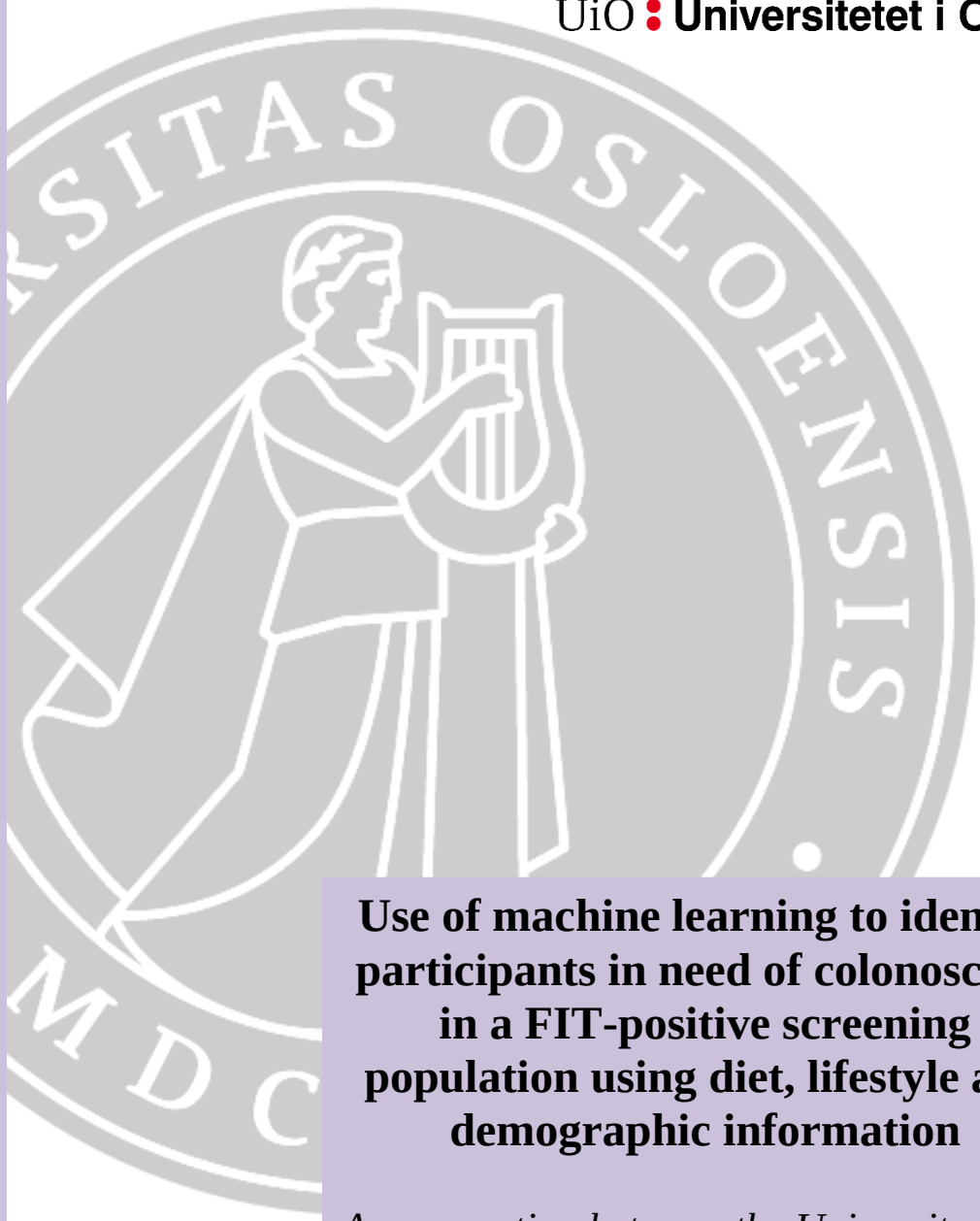
UiO : Universitetet i Oslo



Use of machine learning to identify participants in need of colonoscopy in a FIT-positive screening population using diet, lifestyle and demographic information

A cooperation between the University of Oslo and the Cancer Registry of Norway

Master`s Thesis by
Emilie Syse Jalland

Department of Nutrition
Faculty of Medicine
University of Oslo

May 2022

Use of machine learning to identify participants in need of colonoscopy in a FIT-positive screening population using diet, lifestyle and demographic information

By Emilie Syse Jalland

Supervisors:
Anette Hjartåker
Einar Birkeland
Ane Sørlie Kværner

A cooperation between the Cancer Registry of Norway and the University of Oslo

Department of Nutrition, Faculty of Medicine
UNIVERSITY OF OSLO

May 2022

2022

Use of machine learning to identify participants in need of colonoscopy in a FIT-positive screening population using diet, lifestyle and demographic information

http://www.duo.uio.no

Print production: Reprosentralen, University of Oslo

Acknowledgements

This master thesis was written at the Department of Nutrition, Faculty of Medicine, University of Oslo, between August 2021 and May 2022. The thesis has been done in collaboration with the Cancer Registry of Norway and have contributed to innovative nutrition epidemiological research.

I would like to firstly extend gratitude to three unique supervisors: Einar Anette, and Ane - this thesis would not have been the same without your invaluable help! I am grateful for your close participation, follow-up and constructive feedback throughout the year. Einar, thank you for your patience, knowledge and ability to explain machine learning in a clarifying manner. You have given me insight into the world of programming – an exciting world that few nutrition students take part in. Anette, thank you for sharing your extensive experience and knowledge of nutrition and cancer. I am grateful that you took over the baton after Ane; it has been an honour to have such a renowned professor as supervisor of this master's thesis. Ane, thank you for being an ongoing motivational driver for this thesis. In the months of your participation, you had an invaluable presence and you are a KEF that I look up to.

Further, I must thank the people in the CRCbiome research group and Markus Knudsen, for discussion and advice during this period. Anne-Marte Wetting at the Department of Nutrition, thank you for helping me understand the structure of KBS and the subsequent dietary variables. Further on, I am appreciative to Pål Aam who on several occasions has read through and given me feedback on the thesies; this has been to great help.

I am grateful for my family, friends and Erlend for endless support during these five years of study. Thank you for cheering me on through ups and downs, it has meant a lot to me. And to all my fellow class mates – we finally made it, congratulations!

Abstract

Background: Fecal Immunochemical Test (FIT) is a widely used colorectal cancer (CRC) screening method, which will be used in the new national colorectal cancer screening in Norway. FIT has shown to reduce both CRC incidence and mortality. Nevertheless, the test sensitivity and specificity is suboptimal, resulting in undetected cases (suboptimal sensitivity) and unnecessary references to colonoscopy (suboptimal specificity). Diet and lifestyle are important risk factors for CRC. Incorporation of a risk prediction models, created with risk factors, in the screening setting can contribute to a more accurate screening offer. **Objectives:** To create a risk stratification algorithm using Random Forest to predict participants with the need of colonoscopy in a FIT-positive population. Specifically, we wanted to investigate the difference when using dietary, lifestyle demographic and FIT data in the algorithm compared to using dietary data. **Participants and method:** In this master's thesis, 1476 FIT-positive participants from the CRCbiome study were included. The machine learning algorithm was created with the use of Random Forest and each prediction tree was built on a bootstrapped dataset. We used an 80/20 split, where 1183 participants were used to train the model, and the best models were then evaluated in the test dataset including 293 participants. We created four different datasets with input variables obtained from self-reported questionnaires (a validated food frequency questionnaire and a lifestyle and demographic questionnaire), as well as the screening database (the FIT value). Diagnostic information from a follow-up colonoscopy formed the basis for the outcome classification, and participants were allocated into "true negatives" and "critical to find". Outcome classification was done in four different ways. **Results:** The best performing model included 11 dietary variables known to have an impact on CRC risk, 7 lifestyle and demographic risk variables and the result of the FIT test. This model was created within a cohort only including participants with advanced adenoma or CRC and participants without adenoma or other lesions. The model showed an area under the receiver operating characteristic curve (AUROC) of 0.64 and 0.59 in the training dataset and the test dataset, respectively. Further, a model only including the CRC relevant dietary variables trained within the same cohort gave an AUROC of 0.61 in the training dataset and 0.59 in the test dataset. **Conclusion:** None of the models were able to give a satisfying prediction of who among the screening participants were in need of a colonoscopy. However, our results highlight the potential added benefit of including dietary variables in CRC prediction models.

Table of Contents

List of tables

List of figures

Abbreviations

AICR	American Institute of Cancer Research
ANOVA	Analysis of variance
AUROC	Area under the receiver operating characteristic curve
BMI	Body mass index
BCSN	Bowel Cancer Screening in Norway
CIMP	CpG Island Methylator Phenotype
CIN	Chromosomal instability
CRC	Colorectal cancer
DNN	Deep Neural Network
EPIC	European Prospective Investigation into Cancer and Nutrition
FFN	Feedforward Neural Network
FFQ	Food Frequency Questionnaire
FIT	Fecal Immunochemical Test
gFOBT	Guaiac-based Fecal Occult Blood Test
KBS	Kostberegningssystem
LDQ	Lifestyle and Demography Questionnaire
MDS	Mediterranean Diet Score
ML	Machine Learning
MSI	Microsatellite instability
RF	Random Forest
SD	Standard Deviation
SES	Socioeconomic status
WCRF	World Cancer Research Fund
XGbooster	Extreme Gradient Boosting

1 Introduction

1.1 Colorectal cancer

1.1.1 Incidence, mortality and survival rates

In 2020, colorectal cancer (CRC) was the third most diagnosed type of cancer in the world, responsible for 10 % of all cancers (1). Moreover, nearly 1 million people died from CRC the same year, making it the second most deadly cancer worldwide (1). Norway is one of the countries with the highest incidence rates (1, 2), with CRC as the second most diagnosed cancer across sex (3). The number of cases per 100 000 people was 51 and 54 for colon cancer, and 18 and 21 for rectal cancer, for women and men, respectively (3). Norwegians developing CRC are mostly diagnosed with CRC between the age of 60 and 84 years (3). However, there has been an increase in CRC incidence among adults under the age of 50 years in some high income-countries (2). Despite an increasing incidence of CRC in Norway, more people survive the diagnosis than before (3). Five-year survival rate for colon cancer is 69 % for men and 71 % for women, and for rectal cancer the rate is 71 % and 72 % for men and women, respectively (3).

1.1.2 Pathology of colorectal cancer

CRC refers to cancer in the colon or in the rectum. The growth of a colorectal tumour takes several years (4). The development of CRC can be divided into non-advanced adenoma, advanced adenomas/serrated lesions, and CRC. Advanced adenomas have a size larger than 10 mm, a villous structure, or a high-grade dysplasia; this make them more prone to develop into a cancer (5).

The initiation of carcinogenesis is often caused by alteration in two to eight “driver genes”. Driver genes are important in signalling pathways that regulate genome maintenance, cell fate and cell survival. Hence, alteration in these genes enables favourable cell growth to the cell in which it occurs (6). The cells proliferate and this leads to abnormal growth, a neoplasm (7). As the neoplasm develops progressively into a benign tumour, which may transform into a malign tumour with metastatic potential, the cells accumulate more mutations (7). These genetic alterations include gain of function defects in some oncogenes and loss of function defects in some tumour suppressor genes (8). The biological capabilities a normal cell

acquires when developing into a cancer cell has been described by Hanahan and Weinberg (9). These biological capabilities, or hallmarks, include sustaining proliferative signalling, evading growth suppressors, resisting cell death, including angiogenesis, enabling replicative immortality and activation of invasion and metastasis (9). A metastatic tumour has migrated from the organ of origin to another, through lymph or blood vessels. In a new tissue the tumour cells may settle and potentially form a tumour (9). The most common metastatic site for CRC is in the liver, accompanied by bone and lung tissue (7). It is not easy to estimate the duration of each phase of tumour development and metastasis, in general it takes decades, and it varies between different tumour types (7).

There are three main global epigenetic and genetic abnormalities in colorectal carcinogenesis. The first aberration is chromosomal instability (CIN), which is characterized by abnormalities in chromosomal copy number, for instance resulting in aneuploidy or polyploidy, and in the chromosomal structure. These error commonly occur during mitosis due to defects centrosome number and in mitotic checkpoint proteins (7). The second aberration is CpG island methylator phenotype (CIMP). A type of ocular epigenetic modification that has occurred in repetitive CG dinucleotides in the promoter region of tumour suppressor genes, causing silencing of gene expression. The definition of CIMP varies greatly and the underlying cause is not entirely clear (7). Microsatellite instability (MSI) is the third major aberration, defined as changes in the length of short nucleotide tandem repeats in DNA sequences, microsatellites. MSI is probably caused by gene silencing due to promotor hyper methylation, resulting in loss of function in DNA mismatch repair genes (7).

Approximately 60-65 % of all incident cancers arise sporadically; these are tumours that occur in individuals without any inherited genetic mutation that increases the risk or with family history of CRC (7). There are two precursor subtypes responsible for most of these CRCs, the adenomatous polyps (adenomas) and the serrated polyps (7). Adenomas are the result of CIN and are a part of the adenoma-carcinoma pathway. This pathway is considered the traditional pathway (10) and is estimated to be responsible for 80-95 % of all sporadic CRCs (7). Serrated polyps are estimated to be responsible for 10-15 % of all sporadic CRC (7) and is a group of diverse lesions. This type of polyp belongs to the serrated neoplasia pathway, characterised by CIMP (10). In addition, a carcinogenic pathway driven by chronic bowel inflammation has been proposed (11). A cohort analysis by Jess *et al.* estimated a

nearly two and a half as high chance of CRC in people with inflammatory bowel disease compared to the general population (12). However, this inflammatory pathway is responsible for less than two percent of all cancer incidents (7). Family history is present in 25 % of all cancer cases, however, only 5 % are attributed to hereditary cancer syndromes such as Familial adenomatous polyposis or hereditary nonpolyposis colorectal cancer (Lynch syndrome) (7).

The anatomic site of the tumour or polyp presents a way to classify CRC subtypes into different groups. It is common to separate the intestine into three segments: the proximal colon (the right side), the distal colon (left side) and rectum (7). It is suggested that tumours at different locations in the large intestine have different aetiology (7, 13) and are affected differently by risk factors (13, 14). The explanation to this may be different response and exposures to environmental factors between proximal and distal colon, due to different embryological origin as well as postnatal changes in the mucosa (15). Studies exanimating the heterogeneity in risk factors and anatomical subside are inconclusive (13, 16). The European Investigating into cancer and Nutrition study (EPIC) found an inverse relation between physical activity and proximal and distal cancer, but not for cancer in the rectum (13). Further, current smokers were found to have an increased risk on proximal colon and rectal cancer, but not distal colon cancer (13) . A prospective cohort study by Wei *et al*. found a significant association between consumption of red meat, processed meat, folate and alcohol with colon cancer, but not with rectal cancer (16). Women have a higher prevalence of tumours in the proximal colon compared to men who have a higher prevalence of cancer in the distal colon (7). The serrated neoplasia pathway is more often the cause of a tumour found in the proximal colon whereas the adenoma-carcinoma pathway is more likely to be the cause of a tumour found in the distal colon (7, 17). Tumours found, at an advanced stage, in the right side of the colon are associated with an overall worse prognosis compared to tumours in later stages found at the left side (17).

1.2 Risk factors for colorectal cancer

CRC is a multifactorial disease with many identified risk factors, both modifiable and non-modifiable (10, 18). Many of the modifiable risk factors are related to our way of living, and influence the risk of developing the disease dramatically. It is estimated that nearly half of all CRC cases could have been prevented with a healthier lifestyle (19). Many of the

demographic and medical related risk factors, such as age and family history of CRC are among the non-modifiable risk factors. These are more determinant risk factors, which are difficult to influence (18).

1.2.1 Demographic and medical risk factors

Some of the personal medical history influencing the risk of CRC is sex and age. In Norway, which is a high-incidence country, the cumulative risk for developing CRC before the age of 75 years is 5 % for the general population (10, 20) . Increasing age is a risk factor; hence, most people receive the diagnosis after the age of 50 years (3, 7). Sex also influences the risk of developing the disease, in total more men than women are diagnosed with CRC both in Norway and worldwide (3, 10). Other medical conditions strongly related to CRC risk is Inflammatory Bowel Disease (21), Diabetes Mellitus (22) and previous history of cancer or adenomas of the intestine (23). People with one or more first-degree relatives diagnosed with CRC and people with hereditary cancer syndromes have a higher risk (10, 18). A meta-analysis including 16 studies found a nearly double risk of CRC in participants with first degree relatives diagnosed with CRC compared to those without family history (21). The incidence of various types of cancer is often reflected by socioeconomic status (SES) (e.g. education, income or occupation), both at an individual and a geographic level (24). A body of research has investigated the relationship between SES and CRC, however the results are inconsistent for different countries and ways to measure SES (24-26). A review by Arts *et al.* found a higher incidence of CRC among screening participants with high SES in Europa, South Korea and Australia (26). However, Lynge *et al.* found that the distribution among different socioeconomic classes in Norway has changed over time and that all socioeconomic groups now contribute to the increase in colon cancer incidence (27).

1.2.2 Lifestyle risk factors

World Cancer Research Fund (WCRF) and American Institute of cancer research (AICR) regularly publish a report with the latest research and findings regarding the relationship between lifestyle factors and CRC (19). The aim is to help people make informed choices and avert development of preventable cancers. The latest report from 2017 established strong evidence for physical activity reducing the risk of colon cancer (19). In fact, it is one of the few cancers where absence of physical activity is an established risk factor (28). It is thought

that physical activities reduce the risk of CRC by having a favourable effect on the immune system, metabolic hormones, inflammation and motility in the gut (29).

Further in the report, convincing evidence was obtained for a decreased risk of CRC with regular consumption of wholegrain, dietary fiber, dairy products, and calcium supplement (19). Contrariwise, regular intake of red and processed meat, as well as more than one alcoholic beverage a day was strongly associated with increased risk of CRC (19). The report suggested that regular consumption of vitamin D, foods with vitamin C and regular implementation of fish in the diet might have a protective effect (19). This was suggested for the use of multivitamins as well; however, the different combinations of ingredients in the supplements made it difficult to detect the active ingredient (19). Other dietary habits increasing the risk was low consumption of fruits and vegetables, as well as high consumption of foods with haem-iron (19).

Other lifestyle factors related to increased risk is overweight and obesity (30). Especially the abdominal visceral fat of people with excessive body mass are of greater concern and are shown to increases the risk (30). The adipocytes are thought to cause low grade inflammation (31, 32) and high insulin concentration (33-35) which contributes to CRC development. Smoking has also shown to be an important lifestyle factor that increases the risk of CRC (36). In a meta-analysis the relationship between smoking and CRC, smoking increased the risk of with 15-20 % (36).

Nonetheless, these risk factors are associated with overall CRC, but there is a difference in how strong the association between risk factors and the different types of lesions is (37). He *et al.* found that adenomas are more associated with dietary factors compared to serrated polyps, which has shown a stronger link to alcohol intake, smoking and BMI than adenomas (37).

1.3 Screening for colorectal cancer

In the literature, randomized clinical trials have shown that CRC screening reduces incidence (38-41) and mortality (38-45). Some regard screening as the most powerful tool to reduce incidence and mortality of CRC due to the major difficulties when trying to implement prevention strategies or changing people's way of living (46). With screening it is possible to

detect asymptomatic cancer early in the carcinogenic process and remove precursor lesions and thus prevent disease development (47). Survival of CRC is strongly related to stage at the time of diagnosis (3).

As of today, several different screening techniques are in use. Some methods are more invasive such as colonoscopy and flexible sigmoidoscopy. These methods, along with computed tomography colonography, allows direct visualization of the colon and the rectum (46). Colonoscopy is regarded as the gold standard method (48). However, colonoscopy is seldom used as a primary screening tool. Rather it is used as a second-step approach in screening programs for diagnostic classification after a positive test result from another less invasive screening method has been detected (48).

Both FIT and the guaiac-based fecal occult blood test (gFOBT) are inexpensive and non-invasive screening methods designed to detect blood in feces. The gFOBT is based upon the oxidation of guaiac by hydrogen peroxide. Therefore, dietary and medical restrictions are often needed prior to testing. In addition, to enhance the diagnostic accuracy, it is recommended to collect three stool samples at each screening round (46). In contrast to the gFOBT, the FIT test uses antibodies to detect human haemoglobin in stool and as a consequence no dietary restrictions are needed. In addition, only one stool sample is required at each screening round (46). Because of these advantages and the fact that the FIT test has been shown to have higher sensitivity for detection of adenomas and CRC compared to the gFOBT, FIT is the preferred stool test in screening (47).

Although FIT is widely used, it is suboptimal – both regarding sensitivity and specificity (46). Studies have shown great differences in diagnostic performance when it comes to type, location and severity of the neoplastic lesion (49), but also age (50) and sex (50, 51). Since the test does not have optimal sensitivity and specificity, some neoplasms are missed while some individuals are unnecessarily referred to colonoscopy (46). Even though colonoscopy is performed by skilled professionals, with high quality equipment, adverse events such as bleeding, perforation and pain can occur (47, 52). Results from a large meta-analysis has shown a greater incidence of perforation in colonoscopies following a positive FIT test compared with colonoscopies in average risk populations (53). Further, colonoscopy is a costly procedure and burdensome, hence the resource is limited and should only be

considered for people with the need of the examination (46, 54). There are also concerns about the limited capacity of the FIT to detect all precancerous lesions, following a missed opportunity to prevent cancer by removing these lesions (49).

1.4 Risk prediction models

During the past decades, in response to the growing incidence of CRC, development of risk models with potential to stratify participants into risk categories have been developed (55). These models offer the potential of improving the effectiveness of screening by personalizing the programs (55). Numerous risk factors in various combinations have been used to predict colorectal neoplasia (advanced adenoma or CRC) in average risk populations (55-59). Prediction models have been created with anything from two (60) to 15 variables (55, 61). Most prediction models use easy to collect lifestyle and demographic variables, such as sex, age and family history of CRC (58). However, some also include information from clinical tests and laboratory analyses. Usher-smith *et al*. conducted a systematic review where 52 risk models were compared, including in total 87 different predictor variables in different combinations, ranging from personal characteristics to diet and lifestyle factors, drug use, biomarkers and results from other screening tests (55). Other have also investigated the potential utility of genetic factors alone and in combination with other more commonly used variables (62).

There are some studies that examine the use of risk variables in combination with the FIT result in prediction models (63-71). These models are designed to be used in screening referral decisions (63-71). Stegeman *et al.* created a prediction model in which the result of the FIT test, along with sex, age, BMI, smoking status, calcium intake, NSAIDs and family history of CRC were used. Using this model, they were able to identify five more individuals with advanced neoplasia if 120 underwent colonoscopy, compared with FIT alone (63). A risk prediction model by Li *et al*. combined personalized characteristics such as diarrhea, constipation and bleeding, all potential symptoms of CRC, with the FIT result. The model was found to be better to predict people with advanced neoplasia compared to the FIT test alone (71). Tao *et al.* investigated the use of four blood based inflammatory markers against and in combination with the FIT test result, which did not yield an improved detection of neither advanced adenoma nor CRC (65). However, only three of the studies include FIT-positive participants only (66, 68, 69). A Danish cross-sectional study among FIT-positive

participants, created a prediction model based on age, sex and the result of the FIT test to predict advanced neoplasia and CRC. The model was validated and showed an area under the receiver operating characteristic curve (AUROC) of 0.67 and 0.74 for prediction of advanced neoplasia and CRC, respectively (69). The same positive results have been seen in a Spanish study when combining sex, age and FIT result (66).

1.5 Machine learning

Machine learning (ML) is a branch of artificial intelligence. In ML, algorithms are used to discover patterns in data with limited human input or programming (72-74). Datasets may include a high number of variables as well as a multitude of observations, and it is often less structured. Consequently, it may be difficult to handle for some traditional statistical methods (74). ML algorithms are often classified based on the type of approach used: supervised learning or unsupervised learning (73, 74). In supervised learning, the aim is to build a model for prediction based on a known target or output. This is in contrast to unsupervised learning, which is used to identify patterns in the data without consideration of any outcome variables. Supervised learning can further be divided into regression or classification models depending on the type of output variable. If the model aims to predict a continuous output variable it is referred to as a regression model, whereas classification models indicate prediction of a categorical output variable (75). In this master`s thesis a supervised classification model is used.

ML is becoming increasingly more common in medical and epidemiological research as it provides new tools to handle problems that are difficult to solve with traditional statistics (75). However, there are few studies within nutritional epidemiology who utilize this approach even though it may be beneficial (74). Some studies have used machine learning to investigate the relationship between diet and cardio metabolic risk (76-80). For instance, a study by Rigdon *et al*. found that the inclusion of multilevel dietary data in a ML model, improved cardiovascular risk prediction (76). Another study, on pregnant women, used machine learning to explore the relationship between fruit and vegetable consumption and adverse pregnancy outcomes (81). Further, with inclusion of data on dietary habits, anthropometry, physical activity, blood parameters and gut microbiota, Zeevi *et al.* created a prediction algorithm that correctly predicted personal postprandial glycemic response after meals (82). Others have used classification algorithms to predict obesity, hypertension,

dyslipidemia and type 2 diabetes mellitus (83). There are to our knowledge only a few studies investigating the relation between dietary information and CRC using this approach (84-86).

Although the opportunities are there, few studies as of now are using ML to predict CRC or its precursor lesions (68, 85, 87-91). To the best of our knowledge, only one previous study has examined the ability of ML to aid in the detection of advanced neoplasms (advanced adenoma or CRC) following a positive FIT test (68). Further, no studies have investigated the possible predictive value of dietary information alone on CRC and its precursor lesions in such a high risk population.

2 Aim of the study

This master's thesis is a sub-project of the CRCbiome study, a screening trial where all participants have received a positive FIT test and consequently have been referred for follow-up colonoscopy. The overall objective of the master`s thesis is to distinguish participants with a true positive test (i.e. detection of some kind of neoplastic lesion at colonoscopy) from participants with a false positive test by applying dietary and other questionnaire data in a machine learning approach.

Primary aim:

- To examine to what degree demographic, lifestyle and dietary data can be used to identify participants with a positive FIT screening test that have a need for further colonoscopy examination, by use of machine learning

Secondary aim:

- To examine how the accuracy of the classification algorithm is affected by using a combination of demographic, lifestyle and dietary data as input variables compared to using dietary data only

3 Methods

3.1 The CRCbiome study

The CRCbiome study is a large ongoing prospective cohort study carried out by the Cancer Registry of Norway. The main aim of the study is to use gut metagenome, demographic and lifestyle data to develop a classification algorithm for identification of advanced colorectal lesions (92).

The CRCbiome study is a sub-study of the Bowel Cancer Screening in Norway (BCSN). BCSN is a randomized controlled screening trial, started in 2012 as a pilot for the upcoming national CRC screening program. BCSN is designed to compare FIT tests given every second year (with a maximum of four repetitions) with a single sigmoidoscopy examination. In 2024, the FIT arm of the study is expected to be completed. All the participants in the BCSN trial who received a positive FIT test during the period 2017-2021 were invited to join the CRCbiome study (47). If the haemoglobin content exceeded 15 mcg/g feces it was considered a positive FIT test (92).

The FIT tests were conducted at home, and the stool sample kits were mailed to the participants. In a test, 10 mg of stool was collected using a plastic stick and stored in 2 ml buffer (4-(2-hydroxyethyl)-1-piperazineethanesulfonic, Bovine serum albumin and sodium acid). Further, samples were mailed to a laboratory at Oslo University Hospital for storage at – 80 ˚C and analysis. The OC-Sensor Diana (Eiken Chemical, Tokyo, Japan) was used to conduct the immunochemical testing of haemoglobin in stool. No dietary or medical restrictions were needed prior to the test (92).

Only the FIT positive participants were selected to contribute to the CRCbiome study. These participants were selected as they, according to the BCSN trial protocol, were to be referred for follow-up colonoscopies (47, 92). The colonoscopy examination provides detailed clinic pathological information, information that is not available for the participants with negative FIT tests (92).

Prior to the colonoscopy examination, participants were invited to the CRCbiome study (**Supplementary file 1**). Along with the information letter, participants received two

questionnaires: a lifestyle and demographic questionnaire (LDQ) and a food frequency questionnaire (FFQ). The questionnaires were to be completed before the colonoscopy. Return of at least one of the questionnaires was considered as consent to the CRCbiome study (92).

3.2 Participants and eligibility

Men and women, living in either one of two municipalities in South-East Norway (Bærum or Moss), aged 50-74 years in 2012, were invited to the CRCbiome study. In total, 2698 participants were invited, of which 1653 agreed to participate (92). In total 1616, participants from the CRCbiome study completed the FFQ and were considered for this master project (**Figure 3**). However, 15 participants withdrew from the study after baseline. A number of participants were excluded in the analyses due to not showing up for colonoscopy (n = 39); having a low or medium quality FFQ (n = 21 and 10, respectively); reporting a too low (<600 and <800 kcal/day for women and men, respectively, n = 9) or a too high energy intake (> 3500 and 4200 kcal/day for women and men, respectively, n = 46). In total, 1476 participants were included in this master`s thesis, of this 1183 were randomly selected to be in the training data set and 293 were randomly selected to be in the test data set.

