart, estimatedEnd))
}

# ROYKANT1019-ROYKANT55MM
colNames <- c("ROYKANT1019", "ROYKANT2029", "ROYKANT3039", "ROYKANT4049", "
    ROYKANT50MM")
i <- 0
for (colName in colNames) {
  colValue <- convertStringToNumber(obs[colName])

  if (!is.na(colValue)) {
    if (is.na(estimatedStart)) {
      # First interval after NA is estimated Start age
      # The average age start at 12 and increase by 5 for each interval
      estimatedStart <- 14.5 + i
    }
  } else if (!is.na(estimatedStart)) {
      # The loop continues here, and the previous interval before new NA is
           estimated stop age
      estimatedStop <- 14.5 + (i - 10)

      # Jump out of loop, we have found the value for start and stop age
      break
  }

  i <- i + 10
}

if (!is.na(estimatedStart)) {
  return(c(estimatedStart, estimatedEnd))
}

# ROYKANT1-ROYKANT6
colNames <- colNames <- sprintf("ROYKANT%d", seq(1:6))
i <- 0
for (colName in colNames) {
  colValue <- convertStringToNumber(obs[colName])

  if (!is.na(colValue)) {
     if (is.na(estimatedStart)) {
      # First interval after NA is estimated Start age
      # The average age start at 12 and increase by 5 for each interval
      estimatedStart <- 14.5 + i
    }
  } else if (!is.na(estimatedStart)) {
    # The loop continues here, and the previous interval before new NA is
         estimated stop age
    estimatedStop <- 14.5 + (i - 10)
```

```
      # Jump out of loop, we have found the value for start and stop age
      break
    }

    i <- i + 10
  }

  return(c(estimatedStart, estimatedEnd))
}

estimateStartAgeFromIntervals <- function (obs) {
  return(estimateStartAndStopAgeFromIntervals(obs)[1])
}

estimateStopAgeFromIntervals <- function (obs) {
  return(estimateStartAndStopAgeFromIntervals(obs)[2])
}

# Age at start. Calculates an average if several different answers have been
     given for one woman.
# Returns the floor of number.
startAge <- function (obs) {
  ageAtBloodSample <-  convertStringToNumber(obs["AgeAtBloodSample"])

  if (obs["ClosestQuest"] == "z")
    colNames <- c("SIGALDER", "YSIGALDER", "YSIGALDERB", "ZSIGALDER")
  else if (obs["ClosestQuest"] == "y")
    colNames <- c("SIGALDER", "YSIGALDER", "YSIGALDERB")
  else
    colNames <- c("SIGALDER")

  ageAtStart <- calculateAverageOption(obs, colNames)

  # Even if the closest questionnaire is x or y, we sometimes have a start age in
        z.
  # If we are missing the value for x or y, we can use ZSIGALDER if available.
  # As long as the start age is less than the blood sample date, or else it doesn
        't make sense.

  zAge <-obs["ZSIGALDER"]

  if (is.na(ageAtStart) & !is.na(zAge)) {
    zAge <- convertStringToNumber(zAge)
    if (zAge < ageAtBloodSample)
      ageAtStart <- zAge
  }

  # If it is still not available (NA), inspect the smoking intervals/timeline
  if (is.na(ageAtStart))
    ageAtStart <- estimateStartAgeFromIntervals(obs)

  return(ageAtStart)
}

# Age at stop.  Calculates an average if several different answers have been
     given
# for one woman.
# Returns the floor of number.
stopAge <- function (obs) {
  ageAtBloodSample <-  convertStringToNumber(obs["AgeAtBloodSample"])

  if (obs["ClosestQuest"] == "z")
    colNames <- c("yROYKSTOP", "YRALDSLUTT", "ZROYKSTOP", "ZRALDSLUTT")
  else if (obs["ClosestQuest"] == "y")
    colNames <- c("yROYKSTOP", "YRALDSLUTT")
  else
    colNames <- character()

  ageAtStop <- calculateAverageOption(obs, colNames)

  # Even if the closest questionnaire is x or y, we sometimes have a stop age in
        z.
  # If we are missing the value for x or y, we can use "ZROYKSTOP" or "ZRALDSLUTT
        " if available.
  # As long as the stop age is less than the blood sample date, or else it doesn'
        t make sense.

  colNames <- c("ZROYKSTOP", "ZRALDSLUTT")
  zAge <- calculateAverageOption(obs, colNames)

  if (is.na(ageAtStop) & !is.na(zAge)) {
    zAge <- convertStringToNumber(zAge)
    if (zAge < ageAtBloodSample)
```