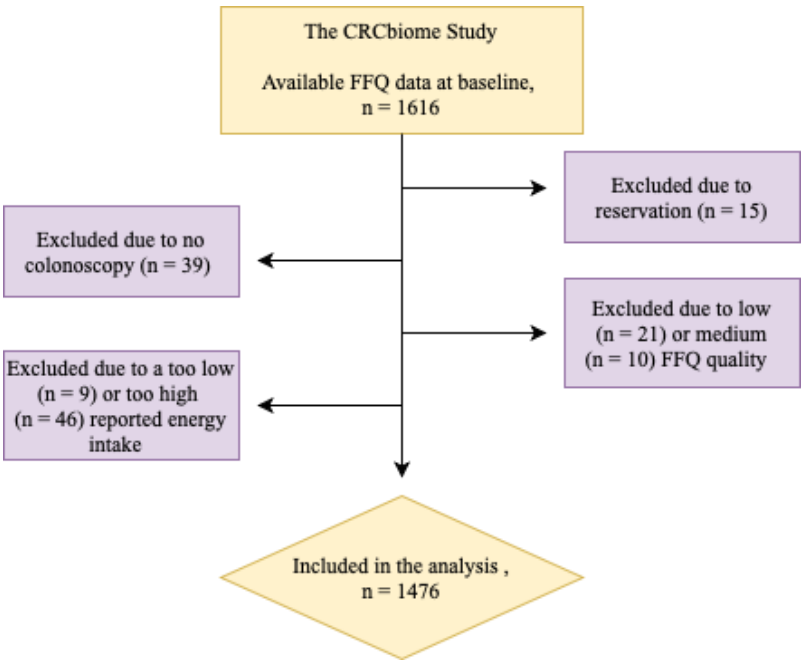


Figure1. Flowchart for the master`s thesis

3.3 Assessment of dietary intake

A semiquantitative FFQ was used to assess the usual dietary intake during the preceding year (**Supplementary file 2**). Participants filled out the FFQ by themselves after receiving a positive FIT result and before colonoscopy. The questionnaire used contains 23 main questions, covering 256 food items. In addition, the questionnaire includes open fields for entry of free text to report food items that the FFQ does not cover. For most of the questions, participants were asked to report on how often they consumed the food item, the answer options ranging from "seldom/never" to "several times a day". Furthermore, the participants were asked to estimate portion sizes, typically given in different household units such as glasses, spoons and deciliters (92). For "preferred cooking fat" there was no question about frequency or portion size. The questionnaire is a modified version of a FFQ developed and validated at the Department of Nutrition, University of Oslo (93-98). The FFQ has been validated for selected food items and food groups (95-98), energy intake (93-98), as well as intake of macro and micronutrients (93, 95, 98). Scanning of the FFQs was done with the use of the Cardiff TeleForm program (Datascan, Oslo, Norway). Calculation of dietary intake (food and nutrient intake) was done using "Kostberegningssystemet" (KBS), developed at the Department of Nutrition, University of Oslo. The database AE-18 was used, which was the latest version available at the time the study was conducted. Missing answers about frequency in the FFQ were imputed as null intake, and missing answers about portion size were imputed as the smallest portion (92). The FFQs were quality controlled and evaluated by trained personnel according to a list of predefined criteria (**Supplementary file 3**).

3.4 Assessment of lifestyle and demographic data

A four-page questionnaire (LDQ) with ten main questions was used to collect information about demographic and lifestyle factors (92) (**Supplementary file 4)**. Information about CRC among first-degree relatives, education, nationality, smoking habits, and the past year`s physical activity level were retrieved from the LDQ. The questionnaire is a modified version of a questionnaire used in previous national surveys (99, 100). The questionnaire was tested on a pilot group prior to study start and adjusted based on participant's feedback. Information about weight and height were obtained from the FFQ and sex, age and the FIT result were obtained from the screening database (92).

The physical activity reported in the LDQ was calculated into a “physical activity score”. This was the sum of physical activity with moderate intensity plus the amount of physical activity with high intensity times two. This recalculation was done to compensate for the fact that the recommendation for physical activity can be achieved through 75 min / week with high-intensity physical activity, instead of the standard recommendation of 150 min / week of moderate physical activity (101). A “family history of CRC” was defined as having a parent, sibling or a child with CRC. The participant’s reported nationality was classified as either “native” or “non-native”, where native was synonymous to Norwegian and non-native included all other countries.

3.5 Assessment of outcome information

From the BCSN database, we received clinic pathological information about the colorectal lesions detected at follow-up colonoscopy. Based on the diagnostic findings, participants were categorized into the groups “no adenoma”, “non-advanced serrated lesions/other lesions”, “non-advanced adenoma <3”, “non-advanced adenomas ≥3”, “advanced serrated lesions”, “advanced adenomas” or “CRC”. For the main analysis, we created the outcome groups “critical to find” and “true negatives” to be used in the prediction models (**Table 1**). Participants allocated to the critical to find group were diagnosed with advanced serrated lesions, non-advanced adenoma (≥3), advanced adenomas and CRC. Participants with no adenoma, non-advanced serrated lesions/other lesions or non-advanced adenoma (<3) were considered to be true negatives. The splitting was based on the European Society of Gastrointestinal Endoscopy guidelines about who would need endoscopic surveillance after polypectomy (102) and can therefore be interpreted as distinguishing those at need for colonoscopy when screened.

In addition to the main split including all study participants, we progressively excluded diagnostic groups, to see if it could improve the prediction models. The first diagnostic group to be excluded was "non-advanced adenoma (<3)". This was done because removing these adenomas during the colonoscopy, considerably limits their potential to progress to cancer. “Advanced serrated lesions” was the next group to be excluded because these lesions do not have the same risk factors or etiology as adenomas (103). For the last analysis, only those

with serious findings (i.e. CRC or advanced adenoma) and those without any findings were included in the model.

Table 1. Division of the participants into different cohorts

Overall cohort	Outcome variables	
No.1 *n=1183*	Critical to find *(n = 436)* • advanced serrated lesions • non-advanced adenomas (≥3) • advanced adenomas • CRC	True negative *(n = 747)* • non-advanced adenoma (<3) • non-advanced serrated lesions/other lesions • no adenoma
Sub-cohort	**Outcome variables for sub-cohort analysis**	
No. 2 *n=872*	Critical to find *(n = 436*) • advanced serrated lesions • non-advanced adenomas (≥3) • advanced adenomas • CRC	True negatives *(n = 436)* • non-advanced serrated lesions/other lesions • no adenoma
No.3 *n=813*	Critical to find *(n = 377)* • non-advanced adenomas (≥3) • advanced adenomas • CRC	True negatives *(n = 436)* • non-advanced serrated lesions/other lesions • no adenoma
No.4 *n=708*	Critical to find *(n = 272)* • advanced adenomas • CRC	True negatives *(n = 436)* • non-advanced serrated lesions/other lesions • no adenoma

Abbreviations: No.; number, n; number of participants, CRC; colorectal cancer

3.6 Data analysis

3.6.1 Random forest

For prediction of "critical to find" participants we used a supervised machine learning technique named Random Forest (RF) (104) (**see Figure 2)**. RF is a ML technique based on random subsampling and decision trees. Before creating the prediction model, the dataset is divided into two; one dataset to train the model, and one dataset to test the model. The training dataset is used to create numerous decision trees that work as an ensemble to generate the best prediction model. Each decision tree is grown with a bootstrapped dataset, this is a dataset made from randomly drawn samples, with replacement, from the training set. Each decision tree consists of a series of nodes which each split the dataset in two (internal nodes) based on the value of a variable. At each split in the decision tree, a given number of variables are randomly selected. The designated variable is the one that best splits the sample, as measured by the impurity of classes in the two resulting datasets. The tree keeps on splitting the samples until there are only a given number of samples left in the node, this is a so-called

leaf node. When all the trees are grown, the test set samples are used to test the classification accuracy of the RF model. Each sample is tested on each tree, and each tree contributes with a prediction for each sample. The prediction given by most of the trees "wins". The proportion of test samples that is correctly classified is used to measure the accuracy of the model (104). Alternatively, the fraction of trees supporting an outcome class can be treated as a continuous variable describing the probability that the tested sample belongs to that class.

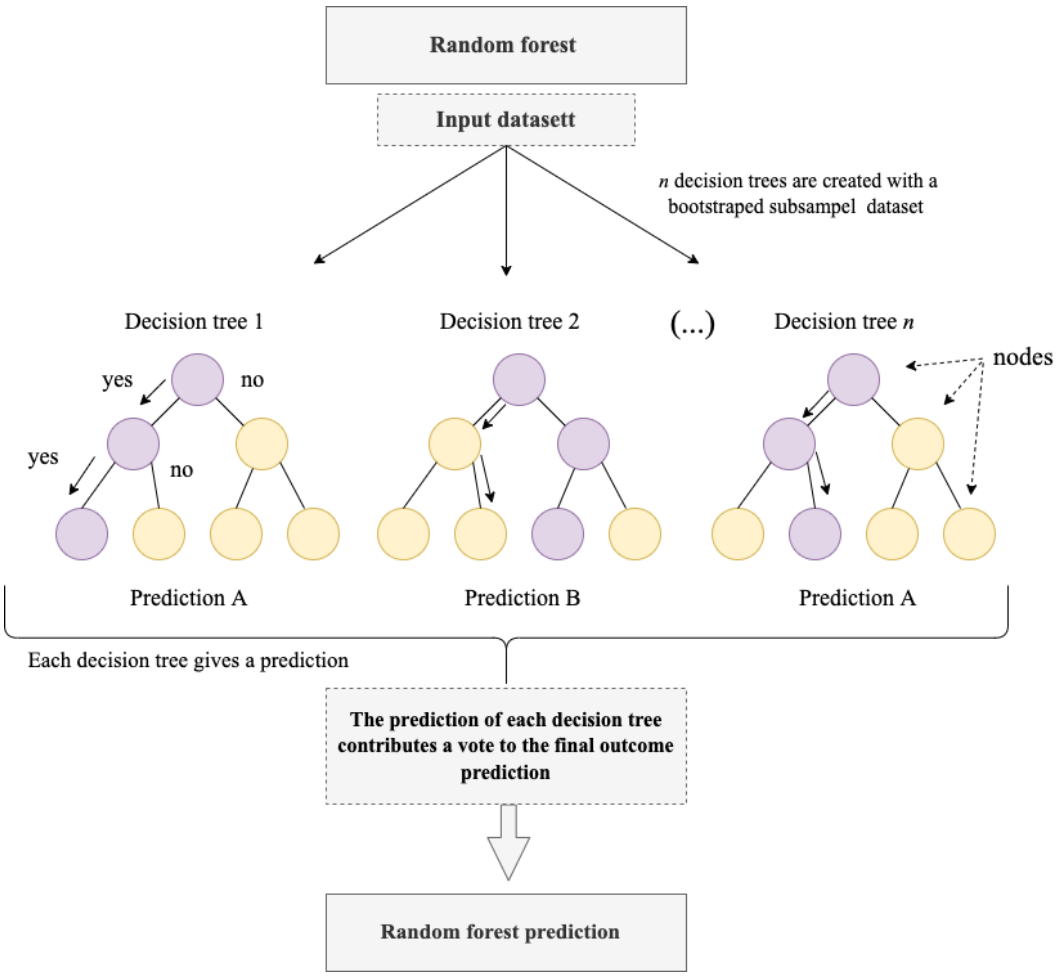


Figure 2.Simplified model of Random Forest

To construct the algorithm, the R package ranger (105) was used, as implemented in Ttidymodels (106), using the packages Parsnip (107), Recipes (108), Tune (109) and Yardstick (110) for model specification, data preprocessing, hyperparameter tuning and evaluation, respectively. The dataset was split in two parts: a training dataset containing 80 % of the original data and a test dataset containing the remaining 20 %. To ensure equality between the training and the test set a stratified split was performed. The stratified split was performed to ensure even distribution between the datasets regarding age, sex and

colonoscopy result. The training dataset was used for initial model performance, enabling comparison of different models before final evaluation in the test dataset. Here, we split the training dataset once more, randomly 80/20 to internal training/test datasets. In each internal training split, we tuned hyperparameters using 5-fold cross validation. The model was trained with different hyperparameter settings, and then validated for each setting. Hyperparameter tuning was used for: 1) the number of trees, 2) the number of variables considered when splitting a node and 3) the number of data samples needed to keep splitting the nodes. Using the area under the receiver operating characteristic curve (AUROC) as a measure of model performance, we found the optimal parameters to use in the model (see **Figure 3**). An AUROC curve is defined as a plot of the sensitivity versus 1-specificity for a diagnostic test (111). It is commonly used when sensitivity and specificity are appropriate measurements but the predictor is continuous or ordinal (111). The model trained in the internal training dataset was further evaluated in the internal test dataset. The models showing the best AUROC in the training dataset was acknowledged as the final model. The final model was then evaluated with the test dataset.

Each model was built with 50 to 500 decision trees. A hyperparameter grid search was implemented to obtain optimal parameters. To select the number of variables randomly sampled as candidates at each split, a range from the square root of all implemented variables divided by four to the square root of all implemented variables times four was used. A range from 5 to 15, was also used to determine the number of data points needed to keep splitting. The script of the Random forest is provided in **Supplementary file 5.**

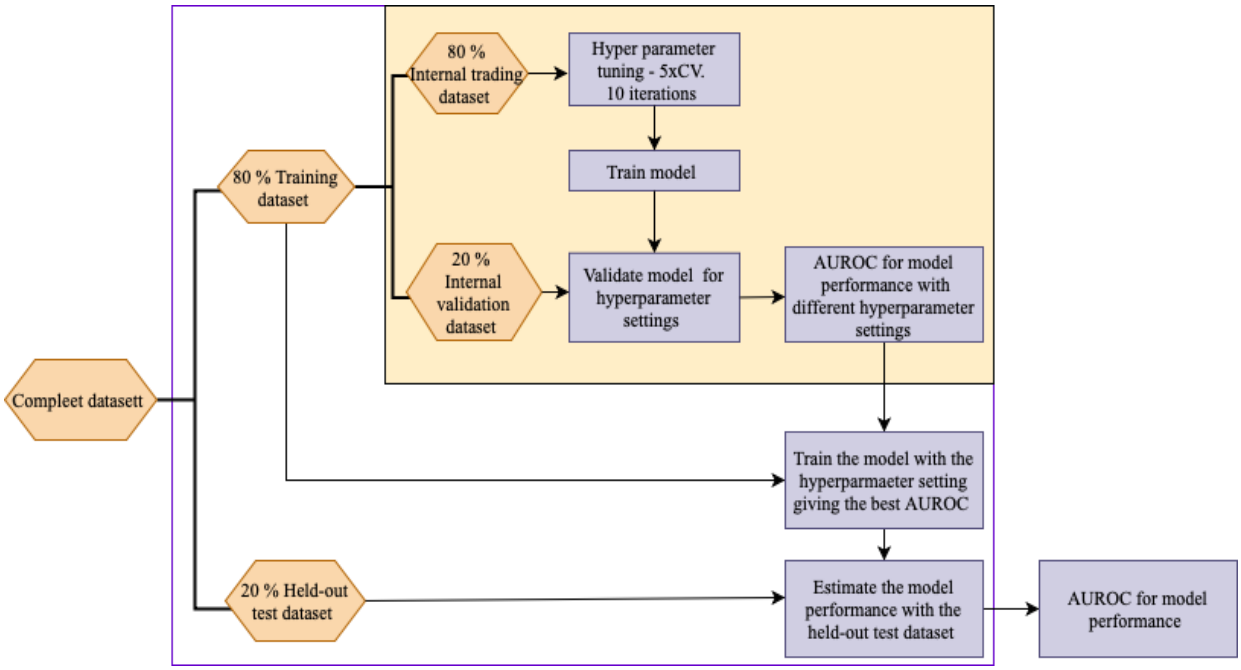


Figure 3. Random Forest pipeline. *An 80/20 split was done on the dataset, creating a training set and a test set. To maintain the distribution between sets we used stratification. To select the best hyperparameter settings to train the models, we used 5 fold cross validation. To evaluate the model we used the held-out test set. The figure is inspired by Topçuoğlu et al. (112).*

3.6.2 Selection of input variables

To answer the research questions, different combinations of input variables were tested. A total of four different variable combinations were tested. The first dataset included the FIT result, demographic and lifestyle information, as well as all dietary data *(Overall diet plus*). The second dataset included all dietary information only (*Overall diet*). The third data set included the FIT result demographic and lifestyle data, as well as selected dietary variables known to have an impact on CRC risk (*CRC relevant dietary factors plus*). The forth dataset contained only the CRC relevant dietary variables (*CRC relevant dietary factors).*

The demographic and lifestyle variables included in *Overall diet plus* and *CRC relevant dietary factors plus* were "sex", "age", "BMI", "physical activity score", "smoking habits", "education" and "family history of CRC" (**Supplementary file 6**). These variables are known predictors of CRC risk. In addition, "FIT value" was implemented in these datasets.

The dietary data used as input variables in *Overall diet plus* and *Overall diet* were selected with the intention of giving an overall impression of the participants' diet (not nutrients). Prior to this master's thesis, dietary information obtained from the FFQ had been calculated in KBS, resulting in a dataset with variables based on standard KBS grouping. The grouping of food items in KBS follows a hierarchical structure and is based on traditional categorization of food items. Each food group is subdivided into smaller categories, finally ending up as food items. An example of this is the variable "Bread" (a category 1 variable), which is divided further into "White bread", "Kneipp bread", "Whole wheat bread/Dark bread" and "Crispbread/Flat bread"(category 2 variables), where for example the variable "White bread" is further divided into smaller categories (category 3 variables, e.g. "White bread or roll from known brand"), which can be further divided into even smaller categories (e.g. "Hamburger bread", "Ciabatta"). In this study, category 2 variables were mostly chosen, nevertheless where a finer division was needed, category 3 variables were used and where a larger group made more sense to use, category 1 was used. Some variables were left to be as they originally were, and some were merged with other variables with similar nutrient content to create more inclusive variables or variables considered more relevant to the study. For the variables "Egg", "Cake" and "Spices and herbs" the category 1 variable was used. An example of where category 3 was used was "fish offal" and "shellfish", as these two were included in the same category 2 variable from standard KBS grouping. An example of different variables merged together was "unprocessed potatoes", consisting of the three category 2 variables "raw potatoes", "boiled potatoes" and "fried potatoes".

The standard KBS grouping of food items is based on foods categories and not nutrient content. To make the content in the different dietary variables most relevant for this theses some regrouping of food items was done. To accomplish this we were given access to a dataset where the smallest grouping was at the food level, hence some regrouping of food items was done if the food item did not fit optimally within the variable category. As an example, "ice tea with sugar" was regrouped from the category 2 variable "TEA" into a new variable named "Energy drinks and ice lolly". In total this was done with 14 different food items.

Two of the included category 2 variables (i.e. pasta dish and stew), consisting of compound dishes, were not further divided in to category 3 variables with the KBS grouping. Therefore they were split, and merged with other variables. As an example "Pasta dish", which only

contained the food item lasagna, was split into three: A third was put together with "Pasta, rice and grains", a third with "Sauce" and a third with "Processed red meat". A variable named “Infant food” containing nutrient drink as the only food item, and “drinking water” was excluded. In total 46 dietary variables were selected, and the grouping of the included dietary variables is accounted for in **Supplementary file 7**.

In the datasets *CRC relevant dietary factors plus* and *CRC relevant dietary factors*, only dietary factors or nutrients suggested to have an impact on CRC risk were included (19, 113). The variables were “fiber”, “wholegrain”, “red meat “, “processed meat”, “alcohol”, “dairy products”, “fruits”, “vegetables”, “fish”, “calcium” and “vitamin D” (**Supplementary file 8)**.

3.6.3 Statistics

R was used for data processing, to perform statistical analyses and machine learning. The baseline characteristics are presented as number (n) and percentage (%) for categorical variables. Continuous variables are presented as median or mean, with 25- and 75 percentile (Q1, Q3) or standard deviation (SD), respectively. Differences between diagnostic groups were investigated with the use of chi-square test for categorical variables, and Kruskal-Wallis one way analysis of variance test or one way ANOVA analysis for continuous variables. The test was performed with an assumption of independence between the different observations. Nevertheless, for chi-square test when the expected count for some cells was below 5 (i.e. people categorized as "missing"), thus the participants categorized as “missing” were excluded from the analysis. The level of significance was defined to be $p < 0.05$.

To ensure complete datasets for the RF models, participants with missing information about BMI, age and physical activity score were allocated the median value in the population. Missing information about family history of CRC, education and smoking habits were set as “Missing”. The performance of each model was assessed by computing the area under the receiver operating characteristic curve (AUROC). Other performance metrics used to evaluate the models include sensitivity and specificity. As all models run in the training dataset were carried out using a 5-fold cross validation with 10 times repetition, the means and SD is provided for all metrics in the training dataset.

To understand how the RF models generate their predictions and which variables were important for the predicative performance, a mean variable importance was calculated. Gini importance/mean decrease in impurity was used for the calculation and conducted for all input variables in a model. As described above, in a tree, each node split is based on the measure of impurity in the two resulting datasets/internal nodes. Gini importance is calculated from the sum a variable has to decrease the impurity across all decision trees in the forest. The decrease in impurity is measured by the difference between the impurity of a node and the weighted impurity of the resulting internal nodes (104, 114). We provided the mean variable importance as all models are trained with 10 iterations.

In nutritional epidemiology, energy adjustment is often used to take into account differences in energy intake (115). As described by Willet *et al*. energy intake are, among other factors, determined by body size. Thus, energy adjustment may be appropriate as a crude intake of some nutrient will be less of an effect for people with larger body sizes compared with a smaller one, due to higher energy consume (115). Accordingly, we wanted to investigate if energy adjustment could improve predictive performance. Dietary variables in the four datasets were energy adjusted into intake per 1000kcal.

Previous literature has shown that males are more prone to developing adenoma compared to females (116). Therefore, we investigated if any of the models created with the overall cohort would perform better for either males or females. This was done by dividing the overall cohort into groups based on gender.

3.6.4 Sample size and statistical power

To provide a sufficient power for development of a classification algorithm, the number of participants included should enable 1) a sufficiently large training set for development of a classification model, and 2) a leave-out test set of sufficient size for validation. The CRCbiome study employs a strategy where the dataset is split 80/20 to a training set and a leave-out test set (92), which is designed to fulfill these aims. To mitigate any issues of data leakage, the master`s thesis uses the split into a training and test set that was defined for the CRCbiome study as a whole.

3.6.5 Ethics

Regional Committee for Medical and Health Research Ethics (REC) has given ethical approval for the CRCbiome study and the BCSN trial (REC protocol Approval no. 63148 and 2011/1272, respectively) **(Supplementary file 9).** Analyses performed as part of this master`s thesis are covered by the REC approval of the CRCbiome study. The BCSN trial is registered at the National Institute of Health Clinical Trails (identified: NCT0153855). All sensitive data are processed in accordance with the General Data Protection Regulation and Norwegian Data Protection Act.

4 Results

4.1 Participant characteristics in the training dataset

The FIT result, lifestyle and demographic characteristics for the participants in the training dataset are presented in **Table 2** by colonoscopy result. Overall, 527 (44.5 %) of the participants in the training dataset were female. Median age at recruitment was 67 years, however, there was a significant difference in age between colonoscopy result groups ($p < 0.01$). Across all groups, most participants were Norwegian (90 %), and had a degree from high school or higher (81.7 %). In total, 16.3 % reported a family history of CRC, with no significant difference between groups. Nonetheless, 30 % of the participants diagnosed with CRC reported a family history of CRC. Most participants had a BMI higher than 25 kg/m^2 (65.5 %), and 19 % of the participants had a BMI > 30 kg/m^2. There was a significant difference in the distribution of BMI categories across the diagnostic groups ($p = 0.03$), with more participants having a BMI < 25 kg/m^2 in the no adenoma group (42 %) and most participants having a BMI > 30 kg/m^2 in the non-advanced serrated/other lesion group (56 %). The median physical activity score for all participants was 135 minutes/week; there were no significant differences between groups.

Table 3 shows the consumption of nutrients and food for participants in each diagnostic group. The average intake of protein, sugar, fiber, vitamin D, calcium, total fat and alcohol was in accordance with the Norwegian Directorate of Health's recommendations (117, 118). The mean consumption of saturated fat was higher than the recommended level for all diagnostic groups (<10 Energy %) (117). A significant difference in alcohol and wholegrain consumption was found between diagnostic groups. Over all, the median intake of red and processed meat in the training dataset was 441 g/week. This is in line with the Norwegian dietary recommendations to minimalize the consumption of red and processed meat to less than 500 g/week (117). The total median intake of fruits and vegetables (when juice contributes with up to 100 g), was 432 g/day, which is lower than the recommendation (> 500 g/day) (117). Median consumption of dairy products was lowest among participants with CRC (250 g/day; 25 and 75 percentile: 119, 495), and highest among participants with advanced serrated lesions (348 g/day, 25 and 75 percentile: 138, 595).

Table 2. Participant characteristics by colonoscopy result in the training dataset

	No adenoma (n = 352)	Non-advanced serrate/other lesion (n = 84)	Non-advanced adenoma (<3) (n = 311)	Non-advanced adenoma (>3) (n = 105)	Advanced serrated lesion (n = 59)	Advanced adenoma (n = 222)	CRC (n = 50)	p-values[1]
Sex								**0.005**
Female	182 (51.7)	43 (51.2)	134 (43.1)	41 (39)	26 (44.1)	78 (35)	23 (46)	
Male	170 (48.3)	41 (48.8)	177 (56)	64 (61)	33 (55.9)	144 (64.9)	27 (54)	
Age (years)	65.8 (60.6, 71.1)	65.1 (60.4, 69.9)	67.2 (62.2, 72.1)	70.1 (64.8, 74)	69.9 (65.1, 72.2)	67.1 (62.9, 71.9)	66.8 (61.6, 72.7)	**<0.001**[±]
Nationality								0.55[c]
Native Norwegian	319 (90.6)	78 (92.9)	278 (89.4)	91 (86.7)	52 (88.1)	202 (91)	43 (86)	
Non-native	25 (7.1)	3 (3.6)	16 (5.1)	7 6.7)	3 (5.1)	9 (4.1)	5 (10)	
Missing	8 (2.3)	3 (3.6)	15 (5.5)	7 (6.7)	4 (6.8)	11 (5.0)	2 (4.0)	
Education								0.52[c]
Primary school	67 (19.0)	7 (8.3)	44 (14.1)	20 (19)	10 (16.9)	39 (17.6)	8 (16)	
High school	132 (37.5)	41 (48.8)	116 (37.3)	41 (39)	23 (39)	85 (38.3)	21 (42)	
University/college	147 (41.8)	35 (41.7)	147 (47.3)	41 (39)	23 (39)	93 (41.9)	21 (42)	
Missing	6 (1.7)	1 (1.2)	4 (1.3)	3 (2.9)	3 (5.159)	5 (2.3)	0	
Family history of CRC								0.26
Yes	52(14.8)	12 (14.3)	48 (15.4)	18 (17.1)	10 (16.9)	38 (17.1)	15 (30)	
No	260 (73.9)	62 (73.8)	242 (77.8)	74 (70.5)	43 (72.9)	166 (74.8)	30 (60)	
Unknown	40 (11.4)	10 (11.9)	21 (6.8)	13 (12.4)	6 (10.25)	18 (8.1)	5 (10)	
BMI (kg/m^2)[b+]	26.6 (4.5)	27.0 (4.0)	26.9 (4.2)	27.6 (4.1)	26.6 (3.2)	27.1 (3.9)	27.3 (3.9)	0.37[∞]
BMI categories								
BMI < 25 kg/m^2	148 (42)	23 (27.4)	102 (32.8)	29 (27.6)	18 (30.5)	65 (29.3)	17 (34)	**0.03**
BMI 25-29.9 kg/m^2	66 (18.8)	12 (15.5)	69 (19.3)	22 (21)	9 (15.3)	42 (18.9)	13 (26)	
BMI >30 kg/m^2	137 (28.9)	47 (56)	148 (47.6)	53 (50.5)	32 (54.2)	114 (51.4)	19 (38)	
Missing	8 (2.3)	1 (1.2)	1 (0.3)	1 (15)	0	1 (0.5)	1 (2)	
Smoking habits								0.08[c]
Smoker	55 (15.6)	20 (23.8)	68 (21.9)	25 (23.8)	16 (27.1)	48 (21.6)	6 (12)	
Non-smoker	291 (82.7)	63 (75)	238 (76.5)	78 (74.3)	41(69.5)	169 (76.1)	44 (88)	
Missing	6 (1.7)	1 (1.2)	5 (1.6)	2 (1.9)	2 (3.45)	5 (2.3)	0	
Physical activity[a]*	135 (15, 315)	180 (0, 300)	135 (7.5, 300)	105 (15, 308)	135 (0,315)	135 (0, 300)	135 (22.5, 292)	0.95[±]
FIT value* **(**mcg haemoglobin /g feces)	32.6 (20.4, 70.2)	33.4 (24, 58.8)	32.2 (21.6, 63)	31.2 (22.4, 58.8)	30.4 (21.2, 58.8)	41.6 (27.6, 102.6)	79.8 (46, 280)	**<0.001**[±]

*Abbreviations: FIT value, Faecal Immunochemical Test value; BMI, Body Mass Index; mcg/g; micrograms per gram. Data are presented as [+]mean with standard deviation (SD) or *median with 25-and 75 percentile (Q1, Q3) for continuous variables, and numbers (%) categorical variables. [1]P-values are collected from [±]Kruskal-Wallis one way analysis of variance test, [∞]One way ANOVA and chi-square test. Information available from [a]1165 participants and [b]1177 participants. [c].Participants categorized as missing were not included in the analysis. BMI are tested as a continuous variable and as a categorical variable.*

Table 3. Nutrient and food intake by colonoscopy result in the training dataset

	No adenoma (n = 352)	Non advanced serrate/other lesion (n = 84)	Non-advanced adenoma (<3) (n = 311)	Non-advanced adenoma (>3) (n =105)	Advanced serrated lesion (n = 59)	Advanced adenoma (n = 222)	CRC (n = 50)	p-values[1]
Energy								
Kcal/d	2210 (639)	2282 (742)	2207 (666)	2128 (670)	2327 (634)	2327(664)	2117 (690)	0.09
KJ/d	9250 (1670)	9550 (3100)	9240 (2780)	8910 (2800)	9700 (2730)	9740 (2780)	8860 (2890)	0.09
Protein (E %)	17 (3)	17 (3)	16 (3)	17 (3)	16 (2)	17 (2)	17(3)	0.80
Carbohydrates (E %)	42 (7)	42 (7)	41 (7)	41 (7)	42 (6)	41 (7)	42 (7)	0.31
Sugar (E %)	5 (4)	5 (4)	5 (4)	5 (5)	6 (4)	5 (4)	5 (4)	0.66
Fiber (g/d)	29 (10)	29 (10)	28(10)	28 (11)	30 (9)	30 (11)	29 (10)	0.43
Wholegrain (g/day)	66 (45, 95)	70 (51, 96)	68 (43, 92)	57 (37, 83)	70 (46, 95)	61 (41, 91)	60 (36, 77)	**0.01**[±]
Fat (E %)	35 (6)	34 (6)	35 (6)	35 (7)	35 (6)	35 (5)	35 (6)	0.93
Saturated fat (E %)	12 (3)	12 (3)	12(3)	12 (3)	12 (3)	12 (3)	12 (2)	0.91
Alcohol (E %)	3 (1, 5)	4 (1, 7)	3 (1, 6)	3 (1, 8)	3 (1, 5)	4 (1, 6)	3 (1, 7)	**0.03**[±]
Alcohol (g/day)	7 (2, 17)	12 (3, 22)	10 (2, 20)	8 (1, 20)	11 (4, 17)	12 (4, 20)	8 (3, 20)	**0.006**[±]
Calcium (mg/day)	960 (720, 1260)	1020 (716, 1380)	933 (714, 1310)	893 (593, 1150)	1010 (696, 1410)	913 (648,1210)	834 (651, 1240)	0.32[±]
Vitamin D (µg/day)	14 (8, 24)	15 (9, 24)	14 (7, 25)	13 (7, 25)	14 (7, 24)	14 (8, 25)	12 (9, 20)	0.95[±]
Unprocessed red meat (g/day)	35 (21, 50)	35 (24, 46)	32 (20, 49)	32 (18, 48)	44 (19, 59)	37 (23, 55)	33 (21, 49)	0.14[±]
Processed meat (g/day)	26 (12, 41)	26 (16, 40)	27 (15, 43)	31 (11, 42)	25 (13, 48)	32 (18, 51)	26 (16, 39)	**0.03**[±]
Fruit (g/day)	206 (117, 300)	197 (99, 328)	184 (105, 299)	178 (109, 290)	219 (137, 320)	196 (115, 338)	218 (129, 368)	0.70[±]
Vegetables (g/day)	246 (149, 365)	221 (150, 325)	228 (146, 326)	232 (153, 349)	230 (149, 360)	235 (155, 355)	272 (182, 366)	0.52[±]
Dairy products (g/day)	331 (214, 574)	341 (160, 589)	322 (167, 643)	279 (137, 503)	348 (138, 595)	268 (127, 527)	250 (119, 495)	0.15[±]
Fish (g/day)	60 (37, 84)	54 (36, 76)	57 (36, 83)	56 (37, 85)	65 (32, 97)	64 (42, 93)	52 (31, 77)	0.17[±]

Abbreviation: Kcal/d, kilocalories per day; KJ/d, kilo Joules per day, E %, energy percent; g/d, gram per day; mg/d, milligram per day; µg/d, microgram per day. Nutrient and food group intake presented as mean with standard deviation (SD) or median with 25-and 75 percentile (Q1, Q3). [1]P-values from One way ANOVA and [±]Kruskal-Wallis One way analysis of variance test. P-values presented in bold font type are statistically significant (<0.05). Definition of unprocessed red meat, processed meat, fruits, vegetables, dairy products and fish are presented in supplementary file 7.

4.2 Model performance within the training dataset

We wanted to investigate if it was possible to create a prediction model that was able to identify participants who were in need of colonoscopy in a FIT-positive population. Which of the participants considered to be critical to find varies among cohort, described in **Table 1, pg. 15**. Four datasets were created: *Overall diet plus*, *Overall diet*, *CRC relevant dietary factors plus* and *CRC relevant dietary factors*. **Table 4** shows an overview of the model names, input dataset and cohort used as outcome.

Table 4. Prediction model names

Dataset input	**Overall cohort** *n=1183*	**Sub-cohort no.2** *n=872*	**Sub-cohort no.3** *n=813*	**Sub-cohort no.4** *n=708*
Overall diet plus *(46 dietary variables, 7 lifestyle and demographic variables and FIT values)*	Model 1	Model 1.2	Model 1.3	Model 1.4
Overall diet *(46 dietary variables)*	Model 2	Model 2.2	Model 2.3	Model 2.4
CRC relevant dietary factors plus *(11 dietary variables with impact on CRC development, 7 lifestyle and demographic variables and FIT values)*	Model 3	Model 3.2	Model 3.3	Model 3.4
CRC relevant dietary factors (*11 dietary variables with impact on CRC development)*	Model 4	Model 4.2	Model 4.3	Model 4.4

4.2.1 Model performance when including the overall cohort

Initially, we created four different models with each of the datasets and the overall cohort. All prediction models included in the results of this thesis were created with down sampling; this is done by only selecting as many samples from the majority class as from the rarest class. In terms of predictive performance within the overall cohort, Model 3.4 was best at distinguishing those who were in greater need of colonoscopy from true negatives with an AUROC of 0.59, see **Table 5**. Further, 3.4 had the highest sensitivity (0.59) and specificity (0.53) thus it has a greater ability to correctly designate participants compared with the other models created with the other datasets, in the overall cohort.

Table 5. Model performance with the overall cohort (n = 1183)

	Sensitivity	Specificity	AUROC
Model 1	0.56 (0.04)	0.51 (0.02)	0.55 (0.02)
Model 2	0.55 (0.05)	0.49 (0.04)	0.51 (0.02)
Model 3	0.59 (0.05)	0.53 (0.02)	**0.59** (0.03)
Model 4	0.54 (0.05)	0.51 (0.02)	0.54 (0.03)

Abbreviations: AUROC, Area under the curve receiver operating characteristic. Models created with the datasets Overall diet plus, Overall diet, CRC relevant dietary factors plus and CRC relevant dietary factors, all trained in the overall cohort. All models are run with 10 iteration. Performance is presented as mean sensitivity, specificity and AUROC, with standard deviation (SD).All models are created with downsampling.

4.2.2 Sub-cohort analysis

Further, we wanted to investigate if it was possible to create a model with better performance training the models in the sub-cohorts (described in **Table 1. Pg. 15**). Overall, Model 3.4 showed the highest predictive performance, see **Table 6**. Moreover, the three models created with the dataset *CRC relevant dietary factors plus* generated the highest accuracy with an AUROC ranging from 0.61 to 0.64 depending on which sub-cohort the model was trained in. Models created in the sub-cohort no.4 (models named Model X.4) had the highest AUROC compared to models trained in the other sub-cohorts. Model 3.2 and Model 3.4 were best at classifying participants with adverse colorectal findings into the critical to find group, with both yielding a sensitivity of 0.62. The highest specificity was obtained for Model 4.3 (0.63), showing the highest accuracy in classification of participants without adverse findings.

Table 6. Model performance for the sub-cohorts, within the training dataset

	Sens	Spec	AUROC		Sens	Spec	AUROC		Sens	Spec	AUROC
Model 1.2	0.59 (0.07)	0.55 (0.04)	0.58 (0.02)	**Model 1.3**	0.58 (0.05)	0.55 (0.06)	0.57 (0.04)	**Model 1.4**	0.57 (0.10)	0.59 (0.06)	0.62 (0.05)
Model 2.2	0.55 (0.06)	0.50 (0.05)	0.53 (0.04)	**Model 2.3**	0.52 (0.04)	0.52 (0.05)	0.52 (0.04)	**Model 2.4**	0.55 (0.05)	0.49 (0.04)	0.51 (0.03)
Model 3.2	0.62 (0.05)	0.57 (0.06)	0.62 (0.03)	**Model 3.3**	0.59 (0.05)	0.56 (0.04)	0.61 (0.04)	**Model 3.4**	0.62 (0.09)	0.60 (0.04)	**0.64** (0.05)
Model 4.2	0.58 (0.05)	0.54 (0.04)	0.57 (0.03)	**Model 4.3**	0.45 (0.04)	0.63 (0.05)	0.56 (0.02)	**Model 4.4**	0.61 (0.08)	0.56 (0.06)	0.61 (0.07)

Abbreviations: Sens, sensitivity; Spec, specificity; AUROC, Area under the curve receiver operating characteristic. Models created with Overall diet plus, Overall diet, CRC relevant dietary factors plus and CRC relevant dietary factors trained within the three sub-cohorts. All models are run with 10 iterations. Performance is presented as mean sensitivity, specificity and AUROC, with standard deviation (SD). All models are created with downsampling.

4.2.3 Variable importance

To understand how the Random Forest models generate their predictions, variable importance is calculated. This was done for all models created with all the four datasets and trained within the different cohorts. We report the mean variable importance for the 10 iterations in the training dataset.

Table 7 lists the 10 most important variables ranked according to their contribution in the model, alongside their mean importance score. The models created with *Overall diet plus* and *CRC relevant dietary factors plus*, including the FIT result, lifestyle, and demographic variables as input only ranked "FIT value", "age" and "BMI" among the non-dietary variables to be among the ten most important. Moreover, it is worth noticing that "FIT value" and "age" were ranked as top two for most models. Both Model 1.4 and Model 3.4 ranked "FIT value" as the most important variable, with a mean importance calculated of 12.56 and 15.50, respectively. Further the models ranked "milk & yoghurt" and "dairy products" with a mean importance of 7.13 and 11.35, respectively. A gap in mean importance was seen between the "FIT value" until the next variable.

Variable importance for models created with *Overall diet plus* or *Overall diet* were homogeneous. Almost all models ranked "milk and yoghurt", "white cheese" and "grain products" among their ten most important variables. Other variables identified as important for models trained with this dataset was "processed red meat", "wine and liquor" and "fresh and frozen fruit". Interestingly, several models additionally ranked "unprocessed potatoes" and "conserved vegetables" to be among the most important variables.

Models created with *CRC relevant dietary factors plus* or *CRC relevant dietary factors* showed similar findings as described above. Almost all models ranked both "dairy products", "wholegrain" and "processed meat" as the most important dietary variables for the prediction. Further "red meat", "alcohol", "vitamin D" and "fish" were ranked among the ten variables of importance for most models.

Table 7. Mean variable importance for models

Model 1		Model 1.2		Model 1.3		Model 1.4	
FIT value	10.98	FIT value	12.36	Age	10.90	FIT value	12.56
Age	9.58	Age	11.86	FIT value	10.60	Milk & yoghurt	7.13
Milk & yoghurt	8.70	Milk & yoghurt	10.77	Milk & yoghurt	9.10	Proc red meat	5.76
Cons veg	7.55	BMI	10.04	BMI	7.72	Wine & liquor	5.38
Grain products	7.41	Wine and liquor	8.69	Proc red meat	6.74	Age	5.28
Butter	7.31	Grain products	8.59	Wine and liquor	6.71	Cons veg	4.95
White cheese	7.29	White cheese	7.85	Conserved fruits	6.69	BMI	4.94
BMI	7.24	Vegetables	7.79	White cheese	6.62	Cream products	4.84
Unproc potato	7.22	Unproc potato	7.70	Grain products	6.42	White cheese	4.81
Sause	7.13	Conserved fruits	7.64	Cream products	6.37	Grain products	4.74
Model 2		**Model 2.2**		**Model 2.3**		**Model 2.4**	
Milk & yoghurt	10.02	Milk & yoghurt	12.23	Milk & yoghurt	11.06	Milk & yoghurt	8.22
Unproc potato	8.77	Unproc potato	9.85	Conserved fruits	8.16	Proc red meat	6.62
Butter	8.69	Grain products	9.51	Proc. read meat	8.05	Cream products	5.92
Grain products	8.53	Wine and liquor	8.92	Cream products	7.96	Wine and liquor	5.86
Cons veg	8.34	Cream products	8.64	Unproc potato	7.67	Cons veg	5.65
Lean fish	8.04	Fatty fish	8.64	Fatty fish	7.51	Grain products	5.60
White cheese	8.03	Proc red meat	8.63	White cheese	7.42	Unproc potato	5.60
Sweetener	7.94	Fre/froz fruits	8.54	Lean fish	7.37	Butter	5.53
Fre/froz. fruits	7.79	Vegetables	8.46	Wine and liquor	7.35	Fre/ froz fruits	5.38
Sause	7.78	Sause	8.40	Fre/froz fruits	7.34	Sauce	5.37
Model 3		**Model 3.2**		**Model 3.3**		**Model 3.4**	
FIT value	17.42	Age	20.80	Age	13.07	FIT value	15.50
Age	16.06	FIT value	19.82	Dairy products	12.93	Dairy products	11.35
Vitamin D	15.24	Wholegrain	17.90	FIT value	12.82	Alcohol	10.78
Wholegrain	14.99	Vitamin D	17.66	Wholegrain	12.37	Processed meat	10.73
Red meat	14.73	Dairy products	17.21	Processed meat	11.72	Wholegrain	10.37
Processed meat	14.36	Red meat	17.17	BMI	11.72	Red meat	10.32
Fish	14.24	Alcohol	17.11	Calcium	11.71	Vitamin D	10.01
Vegetables	14.20	BMI	16.66	Red meat	11.61	Age	9.86
Fruits	14.04	Fish	16.36	Alcohol	11.47	Fruits	9.52
Alcohol	13.98	Processed meat	16.15	Fish	11.30	Fish	9.37
Model 4		**Model 4.2**		**Model 4.3**		**Model 4.4**	
Fish	25.02	Dairy products	27.34	Wholegrain	25.54	Dairy products	19.15
Wholegrain	24.69	Wholegrain	26.20	Dairy products	25.40	Processed meat	18.28
Dairy products	24.28	Processed meat	26.05	Processed meat	24.56	Wholegrain	16.79
Processed meat	23.75	Fish	26.01	Fish	22.87	Fish	16.60
Fruits	23.62	Vitamin D	25.59	Vegetables	22.73	Fruits	16.47
Vegetables	23.57	Red meat	24.92	Fruits	22.53	Alcohol	16.23
Vitamin D	23.05	Fruits	24.68	Vitamin D	22.27	Vitamin D	16.30
Calcium	22.14	Vegetables	24.16	Red meat	22.24	Red meat	15.53
Alcohol	21.71	Alcohol	23.92	Alcohol	21.94	Vegetables	15.25
Red meat	21.36	Calcium	22.33	Fiber	21.26	Calcium	15.01

Abbreviations: Cons veg; conserved vegetables; Unproc potatoes, unprocessed potatoes; Proc red meat, processed red meat; Milk & yoghurt, milk and yoghurt; Fre/froz fruits, fresh and frozen fruit; FIT value; Faecal Immunochemical Test value; BMI, Body Mass Index. Variable importance of models created Overall diet plus, Overall diet, CRC relevant dietary factors plus and CRC relevant dietary factors trained within with all the four cohorts. The ten most important variables with their respective mean importance value, the most influential variables are at the top of every list.

4.2.4 Model performance with energy adjusted variables

Energy adjustment (intake per 1000 kcal) of the dietary variables was performed to further investigate their impact on the prediction models, and whether they were affected by the participant's energy intake. The models created with energy adjusted variables were named according to the models described above as only the energy adjustment differentiate them. However, an "E" is added at the end of the name to enable discrimination between models.

Models created with the energy adjusted *CRC relevant dietary factors plus* as input had a better performance than the models created with crude dietary variables, see **Table 8**. The best predictive performance was shown in Model 3.4 E with an AUROC of 0.65. This model had a sensitivity of 0.65, being the highest of all models. Model 3.2 E generated an AUROC of 0.64, with a sensitivity of 0.64 and specificity of 0.58. Model 3.2 E was trained in a larger cohort (n=872) compared with Model 3.4 E (n=708). The predictive performance of models created with the energy adjusted *Overall diet* or *Overall diet plus* were worse than the models created with the original datasets (that is not energy adjusted). The models created with energy adjusted *CRC relevant dietary factors* showed the same predictive performance as models created with the original dataset.

Table 8. Models with energy adjusted (intake per 1000 kcal) dietary variables

Model name	Sens	Spec	AUROC	Model name	Sens	Spec	AUROC
Model 1E	0.55	0.52	0.56	**Model 2E**	0.53	0.49	0.52
	(0.05)	(0.05)	(0.03)		(0.05)	(0.04)	(0.03)
Model 1.2 E	0.57	0.54	0.57	**Model 2.2 E**	0.52	0.49	0.51
	(0.04)	(0.04)	(0.03)		(0.07)	(0.09)	(0.04)
Model 1.3 E	0.57	0.54	0.56	**Model 2.3 E**	0.51	0.51	0.51
	(0.04)	(0.06)	(0.03)		(0.05)	(0.05)	(0.04)
Model 1.4 E	0.56	0.58	0.59	**Model 2.4 E**	0.49	0.55	0.52
	(0.07)	(0.07)	(0.04)		(0.08)	(0.05)	(0.04)
Model name	Sens	Spec	AUROC	Model name	Sens	Spec	AUROC
Model 3 E	0.62	0.55	0.60	**Model 4 E**	0.56	0.50	0.54
	(0.04)	(0.02)	(0.03)		(0.04)	(0.05)	(0.03)
Model 3.2 E	0.64	0.58	0.64	**Model 4.2 E**	0.59	0.52	0.56
	(0.06)	(0.03)	(0.03)		(0.05)	(0.06)	(0.02)
Model 3.3 E	0.62	0.56	0.63	**Model 4.3 E**	0.54	0.54	0.56
	(0.06)	(0.04)	(0.04)		(0.05)	(0.06)	(0.04)
Model 3.4 E	0.65	0.58	0.65	**Model 4.4 E**	0.62	0.55	0.61
	(0.10)	(0.03)	(0.05)		(0.10)	(0.04)	(0.06)

Abbreviations: Sens, Sensitivity; Spec, Specificity; AUROC, Area under the curve receiver operating characteristic. All the dietary variables used in Overall diet plus, Overall diet, CRC relevant dietary factors plus and CRC relevant dietary factors were energy adjusted into intake per 1000kcal for food items/nutrients, instead of total intake per day. We created models with all the four (energy adjusted) datasets as input and all the four cohorts as output. Each model were run with 10 iterations, hence model performance is presented as mean sensitivity, specificity and AUROC, as well as standard deviation (SD). All models were created with down sample.

Mean variable importance was calculated to examine which variables were of greatest importance to the model prediction and to which degree it differed from the models without energy adjustment, see **Supplementary file 10**. **Table 9** lists mean variable importance of models created with *CRC relevant dietary factors plus* and *CRC relevant dietary factors*. Model 3.2 E generated an improved predictive performance compared to Model 3.2, both models ranked the same five variables as the most important. Model 3.4 E ranked "FIT value" as most important, same as Model 3.4. In both models "FIT value" was assigned a mean importance value subsequently higher than the variable ranged as second best. "FIT value" was assigned a mean importance value of 14.69, whereas "wholegrain" ranked as number two was assigned a mean importance value of 10.82. Model 4.4 E, yielded an equal AUROC as Model 4.4. Both models ranked "dairy products" as the most important variable. Further in Model 4.4 E "calcium", "wholegrain" and "alcohol" were ranked on the top of the list. Interestingly, energy adjusted "calcium" (mg /1000kcal) was ranked as more important in both Model 3.4 E and Model 4.4 E. In Model 3.4 "calcium" was not among the most important variables, and in Model 4.4 "calcium" was ranged as the tenth most important variable out of 11 possible input variables.

Table 9. Mean variable importance of models with energy adjusted dietary variables

Model 3 E		**Model 3.2 E**		**Model 3.3 E**		**Model 3.4 E**	
FIT value	16.96	Age	17.65	Wholegrain	12.69	FIT value	14.69
Age	15.79	FIT value	17.13	FIT value	12.67	Wholegrain	10.82
Wholegrain	15.34	Wholegrain	16.62	Dairy products	12.57	Dairy products	10.37
Vitamin D	15.22	Dairy products	16.52	Age	12.58	Calcium	9.84
Calcium	14.87	Vitamin D	15.63	Calcium	11.51	Alcohol	9.47
Fish	14.43	Fruits	15.43	Processed meat	11.43	Vitamin D	9.11
Fruits	14.39	Calcium	15.36	Red meat	11.40	Vegetables	8.97
Vegetables	14.34	Alcohol	15.31	BMI	11.39	Age	8.91
Red meat	14.32	BMI	15.24	Fruits	11.34	BMI	8.73
Alcohol	14.18	Red meat	14.69	Vitamin D	11.18	Red meat	8.64
Model 4 E		**Model 4.2 E**		**Model 4.3 E**		**Model 4.4 E**	
Fish	24.06	Dairy products	25.28	Dairy products	22.21	Dairy products	15.45
Wholegrain	23.99	Wholegrain	24.20	Wholegrain	22.87	Calcium	14.13
Dairy products	23.62	Alcohol	23.60	Calcium	20.11	Wholegrain	14.08
Fruits	23.01	Fruits	23.56	Vegetables	19.96	Alcohol	14.08
Alcohol	22.85	Vitamin D	23.13	Processed meat	19.79	Processed meat	13.59
Vitamin D	22.62	Fish	23.03	Red meat	19.77	Fruits	13.55
Calcium	22.36	Red meat	22.71	Fruits	19.69	Vegetables	13.31
Vegetables	22.21	Vegetables	22.40	Fiber	19.16	Vitamin D	13.29
Red meat	21.88	Calcium	22.20	Vitamin D	19.02	Red meat	13.02
Processed meat	21.76	Processed meat	21.94	Alcohol	18.75	Fish	12.70

Abbreviations: FIT, Faecal Immunochemical Test value; BMI, Body Mass Index. Variable importance of models created with energy adjusted CRC relevant dietary factors plus and CRC relevant dietary factors, trained in all the four cohorts. The ten most important variables are listed with their respective mean importance value, the most influential variables are at the top of every list.

4.2.5 Sub analysis

There was no difference in model performance between sexes, and stratifying by gender did not improve model performance, see **Table 10**. The stratification was only dun in the overall cohort.

Table 10. Model performance stratified by sex

		Sensitivity	Specificity	AUROC
Model 1	Female	0.51 (0.10)	0.51 (0.07)	0.52 (0.05)
	Male	0.51 (0.03)	0.50 (0.05)	0.51 (0.04)
Model 2	Female	0.47 (0.11)	0.53 (0.04)	0.51 (0.06)
	Male	0.51 (0.03)	0.50 (0.05)	0.51 (0.04)
Model 3	Female	0.53 (0.10)	0.56 (0.05)	0.55 (0.06)
	Male	0.53 (0.08)	0.51 (0.04)	0.53 (0.05)
Model 4	Female	0.51 (0.07)	0.51 (0.07)	0.52 (0.03)
	Male	0.50 (0.05)	0.51 (0.04)	0.52 (0.04)

Abbreviations: AUROC, Area under the receiver operating characteristic curve. Model performance when the overall cohort is spilt into males and female groups. Prediction models are created with the use of Overall diet plus, Overall diet, CRC relevant dietary factors plus and CRC relevant dietary factors. All models are run with 10 iterations, and performance is presented as mean sensitivity, specificity and AUROC, as well as standard deviation (SD).

As the FIT test are commonly used as a screening tool by itself we wanted to investigate how our models work when the result from the FIT (i.e. "FIT value") test was not used as input. Since the only dataset including "FIT value" and showing an interesting predictive performance was the *CRC relevant dietary factors plus*, only this dataset was investigated. **Table 11** shows predictive performance of models created without the "FIT value"**.** A minimal reduction in predictive performance was seen for Model 3 (no FIT value), Model 3.3 (no FIT value) and Model 3.4 (no FIT value), compared to the corresponding models including the FIT result. Moreover, Model 3.2 (no FIT value) was found to classify participants as correctly as the corresponding models with the FIT result, showing an AUROC of 0.62 and a sensitivity of 0.62.

Table 11. Model performance with *CRC relevant dietary factors plus* (no FIT value)

	Sensitivity	Specificity	AUROC
Model 3 (no FIT value)	0.58 (0.05)	0.52 (0.04)	0.57 (0.04)
Model 3.2 (no FIT value)	0.62 (0.06)	0.55 (0.05)	0.62 (0.03)
Model 3.3 (no FIT value)	0.59 (0.05)	0.56 (0.06)	0.60 (0.03)
Model 3.4 (no FIT value)	0.60 (0.09)	0.58 (0.04)	0.61 (0.05)

Abbreviations: AUROC, Area under the receiver operating characteristic curve. Performance of models using CRC relevant dietary factors plus without the FIT value as input, the models are trained in the four cohorts. Models are run with 10 iterations, and performance is presented as mean with standard deviation (SD).

Table 12 reports the mean important variables for the models created with *CRC relevant dietary factors plus* where the variable "FIT value" was not included as input. "Age" was ranked as the most important variable among three of the models. However, in the corresponding models (Model 3, Model 3.2 and Model 3.3) the variable "age" was ranged as the most or second most important variable. Model 3.4 (no FIT value) ranged "dairy products", "alcohol" and "wholegrain" as most important variables, compared to Model 3.4 which ranked "FIT value", "dairy products" and "alcohol" as the most important variable. The exclusion of "FIT result" did not lead to other lifestyle and demographic variables being ranked among the ten most important variables.

Table 12. Mean variable importance (no FIT value)

Model 3 (no FIT value)		**Model 3.2 (no FIT value)**		**Model 3.3 (no FIT value)**		**Model 3.4 (no FIT value)**	
Age	19.89	Age	19.84	Age	13.57	Dairy products	10.58
Vitamin D	18.52	Wholegrain	17.70	Dairy products	13.25	Alcohol	9.86
Red meat	17.95	Dairy products	17.34	Processed meat	12.56	Wholegrain	9.71
Wholegrain	17.94	Vitamin D	17.09	Wholegrain	12.10	Processed meat	9.54
Dairy products	17.90	BMI	17.09	Red meat	11.94	Red meat	9.48
Fish	17.76	Red meat	16.90	Vitamin D	11.83	Age	9.34
Vegetables	17.37	Alcohol	16.74	Calcium	11.65	Calcium	9.19
Processed meat	17.35	Processed meat	16.04	BMI	11.56	Fish	9.14
Fruits	17.22	Fruits	17.72	Alcohol	11.54	BMI	9.5
Alcohol	17.17	Fish	15.51	Fruits	11.42	Fruits	9.10

Abbreviations: BMI, Body Mass Index. Variable importance of models using CRC relevant dietary factors plus without the FIT result as input, trained in all four cohort with 10 iterations. The ten most important variables with their respective mean importance value, the most influential variables are at the top of every list.