```
      ageAtStop <- zAge
  }

  # If it is still not available (NA), inspect the smoking intervals/timeline
  if (is.na(ageAtStop))
    ageAtStop <- estimateStopAgeFromIntervals(obs)

  return(ageAtStop)
}
```

F Find Closest Questionnaire

```
# find_closest_quest.R

findClosestQuest <- function(women) {
  # Comparing the dates
  women$STARTDAT_X <- as.Date(sprintf(”%06d”, women$STARTDAT), ”%d%m%y”)
  women$STARTDAT_Y <- as.Date(sprintf(”%06d”, women$ystartdat), ”%d%m%y”)
  women$STARTDAT_Z <- as.Date(sprintf(”%06d”, women$zstartdat), ”%d%m%y”)
  women$STARTDAT_B <- as.Date(sprintf(”%06d”, women$BPROVEDATO), ”%d%m%y”)

  # Find the closest questionnaire
  women$ClosestQuest = ”x”
  women$ClosestQuest[!is.na(women$STARTDAT_Y) & women$STARTDAT_B >= women$
      STARTDAT_Y] <- ”y”
  women$ClosestQuest[!is.na(women$STARTDAT_Z) & women$STARTDAT_B >= women$
      STARTDAT_Z] <- ”z”

  return(women)
}
```

G Smoking Intensity

```r
# smoking_intensity.R
# Intensity (average no. of cigarettes smoked per day)

optionToIntensityVector <- c(0, 2.5, 7, 12, 17, 22, 27)

# Option for X
calculateOptionX <- function (obs) {
  colNames <- sprintf("ROK%d", seq(1:8))
  option <- calculateAverageOption(obs, colNames)

  if (!is.na(option)) {
    return(option)
  }

  colNames <- c("ROYKANT1019", "ROYKANT2029", "ROYKANT3039", "ROYKANT4049", "
      ROYKANT50MM")
  option <- calculateAverageOption(obs, colNames)

  if (!is.na(option)) {
    return(option)
  }

  colNames <- c("ROYKANT1", "ROYKANT2", "ROYKANT3", "ROYKANT4", "ROYKANT5", "
      ROYKANT6")
  option <- calculateAverageOption(obs, colNames)

  if (!is.na(option)) {
    return(option)
  }

  return(NA)
}

# Option for Y
calculateOptionY <- function (obs) {

  optionX <- calculateOptionX(obs)

  colNames <- c("yROYKANT1", "yROYKANT2", "yROYKANT3", "yROYKANT4", "yROYKANT5",
      "yROYKANT6")
  optionY <- calculateAverageOption(obs, colNames)

  if (!is.na(optionY)) {
    if (!is.na(optionX)) {
      return((optionX + optionY) / 2)
    } else {
      return(optionY)
    }
  }

  return(optionX)
}

# Option for Z
calculateOptionZ <- function (obs) {

  optionY <- calculateOptionY(obs) #includes option for both x and y

  last5years <- obs["ZROKSIST5"]
  last8years <- obs["ZROKSIST8"]
  optionZ <- NA

  if (!is.na(last5years)) {
    last5years <- convertStringToNumber(last5years)
    optionZ <- last5years
  } else if (!is.na(last8years)) {
    last8years <- convertStringToNumber(last8years)
    optionZ <- last8years
  }

  if (!is.na(optionZ)) {
    if (!is.na(optionY)) {
      return((optionZ + optionY) / 2)
    } else {
      return(optionZ)
    }
  }

  return(optionY)
}
```

```
calculateIntensity <- function (obs) {
  option <- NA

  if (obs["ClosestQuest"] == "x") {
    option <- calculateOptionX(obs)
  } else if (obs["ClosestQuest"] == "y") {
    option <- calculateOptionY(obs)
  } else {
    option <- calculateOptionZ(obs)
  }

  option <- trunc(option + 0.5) # Use trunc instead of round, because round
       rounds 0.5 to even.

  return(optionToIntensityVector[option + 1])
}

calculateOption1019 <- function (obs){
  colNames <- c("ROYKANT1014", "ROYKANT1519")
  option <- calculateAverageOption(obs, colNames)

  if (!is.na(option))
    return(option)

  return(NA)
}

smokingIntensity <- function (women) {
  women$ROYKANT1019 <- apply(women, 1, calculateOption1019)
  women$Intensity <- apply(women, 1, calculateIntensity)

  return(women)
}
```

H Smoking Status NFC

```
# smoking_status_nfc.R
# Smoking Status (Never, Former, Current)

# Current smokers from the blood sample questionnaires

smokingStatusNfc <- function (women) {
  # Creating a blank column for smoking status
  women$SmokingStatus = NA