To further investigate how each individual variable in the *CRC relevant dietary factors plus (with both crude and energy adjusted dietary data)* were to correctly classify participants within sub-cohort no.4, we calculated the AUROC (see **Table 13)** associated with each variable in isolation. Only sub-cohort no.4 was tested as the best predictive models was trained within this cohort. The "FIT value" alone provided an AUROC of 0.62, a predictive performance only 0.02 lower than Model 3.4. The variables found to provide the highest accuracy individually, was the same variables found to be among the most important variables for Model 3.4 (i.e. dairy products, alcohol and processed meat) and Model 3.4 E (i.e. whole grain, dairy products and calcium).

Table 13. AUROC of the individual variables used in *CRC relevant dietary factors plus*

Dietary variables (Used in both datasets)	**AUROC** (Crude intake)	**AUROC** (Energy adjusted)	**Other variable** (Only used in the *CRC relevant dietary factors plus)*	**AUROC**
Dairy products	0.55	0.57	FIT value	**0.62**
Alcohol	0.57	0.57	Sex	0.57
Processed meat	0.57	0.56	Family history of CRC	0.45
Wholegrain	0.55	0.58	Smoking habits	0.48
Red meat	0.54	0.52	Education	0.50
Fish	0.54	0.53	Physical activity score	0.51
Fruits	0.48	0.51	Age	0.54
Vegetables	0.51	0.50	BMI	0.53
Vitamin D	0.48	0.49		
Calcium	0.54	0.58		
Fiber	0.50	0.52		

Abbreviations: AUROC, Area under the receiver operating characteristic curve; FIT value, Faecal Immunochemical Test value. AUROC of the variables used in CRC relevant dietary factors plus and CRC relevant dietary factor, both energy adjusted and with original dataset (crude intake).

4.3 Model performance within the test dataset

In ML, the test set is utilized for an objective evaluation of the model. Because of a risk of information leakage in the CRCbiome study there was a desire to minimize the use of test datasets. The minimal use of the test dataset is determined by the Data access committee in the CRCbiome study (92). Therefore, only four of the models were evaluated with the test dataset. Model 3.4, Model 4.4 and Model 3.4 E and Model 4.4 E were the selected models. Model 3.4 and Model 3.4 E were selected due to the overall greater predictive performance in the training dataset. Model 4.4 and Model 4.4 E were selected because of moderate predictive performance in the training dataset, and they uses the same dietary input variables as Model 3.4 and Model 3.4 E, respectively.

The overall predictive performance (i.e. AUROC) for all models were reduced in the test dataset. Model 3.4 and Model 4.4 were slightly reduced in the test dataset compared to the training dataset, both showing an AUROC of 0.59 (**Table 14**). Even so, both models improved their ability to categorize people with advanced adenomas and CRC as critical to find with a sensitivity of 0.63. Model 3.4 E and Model 4.4 E had a substantial reduction in predictive performance showing an AUROC of 0.57 and 0.55, compared to training dataset 0.65 and 0.61, respectively. In addition, there was a reduction in both sensitivity and specificity for both Model 3.4 E and Model 4.4 E in the test dataset.

Table 14. Model performance in the test dataset

	Sensitivity	Specificity	AUROC
Model 3.4	0.63	0.50	0.59
Model 4.4	0.63	0.53	0.59
Model 3.4 E	0.54	0.52	0.57
Model 4.4 E	0.55	0.50	0.55

Abbreviations: AUROC, Area under the curve receiver operating characteristic. Classification evaluation of Model 3.4, Model 4.4, Model 3.4 E and Model 4.4 E with the test dataset.

Table 15 lists the variable importance in the four models evaluated in the test dataset. For Model 3.4, the same four variables were ranked as most important in the test dataset as in the training dataset. "FIT value", "dairy products", "processed meat" and "alcohol" had an importance value of 15.15, 12.26, 12.19 and 11.89, respectively. For Model 4.4, only the first two variables in the ranking were the same in training and the test dataset. These variables, "processed meat" and dairy products", which were assigned an importance value of 28.84 and 28.31. "Alcohol" and "fruit" were further ranked as important, and assigned an importance value of 25.87 and 23.19, respectively.

In Model 3.4 E arranged the same three variables at the top of the variable importance list in both the test and the training dataset. "FIT value" was arranged at the top and given an importance value of 25.48. "Dairy products" and "wholegrain" were assigned an importance value of 16.06 and 16.05. In the test dataset Model 4.4 E, ranked "dairy products", "calcium" and "alcohol" as the most important variables. The variables generated an importance value of 19.46, 18.41 and 18.25. Nevertheless, there was a minimal difference the importance value from the top of the list until the bottom, only ranging from 19.46 to 16.49, showing an almost similar importance of all the included variables.

Table 15. Variable importance of models evaluated in the test dataset

Model 3.4		Model 4.4		Model 3.4 E		Model 4.4 E	
Variable		Variable		Variable		Variable	
FIT value	15.11	Processed meat	28.84	FIT value	25.48	Dairy products	19.46
Dairy products	12.26	Dairy products	28.31	Dairy products	16.06	Calcium	18.41
Processed meat	12.19	Alcohol	25.87	Wholegrain	16.05	Alcohol	18.25
Alcohol	11.80	Fruit	23.19	Alcohol	14.77	Processed meat	17.97
Fish	11.60	Vegetables	23.17	Vegetables	12.85	Wholegrain	17.91
Age	11.50	Wholegrain	22.87	Fruits	12.83	Fruit	17.17
Calcium	11.31	Vitamin D	22.80	Fish	12.75	Fish	17.09
BMI	11.30	Red meat	22.04	Calcium	12.70	Vitamin D	17.05
Wholegrain	11.23	Fish	21.59	Age	12.60	Fiber	16.80
Red meat	11.19	Calcium	19.40	Vitamin D	12.57	Vegetables	16.77
Fruits	11.19	Fiber	19.09	Processed meat	12.30	Red meat	16.49
Vitamin D	11.15			Red meat	11.12		
Fiber	10.89			BMI	10.81		
Vegetables	10.98			Fiber	10.38		
Physical activity score	8.74			Physical activity score	7.99		
Sex	4.49			Sex	3.73		
Family history of CRC	3.99			Education	2.00		
Education	3.89			Family history of CRC	1.57		
Smoking habits	3.10			Smoking habits	1.00		

Abbreviations: FIT value, Faecal Immunochemical Test value; BMI, Body mass index. Variable importance for Model 3.4, Model 4.4, Model 3.4 E and Model 4.4 E in the test dataset.

5 Discussion

In this study, the machine learning method Random Forest was used to create a classification algorithm to detect screening participants with a greater need of colonoscopy. Different models, with various input and outcome variables were created. Further, the best models were evaluated in the test dataset. We found that the best models to distinguish participants were created with the dataset *CRC relevant dietary factors plus*. This dataset includes 11 dietary variables (with or without energy adjustment), 7 lifestyle and demographic variables and the result from the FIT test. In the training dataset Model 3.4 and Model 3.4 E, with the dataset *CRC relevant dietary factors plus* as input variables, were best at distinguishing those with advanced adenomas or CRC from true negative participants, showing an AUROC of 0.64 and 0.65, respectively. Further, these models were evaluated in the test dataset, showing an AUROC of 0.59 and 0.57, respectively.

Dietary information appeared to be important for the prediction of all the models. Even the models created with datasets including more than dietary variables (i.e. *Overall diet plus* and *CRC relevant dietary factors plus*) as input, ranked dietary variables to be important. Only the non-dietary variables "age", "FIT value" and "BMI" were ranked among the ten most important variables for all the models. The remaining important variables consisted of dietary variables. All models showed a consistency in which dietary variables were the most important for the prediction. "Dairy products", "wholegrain", "alcohol" and "processed meat" were listed among the top ten most important in most models.

5.1 Methodological considerations

5.1.1 The study population

All the participants in this master`s thesis were recruited from the CRCbiome study (92), who further recruited their participants from the BCSN trial (47). Only participants with a positive FIT test were asked to participate in the CRCbiome study (92). Participants were invited to the CRCbiome study after being informed about their FIT result, but before the colonoscopy. A complete colonoscopy was required to participate (92). Along with the invitation to be a part of the CRCbiome study, the participants received two questionnaires (LDQ and FFQ), and these were to be filled out prior to the colonoscopy. The worrying news of a positive FIT test may have affected the participants, which may result in an overrepresentation of highly

motivated subjects completing the questionnaires. Since the participants in our study have undergone two different screening methods, as well as completed a comprehensive FFQ and LDQ, it is reasonable to assume that selection bias has been introduced. Previous studies have shown that people with low socioeconomic status (119-122), an unhealthy lifestyle (121, 123), and those with a non-native ethnicity (120, 124) participate to a lesser extent in screening programs. A study by Botteri *et al*., examining the same population as included in this master`s thesis, showed that people using psychotropic drugs or antidiabetics, those with a long driving distance to the screening center and those with immigrant status to a lesser extent showed up for subsequent colonoscopy after a positive FIT test (125). These findings indicate that our population may not be as generalizable as we had hoped. However, we assume that this study represents people that would participate in screening programs for CRC.

5.1.2 Assessment of dietary intake with the FFQ

The FFQ used to assess dietary information in this master`s thesis has previously been validated for food items, micro- and macro nutrients as well as energy intake (93-98). Even though the validity of the questionnaire by most studies has been considered as reasonable, it still has some limitations regarding food components associated with CRC. The FFQ has been found to overreport the consumption of vegetables (97) and fruits (96) and fibre (93) compared to other dietary assessment methods in validation studies. Several validation studies have found underestimation of alcohol (93, 126) and sugar consumption (93, 95, 126). When interpreting the findings from the FFQ, it is important to have the results of the validation studies in mind, as the participants can appear healthier than what they are. In addition, Andersen *et al.* regarded the FFQ as valid to assess the dietary consumption at group level, but limited to assess the dietary habits at individual level (94). This is possibly a weakness in our study, as we want to classify the participants as correctly as possible with the help of, among other things, dietary variables.

The FFQ was self-administered and participants were asked to report the consumption of food the past year. Thus both recall bias and information bias may have been introduced. A potential side effect of participating in screening is the potential psychological stress. One can imagine that fear or anxiety that may occur after a positive FIT test had an impact on participants' answers. However, a sub study of the BCSN pilot, showed that receiving a

positive FIT test did not generate more clinically relevant psychological harm (127). Thus, it is unlikely that fear or other psychological changes caused by a positive screening result, have meaningfully affected the completion of the questionnaires.

5.1.3 Assessment of lifestyle and demographic data

7 lifestyle and demographic variables (**Supplementary file 2**) were used as input in the datasets *Overall diet plus* and *CRC relevant dietary factors plus*. Of these "physical activity", "smoking habits", "education" and "family history of CRC" were collected through a self-administered LDQ. This method may have some uncertainties as the LDQ is not validated. However, it has been used by others (99, 100) in addition to being tested on a pilot group (92). As the LDQ contained questions about the participants' lives, recall error is possible. The use of self-report questioners has by others been found to underestimate smoking habits prevalence (128) and overestimate physical activity (129, 130), hence reporting bias may have been introduced in our study and affected the results.

5.1.4 Input variables

The dietary variables used in all four datasets were created from a dataset with pre-grouping of foods. This was done in the dietary information retrieval system KBS and the categorization was based on standard grouping of foods. To obtain dietary variables most relevant for our study, some regrouping was done, mainly based on nutrient content. It is possible that the grouping could affect the prediction performance of the models created in this master`s thesis. Seen in retrospect, there is a possibility that the prediction models, created with variables including the total diet, would have performed better if the total diet was grouped into more inclusive categories. For instance, it may have been unnecessary to include one variable each for "processed white meat" and "processed red meat". Rather, one common variable could have been used, as research shows the same effect on CRC regardless of the type of meat (19). Using three different variables for grain products (i.e. "Refined grain products", "Grain products" and "Pasta, rice, and grains") may not have been the most appropriate division for further interpretation. It may have been more optimal to organize all grain products into two variables, one containing refined grain products such as French baguettes, and one containing wholegrain products such as rye bread. Nevertheless, this only applies to the dataset *Overall diet* and *Overall diet plus*.

On the other hand, the datasets *CRC relevant dietary factors* and *CRC relevant dietary products plus* contain dietary variables that are composed of dietary variables from the datasets described above. For instance, "dairy products" was created from a merger of “milk and yoghurt”, “cream products”, “white cheese” and “brown cheese”. These datasets were created with the intention of including only dietary factors that have an impact on CRC. To account for potential misclassification, we used total fiber and wholegrain intake instead of dietary variables that contained food items that are the source of these dietary components. In order to take into account the alcohol percentage in various beverages, alcohol in grams per day was used instead of the consumption of wine, beer and other alcohol containing beverages.

The FFQ contains 17 questions and an open field to report intake of various dietary supplements, including multivitamin supplements. The registered use of dietary supplements was calculated into the total nutrient intake of the participants. WCRF and AICR have found the use of multivitamin supplementation to have a possibly protective effect on CRC (19). However, in our study there was no available information about which of the participants used what type of supplement since this already was calculated into the total nutrient consumption. Therefore, supplement use was not included as a specific variable in neither of the datasets.

In *Overall diet plus* and *CRC relevant dietary factors plus* we implemented 7 lifestyle and demographic variables, in addition to the result from the FIT. Of these, “age”, “sex”, “family history of CRC”, “smoking habits” and “BMI” are identified in a systematic review to be the most used risk variables in prediction models for CRC (58). Less implemented, but considered frequently, are the risk variables physical activity and education (55, 58). Wells *et al.* included both these variables in a multi risk factor prediction model for CRC, resulting in a model with high accuracy (131). Most prediction models are developed with the intention of risk stratification with easy to collect information. Hence, information about previous screening history such as the result from a FIT is not implemented in most models. Despite that, Stegeman *et al.* compared combining risk factors with FIT results against FIT results alone, resulting in a significant improvement in discrimination ($p<0.001$) (63).

5.1.5 The machine learning method

In our study only one specific ML algorithm was used to distinguish between those who have a need of colonoscopy and those who are true negatives. RF has previously proven to be a suitable method for classification problems (112) and risk prediction (132, 133). However, different types of ML techniques exist. Some other studies have applied ML in CRC screening (68, 85, 87-91), mostly by using different forms of Neural Networks (68, 87-91). Most studies use more than one approach to create the final model (68, 85, 88, 89, 91). When comparing the results, variance in predictive abilities among the different approaches is often found (68, 85, 88, 89, 91). Which methods work the best is dependent on input, output and study sample. As no other ML method, or traditional method, was tried in our study, we cannot be sure that RF is the most accurate method to predict who among the participants were in greater need of colonoscopy and who were not.

A total of 1476 participants were included in this master's thesis. Of these, 1183 were randomly allocated to the training dataset and 293 to the test dataset. Splitting the dataset is a common method in ML, however it may have reduced statistical power in this study. There is a possibility that the number of participants in the training dataset was too small to generate a good prediction model. The training dataset became even smaller due to downsampling for balancing of outcome groups, and when diagnostic groups were excluded to create the sub-cohorts. Still, models based on sub-cohort no.4, with only 708 participants included produced the most accurate predictions. Power issues also apply for the test dataset. This contained 293 participants, which may have been too small to fully validate the model.

5.2 Discussion of results

5.2.1 Prediction models

This study is unique as it compares the use of dietary variables alone against a combination of traditionally used risk variables with FIT results and several dietary variables to create a prediction model for people with the need of colonoscopy. Several models were created, but only four were tested in the test dataset. The predictive performance decreased for all four models evaluated in the test dataset. This is not common in RF, as the test dataset in theory are supposed to have a better strength for the final model compared to the training dataset. However, coincidences may have resulted in different distribution between the participants in the test and in the training dataset. Which further resulted in the final models, selected to be

evaluated in the test dataset, worked particularly better for the participants in the training dataset compared to participants in the test dataset.

Even though the use of energy adjusted dietary variables showed promising results in the training dataset, models evaluated in the test dataset did not show satisfying abilities to correctly classify the participants. This is appeared in both Model 3.4 E and Model 4.4 E with a decrease in AUROC of 0.08 and 0.06, respectively. To compare, the predictive performance of Model 3.4 and Model 4.4 also decreased in the test dataset, however both showing an AUROC of 0.59 in the test dataset.

All the models run in the test dataset were trained with the variables "red meat", "processed meat", "fruits", "vegetables", "fish", "dairy products", "calcium", "alcohol", "vitamin D", "wholegrain" and "fiber". Additionally, in Model 3.4 and Model 3.4 E, "sex", "age", "family history of CRC", "BMI", "smoking habits", "physical activity", "education" and "FIT value" were included. Moreover, models were trained within sub-cohort no.4. In this cohort the critical to find group consisted only of people diagnosed with advanced adenoma and CRC and the true negative groups consisted of participants with no adenoma or non-advanced serrated/other lesions. By using this cohort, we excluded participants with non-advanced adenoma and advanced serrated lesions. The exclusion may have led to an increase in the difference between participants critical to find and true negatives. For instance, the participants categorized as true negatives did have a higher median consumption of wholegrain and dairy products, and the participants were more likely to be female and at a normal BMI. In contract, critical to find participants had a higher median consume of processed meat, were more likely to be obese or overweight and male than the true negatives. Besides, the results from the FIT test was nearly two and a half times higher in participants with CRC compared true negatives.

In general, prediction models for CRC have been created based on information about lifestyle and demographic factors (55-59), genes (55, 62) and blood parameters (55, 58). Only a few models incorporate dietary variables beyond alcohol and red meat (55, 59), despite the evidence in the literature, showing a strong relation between diet and CRC (19). An exception is Aleksandrova *et al.* who created a lifestyle-based model for CRC prediction (134). Participant information was drawn from the EPIC. In the final model, consumption of alcohol, vegetables, dairy products, processed meat and sugar and confectionary, were used in combination with age, waist circumference, height, smoking habits and physical activity. The

model generated an AUROC of 0.71 in the derivation cohort, and 0.71 in the validation cohort, also collected from the EPIC study (134).

Most risk prediction models are developed by the use of logistic regression (55, 58). Min *et al.* investigated the performance of two Deep Neural Networks (DNN) models against the prediction performance of two simple scoring models (87). The DNN models were created with the same few input variables as used in the scores, which were age, sex, family history of CRC and smoking. BMI was implemented in one DNN model and in one score. No difference was shown between methods when tested in an external dataset (87). DNN models are able to capture the complex associations between a large numbers of input variables, however discrimination power may be minimal when only a few input variables are used (87). Yang *et al*. illustrated this point by creating a DNN model, including 26 clinical and laboratory parameters, showing a significantly better predictive performance compared to a model created with linear regression using nine of the same input variables (88). Other ML methods such as extreme gradient boosting (XGboost) and RF were additionally applied. The XGboost model performed similar to the DNN model showing an AUROC of 0.76, the RF model generated an AUROC of 0.67 (88). In similarity to our study, Cooper *et al*. used a FIT-positive cohort, participating in a British screening study to create the risk prediction algorithm both feedforward neural network (FNN) and logistics regression was used (68). Further, they only included available routine predictors, and the model was built with an index of multiple deprivation, information on previous screening history, age and gender. The predictive performance of the models yielded AUROC of 0.69 and 0.66 for FNN and logistics regression, respectively (68). Nonetheless, the model was not validated in an internal leave out test set or external cohort (68), therefore one cannot be sure that the predictions are generalizable to the population. Interestingly, an Australian cross-sectional study by Semmler *et al.* investigated the use of LR and XGboost to predict colorectal neoplasia (85). In the "mother model", 50 laboratory, clinical and diet variables were used, including the consumption of alcohol, coffee, red meat, sugar sweetened beverages, fast food, fruit, and vegetables. Both the logistic regression and XGboost model generated an AUROC of 0.66. A similar predictive performance was shown when only 10 of the included variables were used, among them the only dietary variables used was alcohol consumption. The latter model was further tested in a sub-cohort where all participants with adenoma were excluded, resulting in a population consisting of "truly healthy" participants and participants with advanced neoplasm. This sub-cohort may be comparable with our sub-cohort no.4. Analogous to our

results, the models created in this cohort perform the best, showing an AUROC of 0.70 (85). A drawback in the study by Semmler *et al*. is the lack of transparency in the importance of the variables used in the prediction algorithm, making it difficult to interpret how different variables influence the prediction of advanced neoplasia (85). The lack of studies utilizing ML to create prediction models results in few studies to compare our results with.

In the current study, few of the models created with the datasets *Overall diet plus* or *Overall diet* yielded a high prediction. An explanation may be the implementation of a large number of dietary variables, which may not be relevant. At each split the chances for non-relevant variables to be selected were greater, causing a reduced performance (135). Even though our models did not show a great prediction when the total diet was implemented, Morgenstern *et al*. suggest that the use of rich dietary data, collected in an appropriate manner, in combination with ML methods, might improve prediction of chronic diseases by discovering complex, nonlinear dietary exposures and taking advantage of small associations found between dietary variables (74).

5.2.2 Variable importance

It is worth noticing that Model 4.4, only containing dietary variables, gave an equal predictive performance in the test dataset as Model 3.4, which additionally to the dietary variables include lifestyle and demographic variables, as well as the "FIT value" in the model. The fact that using only dietary data in a model yielded an equal performance as a model created with additional risk factors commonly used, underlines the potential of including dietary variables in prediction models. This was further emphasized by the high ranking of dietary variables in the variable importance calculation of all models.

Both Model 3.4 and Model 4.4 ranked the dietary variables "processed meat", "alcohol" and "dairy products" as the top three most important dietary variables in the test dataset, and the same ranking was seen in Model 3.4 in the training dataset. Nevertheless, Model 4.4 ranked "wholegrain" as the most important variable in the training dataset. The importance of "processed meat", "alcohol", "dairy products" and "wholegrain" is consistent with the individual variables' ability to classify participants within sub-cohort no.4 (Table 13). However, due to the small differences observed between variable importance measures, their relative rankings should be interpreted with caution.

In this study, a significant difference in alcohol and processed meat consumption was found between the different diagnostic groups. Nevertheless, the median difference between groups was minimal and only accounted for a few grams. On the contrary, a difference between groups may have been sees in the RF algorithm as both alcohol and processed meat was found to be an important variable for the predictions among most variables. Both alcohol and high intake of processed meat are known to have a harmful effect on the colonocytes (19). Alcohol has been found to induce carcinogenesis through several pathways, including through the metabolism of the active ingredient, ethanol, to harmful metabolites (e.g. acetaldehyde) (136). The harmful metabolites induce production of ROS and DNA-adducts, epigenetic changes and epithelial barrier dysfunction, all of which increase the risk of cancer (136). According to WCRF and AICR, one alcoholic drink (e.g. a 0.33 L beer) per day increases the risk of CRC with 6 % (19). The same regards processed meat as it may contain carcinogenic substances such as N-nitroso compounds, heterocyclic amines and polycyclic aromatic hydrocarbons from cooking at high temperature (137). A higher intake is therefore associated with increased risk (19). Approximately half of the population in the current study followed the recommendations to limit the intake to less the 500 grams of red and processed meat a week and limit the consumption of alcohol to less than 5 E % a day (113).

Among the models, nearly all ranked variables with wholegrain and dairy products among the top ten most important variables. This was seen irrespective of input dataset and cohort. The importance of these variables are not unexpected given the inverse associations previously observed between consumption of both dairy and wholegrain products and CRC risk has shown (19). In fact, consumption of 400 gram of dairy products per day or 200 gram of milk per day, has been observed to account for a risk reductions of 13 % and 6 % for CRC. respectively (19). For models created with energy adjusted dietary variables, calcium was also ranked as highly important for the model performance. One of the preventive effects of calcium is its ability to bind fatty acids as well as unconjugated bile acids, reducing their carcinogenic effect on the epithelial cells (138). Studies have also suggested that calcium influences cell signaling pathways, enhancing cell differentiation and reduces cell proliferation (139). Much of the effect of dairy products on CRC risk can be assigned to the influence of calcium, however butyrate, lactic acid-producing bacteria and lactoferrin may also have protective effects (140). Regarding wholegrain, a consumption of 90 grams per day has been found to reduce CRC risk with 17 % (19). The protective properties of wholegrain are thought to be caused by the content of fiber (i.e. shortening transit time, enhancing

production of short chain fatty acids by microbiota and preventing insulin resistance) and nutrients with anti-carcinogenic effect (141).

Only the non-dietary variables "age", "FIT value" and "BMI" were ranked among the ten most important variables for the performance in models. However, the "FIT value" was ranked at the top for most models where the variables was included in the dataset. A "gap" in the importance value from "FIT value" when ranked at the top to the next variable was seen among most models. Inclusive of Model 3.4 in both the training and the test dataset where "FIT value" was assigned an importance value of 15.50 and 15.11, respectively. The respective next variable which was "dairy products" had a mean importance value of 11.35 in the training dataset and 12.26 in the test dataset. The "FIT values" importance in the models are in alignment with the variable's ability alone to classify participants, showing an AUROC of 0.62.

However, excluding the "FIT value" from the prediction model only led to a small reduction in the AUROC value for Model 3.4 (from 0.64 to 0.61). The higher predictive performance when the result from the FIT was included in the models are in line with previous studies (63, 71, 142). Further, Model 3.4 (no FIT value) ranked "dairy products", "alcohol", "wholegrain" and "processed meat" as the most important variables for the prediction. Although the models without the results from the FIT were not evaluated in the test dataset, the ranking may indicate the importance of dietary variables being incorporated in prediction models of advanced neoplasia.

Interestingly, Model 3.2 (no FIT value) showed an equal AUROC and sensitivity as Model 3.2. This suggests that the models created with *CRC relevant dietary factors plus* within sub-cohort no.2 (i.e. critical to detect: advanced serrated lesions, non-advanced adenoma (>3) in addition to advanced adenoma and CRC; true negatives: no adenoma, non-advanced serrated lesions/other lesions) were not dependent on the "FIT value" to obtain the AUROC. Model 3.2 (no FIT value) ranked "age" as the most important variable, followed by dietary variables.

The lifestyle and demographic variables found to be most important for the models were "BMI" and "age". Worth noticing is that none of the other lifestyle and demographic variables were ranked among the ten most important variables, which may be an artefact introduced by how the importance of the variables was measured. Variable importance when measured

using change in impurity is known to be biased against categorical variables (143, 144), so while our results indicated limited importance, using a permutation approach could have provided a more accurate measure of importance for these lifestyle and demographic variables in our prediction models. Thus, the results cannot be interpreted as lifestyle and demographic variables not having an impact on CRC. Further, the variable importance is in line with the predictive performance (i.e. AUROC shown in Table 13) for each variable alone, except for "sex", in sub-cohort no.4. It is possible that the population is too similar regarding lifestyle and demographic factors (Table 2), which may be a consequence of a too small population where the model is trained. Furthermore, the LDQ used was not validated and may not have captured the correct answers.

Variable importance must be interpreted with caution as it does not state anything more than that the variable itself is important for prediction performance. Variable importance provides the mean decrease in impurity, meaning how well the variable can divide the population into two as "pure" (i.e. case or control) daughter groups as possible. However, the variable importance does not tell anything about where the split was performed (e.g. at what level of intake in grams/day).

5.2.3 Strengths and limitation

A major strength of this master's thesis is that it had its origin in a well-organized, large, prospective cohort study (92). Another strength is that participants are collected from two different centers located in two different municipalities. It has resulted in access to a population with only FIT-positive participants and provided high quality data, and a greater sample-size compared to other nutrition studies. For instance, it provided clinically verified outcome information, lifestyle and demographic data, and detailed dietary information collected through a validated FFQ. As the use of any instrument to evaluate dietary habits, FFQs are prone to errors. To mitigate errors exclusion of participants providing low quality dietary data or unrealistic energy intake was done prior to the analysis. A further strength is the strict definition of the outcome variables, both those who were considered as critical to find and as true negatives, we believe that this makes it easier to interpret which of the variables are important for correct classification of participants. Further, we practice transparency of the RF models by reporting variable importance. Even though they do not

provide a direction of the variable (i.e. more or less associated with critical to find group) they do say something about the importance of implementing the variable in the model.

A limitation to the study is the cross-sectional design, which prevents causal interpretation of the relationship between selected input variables and outcome variables. Another limitation to the study is the representativeness, as people participating in this screening trial have agreed to conduct FIT repeatedly in addition to answer questionnaires. A third limitation of this study is that we only used one approach to create the prediction models, there was no other ML or conventional method used in addition to RF. The implementation of another method could have made it possible to compare results, and even provide a better discrimination.

It may be a limitation that the final models were selected manually based on best predictive performers (i.e. AUROC) instead of predetermined theories. As the final models may have been better fitted to training dataset than to the test dataset. Further, the models were not validated in an external cohort. Since no other study has used both Random Forest and a bouquet of dietary variables to predict CRC, the possibilities of future studies are many.

Our findings highlight the importance of implementing dietary variables known to have an impact on CRC development when creating a prediction model for CRC screening. Furthermore, our study shows the potential of using ML methods, instead of traditional methods, to detect who are in greater need of colonoscopy and who are not.

5.3 The usability of prediction models and further studies

Our models did not yield particularly accurate predictions, hence the implementation into a screening program for CRC probably would not lead to a great improvement. Firstly, the best performing models in the training dataset performed worse in the test dataset. This is not usually the case for RF models as they are known to have minimal overfitting (135). Secondly, the dietary variables used in the model are not easily measured. In our study, they are collected by a comprehensive FFQ, which may be a burden for the participants, and it is more demanding for the clinician as the method requires further processing of the data

Further studies should continue to investigate the inclusion of dietary variables in prediction models as they were shown to be more important for model accuracy. For instance, an idea

could be to use dichotomous dietary variables (e.g. above or below five portions of fruits and vegetables per day). Furthermore, researchers should investigate the use of dietary information collected from a simpler FFQ which can be filled out digitally, with targeted questions, where the answers can be used directly into an algorithm. With these methods it might be easier to interpret the variable importance, it may also be easier to collect and use dietary information in a screening setting. Moreover, studies with larger sample sizes, where several different ML techniques are tried out at the same time should be conducted to review whether this can provide a better prediction of who are in greater need of colonoscopy and who are not.

6 Conclusion

In this cross-sectional study among FIT positive participants, we were unable to create a prediction model with satisfactory ability to detect participants in need of colonoscopy, using detailed dietary, lifestyle and demographic data in combination with FIT data. However, our results showed that the models using dietary variables known to influence CRC risk were better at identifying people in need of colonoscopy than the models incorporating more general dietary variables. Except for a few variables (FIT value, age and BMI), the dietary variables known to influence CRC were in general more important for the predictive performance than the other variables. This was especially the case for the variables “dairy products”, “alcohol”, “wholegrain”, and “processed meat”. Implementation of the FIT result in the model gave, not surprisingly, a more accurate prediction compared to models where this variable was omitted.

In terms of performance, our best models (Model 3.4 and Model 4.4) generated a higher AUROC in the training dataset than in the test dataset. Even though Model 3.4 differed from Model 4.4, by including the FIT result, as well as lifestyle, and demographic variables, the models showed an equal AUROC in the test dataset; thus showing the importance of implementing dietary variables, even if predictive performance was not optimal.

References

1. World Health Organization. Colorectal cancer 2020 [Available from: https://gco.iarc.fr/today/data/factsheets/cancers/10_8_9-Colorectum-fact-sheet.pdf
2. Araghi M, Soerjomataram I, Bardot A, Ferlay J, Cabasag CJ, Morrison DS, et al. Changes in colorectal cancer incidence in seven high-income countries: a population-based study. Lancet Gastroenterol Hepatol. 2019;4(7):511-8.
3. Cancer Registry of Norway. Cancer in Norway 2020 - Cancer incidence, mortality, survival and prevalence in Norway. Oslo: Cancer Registry of Norway; 2021.
4. Lao VV, Grady WM. Epigenetics and colorectal cancer. Nat Rev Gastroenterol Hepatol. 2011;8(12):686-700.
5. Winawer SJ, Zauber AG. The advanced adenoma as the primary target of screening. Gastrointest Endosc Clin N Am. 2002;12(1):1-9, v.
6. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546-58.
7. Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. Nat Rev Gastroenterol Hepatol. 2019;16(12):713-32.
8. Carethers JM, Jung BH. Genetics and Genetic Biomarkers in Sporadic Colorectal Cancer. Gastroenterology. 2015;149(5):1177-90.e3.
9. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.
10. Dekker E, Tanis PJ, Vleugels JLA, Kasi PM, Wallace MB. Colorectal cancer. Lancet. 2019;394(10207):1467-80.
11. Lucafò M, Curci D, Franzin M, Decorti G, Stocco G. Inflammatory Bowel Disease and Risk of Colorectal Cancer: An Overview From Pathophysiology to Pharmacological Prevention. Front Pharmacol. 2021;12:772101.
12. Jess T, Rungoe C, Peyrin-Biroulet L. Risk of colorectal cancer in patients with ulcerative colitis: a meta-analysis of population-based cohort studies. Clin Gastroenterol Hepatol. 2012;10(6):639-45.
13. Murphy N, Ward HA, Jenab M, Rothwell JA, Boutron-Ruault MC, Carbonnel F, et al. Heterogeneity of Colorectal Cancer Risk Factors by Anatomical Subsite in 10 European Countries: A Multinational Cohort Study. Clin Gastroenterol Hepatol. 2019;17(7):1323-31.e6.
14. Yang J, Du XL, Li ST, Wang BY, Wu YY, Chen ZL, et al. Characteristics of Differently Located Colorectal Cancers Support Proximal and Distal Classification: A Population-Based Study of 57,847 Patients. PLoS One. 2016;11(12):e0167540.
15. Glebov OK, Rodriguez LM, Nakahara K, Jenkins J, Cliatt J, Humbyrd CJ, et al. Distinguishing right from left colon by the pattern of gene expression. Cancer Epidemiol Biomarkers Prev. 2003;12(8):755-62.
16. Wei EK, Giovannucci E, Wu K, Rosner B, Fuchs CS, Willett WC, et al. Comparison of risk factors for colon and rectal cancer. Int J Cancer. 2004;108(3):433-42.
17. Baran B, Mert Ozupek N, Yerli Tetik N, Acar E, Bekcioglu O, Baskin Y. Difference Between Left-Sided and Right-Sided Colorectal Cancer: A Focused Review of Literature. Gastroenterology Res. 2018;11(4):264-73.
18. Hossain MS, Karuniawati H, Jairoun AA, Urbi Z, Ooi J, John A, et al. Colorectal Cancer: A Review of Carcinogenesis, Global Epidemiology, Current Challenges, Risk Factors, Preventive and Treatment Strategies. Cancers (Basel). 2022;14(7).
19. World Cancer Research Fund /American Institute for Cancer Research. Continuous Update Project Expert Report 2018. Diet, nutrition, physical activity and colorectal cancer 2018 [Available from: http://dietandcancerreport.org/.

20. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-49.
21. Johnson CM, Wei C, Ensor JE, Smolenski DJ, Amos CI, Levin B, et al. Meta-analyses of colorectal cancer risk factors. Cancer Causes Control. 2013;24(6):1207-22.
22. Krämer HU, Schöttker B, Raum E, Brenner H. Type 2 diabetes mellitus and colorectal cancer: meta-analysis on sex-specific differences. Eur J Cancer. 2012;48(9):1269-82.
23. Cottet V, Jooste V, Fournel I, Bouvier AM, Faivre J, Bonithon-Kopp C. Long-term risk of colorectal cancer after adenoma removal: a population-based cohort study. Gut. 2012;61(8):1180-6.
24. Mihor A, Tomsic S, Zagar T, Lokar K, Zadnik V. Socioeconomic inequalities in cancer incidence in Europe: a comprehensive review of population-based epidemiological studies. Radiol Oncol. 2020;54(1):1-13.
25. Manser CN, Bauerfeind P. Impact of socioeconomic status on incidence, mortality, and survival of colorectal cancer patients: a systematic review. Gastrointest Endosc. 2014;80(1):42-60.e9.
26. Aarts MJ, Lemmens VE, Louwman MW, Kunst AE, Coebergh JW. Socioeconomic status and changing inequalities in colorectal cancer? A review of the associations with risk, treatment and outcome. Eur J Cancer. 2010;46(15):2681-95.
27. Lynge E, Martinsen JI, Larsen IK, Kjærheim K. Colon cancer trends in Norway and Denmark by socio-economic group: A cohort study. Scand J Public Health. 2015;43(8):890-8.
28. Rezende LFM, Sá TH, Markozannes G, Rey-López JP, Lee IM, Tsilidis KK, et al. Physical activity and cancer: an umbrella review of the literature including 22 major anatomical sites and 770 000 cancer cases. Br J Sports Med. 2018;52(13):826-33.
29. Ruiz-Casado A, Martín-Ruiz A, Pérez LM, Provencio M, Fiuza-Luces C, Lucia A. Exercise and the Hallmarks of Cancer. Trends Cancer. 2017;3(6):423-41.
30. Bardou M, Barkun AN, Martel M. Obesity and colorectal cancer. Gut. 2013;62(6):933-47.
31. Ho GY, Wang T, Gunter MJ, Strickler HD, Cushman M, Kaplan RC, et al. Adipokines linking obesity with colorectal cancer risk in postmenopausal women. Cancer Res. 2012;72(12):3029-37.
32. Zhou B, Shu B, Yang J, Liu J, Xi T, Xing Y. C-reactive protein, interleukin-6 and the risk of colorectal cancer: a meta-analysis. Cancer Causes Control. 2014;25(10):1397-405.
33. Murphy N, Cross AJ, Abubakar M, Jenab M, Aleksandrova K, Boutron-Ruault MC, et al. A Nested Case-Control Study of Metabolically Defined Body Size Phenotypes and Risk of Colorectal Cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC). PLoS Med. 2016;13(4):e1001988.
34. Ma J, Giovannucci E, Pollak M, Leavitt A, Tao Y, Gaziano JM, et al. A prospective study of plasma C-peptide and colorectal cancer risk in men. J Natl Cancer Inst. 2004;96(7):546-53.
35. Jenab M, Riboli E, Cleveland RJ, Norat T, Rinaldi S, Nieters A, et al. Serum C-peptide, IGFBP-1 and IGFBP-2 and risk of colon and rectal cancers in the European Prospective Investigation into Cancer and Nutrition. Int J Cancer. 2007;121(2):368-76.
36. Botteri E, Borroni E, Sloan EK, Bagnardi V, Bosetti C, Peveri G, et al. Smoking and Colorectal Cancer Risk, Overall and by Molecular Subtypes: A Meta-Analysis. Am J Gastroenterol. 2020;115(12):1940-9.
37. He X, Wu K, Ogino S, Giovannucci EL, Chan AT, Song M. Association Between Risk Factors for Colorectal Cancer and Risk of Serrated Polyps and Conventional Adenomas. Gastroenterology. 2018;155(2):355-73.e18.

38. Atkin W, Wooldrage K, Parkin DM, Kralj-Hans I, MacRae E, Shah U, et al. Long term effects of once-only flexible sigmoidoscopy screening after 17 years of follow-up: the UK Flexible Sigmoidoscopy Screening randomised controlled trial. Lancet. 2017;389(10076):1299-311.
39. Schoen RE, Pinsky PF, Weissfeld JL, Yokochi LA, Church T, Laiyemo AO, et al. Colorectal-cancer incidence and mortality with screening flexible sigmoidoscopy. N Engl J Med. 2012;366(25):2345-57.
40. Segnan N, Armaroli P, Bonelli L, Risio M, Sciallero S, Zappa M, et al. Once-only sigmoidoscopy in colorectal cancer screening: follow-up findings of the Italian Randomized Controlled Trial--SCORE. J Natl Cancer Inst. 2011;103(17):1310-22.
41. Holme Ø, Løberg M, Kalager M, Bretthauer M, Hernán MA, Aas E, et al. Effect of flexible sigmoidoscopy screening on colorectal cancer incidence and mortality: a randomized clinical trial. Jama. 2014;312(6):606-15.
42. Mandel JS, Bond JH, Church TR, Snover DC, Bradley GM, Schuman LM, et al. Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota Colon Cancer Control Study. N Engl J Med. 1993;328(19):1365-71.
43. Kronborg O, Fenger C, Olsen J, Jørgensen OD, Søndergaard O. Randomised study of screening for colorectal cancer with faecal-occult-blood test. Lancet. 1996;348(9040):1467-71.
44. Lindholm E, Brevinge H, Haglind E. Survival benefit in a randomized clinical trial of faecal occult blood screening for colorectal cancer. Br J Surg. 2008;95(8):1029-36.
45. Hardcastle JD, Chamberlain JO, Robinson MH, Moss SM, Amar SS, Balfour TW, et al. Randomised controlled trial of faecal-occult-blood screening for colorectal cancer. Lancet. 1996;348(9040):1472-7.
46. Ladabaum U, Dominitz JA, Kahi C, Schoen RE. Strategies for Colorectal Cancer Screening. Gastroenterology. 2020;158(2):418-32.
47. Randel KR, Schult AL, Botteri E, Hoff G, Bretthauer M, Ursin G, et al. Colorectal Cancer Screening With Repeated Fecal Immunochemical Test Versus Sigmoidoscopy: Baseline Results From a Randomized Trial. Gastroenterology. 2021;160(4):1085-96.e5.
48. Schreuders EH, Ruco A, Rabeneck L, Schoen RE, Sung JJ, Young GP, et al. Colorectal cancer screening: a global overview of existing programmes. Gut. 2015;64(10):1637-49.
49. Haug U, Kuntz KM, Knudsen AB, Hundt S, Brenner H. Sensitivity of immunochemical faecal occult blood testing for detecting left- vs right-sided colorectal neoplasia. Br J Cancer. 2011;104(11):1779-85.
50. Brenner H, Qian J, Werner S. Variation of diagnostic performance of fecal immunochemical testing for hemoglobin by sex and age: results from a large screening cohort. Clin Epidemiol. 2018;10:381-9.
51. Holme Ø, Løberg M, Kalager M, Bretthauer M, Hernán MA, Aas E, et al. Long-Term Effectiveness of Sigmoidoscopy Screening on Colorectal Cancer Incidence and Mortality in Women and Men: A Randomized Trial. Ann Intern Med. 2018;168(11):775-82.
52. Lin JS, Perdue LA, Henrikson NB, Bean SI, Blasi PR. Screening for Colorectal Cancer: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. Jama. 2021;325(19):1978-98.
53. Lin JS, Piper MA, Perdue LA, Rutter C, Webber EM, O'Connor E, et al. U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews. Screening for Colorectal Cancer: A Systematic Review for the US Preventive Services Task Force. Rockville (MD): Agency for Healthcare Research and Quality (US); 2016.

54. van Turenhout ST, Terhaar sive Droste JS, Meijer GA, Masclée AA, Mulder CJ. Anticipating implementation of colorectal cancer screening in The Netherlands: a nation wide survey on endoscopic supply and demand. BMC Cancer. 2012;12:46.
55. Usher-Smith JA, Walter FM, Emery JD, Win AK, Griffin SJ. Risk Prediction Models for Colorectal Cancer: A Systematic Review. Cancer Prev Res (Phila). 2016;9(1):13-26.
56. Ma GK, Ladabaum U. Personalizing colorectal cancer screening: a systematic review of models to predict risk of colorectal neoplasia. Clin Gastroenterol Hepatol. 2014;12(10):1624-34.e1.
57. Smith T, Muller DC, Moons KGM, Cross AJ, Johansson M, Ferrari P, et al. Comparison of prognostic models to predict the occurrence of colorectal cancer in asymptomatic individuals: a systematic literature review and external validation in the EPIC and UK Biobank prospective cohort studies. Gut. 2019;68(4):672-83.
58. Peng L, Weigl K, Boakye D, Brenner H. Risk Scores for Predicting Advanced Colorectal Neoplasia in the Average-risk Population: A Systematic Review and Meta-analysis. Am J Gastroenterol. 2018;113(12):1788-800.
59. Peng L, Balavarca Y, Weigl K, Hoffmeister M, Brenner H. Head-to-Head Comparison of the Performance of 17 Risk Models for Predicting Presence of Advanced Neoplasms in Colorectal Cancer Screening. Am J Gastroenterol. 2019;114(9):1520-30.
60. Taylor DP, Stoddard GJ, Burt RW, Williams MS, Mitchell JA, Haug PJ, et al. How well does family history predict who will get colorectal cancer? Implications for cancer screening and counseling. Genet Med. 2011;13(5):385-91.
61. Colditz GA, Atwood KA, Emmons K, Monson RR, Willett WC, Trichopoulos D, et al. Harvard report on cancer prevention volume 4: Harvard Cancer Risk Index. Risk Index Working Group, Harvard Center for Cancer Prevention. Cancer Causes Control. 2000;11(6):477-88.
62. McGeoch L, Saunders CL, Griffin SJ, Emery JD, Walter FM, Thompson DJ, et al. Risk Prediction Models for Colorectal Cancer Incorporating Common Genetic Variants: A Systematic Review. Cancer Epidemiol Biomarkers Prev. 2019;28(10):1580-93.
63. Stegeman I, de Wijkerslooth TR, Stoop EM, van Leerdam ME, Dekker E, van Ballegooijen M, et al. Combining risk factors with faecal immunochemical test outcome for selecting CRC screenees for colonoscopy. Gut. 2014;63(3):466-71.
64. Omata F, Shintani A, Isozaki M, Masuda K, Fujita Y, Fukui T. Diagnostic performance of quantitative fecal immunochemical test and multivariate prediction model for colorectal neoplasms in asymptomatic individuals. Eur J Gastroenterol Hepatol. 2011;23(11):1036-41.
65. Tao S, Haug U, Kuhn K, Brenner H. Comparison and combination of blood-based inflammatory markers with faecal occult blood tests for non-invasive colorectal cancer screening. Br J Cancer. 2012;106(8):1424-30.
66. Auge JM, Pellise M, Escudero JM, Hernandez C, Andreu M, Grau J, et al. Risk stratification for advanced colorectal neoplasia according to fecal hemoglobin concentration in a colorectal cancer screening program. Gastroenterology. 2014;147(3):628-36.e1.
67. Yen AM, Chen SL, Chiu SY, Fann JC, Wang PE, Lin SC, et al. A new insight into fecal hemoglobin concentration-dependent predictor for colorectal neoplasia. Int J Cancer. 2014;135(5):1203-12.
68. Cooper JA, Parsons N, Stinton C, Mathews C, Smith S, Halloran SP, et al. Risk-adjusted colorectal cancer screening using the FIT and routine screening data: development of a risk prediction model. Br J Cancer. 2018;118(2):285-93.
69. Thomsen MK, Pedersen L, Erichsen R, Lash TL, Sørensen HT, Mikkelsen EM. Risk-stratified selection to colonoscopy in FIT colorectal cancer screening: development and temporal validation of a prediction model. Br J Cancer. 2022.

70. Sekiguchi M, Kakugawa Y, Ikematsu H, Hotta K, Konda K, Tanaka Y, et al. Risk Stratification Score Improves Sensitivity for Advanced Colorectal Neoplasia in Colorectal Cancer Screening: The Oshima Study Workgroup. Clin Transl Gastroenterol. 2021;12(3):e00319.
71. Li W, Zhao LZ, Ma DW, Wang DZ, Shi L, Wang HL, et al. Predicting the risk for colorectal cancer with personal characteristics and fecal immunochemical test. Medicine (Baltimore). 2018;97(18):e0529.
72. Sirsat MS, Fermé E, Câmara J. Machine Learning for Brain Stroke: A Review. J Stroke Cerebrovasc Dis. 2020;29(10):105162.
73. Deo RC. Machine Learning in Medicine. Circulation. 2015;132(20):1920-30.
74. Morgenstern JD, Rosella LC, Costa AP, de Souza RJ, Anderson LN. Perspective: Big Data and Machine Learning Could Help Advance Nutritional Epidemiology. Adv Nutr. 2021;12(3):621-31.
75. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A Primer for the Epidemiologist. Am J Epidemiol. 2019;188(12):2222-39.
76. Rigdon J, Basu S. Machine learning with sparse nutrition data to improve cardiovascular mortality risk prediction in the USA using nationally randomly sampled data. BMJ Open. 2019;9(11):e032703.
77. Valavanis IK, Mougiakakou SG, Grimaldi KA, Nikita KS. A multifactorial analysis of obesity as CVD risk factor: use of neural network based methods in a nutrigenetics context. BMC Bioinformatics. 2010;11:453.
78. Panaretos D, Koloverou E, Dimopoulos AC, Kouli GM, Vamvakari M, Tzavelas G, et al. A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002-2012): the ATTICA study. Br J Nutr. 2018;120(3):326-34.
79. Shang X, Li Y, Xu H, Zhang Q, Liu A, Du S, et al. Leading dietary determinants identified using machine learning techniques and a healthy diet score for changes in cardiometabolic risk factors in children: a longitudinal analysis. Nutr J. 2020;19(1):105.
80. Biesbroek S, van der AD, Brosens MC, Beulens JW, Verschuren WM, van der Schouw YT, et al. Identifying cardiovascular risk factor-related dietary patterns with reduced rank regression and random forest in the EPIC-NL cohort. Am J Clin Nutr. 2015;102(1):146-54.
81. Bodnar LM, Cartus AR, Kirkpatrick SI, Himes KP, Kennedy EH, Simhan HN, et al. Machine learning as a strategy to account for dietary synergy: an illustration based on dietary intake and adverse pregnancy outcomes. Am J Clin Nutr. 2020;111(6):1235-43.
82. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized Nutrition by Prediction of Glycemic Responses. Cell. 2015;163(5):1079-94.
83. Kim H, Lim DH, Kim Y. Classification and Prediction on the Effects of Nutritional Intake on Overweight/Obesity, Dyslipidemia, Hypertension and Type 2 Diabetes Mellitus Using Deep Learning Model: 4-7th Korea National Health and Nutrition Examination Survey. Int J Environ Res Public Health. 2021;18(11).
84. Shiao SPK, Grayson J, Lie A, Yu CH. Personalized Nutrition-Genes, Diet, and Related Interactive Parameters as Predictors of Cancer in Multiethnic Colorectal Cancer Families. Nutrients. 2018;10(6).
85. Semmler G, Wernly S, Wernly B, Mamandipoor B, Bachmayer S, Semmler L, et al. Machine Learning Models Cannot Replace Screening Colonoscopy for the Prediction of Advanced Colorectal Adenoma. J Pers Med. 2021;11(10).
86. van Zutphen M, van Duijnhoven FJB, Wesselink E, Schrauwen RWM, Kouwenhoven EA, van Halteren HK, et al. Identification of Lifestyle Behaviors Associated with Recurrence

and Survival in Colorectal Cancer Patients Using Random Survival Forests. Cancers (Basel). 2021;13(10).
87. Min JK, Yang HJ, Kwak MS, Cho CW, Kim S, Ahn KS, et al. Deep Neural Network-Based Prediction of the Risk of Advanced Colorectal Neoplasia. Gut Liver. 2021;15(1):85-91.
88. Yang HJ, Cho CW, Jang J, Kim SS, Ahn KS, Park SK, et al. Application of deep learning to predict advanced neoplasia using big clinical data in colorectal cancer screening of asymptomatic adults. Korean J Intern Med. 2021;36(4):845-56.
89. Liu F, Long Q, He H, Dong S, Zhao L, Zou C, et al. Combining the Fecal Immunochemical Test with a Logistic Regression Model for Screening Colorectal Neoplasia. Front Pharmacol. 2021;12:635481.
90. Nartowt BJ, Hart GR, Roffman DA, Llor X, Ali I, Muhammad W, et al. Scoring colorectal cancer risk with an artificial neural network based on self-reportable personal health data. PLoS One. 2019;14(8):e0221421.
91. Nartowt BJ, Hart GR, Muhammad W, Liang Y, Stark GF, Deng J. Robust Machine Learning for Colorectal Cancer Risk Prediction and Stratification. Front Big Data. 2020;3:6.
92. Kværner AS, Birkeland E, Bucher-Johannessen C, Vinberg E, Nordby JI, Kangas H, et al. The CRCbiome study: a large prospective cohort study examining the role of lifestyle and the gut microbiome in colorectal cancer screening participants. BMC Cancer. 2021;21(1):930.
93. Andersen LF, Solvoll K, Johansson LRK, Salminen I, Aro A, Drevon CA. Evaluation of a Food Frequency Questionnaire with Weighed Records, Fatty Acids, and Alpha-Tocopherol in Adipose Tissue and Serum. American Journal of Epidemiology. 1999;150(1):75-87.
94. Andersen LF, Tomten H, Haggarty P, Løvø A, Hustvedt BE. Validation of energy intake estimated from a food frequency questionnaire: a doubly labelled water study. Eur J Clin Nutr. 2003;57(2):279-84.
95. Carlsen MH, Lillegaard IT, Karlsen A, Blomhoff R, Drevon CA, Andersen LF. Evaluation of energy and dietary intake estimates from a food frequency questionnaire using independent energy expenditure measurement and weighed food records. Nutr J. 2010;9:37.
96. Andersen LF, Veierød MB, Johansson L, Sakhi A, Solvoll K, Drevon CA. Evaluation of three dietary assessment methods and serum biomarkers as measures of fruit and vegetable intake, using the method of triads. Br J Nutr. 2005;93(4):519-27.
97. Carlsen MH, Karlsen A, Lillegaard IT, Gran JM, Drevon CA, Blomhoff R, et al. Relative validity of fruit and vegetable intake estimated from an FFQ, using carotenoid and flavonoid biomarkers and the method of triads. Br J Nutr. 2011;105(10):1530-8.
98. Brunvoll SH, Thune I, Frydenberg H, Flote VG, Bertheussen GF, Schlichting E, et al. Validation of repeated self-reported n-3 PUFA intake using serum phospholipid fatty acids as a biomarker in breast cancer patients during treatment. Nutr J. 2018;17(1):94.
99. Markussen MS, Veierød MB, Kristiansen AL, Ursin G, Andersen LF. Dietary patterns of women aged 50-69 years and associations with nutrient intake, sociodemographic factors and key risk factors for non-communicable diseases. Public Health Nutr. 2016;19(11):2024-32.
100. Knudsen MD, Berstad P, Hjartåker A, Gulichsen EH, Hoff G, de Lange T, et al. Lifestyle predictors for non-participation and outcome in the second round of faecal immunochemical test in colorectal cancer screening. Br J Cancer. 2017;117(4):461-9.
101. Helsedirektoratet. Fysisk aktivitet for barn, unge, voksne, eldre og gravide Norway 2019 [Available from: https://www.helsedirektoratet.no/faglige-rad/fysisk-aktivitet-for-barn-unge-voksne-eldre-og-gravide.
102. Hassan C, Antonelli G, Dumonceau JM, Regula J, Bretthauer M, Chaussade S, et al. Post-polypectomy colonoscopy surveillance: European Society of Gastrointestinal Endoscopy (ESGE) Guideline - Update 2020. Endoscopy. 2020;52(8):687-700.

103. Haque TR, Bradshaw PT, Crockett SD. Risk factors for serrated polyps of the colorectum. Dig Dis Sci. 2014;59(12):2874-89.
104. Breiman L. Random Forests. Machine learning. 2001;45(1):5-32.
105. Marvin N. Wright AZ. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software. 2017;77(1):1-17.
106. Kuhn M, Wickham H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. 2020.
107. Kuhn M, Vaughan D. Parsnip: A Commom API to Modeling and Analysis Functions. 2021.
108. Kuhn M, Wickham H. recipes: Preprocessing and Feature Engineering Steps for Modeling. R 2021.
109. Kuhn M. tune: Tidy Tuning Tools. 2021.
110. Kuhn M, Vaughan D. Yardstick: Tidy Characterizations of Model Performance. 2021.
111. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol. 2010;5(9):1315-6.
112. Topçuoğlu BD, Lesniak NA, Ruffin MTt, Wiens J, Schloss PD. A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems. mBio. 2020;11(3).
113. Kostråd for å fremme folkehelsen og forebygge kroniske sykdommer : metodologi og vitenskapelig kunnskapsbidrag. Oslo: Helsedirektoratet; 2011.
114. Nembrini S, König IR, Wright MN. The revival of the Gini importance? Bioinformatics. 2018;34(21):3711-8.
115. Willett W. Nutritional Epidemiology. 3rd ed. ed. Oxford: Oxford: Oxford University Press; 2012.
116. Waldmann E, Heinze G, Ferlitsch A, Gess II, Sallinger D, Jeschek P, et al. Risk factors cannot explain the higher prevalence rates of precancerous colorectal lesions in men. Br J Cancer. 2016;115(11):1421-9.
117. Nordic nutrition recommendations 2012 : integrating nutrition and physical activity. Copenhagen: Nordic Council of Ministers; 2014.
118. Nordic nutrition recommendations 2012 : part 1 : summary, principles and use. Copenhagen: Nordic council of ministers; 2013.
119. Blanks RG, Benson VS, Alison R, Brown A, Reeves GK, Beral V, et al. Nationwide bowel cancer screening programme in England: cohort study of lifestyle factors affecting participation and outcomes in women. Br J Cancer. 2015;112(9):1562-7.
120. Jäntti M, Heinävaara S, Malila N, Sarkeala T. Sociodemographic features and patterns of non-participation in colorectal cancer screening in Finland. Eur J Public Health. 2021;31(4):890-4.
121. He E, Lew JB, Egger S, Banks E, Ward RL, Beral V, et al. Factors associated with participation in colorectal cancer screening in Australia: Results from the 45 and Up Study cohort. Prev Med. 2018;106:185-93.
122. Deding U, Henig AS, Salling A, Torp-Pedersen C, Boggild H. Sociodemographic predictors of participation in colorectal cancer screening. Int J Colorectal Dis. 2017;32(8):1117-24.
123. Knudsen MD, Kvaerner AS, Botteri E, Holme Ø, Hjartåker A, Song M, et al. Lifestyle predictors for inconsistent participation to fecal based colorectal cancer screening. BMC Cancer. 2022;22(1):172.
124. Deding U, Henig AS, Salling A, Torp-Pedersen C, Bøggild H. Sociodemographic predictors of participation in colorectal cancer screening. Int J Colorectal Dis. 2017;32(8):1117-24.