  # Current smokers

  women$SmokingStatus[women$BROYK == 0] <- ”Current”
  women$SmokingStatus[women$BROYKANTGAR > 0 & women$BROYKANTDAG > 0] <- ”Current”

  # Never smokers

  women$SmokingStatus[women$ClosestQuest == ”z” & is.na(women$SmokingStatus)
                              & (women$ZSIGROYK == 1 |women$ZROYKNAA ==1)
                              & (women$yEVERROK == 1 | is.na(women$yEVERROK))
                              & (women$yEVERROK==1 | women$yROYKNAA ==1 | women$
                                   ySIGROYK == 1 )
                              & (women$EVERROK == 1 | is.na(women$EVERROK))
                              & (women$EVERROK==1 | women$ROYKNAA ==1 | women$
                                   ROYKNAAB ==1 | women$SIGROYK ==1)
                              & (is.na(women$Intensity) | women$Intensity <= 0)] <-
                                   ”Never”

  women$SmokingStatus[women$ClosestQuest == ”y” & is.na(women$SmokingStatus)
                              & (women$yEVERROK == 1 | is.na(women$yEVERROK))
                              & (women$yEVERROK==1 | women$yROYKNAA ==1 |women$
                                   ySIGROYK ==1 )
                              & (women$EVERROK == 1 | is.na(women$EVERROK))
                              & (women$EVERROK==1 | women$ROYKNAA ==1 | women$
                                   ROYKNAAB ==1 | women$SIGROYK ==1)
                              & (is.na(women$Intensity) | women$Intensity <= 0)] <-
                                   ”Never”

  women$SmokingStatus[women$ClosestQuest == ”x” & is.na(women$SmokingStatus)
                              & (women$EVERROK == 1 | is.na(women$EVERROK))
                              & (women$EVERROK==1 | women$ROYKNAA ==1 | women$
                                   ROYKNAAB ==1 | women$SIGROYK ==1)
                              & (is.na(women$Intensity) | women$Intensity <= 0)] <-
                                   ”Never”

  # Former smokers

  women$SmokingStatus[women$ClosestQuest == ”z” & is.na(women$SmokingStatus)
                              & (women$ZSIGROYK ==0 |women$ZROYKNAA ==0
                                 | women$yEVERROK ==0 |women$yROYKNAA ==0 |women$
                                     ySIGROYK ==0
                                 | women$EVERROK==0  | women$ROYKNAA ==0 | women$
                                     ROYKNAAB ==0 | women$SIGROYK ==0)] <- ”Former”

  women$SmokingStatus[women$ClosestQuest == ”y” & is.na(women$SmokingStatus)
                              & (women$yEVERROK ==0 |women$yROYKNAA ==0 |women$
                                   ySIGROYK ==0
                                 | women$EVERROK==0  | women$ROYKNAA ==0 | women$
                                     ROYKNAAB ==0 | women$SIGROYK ==0)] <- ”Former”

  women$SmokingStatus[women$ClosestQuest == ”x” & is.na(women$SmokingStatus)
                              & (women$EVERROK==0  | women$ROYKNAA ==0 | women$
                                   ROYKNAAB ==0 | women$SIGROYK ==0)] <- ”Former”


  return(women)
}
```

I Passive Smoking Status

```
# passive_smoking_status.R
# Passive smoking status

passiveSmokingStatus <- function (women) {
  # Creating a blank column for passive smoking status
  women$PassiveSmoking = NA

  # Passive_childhood

  women$PassiveSmoking[women$ClosestQuest == "z"
                              & (women$ZROKBARN ==0 |women$ZBARNROYK == 0
                                 | women$yROKBARN == 0
                                 | women$ROKBARN ==0)] <- "Passive_childhood"

  women$PassiveSmoking[women$ClosestQuest == "y"
                              & (women$yROKBARN == 0
                                 | women$ROKBARN ==0)] <- "Passive_childhood"

  women$PassiveSmoking[women$ClosestQuest == "x"
                              & women$ROKBARN ==0] <- "Passive_childhood"

  # Passive_adulthood

  women$PassiveSmoking[is.na(women$PassiveSmoking)
                              & (women$ClosestQuest == "z" | women$ClosestQuest ==
                                  "y")
                              & (women$yROKBOR == 0 | women$yROKBOR == 3 | women$
                                  yROKBOR == 4 | women$yROYKARB > 0
                                 | women$ROKBOR ==0)] <- "Passive_adulthood"

  women$PassiveSmoking[is.na(women$PassiveSmoking)
                              & women$ClosestQuest == "x"
                              & women$ROKBOR ==0] <- "Passive_adulthood"