125. Botteri E, Hoff G, Randel KR, Holme Ø, de Lange T, Bernklev T, et al. Characteristics of non-participants in a randomized colorectal cancer screening trial comparing sigmoidoscopy and faecal immunochemical testing. Int J Cancer. 2022.
126. Brunvoll SH, Thune I, Bertheussen GF, Fjeldheim F, Flote VG, Frydenberg H, et al. Dietary changes in early-stage breast cancer patients from pre-surgery and over the 12 months post-surgery. Br J Nutr. 2021;125(2):172-82.
127. Kirkøen B, Berstad P, Botteri E, Åvitsland TL, Ossum AM, de Lange T, et al. Do no harm: no psychological harm from colorectal cancer screening. Br J Cancer. 2016;114(5):497-504.
128. Connor Gorber S, Schofield-Hurwitz S, Hardt J, Levasseur G, Tremblay M. The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. Nicotine Tob Res. 2009;11(1):12-24.
129. Henriksen HB, Berntsen S, Paur I, Zucknick M, Skjetne AJ, Bøhn SK, et al. Validation of two short questionnaires assessing physical activity in colorectal cancer patients. BMC Sports Sci Med Rehabil. 2018;10:8.
130. Sallis JF, Saelens BE. Assessment of physical activity by self-report: status, limitations, and future directions. Res Q Exerc Sport. 2000;71(2 Suppl):S1-14.
131. Wells BJ, Kattan MW, Cooper GS, Jackson L, Koroukian S. Colorectal cancer predicted risk online (CRC-PRO) calculator using data from the multi-ethnic cohort study. J Am Board Fam Med. 2014;27(1):42-55.
132. Ming C, Viassolo V, Probst-Hensch N, Chappuis PO, Dinov ID, Katapodi MC. Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. Breast Cancer Res. 2019;21(1):75.
133. Yang L, Wu H, Jin X, Zheng P, Hu S, Xu X, et al. Study of cardiovascular disease prediction model based on random forest in eastern China. Sci Rep. 2020;10(1):5245.
134. Aleksandrova K, Reichmann R, Kaaks R, Jenab M, Bueno-de-Mesquita HB, Dahm CC, et al. Development and validation of a lifestyle-based model for colorectal cancer risk prediction: the LiFeCRC score. BMC Med. 2021;19(1):1.
135. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition. New York, NY: Springer New York : Imprint: Springer; 2009.
136. Rossi M, Jahanzaib Anwar M, Usman A, Keshavarzian A, Bishehsari F. Colorectal Cancer and Alcohol Consumption-Populations to Molecules. Cancers (Basel). 2018;10(2).
137. Santarelli RL, Pierre F, Corpet DE. Processed meat and colorectal cancer: a review of epidemiologic and experimental evidence. Nutr Cancer. 2008;60(2):131-44.
138. Newmark HL, Wargovich MJ, Bruce WR. Colon cancer and dietary fat, phosphate, and calcium: a hypothesis. J Natl Cancer Inst. 1984;72(6):1323-5.
139. Fedirko V, Bostick RM, Flanders WD, Long Q, Sidelnikov E, Shaukat A, et al. Effects of vitamin d and calcium on proliferation and differentiation in normal colon mucosa: a randomized clinical trial. Cancer Epidemiol Biomarkers Prev. 2009;18(11):2933-41.
140. Norat T, Riboli E. Dairy products and colorectal cancer. A review of possible mechanisms and epidemiological evidence. Eur J Clin Nutr. 2003;57(1):1-17.
141. Slavin JL. Mechanisms for the impact of whole grain foods on cancer risk. J Am Coll Nutr. 2000;19(3 Suppl):300s-7s.
142. Aniwan S, Rerknimitr R, Kongkam P, Wisedopas N, Ponuthai Y, Chaithongrat S, et al. A combination of clinical risk stratification and fecal immunochemical test results to prioritize colonoscopy screening in asymptomatic participants. Gastrointest Endosc. 2015;81(3):719-27.
143. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics. 2007;8:25.

144. Nicodemus KK. Letter to the editor: on the stability and ranking of predictors from random forest variable importance measures. Brief Bioinform. 2011;12(4):369-73.

Supplementary file 1-10

Supplementary file 1: Invitation letter to the CRCbiome study

Supplementary file 2: The FFQ used in the CRCbiome study

Supplementary file 3: FFQ quality control flowchart for the CRCbiome study

Supplementary file 4: The LDQ used in the CRCbiome study

Supplementary file 5: R-script of the Random Forest model

Supplementary file 6: Lifestyle and demographic variables explanation

Supplementary file 7: Dietary variables explanation

Supplementary file 8: CRC relevant dietary variables explanation

Supplementary file 9: Ethical approval for the CRCbiome study

Supplementary table 10: Mean variable importance with energy adjusted dietary variables

Supplementary file 1: Invitation letter to the CRCbiome study

lopenummer / _ref_nr_
navn
adresse

Oslo, _kort_dato_

DELTAKELSE I FORSKNINGSPROSJEKTET

STUDIE AV TARMBAKTERIER OG LIVSSTIL VED TARMSCREENING

Du mottar dette brevet fordi du har levert en avføringsprøve med blod og er invitert til en koloskopiundersøkelse i forbindelse med screening. I forbindelse med dette ønsker vi å invitere deg til å delta i et forskningsprosjekt for å studere om det er en forbindelse mellom tarmbakterier (tarmfloraen), livsstil og forekomst av polypper.

Dette er et tilleggsprosjekt til selve screeningen og din eventuelle deltakelse har ingen betydning for det tilbudet du får i screeningundersøkelsen. Målsettingen med dette tilleggsprosjektet er å finne ut hvilken betydning tarmbakteriene kan ha på tarmkreftrisikoen. Vi vil også undersøke om det er sammenheng mellom kosthold og livsstil, tarmflora og tarmkreftutvikling. Da kan vi forbedre råd om forebygging av kreft samt øke nøyaktigheten på testene.

Mer informasjon om prosjektet finner du på vår hjemmeside **kreftregisteret.no/crc-biome.**

Ved spørsmål ta kontakt via e-post tarmscreening@kreftregisteret.no eller telefon **22 45 13 00** (telefontid fra kl. 8.30 til 11.30).

HVA INNEBÆRER PROSJEKTET?

Deltagelse innebærer at du fyller ut to spørreskjemaer om kosthold og livsstil, før din koloskopiundersøkelse, og tar to avføringsprøver i løpet av året som kommer.

Vi ber om at du fyller ut de to vedlagte spørreskjemaene, og returnerer dem i den frankerte svarkonvolutten eller tar dem med deg når du kommer til koloskopiundersøkelsen. Vi vil kontakte enkelte deltagere per telefon ved behov for utfyllende informasjon. Skjemaene tar totalt ca. en time å fylle ut.

Avføringsprøvene skal tas og sendes på samme måte som du gjorde i screeningundersøkelsen. Den første prøven skal tas ca. to måneder, og den andre ca. et år etter din koloskopiundersøkelse. Prøvetakingsutstyret vil bli sendt til deg i posten.

I prosjektet vil vi innhente og registrere opplysninger om deg. Vi vil registrere funn fra koloskopiundersøkelsen, avføringsprøvene og svar fra spørreskjemaene, og sammenstille disse med data fra hovedundersøkelsen Screening mot tarmkreft - forprosjekt. Opplysningene vil kobles mot sentrale helseregister slik som Kreftregisteret og Reseptregisteret.

MULIGE FORDELER OG ULEMPER

Du vil ikke ha noen direkte fordeler av å delta i studien. Resultater fra studien kan lede frem til ny og viktig kunnskap som kan gi bedre screeningverktøy i fremtiden.

Studien innebærer ingen ulemper for deg som deltager utover medgått tid til å fylle ut spørreskjemaene og avgi avføringsprøvene.

FRIVILLIG DELTAKELSE OG MULIGHET FOR Å TREKKE SITT SAMTYKKE

Det er frivillig å delta i prosjektet. Du kan når som helst og uten å oppgi noen grunn trekke ditt samtykke. Dette vil ikke få konsekvenser for din koloskopiundersøkelse. Dersom du trekker deg fra prosjektet, kan du kreve å få slettet innsamlede prøver og opplysninger, med mindre opplysningene allerede er inngått i analyser eller brukt i vitenskapelige publikasjoner.

Dersom du senere ønsker å trekke deg eller har spørsmål til prosjektet, kan du kontakte sekretariatet for tarmkreftscreening på Kreftregisteret med e-post: **tarmscreening@kreftregisteret.no** eller telefon nr. **22 45 13 00** (sentralbordet, telefontid ved tarmscreeningseksjonen er fra kl. 8.30 til 11.30).

HVA SKJER MED INFORMASJONEN OM DEG?

Informasjonen som registreres om deg skal kun brukes slik som beskrevet i hensikten med studien. Du har rett til innsyn i hvilke opplysninger som er registrert om deg og rett til å få korrigert eventuelle feil i de opplysningene som er registrert.

Alle opplysningene vil bli behandlet uten navn og fødselsnummer eller andre direkte gjenkjennende opplysninger. En kode knytter deg til dine opplysninger gjennom en navneliste.

Prosjektleder har ansvar for den daglige driften av forskningsprosjektet og at opplysninger om deg blir behandlet på en sikker måte. Informasjon om deg vil bli anonymisert eller slettet senest fem år etter prosjektslutt.

HVA SKJER MED PRØVER SOM BLIR TATT AV DEG?

Avføringsprøvene du sender inn skal oppbevares i biobanken *Pilot på et kolorektalcancer screeningprogram*, sammen med resten av prøvene fra tarmscreeningen. Det er Kreftregisteret som er ansvarlig for denne biobanken.

Avføringsprøvene fryses og lagres slik at de kan brukes til å teste om det er andre substanser i avføringen som kan brukes til å påvise kreft eller kreftrisiko.

Disse analysene vil bli utført av våre samarbeidspartner. Informasjon om prosjektet vil publiseres på vår hjemmeside **kreftregisteret.no/crc-biome.**

HVA SLAGS INFORMASJON KAN UNDERSØKELSENE I PROSJEKTET GI?

Avføringsprøvene og funn i koloskopiundersøkelsen skal, sammen med informasjonen fra spørreskjemaene, brukes til å undersøke bakterier og andre biomarkører (mikroRNA). Studien inneholder ikke analyser av arvemateriale (DNA).

FORSIKRING

Som deltaker i studien er du forsikret som enhver vanlig pasient i det offentlige helsevesen (pasientskadeerstatningsordningen).

OPPFØLGINGSPROSJEKT

Som deltakere i denne studien vil du kunne bli kontaktet igjen for å delta i oppfølgningsprosjekter.

GODKJENNING

Prosjektet er godkjent av Regional komite for medisinsk og helsefaglig forskningsetikk (saksnr. 63148)

Tusen takk for hjelpen!

Med vennlig hilsen

Anita Jørgensen, fungerende leder
Pilotprosjektet for tarmscreening

Trine B. Rounge, forsker
Kreftregisteret

Paula Berstad, forsker
Kreftregisteret

Ved spørsmål ta kontakt via e-post **tarmscreening@kreftregisteret.no** eller telefon **22 45 13 00** (telefontid ved tarmscreeningseksjonen er fra kl. 8.30 til 11.30).

Besøk vår hjemmeside **kreftregisteret.no/crc-biome**

Skann QR-koden for å komme direkte til nettsiden

Supplementary file 2: The FFQ used in the CRCbiome study

Prøv så godt du kan å gi et "gjennomsnitt" av dine spisevaner.
Ha det siste året i tankene når du fyller ut.

1. Hvor mye brød pleier du å spise?

Legg sammen det du bruker til alle måltider i løpet av en dag.
(1/2 rundstykke = 1 skive, 1 baguett = 4 skiver, 1 ciabatta = 2 skiver)

	Aldri/ sjelden	Antall skiver pr. dag ½	1	2	3	4	5	6	7	8	9	10	11	12+
Fint brød (loff, baguetter, fine rundstykker, ciabatta)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Mellomgrovt brød (helkornbrød, kneipp, grove rundstykker)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Grovt brød (mer enn 50 % sammalt, mørkt rugbrød)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Fint knekkebrød (kavring)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Grovt knekkebrød (grov skonrok)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Sum skiver pr. dag = _____

Antall skiver pr. uke: _____ x 7 = _____. Tallet brukes i spørsmål 4.
(sum skriver pr. dag)

2. Hva pleier du å smøre på brødet?

Legg sammen det du bruker på skivene i løpet av en uke.
(1/2 rundstykke = 1 skive, 1 baguett = 4 skiver, 1 ciabatta = 2 skiver)

	Aldri/ sjelden	Antall skiver pr. uke 1-5	6-14	15-21	22-28	29-35	36-42	43-49	50-56	57+
Smør (meierismør)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Bremykt	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Brelett	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Myk margarin (Soft Flora, Soft Ekstra)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Soft Oliven	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Vita	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Soft Light, Vita Lett	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Melange	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Annen margarin	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Olivenolje, annen olje på brød	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Majones, remulade på brød	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

3. Hvis du bruker smør/margarin på brødet, hvor mye bruker du?

	Antall skiver ½	1	2	3	4	5 eller flere
En porsjonspakke smør/margarin på 12 g rekker til antall skiver:	☐	☐	☐	☐	☐	☐

60873

4. Hvilke typer pålegg spiser du?

	Aldri/ sjelden	Antall skiver pr. uke 1	2-3	4-5	6-7	8-12	13-18	19-24	25-30	31+
Brunost/prim	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Lett/mager brunost/prim	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Hvitost (eks. Norvegia, Gulost)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Lett/mager hvitost	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Dessertost (eks. Brie, Gräddost, blåmuggoster)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Smøreost (eks. kremost, Philadelfia)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Lett/mager smøreost	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Leverpostei	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Mager leverpostei	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Servelat	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Kokt skinke, lettservelat, kalkunpålegg	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Salami, fårepølse, spekepølse	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Kaviar	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Svolværpostei, Lofotpostei	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Makrell i tomat	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Røkt, gravet laks/ørret	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Sardiner, sursild, ansjos	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Tunfisk	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Reker, krabbe	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Egg (kokt, stekt, eggerøre)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Syltetøy, marmelade	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Lett syltetøy, frysetøy	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Peanøttsmør	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Sjokolade-, nøttepålegg	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Annet søtt pålegg (eks. honning, Sunda, sirup)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Cottage cheese	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Majonessalat (eks. italiensk salat)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Majonessalat lett (eks. lett italiensk salat)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Frukt som pålegg (eks. banan, eple)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Grønnsaker som pålegg (eks. agurk, tomat)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Prøv så godt du kan å gi et "gjennomsnitt" av dine spisevaner.
Ha det siste året i tankene når du fyller ut.

5. Frokostgryn

Svar enten per måned eller per uke.

	Aldri/ sjelden	Gang pr. måned 1	2	3	eller	Gang pr. uke 1	2-3	4-5	6-7	8+		Mengde pr. gang 1	1½	2	3+
Havregrøt	☐	☐	☐	☐		☐	☐	☐	☐	☐	(dl)	☐	☐	☐	☐
Havregryn, 4-korn	☐	☐	☐	☐		☐	☐	☐	☐	☐	(dl)	☐	☐	☐	☐
Mysli, søtet (eks. Solfrokost)	☐	☐	☐	☐		☐	☐	☐	☐	☐	(dl)	☐	☐	☐	☐
Mysli, usøtet (eks. Go'Dag)	☐	☐	☐	☐		☐	☐	☐	☐	☐	(dl)	☐	☐	☐	☐
Cornflakes	☐	☐	☐	☐		☐	☐	☐	☐	☐	(dl)	☐	☐	☐	☐
Honnikorn/Frosties/Chocofrokost	☐	☐	☐	☐		☐	☐	☐	☐	☐	(dl)	☐	☐	☐	☐
All Bran, Weetabix, Havrefras o.l.	☐	☐	☐	☐		☐	☐	☐	☐	☐	(dl)	☐	☐	☐	☐
Puffet ris, havrenøtter	☐	☐	☐	☐		☐	☐	☐	☐	☐	(dl)	☐	☐	☐	☐
	Aldri/ sjelden	**Gang pr. måned 1**	**2**	**3**	**eller**	**Gang pr. uke 1**	**2-3**	**4-5**	**6-7**	**8+**		**Mengde pr. gang 1**	**1½**	**2**	**3+**
Syltetøy til frokostgryn, grøt	☐	☐	☐	☐		☐	☐	☐	☐	☐	(ss)	☐	☐	☐	☐
Sukker til frokostgryn, grøt	☐	☐	☐	☐		☐	☐	☐	☐	☐	(ts)	☐	☐	☐	☐

6. Melk (Husk også å ta med melk du bruker på frokostgryn, grøt og dessert)

(1 glass = 2 dl)

	Aldri/ sjelden	Antall glass pr. dag ½	1	2	3	4	5	6	7+
Helmelk, kefir, kultur	☐	☐	☐	☐	☐	☐	☐	☐	☐
Lettmelk	☐	☐	☐	☐	☐	☐	☐	☐	☐
Ekstra lettmelk	☐	☐	☐	☐	☐	☐	☐	☐	☐
Skummet melk, skummet kultur	☐	☐	☐	☐	☐	☐	☐	☐	☐
Biola/Cultura naturell	☐	☐	☐	☐	☐	☐	☐	☐	☐
Biola/Cultura med bær/frukt	☐	☐	☐	☐	☐	☐	☐	☐	☐
Sjokolademelk, jordbærmelk	☐	☐	☐	☐	☐	☐	☐	☐	☐
Drikkeyoghurt	☐	☐	☐	☐	☐	☐	☐	☐	☐

7. Yoghurt (Husk å ta med yoghurt du bruker til frokostgryn)

Svar enten per måned eller per uke.

	Aldri/ sjelden	Gang pr. måned 1	2	3	eller	Gang pr. uke 1	2-3	4-5	6-7	8+	Beger pr. gang ½	1	2	3+
Yoghurt naturell (125 g)	☐	☐	☐	☐		☐	☐	☐	☐	☐	☐	☐	☐	☐
Yoghurt med frukt (125 g)	☐	☐	☐	☐		☐	☐	☐	☐	☐	☐	☐	☐	☐
Go'morgen yoghurt m/mysli	☐	☐	☐	☐		☐	☐	☐	☐	☐	☐	☐	☐	☐
Lettyoghurt med frukt (125 g)	☐	☐	☐	☐		☐	☐	☐	☐	☐	☐	☐	☐	☐
Lettyoghurt m/mysli	☐	☐	☐	☐		☐	☐	☐	☐	☐	☐	☐	☐	☐

8. Kalde drikker

Svar enten per uke eller per dag, <1 betyr sjeldnere enn 1 gang. Merk at porsjonsenhetene er forskjellige, 1/5 liter tilsvarer ett glass (2 dl), mens 1/3 liter tilsvarer 0,33 liter glassflaske/boks.

	Aldri/ sjelden	Gang pr. uke <1	1-2	3-4	5-6	eller	Gang pr. dag 1	2	3	4+	Mengde pr. gang				
Vann (springvann)	☐	☐	☐	☐	☐		☐	☐	☐	☐	(glass)	1 ☐	2 ☐	3 ☐	4+ ☐
Flaskevann med/uten kullsyre (eks. Farris, Imsdal)	☐	☐	☐	☐	☐		☐	☐	☐	☐	(liter)	1/5 ☐	1/3 ☐	½ ☐	1+ ☐
Appelsinjuice	☐	☐	☐	☐	☐		☐	☐	☐	☐	(glass)	1 ☐	2 ☐	3 ☐	4+ ☐
Eplejuice, annen juice	☐	☐	☐	☐	☐		☐	☐	☐	☐	(glass)	1 ☐	2 ☐	3 ☐	4+ ☐
Eplenektar, annen nektar	☐	☐	☐	☐	☐		☐	☐	☐	☐	(glass)	1 ☐	2 ☐	3 ☐	4+ ☐
Saft med sukker	☐	☐	☐	☐	☐		☐	☐	☐	☐	(glass)	1 ☐	2 ☐	3 ☐	4+ ☐
Saft, kunstig søtet	☐	☐	☐	☐	☐		☐	☐	☐	☐	(glass)	1 ☐	2 ☐	3 ☐	4+ ☐
Brus med sukker	☐	☐	☐	☐	☐		☐	☐	☐	☐	(liter)	1/5 ☐	1/3 ☐	½ ☐	1+ ☐
Brus, kunstig søtet	☐	☐	☐	☐	☐		☐	☐	☐	☐	(liter)	1/5 ☐	1/3 ☐	½ ☐	1+ ☐
Iste med sukker	☐	☐	☐	☐	☐		☐	☐	☐	☐	(liter)	1/5 ☐	1/3 ☐	½ ☐	1+ ☐
Iste, kunstig søtet	☐	☐	☐	☐	☐		☐	☐	☐	☐	(liter)	1/5 ☐	1/3 ☐	½ ☐	1+ ☐
Alkoholfritt øl (eks. Vørterøl, Munkholm)	☐	☐	☐	☐	☐		☐	☐	☐	☐	(liter)	1/5 ☐	1/3 ☐	½ ☐	1+ ☐

9. Alkoholholdige drikker

Svar enten pr. måned eller pr. uke. Merk at porsjonsenhetene er forskjellige, 1/5 liter tilsvarer ett glass (2 dl), mens 1/3 liter tilsvarer 0,33 liter glassflaske/boks.

	Aldri/ sjelden	Gang pr. måned 1	2	3	eller	Gang pr. uke 1	2-3	4-5	6-7	Mengde pr. gang						
Øl, sterk øl, pils	☐	☐	☐	☐		☐	☐	☐	☐	(liter)	1/3 ☐	½ ☐	1 ☐	2 ☐	3 ☐	4+ ☐
Lettøl	☐	☐	☐	☐		☐	☐	☐	☐	(liter)	1/3 ☐	½ ☐	1 ☐	2 ☐	3 ☐	4+ ☐
Rusbrus, Cider m/alkohol	☐	☐	☐	☐		☐	☐	☐	☐	(liter)	1/5 ☐	1/3 ☐	½ ☐	1 ☐	1½ ☐	2+ ☐
Rødvin	☐	☐	☐	☐		☐	☐	☐	☐	(vinglass)	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6+ ☐
Hvitvin	☐	☐	☐	☐		☐	☐	☐	☐	(vinglass)	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6+ ☐
Hetvin (portvin, sherry o.l.)	☐	☐	☐	☐		☐	☐	☐	☐	(1 glass = 4cl)	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6+ ☐
Brennevin, likør	☐	☐	☐	☐		☐	☐	☐	☐	(1 dram = 4cl)	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6+ ☐
Blandede drinker, cocktail	☐	☐	☐	☐		☐	☐	☐	☐	(drink)	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6+ ☐

Prøv så godt du kan å gi et "gjennomsnitt" av dine spisevaner.
Ha det siste året i tankene når du fyller ut.

10. Varme drikker

Svar enten per uke eller per dag, < 1 betyr sjeldnere enn 1 gang.

	Aldri/ sjelden	Gang pr. uke <1	1-2	3-4	5-6	eller	Gang pr. dag 1	2	3	4+	Mengde pr. gang (kopp)	1	2	3-4	5-6	7-8	9+
Kaffe - kokt og presskanne *1 kopp = 2 dl*	☐	☐	☐	☐	☐		☐	☐	☐	☐	(kopp)	☐	☐	☐	☐	☐	☐
Kaffe - traktet, filter *1 kopp = 2 dl*	☐	☐	☐	☐	☐		☐	☐	☐	☐	(kopp)	☐	☐	☐	☐	☐	☐
Kaffe - pulver (instant) *1 kopp = 2 dl*	☐	☐	☐	☐	☐		☐	☐	☐	☐	(kopp)	☐	☐	☐	☐	☐	☐
Espresso *1 kopp = 0,3 dl*	☐	☐	☐	☐	☐		☐	☐	☐	☐	(kopp)	☐	☐	☐	☐	☐	☐
Caffe latte *1 kopp = 3 dl*	☐	☐	☐	☐	☐		☐	☐	☐	☐	(kopp)	☐	☐	☐	☐	☐	☐
Cappucino *1 kopp = 3 dl*	☐	☐	☐	☐	☐		☐	☐	☐	☐	(kopp)	☐	☐	☐	☐	☐	☐
Kakao/varm sjokolade *1 kopp = 2 dl*	☐	☐	☐	☐	☐		☐	☐	☐	☐	(kopp)	☐	☐	☐	☐	☐	☐
Sort te (eks. Earl Grey, solbær) *1 kopp = 2 dl*	☐	☐	☐	☐	☐		☐	☐	☐	☐	(kopp)	☐	☐	☐	☐	☐	☐
Grønn te *1 kopp = 2 dl*	☐	☐	☐	☐	☐		☐	☐	☐	☐	(kopp)	☐	☐	☐	☐	☐	☐
Urtete (eks. nype, kamille, Rooibos) *1 kopp = 2 dl*	☐	☐	☐	☐	☐		☐	☐	☐	☐	(kopp)	☐	☐	☐	☐	☐	☐

	Bruker ikke	Antall pr. kopp ½	1	2	3	4+
Sukker til te (ts/sukkerbit)	☐	☐	☐	☐	☐	☐
Sukker til kaffe (ts/sukkerbit)	☐	☐	☐	☐	☐	☐
Sukketter til te (stk)	☐	☐	☐	☐	☐	☐
Sukketter til kaffe (stk)	☐	☐	☐	☐	☐	☐
Melk/fløte til te (ss)	☐	☐	☐	☐	☐	☐
Melk/fløte til kaffe (ss)	☐	☐	☐	☐	☐	☐

11. Middagsretter

Vi spør både om middagsmåltidene og det du spiser til andre måltider. Legg til slutt sammen hvor mange retter per måned du har merket av for å se om summen virker sannsynlig.

	Aldri/ sjelden	Gang pr. måned 1	2	3	4	5-6	7-8	9+	Mengde pr. gang					
Kjøtt/kjøttretter														
Kjøttpølse av storfe/svin	☐	☐	☐	☐	☐	☐	☐	☐	(pølse)	½ ☐	1 ☐	1½ ☐	2 ☐	3+ ☐
Kjøttpølse av storfe/svin, lett/mager	☐	☐	☐	☐	☐	☐	☐	☐	(pølse)	½ ☐	1 ☐	2 ☐	3 ☐	4+ ☐
Kjøttpølse av kylling/kalkun	☐	☐	☐	☐	☐	☐	☐	☐	(pølse)	½ ☐	1 ☐	2 ☐	3 ☐	4+ ☐
Grillpølse/wienerpølse av storfe/svin	☐	☐	☐	☐	☐	☐	☐	☐	(pølse)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Grillpølse/wienerpølse av kylling/kalkun	☐	☐	☐	☐	☐	☐	☐	☐	(pølse)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Hamburger (m/brød)	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Karbonade	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Kjøttkaker, medisterkaker, kjøttpudding	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Kjøttsaus, gryterett med kjøttdeig	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Taco (tacoskjell med kjøtt og salat)	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Tortilla lefse (med kjøtt og salat)/ wrap	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Kebab	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	1½ ☐	2 ☐	3+ ☐
Lasagne, moussaka	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Pizza (en Grandiosa = ca 550 g)	☐	☐	☐	☐	☐	☐	☐	☐	(pizza)	1/8 ☐	1/4 ☐	½ ☐	3/4 ☐	1+ ☐
Calzone (1 stk = 250-300 g)	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	1½ ☐	2 ☐	2½+ ☐
Pai/quiche	☐	☐	☐	☐	☐	☐	☐	☐	(bit)	1-2 ☐	3-4 ☐	5-6 ☐	7-8 ☐	9+ ☐
Vårruller	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Biff (svin, okse, lam)	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	1½ ☐	2 ☐	2½+ ☐
Koteletter (svin, okse, lam)	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	1½ ☐	2 ☐	2½+ ☐
Stek (svin, okse, lam)	☐	☐	☐	☐	☐	☐	☐	☐	(skive)	1-2 ☐	3-4 ☐	5-6 ☐	7-8 ☐	9+ ☐
Stek (elg, hjort, reinsdyr, rådyr)	☐	☐	☐	☐	☐	☐	☐	☐	(skive)	1-2 ☐	3-4 ☐	5-6 ☐	7-8 ☐	9+ ☐
Gryterett med helt kjøtt, frikassé, fårikål	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1-2 ☐	3-4 ☐	5-6 ☐	7-8 ☐	9+ ☐
Lapskaus, suppelapskaus, betasuppe	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1-2 ☐	3-4 ☐	5-6 ☐	7-8 ☐	9+ ☐

Middagsretter fortsetter neste side.....

Prøv så godt du kan å gi et "gjennomsnitt" av dine spisevaner.
Ha det siste året i tankene når du fyller ut.

Middagsretter forts...

	Aldri/ sjelden	Gang pr. måned 1	2	3	4	5-6	7-8	9+	Mengde pr. gang					
Kjøtt/kjøttretter forts...										1-2	3-4	5-6	7-8	9+
Bacon, stekt flesk	☐	☐	☐	☐	☐	☐	☐	☐	(skive)	☐	☐	☐	☐	☐
Grillet kylling	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1/4 ☐	1/3 ☐	1/2 ☐	3/4 ☐	1+ ☐
Kyllingfilet	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	1½ ☐	2 ☐	3+ ☐
Wok med kjøtt/kylling og grønnsaker	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Kyllinggryte	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1-2 ☐	3-4 ☐	5-6 ☐	7-8 ☐	9+ ☐
Fisk/fiskeretter										1	2	3	4	5+
Fiskekaker, fiskepudding	☐	☐	☐	☐	☐	☐	☐	☐	(kake)	☐	☐	☐	☐	☐
Fiskeboller	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1-2 ☐	3-4 ☐	5-6 ☐	7-9 ☐	10+ ☐
Torsk, sei, hyse, steinbit, uer (kokt)	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Torsk, sei, hyse, steinbit, uer (stekt, panert)	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Fiskepinner	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1-2 ☐	3-4 ☐	5-6 ☐	7-9 ☐	10+ ☐
Sild (fersk, speket, røkt)	☐	☐	☐	☐	☐	☐	☐	☐	(filet)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Makrell (fersk, røkt)	☐	☐	☐	☐	☐	☐	☐	☐	(filet)	½ ☐	1 ☐	1½ ☐	2 ☐	3+ ☐
Laks, ørret (kokt, stekt)	☐	☐	☐	☐	☐	☐	☐	☐	(skive)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Fiskegryte, fiskesuppe	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1-2 ☐	3-4 ☐	5-6 ☐	7-8 ☐	9+ ☐
Fiskegrateng	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1-2 ☐	3-4 ☐	5-6 ☐	7-8 ☐	9+ ☐
Reker, krabbe	☐	☐	☐	☐	☐	☐	☐	☐	(dl, renset)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Wok med sjømat og grønnsaker	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1-2 ☐	3-4 ☐	5-6 ☐	7-8 ☐	9+ ☐
Annet										1-2	3-4	5-6	7-8	9+
Rømmegrøt	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	☐	☐	☐	☐	☐
Risengrynsgrøt, annen melkegrøt	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1-2 ☐	3-4 ☐	5-6 ☐	7-8 ☐	9+ ☐
Pannekaker	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1-2 ☐	3-4 ☐	5-6 ☐	7-8 ☐	9+ ☐
Suppe (tomat, blomkål, ertesuppe)	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1-2 ☐	3-4 ☐	5-6 ☐	7-8 ☐	9+ ☐
Vegetarrett, vegetarpizza, grønnsaksgrateng	☐	☐	☐	☐	☐	☐	☐	☐	(bit/dl)	1-2 ☐	3-4 ☐	5-6 ☐	7-8 ☐	9+ ☐
Hurtignudler (eks. Mr Lee)	☐	☐	☐	☐	☐	☐	☐	☐	(pakke)	½ ☐	1 ☐	1½ ☐	2 ☐	3+ ☐
Omelett	☐	☐	☐	☐	☐	☐	☐	☐	(av antall egg)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐

12. Poteter, ris, spagetti, grønnsaker

Svar enten per måned eller per uke.
Disse spørsmålene dreier seg først og fremst om tilbehør til middagsretter, men spiser du for eksempel en rå gulrot eller salat til lunsj, skal det tas med her.

	Aldri/ sjelden	Gang pr. måned 1	2	3	eller 1	Gang pr. uke 2-3	4-5	6-7	8+		Mengde pr. gang				
Poteter, kokte og bakte	☐	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Potetmos	☐	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Potetsalat m/majones	☐	☐	☐	☐	☐	☐	☐	☐	☐	(ss)	1 ☐	2-3 ☐	4-5 ☐	6-7 ☐	8+ ☐
Fløtegratinerte poteter	☐	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Stekte poteter	☐	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Pommes frites (gatekjøkken, frityrstekt)	☐	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Pommes frites, varmet i ovn	☐	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Bønner/linser	☐	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Ris	☐	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Spagetti, makaroni, pasta	☐	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1-2 ☐	3-4 ☐	5-6 ☐	7-8 ☐	9+ ☐
Pølsebrød, lomper	☐	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Gulrot	☐	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Hodekål	☐	☐	☐	☐	☐	☐	☐	☐	☐	(skalk)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Kålrot	☐	☐	☐	☐	☐	☐	☐	☐	☐	(skive)	1/2 ☐	1 ☐	2 ☐	3 ☐	4+ ☐
Blomkål	☐	☐	☐	☐	☐	☐	☐	☐	☐	(hode)	1/8 ☐	1/6 ☐	1/4 ☐	1/3 ☐	1/2+ ☐
Brokkoli	☐	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1/8 ☐	1/4 ☐	1/2 ☐	3/4 ☐	1+ ☐
Rosenkål	☐	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1-2 ☐	3-4 ☐	5-6 ☐	7-8 ☐	9+ ☐
Løk, rå og stekt	☐	☐	☐	☐	☐	☐	☐	☐	☐	(ss)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Salat (eks. issalat, ruccola)	☐	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1/2 ☐	1 ☐	1 1/2 ☐	2 ☐	2 1/2+ ☐
Paprika	☐	☐	☐	☐	☐	☐	☐	☐	☐	(ring)	1-2 ☐	3-4 ☐	5-6 ☐	7-8 ☐	9+ ☐
Avokado	☐	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1/4 ☐	1/2 ☐	3/4 ☐	1 ☐	1 1/2+ ☐
Tomat	☐	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1/2 ☐	1 ☐	1 1/2 ☐	2 ☐	2 1/2+ ☐
Mais	☐	☐	☐	☐	☐	☐	☐	☐	☐	(ss)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Frosne grønnsakblandinger	☐	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐
Blandet salat (eks. salat, tomat, agurk, mais)	☐	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	1 ☐	2 ☐	3 ☐	4 ☐	5+ ☐

Prøv så godt du kan å gi et "gjennomsnitt" av dine spisevaner.
Ha det siste året i tankene når du fyller ut.

13. Saus og dressing

	Aldri/ sjelden	Gang pr. måned 1	2	3	4	5-6	7-8	9+		Mengde pr. gang				
Brun/hvit saus	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	½ ☐	1 ☐	1½ ☐	2 ☐	3+ ☐
Bearnéssaus, hollandés	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	½ ☐	1 ☐	1½ ☐	2 ☐	3+ ☐
Smeltet margarin/smør	☐	☐	☐	☐	☐	☐	☐	☐	(ss)	½ ☐	1 ☐	1½ ☐	2 ☐	3+ ☐
Kryddersmør	☐	☐	☐	☐	☐	☐	☐	☐	(ts)	½ ☐	1 ☐	1½ ☐	2 ☐	3+ ☐
Majones/remulade vanlig	☐	☐	☐	☐	☐	☐	☐	☐	(ss)	½ ☐	1 ☐	2 ☐	3 ☐	4+ ☐
Majones/remulade lett	☐	☐	☐	☐	☐	☐	☐	☐	(ss)	½ ☐	1 ☐	2 ☐	3 ☐	4+ ☐
Seterrømme (35 % fett)	☐	☐	☐	☐	☐	☐	☐	☐	(ss)	½ ☐	1 ☐	2 ☐	3 ☐	4+ ☐
Lettrømme (20 % fett)	☐	☐	☐	☐	☐	☐	☐	☐	(ss)	½ ☐	1 ☐	2 ☐	3 ☐	4+ ☐
Ekstra lett rømme (10 % fett)	☐	☐	☐	☐	☐	☐	☐	☐	(ss)	½ ☐	1 ☐	2 ☐	3 ☐	4+ ☐
Dressing (eks. Thousand Island)	☐	☐	☐	☐	☐	☐	☐	☐	(ss)	½ ☐	1 ☐	2 ☐	3 ☐	4+ ☐
Lett dressing (eks. lett Thousand Island)	☐	☐	☐	☐	☐	☐	☐	☐	(ss)	½ ☐	1 ☐	2 ☐	3 ☐	4+ ☐
Oljedressing, vinagrette	☐	☐	☐	☐	☐	☐	☐	☐	(ss)	½ ☐	1 ☐	2 ☐	3 ☐	4+ ☐
Soyasaus	☐	☐	☐	☐	☐	☐	☐	☐	(ss)	½ ☐	1 ☐	2 ☐	3 ☐	4+ ☐
Pesto	☐	☐	☐	☐	☐	☐	☐	☐	(ss)	½ ☐	1 ☐	2 ☐	3 ☐	4+ ☐
Tomatsaus, salsa	☐	☐	☐	☐	☐	☐	☐	☐	(ss)	1-2 ☐	3-4 ☐	5-6 ☐	7-8 ☐	9+ ☐
Ketchup	☐	☐	☐	☐	☐	☐	☐	☐	(ss)	½ ☐	1 ☐	2 ☐	3 ☐	4+ ☐
Sennep	☐	☐	☐	☐	☐	☐	☐	☐	(ss)	½ ☐	1 ☐	2 ☐	3 ☐	4+ ☐

14. Hvilken type smør/margarin/olje bruker du mest til matlaging?

(Velg en eller to typer)

Smør/margarin	Oljer
☐ Smør (meierismør)	☐ Olivenolje
☐ Bremykt	☐ Soyaolje
☐ Melange	☐ Maisolje
☐ Soft Flora, Soft Ekstra	☐ Solsikkeolje
☐ Vita	☐ Valnøttolje
☐ Soft Oliven	☐ Rapsolje
☐ Flytende margarin på flaske (Vita, Melange, Bremykt o.l.)	☐ Vita hjertego
☐ Annen margarin	☐ Andre oljer

15. Frukt

Svar enten per måned eller per uke.

	Aldri/ sjelden	Gang pr. måned 1	2	3	eller 1	Gang pr. uke 2-3	4-5	6-7	8+		Mengde pr. gang			
Eple	☐	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	2 ☐	3+ ☐
Pære	☐	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	2 ☐	3+ ☐
Banan	☐	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	2 ☐	3+ ☐
Appelsin	☐	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	2 ☐	3+ ☐
Klementiner	☐	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1 ☐	2 ☐	3 ☐	4+ ☐
Grapefrukt	☐	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	2 ☐	3+ ☐
Fersken, nektarin	☐	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1 ☐	2 ☐	3 ☐	4+ ☐
Kiwi	☐	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1 ☐	2 ☐	3 ☐	4+ ☐
Druer	☐	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1-10 ☐	11-20 ☐	21-40 ☐	41+ ☐
Melon	☐	☐	☐	☐	☐	☐	☐	☐	☐	(skive)	1 ☐	2 ☐	3 ☐	4+ ☐
Jordbær (friske, frosne)	☐	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	½ ☐	1 ☐	2 ☐	3+ ☐
Bringebær (friske, frosne)	☐	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	½ ☐	1 ☐	2 ☐	3+ ☐
Blåbær	☐	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	½ ☐	1 ☐	2 ☐	3+ ☐
Multer	☐	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	½ ☐	1 ☐	2 ☐	3+ ☐
Rosiner	☐	☐	☐	☐	☐	☐	☐	☐	☐	(dl)	½ ☐	1 ☐	2 ☐	3+ ☐
Tørket frukt (eks. aprikos, fiken)	☐	☐	☐	☐	☐	☐	☐	☐	☐	(stk)	1-5 ☐	6-10 ☐	11-15 ☐	16+ ☐
Frukt- og nøtteblanding	☐	☐	☐	☐	☐	☐	☐	☐	☐	(neve)	1 ☐	2 ☐	3 ☐	4+ ☐

16. Grønnsaker og frukt

Hvor mange porsjoner grønnsaker (utenom potet) spiser du vanligvis pr. dag? (En porsjon er f. eks. 1 gulrot, 1 bolle salat)

Mindre enn 1 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5+ ☐

Hvor mange frukt spiser du vanligvis pr. dag?

Mindre enn 1 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5+ ☐

Prøv så godt du kan å gi et "gjennomsnitt" av dine spisevaner.
Ha det siste året i tankene når du fyller ut.

17. Desserter, kaker, godteri

Svar enten per måned eller per uke.

	Aldri/ sjelden	Gang pr. måned 1	2	3	eller	Gang pr. uke 1	2-3	4-5	6-7	8+		Mengde pr. gang			
Iskrem (1 dl=1 pinne=1 kremmerhus)	☐	☐	☐	☐		☐	☐	☐	☐	☐	(dl)	½ ☐	1 ☐	2 ☐	3+ ☐
Saftis/sorbet (1 dl=1 pinne)	☐	☐	☐	☐		☐	☐	☐	☐	☐	(dl)	½ ☐	1 ☐	2 ☐	3+ ☐
Hermetisk frukt, fruktgrøt	☐	☐	☐	☐		☐	☐	☐	☐	☐	(dl)	1 ☐	2 ☐	3 ☐	4+ ☐
Frisk fruktsalat	☐	☐	☐	☐		☐	☐	☐	☐	☐	(dl)	1 ☐	2 ☐	3 ☐	4+ ☐
Pudding (eks. sjokolade, karamell)	☐	☐	☐	☐		☐	☐	☐	☐	☐	(dl)	1 ☐	2 ☐	3 ☐	4+ ☐
Vaniljesaus	☐	☐	☐	☐		☐	☐	☐	☐	☐	(dl)	½ ☐	1 ☐	2 ☐	3+ ☐
Pisket krem	☐	☐	☐	☐		☐	☐	☐	☐	☐	(ss)	1 ☐	2 ☐	3 ☐	4+ ☐
Boller, julekake, kringle	☐	☐	☐	☐		☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	2 ☐	3+ ☐
Skolebrød, skillingsbolle	☐	☐	☐	☐		☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	2 ☐	3+ ☐
Wienerbrød, -kringle	☐	☐	☐	☐		☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	2 ☐	3+ ☐
Muffins, formkake	☐	☐	☐	☐		☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	2 ☐	3+ ☐
Vafler	☐	☐	☐	☐		☐	☐	☐	☐	☐	(plate)	½ ☐	1 ☐	2 ☐	3+ ☐
Lefse, påsmurt	☐	☐	☐	☐		☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	2 ☐	3+ ☐
Sjokoladekake, brownie	☐	☐	☐	☐		☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	2 ☐	3+ ☐
Marsipankake, bløtkake	☐	☐	☐	☐		☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	2 ☐	3+ ☐
Søt kjeks, kakekjeks (eks. Cookies, Bixit, Hob Nobs)	☐	☐	☐	☐		☐	☐	☐	☐	☐	(stk)	1-2 ☐	3-4 ☐	5-6 ☐	7+ ☐
Kokosbolle	☐	☐	☐	☐		☐	☐	☐	☐	☐	(stk)	1 ☐	2 ☐	3 ☐	4+ ☐
Sjokolade (60 g) (eks. melkesjokolade, snickers)	☐	☐	☐	☐		☐	☐	☐	☐	☐	(stk)	½ ☐	1 ☐	2 ☐	3+ ☐
Mørk sjokolade (70% kakao)	☐	☐	☐	☐		☐	☐	☐	☐	☐	(biter)	1-3 ☐	4-6 ☐	7-9 ☐	10+ ☐
Sjokoladebiter/konfekt	☐	☐	☐	☐		☐	☐	☐	☐	☐	(stk)	1-3 ☐	4-6 ☐	7-9 ☐	10+ ☐
Pastiller uten sukker	☐	☐	☐	☐		☐	☐	☐	☐	☐	(stk)	1-3 ☐	4-6 ☐	7-9 ☐	10+ ☐
Drops, pastiller, lakris, seigmenn	☐	☐	☐	☐		☐	☐	☐	☐	☐	(stk)	1-3 ☐	4-6 ☐	7-9 ☐	10+ ☐
Smågodt (1 hg = 100g)	☐	☐	☐	☐		☐	☐	☐	☐	☐	(hg)	½ ☐	1 ☐	2 ☐	3+ ☐
Potetgull	☐	☐	☐	☐		☐	☐	☐	☐	☐	(neve)	1-2 ☐	3-5 ☐	6-10 ☐	11+ ☐
Annen snacks (skruer, crisp, saltstenger, lettsnacks o.l.)	☐	☐	☐	☐		☐	☐	☐	☐	☐	(neve)	1-2 ☐	3-5 ☐	6-10 ☐	11+ ☐
Peanøtter, cashewnøtter (1 neve = 25 gram)	☐	☐	☐	☐		☐	☐	☐	☐	☐	(neve)	1-2 ☐	3-4 ☐	5-6 ☐	7+ ☐
Mandler, hasselnøtter, valnøtter (1 neve = 25 gram)	☐	☐	☐	☐		☐	☐	☐	☐	☐	(neve)	1-2 ☐	3-4 ☐	5-6 ☐	7+ ☐

60873

18. Kosttilskudd (ts = teskje, bs = barneskje)

	Aldri/ sjelden	Gang pr. uke 1	2-3	4-5	6-7	Mengde pr. gang			
Tran	☐	☐	☐	☐	☐	1 ts ☐	1 bs ☐	1 ss ☐	
Trankapsler	☐	☐	☐	☐	☐	(kapsler) 1 ☐	2 ☐	3 ☐	4+ ☐
Fiskeoljekapsler, omega-3 tilskudd	☐	☐	☐	☐	☐	(kapsler) 1 ☐	2 ☐	3 ☐	4+ ☐
Seloljekapsler	☐	☐	☐	☐	☐	(kapsler) 1 ☐	2 ☐	3 ☐	4+ ☐
Multipreparater	**Aldri/ sjelden**	**Gang pr. uke 1**	**2-3**	**4-5**	**6-7**	**Mengde pr. gang 1**	**2**	**3**	**4+**
Sana-sol	☐	☐	☐	☐	☐	(bs) ☐	☐	☐	☐
Biovit	☐	☐	☐	☐	☐	(bs) ☐	☐	☐	☐
Mulitvitamin og mineral (eks. Vitamineral)	☐	☐	☐	☐	☐	(tablett) ☐	☐	☐	☐
Multivitaminer (uten mineraler)	☐	☐	☐	☐	☐	(tablett) ☐	☐	☐	☐
Jernpreparater	**Aldri/ sjelden**	**Gang pr. uke 1**	**2-3**	**4-5**	**6-7**	**Mengde pr. gang 1**	**2**	**3**	**4+**
Duroferon Duretter, Ferromax	☐	☐	☐	☐	☐	(tablett) ☐	☐	☐	☐
Hemofer, hemjern	☐	☐	☐	☐	☐	(tablett) ☐	☐	☐	☐
Amino Jern	☐	☐	☐	☐	☐	(tablett) ☐	☐	☐	☐
Jernmikstur (eks. Floradix)	☐	☐	☐	☐	☐	(bs) ☐	☐	☐	☐
Annet	**Aldri/ sjelden**	**Gang pr. uke 1**	**2-3**	**4-5**	**6-7**	**Mengde pr. gang 1**	**2**	**3**	**4+**
B-vitaminer (flere b-vitaminer i samme tablett)	☐	☐	☐	☐	☐	(tablett) ☐	☐	☐	☐
C-vitamin (60 mg/tablett)	☐	☐	☐	☐	☐	(tablett) ☐	☐	☐	☐
D-vitamin (10 µg/tablett)	☐	☐	☐	☐	☐	(tablett) ☐	☐	☐	☐
E-vitamin (30 mg/tablett)	☐	☐	☐	☐	☐	(tablett) ☐	☐	☐	☐
Folat (folsyre) (200 µg/tablett)	☐	☐	☐	☐	☐	(tablett) ☐	☐	☐	☐

Annet (inkludert helsekostpreparater). Noter navn på preparatet, hvor ofte og hvor mye du tar pr. gang.

19. Måltider

Hvor ofte pleier du å spise følgende måltider i løpet av en uke? (Sett ett kryss for hvert måltid)

	Aldri/ sjelden	1 gang i uken	2 ganger i uken	3 ganger i uken	4 ganger i uken	5 ganger i uken	6 ganger i uken	Hver dag
Frokost	☐	☐	☐	☐	☐	☐	☐	☐
Formiddagsmat/lunsj	☐	☐	☐	☐	☐	☐	☐	☐
Middag	☐	☐	☐	☐	☐	☐	☐	☐
Kveldsmat	☐	☐	☐	☐	☐	☐	☐	☐

Hvor mange ganger i løpet av dagen pleier du å spise et eller annet utenom hovedmåltidene? (eks. godteri, frukt, brødskive)

Sjelden	1 gang om dagen	2 ganger om dagen	3 ganger om dagen	4 ganger om dagen	Mer enn 4 ganger om dagen
☐	☐	☐	☐	☐	☐

20. Kjønn

Mann ☐

Kvinne ☐

21. Alder

Alder: ☐☐ år

22. Vekt og høyde

Høyde: ☐☐☐ cm

Vekt: ☐☐☐ kg

23. Eventuelle andre matvarer

Bruker du regelmessig matvarer, drikker eller andre produkter som ikke er nevnt i spørreskjemaet? Skriv ned dette så detaljert som mulig. Skriv også hvor ofte du spiser/drikker dette (ganger per måned eller uke) og hvor mye du spiser av dette per gang.

BRUK BLOKKBOKSTAVER

Tusen takk for innsatsen!

Supplementary file 3: FFQ quality control flowchart for the CRCbiome study

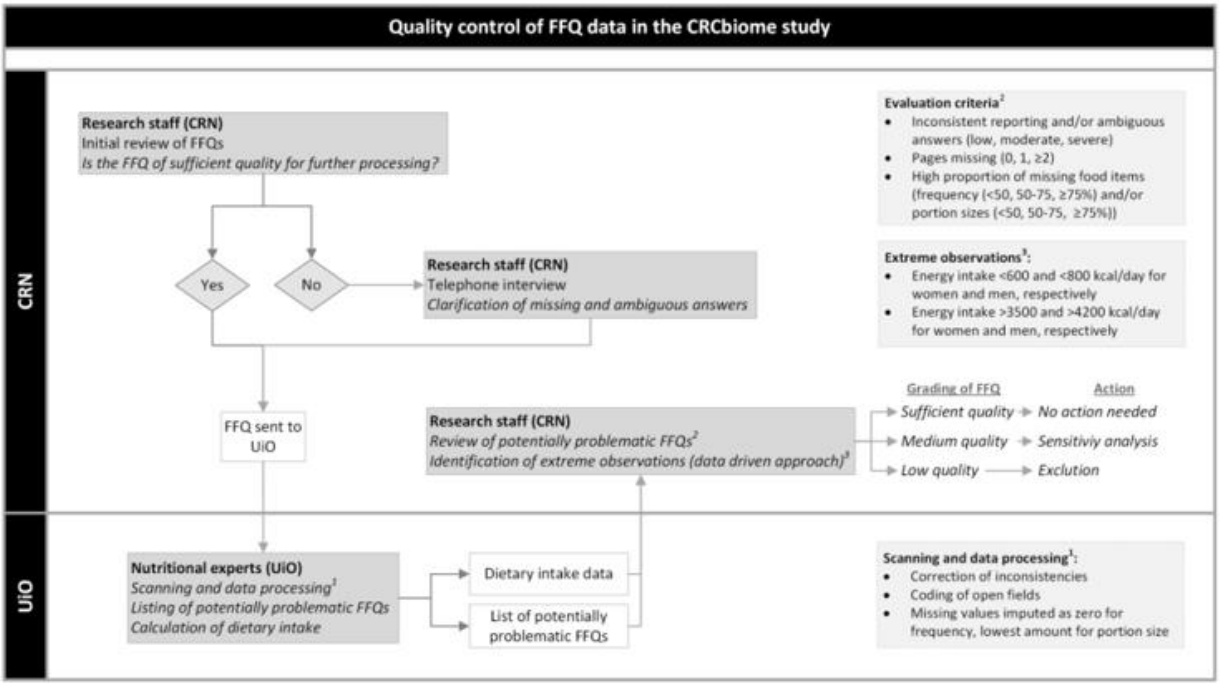


Figure 1. Upon receiving food frequency questionnaires (FFQs) from CRCbiome participants, completion is reviewed by researchers with expertise in nutritional epidemiology. Participants with FFQs of insufficient quality are contacted for clarification of inconsistencies and missing data. Reviewed questionnaires are then scanned using the Cardiff TeleForm program at the University of Oslo (UiO). Food and nutrient calculations are conducted using the software system KBS ("**K**ost**b**eregnings**s**ystem"/Dietary Calculation System) with the latest version of the food database, largely based on the Norwegian Food Composition Table (1). Missing answers are imputed as zero in line with common practice (2–5). Any FFQs regarded as potentially problematic during the data handling process are listed. Dietary intake data and the list of potentially problematic FFQs are then returned to the Cancer Registry of Norway (CRN). Potentially problematic FFQs are reviewed according to a set of predefined criteria, including inconsistency in reporting, number of missing pages and amount of missing food items. Based on these criteria, FFQs are graded as being of low, medium or sufficient quality. Whereas low quality FFQs will be excluded from all analysis where diet is the primary exposure, medium quality FFQs will be included unless sensitivity analysis indicates substantial attenuation of effect estimates. Lastly, in line with common practice in nutrition studies (6), observations with extreme energy intake levels in both the upper and lower range will be excluded.

1. Norwegian Food Safety Authority. Norwegian Food Composition Database 2019. [cited 2020 Jun 16]. http://www.matvaretabellen.no

2. Johansson I, Hallmans G, Wikman Å, Biessy C, Riboli E, Kaaks R. Validation and calibration of food-frequency questionnaire measurements in the Northern Sweden Health and Disease cohort. Public Health Nutr. 2002;5:487–96.

3. Carlsen MH, Lillegaard IT, Karlsen A, Blomhoff R, Drevon CA, Andersen LF. Evaluation of energy and dietary intake estimates from a food frequency questionnaire using independent energy expenditure measurement and weighed food records. Nutr J. 2010;9:1–9.

4. Carlsen MH, Karlsen A, Lillegaard ITL, Gran JM, Drevon CA, Blomhoff R, Andersen LF. Relative validity of fruit and vegetable intake estimated from an FFQ, using carotenoid and flavonoid biomarkers and the method of triads. Br J Nutr. 2011;105:1530–8.

5. Holmberg L, Ohlander EM, Byers T, Zack M, Wolk A, Bruce Å, Bergstrom R, Bergkvist L, Adami HO. A Search for Recall Bias in a Case-Control Study of Diet and Breast Cancer. Int J Epidemiol. 1996;25:235–44.

6. Willett W. Nutritional epidemiology. Oxford; New York: Oxford University Press; 2013.

Supplementary file 4: The LDQ used in the CRCbiome study

STUDIE AV TARMBAKTERIER OG LIVSSTIL VED SCREENING MOT TARMKREFT

Livsstil og andre opplysninger

1279093124

Til denne studien trenger vi noen opplysninger om din bakgrunn og livsstil slik den vanligvis er. Vi er klar over at levevaner varierer over tid. Prøv derfor å angi gjennomsnittet av vanene dine når du svarer på spørsmål om røyking, snus, fysisk aktivitet og melketyper. Angi det du mener gjelder for det siste året når du fyller ut skjemaet. Der du er usikker, angir du svaret så godt du kan.

Riktig markering er: ☒ for svaralternativer.

For tallverdier, skriv et tall i hver rute, for eksempel: | |5| Glass pr. uke

Dato for utfylling: **Dag** ☐☐ **Måned** ☐☐ **20** **År** ☐☐

1. Personlige opplysninger

Nasjonale bakgrunn *(dine foreldres fødeland) (Sett bare ett kryss)*

Hvis dine foreldre har ulike fødeland, kryss av for det området som du føler mest tilhørighet til.

☐ Norge
☐ Sør-Europa, Sør- eller Sentral-Amerika
☐ Afrika
☐ Nord- eller Sentral-Europa (utenom Norge), Nord-Amerika, Australia
☐ Asia

Sivilstatus *(Sett bare ett kryss)*

☐ Enslig
☐ Enke/ enkemann/ gjenlevende partner
☐ Gift/ registrert partner/ samboende
☐ Skilt/ skilt partner/ separert/ separert partner

Høyeste fullførte utdanning *(Sett bare ett kryss)*

☐ Grunnskole/ folkeskole
☐ Universitet/ høgskole (fullført minst 2 år)
☐ Videregående skole

Yrkesstatus Er du for tiden: *(Sett bare ett kryss)*

☐ Yrkesaktiv
☐ Hjemmeværende
☐ Pensjonist
☐ Arbeidsledig
☐ På uførepensjon, ev. kombinert med arbeid eller andre ytelser (f.eks. alderspensjon)
☐ På attføring/rehabilitering/ arbeidsavklaringspenger/ langtidssykemeldt (mer enn 3 mnd)

2. Røyking

Røyker du nå? *(Sett bare ett kryss)*

Ta med både fabrikklagde og hjemmerullede sigaretter. Hvis du har sluttet eller trappet ned antallet sigaretter flere ganger, prøv så godt du kan å gi et gjenommsnitt.