  # Passive_both (here it overwrites passive childhood)

  women$PassiveSmoking[women$PassiveSmoking == "Passive_childhood"
                              & (women$ClosestQuest == "z" | women$ClosestQuest ==
                                  "y")
                              & (women$yROKBOR == 0 | women$yROKBOR == 3 | women$
                                  yROKBOR == 4 | women$yROYKARB > 0
                                 | women$ROKBOR == 0)] <- "Passive_both"

  women$PassiveSmoking[women$PassiveSmoking == "Passive_childhood"
                              & women$ClosestQuest == "x"
                              & women$ROKBOR ==0] <- "Passive_both"

  # Not_PassiveSmoker

  women$PassiveSmoking[women$ClosestQuest == "z" & is.na(women$PassiveSmoking)
                              & (women$ZROKBARN ==1 |women$ZBARNROYK == 1
                                 | women$yROKBARN == 1
                                 | women$ROKBARN ==1)
                              & (women$yROKBOR == 1 | women$ROKBOR ==1)
                              & (women$yROYKARB == 0 | is.na(women$yROYKARB))] <- "
                                  Not_PassiveSmoker"


  women$PassiveSmoking[women$ClosestQuest == "y" & is.na(women$PassiveSmoking)
                              & (women$yROKBARN == 1
                                 | women$ROKBARN ==1)
                              & (women$yROKBOR == 1 | women$ROKBOR ==1)
                              & (women$yROYKARB == 0 | is.na(women$yROYKARB))] <- "
                                  Not_PassiveSmoker"

  women$PassiveSmoking[women$ClosestQuest == "x" & is.na(women$PassiveSmoking)
                              & women$ROKBARN ==1
                              &  women$ROKBOR ==1] <- "Not_PassiveSmoker"

  return(women)
}
```

J Age at Blood Sample

```
# age_at_blood_sample.R

# Function for calculating age at blood sample
calculateAgeAtBloodSample <- function (obs) {
  bloodSampleYear <- extractYearFromIntegerDate(convertStringToNumber(obs["
      BPROVEDATO"]))
  fourDigitBirthYear <- convertStringToNumber(obs["FAAR"]) + 1900

  age <- bloodSampleYear - fourDigitBirthYear

  return(age)
}

ageAtBloodSample <- function (women) {
  # To make a new column with the defined function
  women$AgeAtBloodSample <- apply(women, 1, calculateAgeAtBloodSample)

  return(women)
}
```

K Duration of Smoking

```
# duration_of_smoking.R
# Duration of smoking (only for current and former Smokers)

# Duration of smoking for current smokers
smokingDurationCurrent <- function (obs) {
  ageAtStart <- startAge(obs)

  duration <- convertStringToNumber(obs["AgeAtBloodSample"]) - ageAtStart

  return(duration)
}

# Duration of smoking for formers smokers
smokingDurationFormer <- function (obs) {
  ageAtStart <- startAge(obs)
  ageAtStop <- stopAge(obs)

  return(ageAtStop - ageAtStart)
}

calculateSmokingDuration <- function (obs) {
  smokingStatus <- obs["SmokingStatus"]

  if (is.na(smokingStatus)) return(NA)
  else if (smokingStatus == "Current") return(smokingDurationCurrent(obs))
  else if (smokingStatus == "Former") return(smokingDurationFormer(obs))
  # Just to check if there are any mistakes in the "Never"-status
  else if (smokingStatus == "Never") return(smokingDurationFormer(obs))
  else return(NA)
}

smokingDuration <- function (women) {
  women$SmokingDuration <- apply(women, 1, calculateSmokingDuration)

  return(women)
}
```

L Time Since Cessation

```
# time_since_cessation.R
# Time since cessation (TSC) (only needed for former smokers)

calculateTimeSinceCessation <- function (obs) {
  ageAtStop <- stopAge(obs)

  tsc <- convertStringToNumber(obs["AgeAtBloodSample"]) - ageAtStop

  return(tsc)
}

calculateTSC <- function (obs) {
  smokingStatus <- obs["SmokingStatus"]

  if (is.na(smokingStatus)) return(NA)
  else if (smokingStatus == "Former") return(calculateTimeSinceCessation(obs))

  # Just to check if there are any mistakes in the "Current" and "Never"-status.
  # If the answer is 0 or NA, then it is correct.

  else if (smokingStatus == "Current") return(calculateTimeSinceCessation(obs))
  else if (smokingStatus == "Never") return(calculateTimeSinceCessation(obs))

  else return(NA)
}

timeSinceCessation <- function (women) {
  women$TSC <- apply(women, 1, calculateTSC)

  return(women)
}
```