☐ Ja, daglig ☐ Ja, av og til

☐ Nei, ikke nå ☐ Nei, har aldri røykt

Hvis ja, hvor mye? ☐☐ Sigaretter pr. uke **eller** ☐☐ Sigaretter pr. dag

---

Hvis du har røykt tidligere og sluttet

Hvor mye pleide du å røyke? ☐☐ Sigaretter pr. uke **eller** ☐☐ Sigaretter pr. dag

Hvor mange år eller måneder er det siden du sluttet å røyke siste gang? ☐☐ år **eller** ☐☐ mnd

Hvor mange år eller måneder har du røykt totalt? ☐☐ år **eller** ☐☐ mnd

3. Snus

Bruker du snus? *(Sett bare ett kryss)*

Ta med både posesnus og snus i løsvekt. Hvis du har sluttet eller trappet ned antallet snusporsjoner flere ganger, prøv så godt du kan å gi et gjennomsnitt.

☐ Ja, daglig ☐ Ja, av og til

☐ Nei, ikke nå ☐ Nei, har aldri snust

Hvis ja, hvor mye? ☐☐ Porsjoner pr. uke **eller** ☐☐ Porsjoner pr. dag

---

Hvis du har brukt snus tidligere og sluttet

Hvor mye pleide du å snuse? ☐☐ Porsjoner pr. uke **eller** ☐☐ Porsjoner pr. dag

Hvor mange år eller måneder er det siden du sluttet å bruke snus siste gang? ☐☐ år **eller** ☐☐ mnd

Hvor mange år eller måneder har du brukt snus totalt? ☐☐ år **eller** ☐☐ mnd

4. Fysisk aktivitet

Har du noen kroniske sykdommer eller tilstander som gjør at du ikke kan utføre fysisk aktivitet?

☐ Nei ☐ Ja, angi grunn | ☐ leddgikt ☐ ryggplager ☐ hofte/kneplager ☐ annet

Tenk gjennom hvor lang tid i løpet av en vanlig uke du tilbringer i fysisk aktivitet?
Ta bare med episoder som varer i minst 10 minutter.
Hvor lang tid tilbringer du hver uke på:

Lett anstrengende aktiviteter som krever lite innsats (rolig gange, rolig sykling, hus- og hagearbeid):	Middels anstrengende aktiviteter som krever moderat innsats og får deg til å puste litt mer enn vanlig (sykle/svømme/gå på ski i moderat tempo, jogge, danse, styrketrening):	Meget anstrengende aktiviteter som krever hard innsats, får deg til å puste mye mer enn vanlig (aerobics, løpe/sykle/svømme/gå på ski i rask tempo):
timer per uke	timer per uke	timer per uke
☐ ingenting	☐ ingenting	☐ ingenting
☐ mindre enn 0,5 time	☐ mindre enn 0,5 time	☐ mindre enn 0,5 time
☐ 0,5 til 1 time	☐ 0,5 til 1 time	☐ 0,5 til 1 time
☐ 1,5 - 2 timer	☐ 1,5 - 2 timer	☐ 1,5 - 2 timer
☐ 2,5 - 3,5 timer	☐ 2,5 - 3 ,5 timer	☐ 2,5 - 3,5 timer
☐ 4-6 timer	☐ 4-6 timer	☐ 4-6 timer
☐ 7 eller flere timer	☐ 7 eller flere timer	☐ 7 eller flere timer

5. Bruk av melk vs. surmelk

Hvis du bruker melk hvor mye bruker du av hver type?

Som surmelk regnes alle typer kulturmelk, Cultura, Kefir, drikkbar Biola og tykkmelk. Mengden melk til en porsjon kornblanding regnes som et glass.

Hvor mye melk bruker du? ☐☐ Glass pr. uke **eller** ☐☐ Glass pr. dag

Hvor mye surmelk bruker du? ☐☐ Glass pr. uke **eller** ☐☐ Glass pr. dag

6. Keisersnitt

Ble du født med keisersnitt? (*Sett bare ett kryss)*

☐ Nei ☐ Ja ☐ Vet ikke

7. Fjerning av blindtarm

Er din blindtarm fjernet? (*Sett bare ett kryss)*

☐ Nei ☐ Ja ☐ Vet ikke **4135093128**

8. Medisiner

Har du brukt noen av de følgende medisiner de siste 3 månedene?
Ta med både medisiner kjøpt med og uten resept.

Antibiotika

☐ Ja ☐ Nei ☐ Vet ikke

Syrenøytraliserende legemidler
F.eks. Nexium, Somac

☐ Ja ☐ Nei ☐ Vet ikke

9. Kroniske sykdommer og matintoleranse

Har du en kronisk mage-tarmlidelse påvist av lege?

☐ Nei ☐ Ja, hvilken? _____

☐ Irritabel tarmsyndrom ☐ Ulcerøs kolitt ☐ Cøliaki ☐ Crohns sykdom

☐ Annet

Har du intoleranse mot enkelte matvarer eller matkomponenter?

☐ Nei ☐ Ja ☐ Vet ikke

Hvis ja, oppgi hvilken:

10. Familiehistorie for tarmkreft

Har noen av dine nærmeste slektninger hatt tarmkreft, eller har det nå?
Med nærmeste slektninger menes mor, far, bror, søster eller egne barn.

☐ Ja, mor ☐ Ja, far ☐ Ja, søster/bror ☐ Ja, barn ☐ Nei ☐ Vet ikke

Vi ber om ditt telefonnummer slik at vi kan kontakte deg hvis nødvendig.

Ditt telefonnummer ☐☐☐☐☐☐☐☐

Det er i orden at vi ringer deg mellom klokken (f.eks 0830) ☐☐☐☐ og ☐☐☐☐

Tusen takk for innsatsen!

Supplementary file 5: R-script of the Random Forest model

```
tidy_rf <- function(dat,
                    target_var,
                    training_prop = .8,
                    n_CV = 5,
                    hyperparams,
                    iter = 1,
                    hypopt_cores = 10,
                    downsample = TRUE,
                    pca = FALSE,
                    pca_vars = "",
                    pca_remove_cor_features = TRUE,
                    importance_measure = "permutation",
                    test_dat = NULL) {

  dat <- dat %>%
    rename(target = all_of(target_var)) %>%
    mutate(target = factor(target))

  set.seed(iter)
  if (is.null(test_dat)) {
    tr_te_set <-  dat %>%
      initial_split(prop = training_prop, strata = target)
  } else {
    test_dat <- test_dat %>%
      rename(target = all_of(target_var)) %>%
      mutate(target = factor(target))

    combined <- bind_rows(dat, test_dat)
    ind <- list(analysis = seq(nrow(dat)), assessment = nrow(dat) + seq(nrow(test_dat)))
    tr_te_set <- make_splits(ind, combined)
  }
```

Supplementary file 6: Lifestyle and demographic variables explanation

Abbreviation	Variable name	Variable explanation
Kjonn.x	Sex	Includes male and female.
Age invitation	Age	Participants age at the time the baseline of the CRCbiome study.
BMI	BMI	Body Mass Index. Calculated as weight (in kilograms) divided by the square of the height (in meter). Both height and weight are self-reported. BMI was used as a continuous variable.
PhysAct_Score	Physical activity score	Calculated as moderate intensity physical activity in minutes, plus minutes with high intensity physical activity times two.
Smoking	Smoking habits	Reporting of daily or occasional smoking was considered as smokers. In addition people missing answer about usual smoking habits, but answered that they were formerly smoking for > 30 years or quitted within 5 years/60 months was categorised as smokers. Reporting of quitting smoking for more the 5 years/ 60 months, missing information on smoking information, but usually smoking <30 years and never smoked was considered non-smokers.
Utdanning	Education	Reported as the highest level of education: primary school, high school, university/college and missing.
Tarmkreft_Familie	Family history of CRC	A positive family history of CRC includes a parent, sibling or child with CRC.

Supplementary file 7: Dietary variables explanation

Abbreviation	Variable names	Variable explanation
LOFF+ LCT + SMBR + (FROK_S - V807) + (GRYTE/3) = ref_grain_product	Refined grain products	Includes all kind of bread made with 100 % sifted flour (e.g., white bread, hamburger buns, pizza doughs without topping, phyllo dough, wraps). Does also include sweetened cereals (e.g., Honey corn, cornflake's) and crackers.
GROVBR + KNEIP + KNFL + FROK_U + V807 + V5456 = grian_product	Grain products	Includes all kind of bread, crisp bread and flat bread made with whole-wheat flour (< 50 % or > 50 %). Does also include unsweetened cereals (oats, Bran flakes) and porridge (oatmeal).
MELRIS + (SUSHI/2) + (PASTAR/3) = pasta_rice_grain	Pasta, rice, and grains	Includes all kind of pasta (e.g., macaroni, spaghetti), rice, flour, groats and wholegrain (e.g., couscous, quinoa).
KAKER	Cake	Includes all kind of cakes and sweet baking goods. E.g., waffles, cinnamon buns, chocolate cake, cookies, and soft cake.
POTETRA + POTETKOKT + POTETSTEKT = unproc_pot	Unprocessed potatoes	Includes raw, boiled and fried potatoes.
POTGRAT + POTSAL + POTRET + POMFRIT + POTMOS = proc_pot	Processed potatoes	Includes potatoes au gratin, potato salad (with cream or oil dressing), mashed potatoes and French fries.
GRS_FF - V1129 = Vegetables	Vegetable	Includes all fresh, frozen or cooked vegetables
GRS_K	Conserved vegetables	Includes all canned and pickled vegetables (e.g., canned tomatoes, canned corn, and ketchup).
GRSRETT + (GRYTE/3) + GRUP = veg_dish	Vegetable dish	Includes all dishes made of vegetables (e.g., vegetable stew, wok) and all kind of vegetable soups.
FRU_F	Fresh and frozen fruit	Includes all kinds of fruits and berries, fresh or frozen.
FRU_K + V2652 = cons_fruit	Conserved fruits	All kinds of canned, pickled, and dried fruits and berries. Including jam, marmalade, raisins, and fruit cocktail.
JUICE	Juice	Includes all juices made from fruits and berries.
FETFRU + V1129 +V9805 = fatty_fruits	Fatty fruits	All kinds of olives, nuts (with and without salt), seeds and avocado.
KJOT_H + KJO_AP = unproc_red_meat	Unprocessed red meat	All kind of red meat and venison, unprocessed.
LJOT_P + KJOPL +(PASTAR/3) + (GRYTE/3) = proc_red_meat	Processed read meat	Includes all kind of processed read meat (e.g., minced meat, sausages, meatloaf, ham, salami).
BLODIN	Meat offal	Includes products made from blood and offal from mammals (e.g., lung puree and blood pudding).
HVIKJO_R + HVIKJO_V = unproc_white_meat	Unprocessed white meat	Includes unprocessed meat from poultry.
HVIKJO_P	Processed white meat	Includes processed meat form poultry (e.g. sausages, turkey ham).
FISK_MH + FISK_U + (SUSHI/4) + (SUSHI/4) = lean_fish	Lean fish	Includes filet from lean and semi fatty fish. (e.g., cod, halibut). Also includes clipfish and lutefish.
FISK_F + (SUSHI/4) = fatty_fish	Fatty fish	Includes filet from fatty fish (e.g. salmon, rout)
FISK_P	Processed fish	Includes products made from fish (e.g., fish finger, breaded fish filet, fish cakes).
SKALDY	Shellfish	Includes all kind of shellfish.
IMAT_F	Fish offal	Includes all kind of fish offal (e.g. fish liver and roe).

FISKPA	Fish spread	Spread made from fish (e.g., mackerel in tomato).
EGG	Egg	Includes all egg dishes. Including boiled egg, scrambled egg, omelette, and fried egg.
MELKYO + V5450 + (0.8*V8406) + (0.8 V8407) + (0.8 V8433) = milk_yogh	Milk and yoghurt	Includes all kinds of milk, both high fat, low fat, skimmed, flowered and sour milk. Includes yoghurt, both natural, flavoured, and low fat and rice porridge.
FLOTIS + V5368 = cream_prod	Cream products	Includes cream and sour cream, both high fat and low fat, and cream products. E.g., ice cream, sour cream porridge, custard, and puddings.
OST_H	White cheese	Includes hard white cheese, cream cheese, dessert cheeses and low-fat cheeses (e.g., cottage cheese).
OST_B	Brown cheese	Includes all kinds of brown cheese and prim.
MARG + MARG_L=margarine	Margarine	Includes all kind of margarine, including light margarine.
SMOR + SMOR_U = butter	Butter	Includes all kind of butter, including butter-margarine blend products (e.g., Bremykt) and diet butter.
OLJE_A	Oil	Includes all kind of oil used in cooking or /and in dishes. E.g., olive oil and sunflower oil.
MAJODR	Dressing	Includes sour cream dressings, mayonnaise, mayonnaise like products (e.g., aioli and rémoulade) and mayonnaise salads (e.g., Italian salad).
SUK_MV + HONSIR = sweetner	Sweetener	Includes sugar, artificial sweetener, honey, syrup, and other sweeteners used to sweeten food or drinks.
SJOK + SOTPA = sweets	Sweets	Include all kind of candy and chocolate, also chocolate spread (e.g. Nugatti).
KAFFE – (0.8 V8406) – (0.8 V8407) – (0.8 V8433) = coffee	Coffee	Includes coffee (does not include milk and sweetener used in coffee)
TE – (V8420 + V8421) = tea	Tea	Includes tea (does not include iced tea, milk or sweetener).
VINBR	Wine and liquor	Include all kind of wine, liquor, and cocktails with alcohol.
OL	Beer	Includes beer with alcohol.
MELKERS + V5277 = milk_sub	Milk substitute	Includes non-dairy beverages (e.g. oat milk, almond milk, soy drink, coconut milk)
SAFTK + SAFTIS + SABR_S + OLV_AF + V8421= energy_drinks	Energy drinks and ice lolly	Include soft drinks with sugar (E.g., soda with sugar, nectar, squash, and ice tea) and non-alcoholic beer. Does also include sorbet and ice lolly (e.g. Lollypop).
SABR_L + V8420 = no_energy_drinks	No energy drinks	Include artificial sweetened soda, squash and iced tea.
SNACKS	Snacks	Includes all kind of snacks products (e.g. potato chips, tortilla chips and popcorn)
SAUS + (PASTAR/3) = sause	Sauce	Includes all type of sauces. E.g., mustard, béarnaise sauce, gravy, cream sauces, salsa, pesto.
PULVER	Powder	Includes broth powder and powder in instant soups.
KRYDDERE	Spices and herbs	Includes all kind of spices (e.g., salt) and herbs (e.g., basil) used in dishes and in cooking.

Abbreviations: V807, havrefras; V5456,havregrøt; V5450, risgrøt; V1129, avocado: V2652, A-frukt herm+grøt; V5368, N-pudding; V8421, iste med sukker; V8406, caffe latte enkel H melk; V8407, cappuccino enkel H melk; V8433, iskaffe kunstigsøtet; V8420, iste lett tine; V8421, iste med sukker Tine; V5277, Yofu Soyayoghurt; V9805, sesampostei.

Supplementary file 8: CRC relevant dietary variables explanation

Abbreviation	Variable names	Variable explanation
Fruit_wcrf	Fruits	Includes all kind of fruits, both fresh and frozen. However, it only includes 50 % of conserved fruits, jam, and dried fruits consumed and only 100g of all juice consumed.
Vegetables_wcrf	Vegetables	Includes all kind of vegetables fresh, frozen or prepared in any way. Includes pickled and conserved vegetables. Includes 50 % of vegetable dishes consumed (e.g. vegetable stew and vegetable soup).
Red_meat_wcrf	Red meat	Includes all kind of red meat, except from venison. Duse not include processed meat products.
Proc_meat_wcrf	Processed meat	Include all kind and products of red and white meat.
Dairy_wcrf	Dairy products	Includes all products made from animal milk (e.g. Milk, cream, cheese, ice cream).
Fish_wcrf	Fish	Includes all fish, but only 90 % of fish spread (e.g. marcel in tomato sauce), 60% of fish products (e.g. fish cakes, fish fingers). Shellfish and fish offal are not included.
Alko.x	Alcohol	Total intake of alcohol in the diet.
Fiber.x	Fiber	Total intake of fiber in the diet.
Fullk.x	Wholegrain	Total intake of wholegrain in the diet.
VitD	Calcium	Total intake of calcium in the diet.
Ca.x	Vitamin D	Total intake of vitamin D in the diet.

Supplementary file 9: Ethical approval for the CRCbiome study

REK
REGIONALE KOMITEER FOR MEDISINSK OG HELSEFAGLIG FORSKNINGSETIKK

Region:	Saksbehandler:	Telefon:	Vår dato:	Vår referanse:
REK sør-øst D	Finn Skre Fjordholm	+47 22 84 58 21	18.12.2019	63148
			Deres referanse:	

Trine Ballestad B Rounge

63148 Tarmbakterier og livsstil ved screening mot tarmkreft

Forskningsansvarlig: Kreftregisteret - Institutt for populasjonsbasert kreftforskning

Søker: Trine Ballestad B Rounge

Søkers beskrivelse av formål:

Tarmkreftsymptomer er ofte uspesifikke og sykdommen oppdages ofte for sent til at behandlingen kan forlenge livet. Dagens screeningtester er enten omfattende og ubehagelige eller unøyaktige. Det er et behov for bedre tester.
Det er sammenheng mellom den enkeltes tarmflora og tarmkreftutvikling. Livsstil kan påvirke tarmens bakterieflora og kreftrisiko, men dette samspillet er lite kjent. Ved å kartlegge alle bakterier som finnes i tarmen kan man utvikle tester som kan brukes til å oppdage forstadier og kreft tidlig.
Vårt hovedmål er å utvikle nye tester for tarmbakterier som kan brukes i fremtidige screeningprogram slik at prøvetagning forenkles og resultatet blir sikrere. Vi vil også undersøke om det er sammenheng mellom kosthold og livsstil, tarmflora og tarmkreftutvikling. Da kan vi forbedre råd om forebygging av kreft samt øke nøyaktigheten på testene.

REKs vurdering

Vi viser til søknad om forhåndsgodkjenning av ovennevnte forskningsprosjekt. Søknaden ble behandlet av Regional komité for medisinsk og helsefaglig forskningsetikk (REK sør-øst D) i møtet 04.12.2019. Vurderingen er gjort med hjemmel i helseforskningsloven § 10.

Prosjektet er en samling av to nåværende delprosjekter under REK 2011/1272 D «Pilot på et kolorektalcancer screeningprogram» og REK 2010/3087 A «S-98052a NORCCAP».

Det fremgår at omfanget av prosjektet tidligere er godkjent av REK, og i søknaden vises det til to vedtak på endringssøknader i REK 2011/1272 datert den 17.3.2017 og 06.03.2018, og ett vedtak i REK 2010/3087 datert den 07.04.2016.

Det er en sammenheng mellom tarmens bakterieflora og risiko for kreft, og formålet med prosjektet er å undersøke denne sammenhengen nærmere. Deltagerne skal fylle ut to spørreskjema før gjennomføring av koloskopiundersøkelse. Denne undersøkelsen inngår i de to tidligere godkjente prosjektene, og data fra denne undersøkelsen blir tatt i bruk i dette prosjektet.

Det blir gjort analyser av en avføringsprøve fra REK 2011/1272. Videre skal det avleveres to avføringsprøver i løpet av et år.

Det hentes inn summariske opplysninger fra Kreftregisteret og Dødsårsaksregisteret. Fra Reseptregisteret hentes det inn opplysninger om bruk av antibiotika og medisiner som påvirker tarmen.

Komiteen har vurdert søknaden og har ingen innvendinger til studien som sådan. Komiteen har imidlertid flere merknader til informasjonsskrivet og godkjenner prosjektet på vilkår om at dette endres i henhold til disse.

Vilkår

- Det står i informasjonsskrivet at prøvene lagres «i en forskningsbiobank, sammen med resten av prøvene fra *Screening mot tarmkreft – forprosjekt*». Komiteen legger til grunn at det her er snakk om biobanken som er tilknyttet REK 2011/1272. Det bes om at det avklares hvilken biobank prøven skal lagres i, og at informasjonsskrivet oppdateres slik at navn på biobanken og ansvarshavende fremgår av informasjonsskrivet.

- Informasjonsskrivet må inneholde mer informasjon om prosjektet.

- I innledningen av skrivet bør også sammenhengen mellom prosjektet og REK 2011/1272 og REK 2010/3087 forklares nærmere.

Vedtak

Godkjent med vilkår

REK har gjort en helhetlig forskningsetisk vurdering av alle prosjektets sider. Prosjektet godkjennes med hjemmel i helseforskningsloven § 10, under forutsetning av at ovennevnte vilkår er oppfylt.

Vi gjør samtidig oppmerksom på at etter ny personopplysningslov må det også foreligge et behandlingsgrunnlag etter personvernforordningen. Det må forankres i egen institusjon.

I tillegg til vilkår som fremgår av dette vedtaket, er godkjenningen gitt under forutsetning av at prosjektet gjennomføres slik det er beskrevet i søknad og protokoll, og de bestemmelser som følger av helseforskningsloven med forskrifter.

Tillatelsen gjelder til 01.01.2034. Av dokumentasjonshensyn skal opplysningene likevel bevares inntil 01.01.2039. Forskningsfilen skal oppbevares atskilt i en nøkkel- og en opplysningsfil. Opplysningene skal deretter slettes eller anonymiseres, senest innen et halvt år fra denne dato.

Forskningsprosjektets data skal oppbevares forsvarlig, se personopplysningsforskriften kapittel 2, og Helsedirektoratets veileder for «Personvern og informasjonssikkerhet i forskningsprosjekter innenfor helse og omsorgssektoren».

Dersom det skal gjøres vesentlige endringer i prosjektet i forhold til de opplysninger som er gitt i søknaden, må prosjektleder sende endringsmelding til REK.

Prosjektet skal sende sluttmelding på eget skjema, senest et halvt år etter prosjektslutt.

Komiteens avgjørelse var enstemmig

Med vennlig hilsen

Finn Wisløff
Professor em. dr. med.
Leder

Finn Skre Fjordholm
Rådgiver

Kopi: Kreftregisteret ved øverste administrative ledelse: kreftregisteret@kreftregisteret.no; giske.ursin@kreftregisteret.no

Sluttmelding
Søker skal sende sluttmelding til REK sør-øst D på eget skjema senest seks måneder etter godkjenningsperioden er utløpt, jf. hfl. § 12.

Søknad om å foreta vesentlige endringer
Dersom man ønsker å foreta vesentlige endringer i forhold til formål, metode, tidsløp eller organisering, skal søknad sendes til den regionale komiteen for medisinsk og helsefaglig forskningsetikk som har gitt forhåndsgodkjenning. Søknaden skal beskrive hvilke endringer som ønskes foretatt og begrunnelsen for disse, jf. hfl. § 11.

Klageadgang
Du kan klage på komiteens vedtak, jf. forvaltningsloven § 28 flg. Klagen sendes til REK sør-øst D. Klagefristen er tre uker fra du mottar dette brevet. Dersom vedtaket opprettholdes av REK sør-øst D, sendes klagen videre til Den nasjonale forskningsetiske komité for medisin og helsefag (NEM) for endelig vurdering.

Alle skriftlige henvendelser om saken må sendes via REK-portalen
Du finner informasjon om REK på våre hjemmesider rekportalen.no

Supplementary file 10: Mean variable importance with energy adjusted dietary variables

Model 1 E		**Model 1.2 E**		**Model 1.3 E**		**Model 1.4 E**	
FIT value	11.96	Age	10.07	FIT value	10.24	FIT value	10.16
Age	10.20	FIT value	9.69	Age	8.71	Milk & yoghurt	5.83
Milk & yoghurt	8.46	Milk & yoghurt	9.38	Milk & yoghurt	8.66	Grain products	4.83
Butter	8.30	Grain products	8.54	BMI	7.63	Proc red meat	4.75
Grain products	8.06	BMI	8.46	Grain products	7.57	Butter	4.60
White cheese	8.01	White cheese	7.46	White cheese	7.29	Age	4.57
Unproc potato	7.54	Fre/froz fruit	7.13	Cons fruits	7.17	Cream products	4.50
Vegetable	7.54	Vegetable	7.11	Vine and liquor	6.78	Cons vegetables	4.45
BMI	7.40	Unproc potato	7.00	Proc red meat	6.52	BMI	4.33
Lean fish	7.25	Conserved fruits	6.95	Fre/froz fruit	6.21	Fatty fish	4.20
Model 2 E		**Model 2.2 E**		**Model 2.3 E**		**Model 2.4 E**	
Milk & yoghurt	9.17	Milk & yoghurt	9.82	Milk & yoghurt	9.39	Milk & yoghurt	7.31
White cheese	8.42	Grain products	9.30	Grain products	8.23	Grain products	5.93
Grain products	8.41	White cheese	8.62	Cons. fruits	7.71	Proc. red meat	5.84
Unproc potato	8.14	Unproc potato	8.07	Wine and liquor	7.47	Butter	5.69
Butter	8.07	Fre/froz fruit	8.04	Unproc potato	7.24	Cream products	5.56
Fre/froz fruit	7.81	Vegetable	7.93	Proc red meat	7.15	Fre/froz fruit	4.83
Lean fish	7.78	Cons. vegetables	7.86	White cheese	7.13	White cheese	4.78
Cream products	7.70	Cons fruits	7.856	Cream products	6.89	Cons. fruits	4.72
Vegetable	7.61	Butter	7.64	Sweetener	6.85	Cons vegetables	4.71
Sweetener	7.47	Fatty fish	7.54	Cons vegetables	6.82	Sweetener	4.68
Model 3 E		**Model 3.2 E**		**Model 3.3 E**		**Model 3.4 E**	
FIT value	16.96	Age	17.65	Whole grain	12.69	FIT value	14.69
Age	15.79	FIT value	17.13	FIT value	12.67	Whole grain	10.82
Whole grain	15.34	Whole grain	16.62	Dairy products	12.57	Dairy products	10.37
Vitamin D	15.22	Dairy products	16.52	Age	12.58	Calcium	9.84
Calcium	14.87	Vitamin D	15.63	Calcium	11.51	Alcohol	9.47
Fish	14.43	Fruits	15.43	Processed meat	11.43	Vitamin D	9.11
Fruits	14.39	Calcium	15.36	Red meat	11.40	Vegetables	8.97
Vegetables	14.34	Alcohol	15.31	BMI	11.39	Age	8.91
Red meat	14.32	BMI	15.24	Fruits	11.34	BMI	8.73
Alcohol	14.18	Red meat	14.69	Vitamin D	11.18	Red meat	8.64
Model 4 E		**Model 4.2 E**		**Model 4.3 E**		**Model 4.4 E**	
Fish	24.06	Dairy products	25.28	Dairy products	22.21	Dairy products	15.45
Whole grain	23.99	Whole grain	24.2	Whole grain	22.87	Calcium	14.13
Dairy products	23.62	Alcohol	23.60	Calcium	20.11	Whole grain	14.08
Fruits	23.01	Fruits	23.56	Vegetables	19.96	Alcohol	14.08
Alcohol	22.85	Vitamin D	23.13	Processed meat	19.79	Processed meat	13.59
Vitamin D	22.62	Fish	23.03	Red meat	19.77	Fruits	13.55
Calcium	22.36	Red meat	22.71	Fruits	19.69	Vegetables	13.31
Vegetables	22.21	Vegetables	22.40	Fiber	19.16	Vitamin D	13.29
Red meat	21.88	Calcium	22.20	Vitamin D	19.02	Red meat	13.02
Processed meat	21.76	Processed meat	21.94	Alcohol	18.75	Fish	12.70

Abbreviations: Cons vegetables, conserved vegetables; Unproc potato, unprocessed potatoes; Proc red meat, processed red met; Proc Meat, Processed meat; Milk & yoghurt, milk and yoghurt, Fre/ froz fruit, fresh and frozen fruits; FIT, Faecal Immunochemical Test; BMI, Body Mass Index. Variable importance of models created with energy adjusted Overall diet plus, Overall diet, CRC relevant dietary factors plus and CRC relevant dietary factors, with all the four cohorts as output. The ten most important variables are listed with their respective importance ranking, the most influential are at the top of every list.