UiT The Arctic University of Norway

Faculty of Science and Technology

Associations between dietary patterns and risk factors for cardiovascular disease

The seventh Tromsø study

—

Åse Mari Moe

A dissertation for the degree of Philosophiae Doctor – November 2024

Associations between dietary patterns and risk factors for cardiovascular disease

The seventh Tromsø study

Åse Mari Moe

Thesis submitted for the degree of Philosophiae Doctor
November 2024

Front cover: Illustrations by Sigrid Moe.

ISBN (printed): 978-82-8236-612-0
ISBN (electronic): 978-82-8236-613-7

Abstract

In this thesis, the association between dietary patterns and risk factors for cardiovascular diseases was investigated using data from the seventh Tromsø Study (Tromsø7). A total of 21,083 individuals aged 40 years and older participated in Tromsø7. Approximately 10,000 of these participants completed 90% or more of the food frequency questionnaire. Their responses formed the foundation of our analysis.

Dietary patterns were identified using hierarchal clustering of food variables, factor analysis, the treelet transform and exploratory structural equation models. All four methods identified a dietary pattern high in sweets and meat. This dietary pattern was more commonly consumed by the younger participants. Participants with a high score on this dietary pattern usually had lower physical activity. Additionally, a health-conscious pattern was identified by all methods. A high intake of this pattern was associated with higher physical activity and higher education.

The association between dietary patterns and metabolic syndrome, including its defining components, was modelled using logistic regression. The components of metabolic syndrome include elevated waist circumference, low HDL-cholesterol, elevated triglycerides, insulin resistance and hypertension. Models with and without dietary patterns were compared to investigate how dietary patterns affected the prediction of metabolic syndrome and its components. Including dietary patterns in the models resulted in a clear improvement in prediction for metabolic syndrome and elevated waist circumference. The dietary pattern high in sweets and meat was significantly associated with both metabolic syndrome and elevated waist circumference. Among men, the health-conscious pattern and the traditional pattern were associated with lower level of triglycerides.

An integrated analysis of food variables and risk factors for cardiovascular disease was also conducted using exploratory structural equation models. Our model included analysis of dietary patterns and mediation analysis, using obesity as a mediator between dietary patterns and the other risk factors for cardiovascular disease. Similar to using the logistic regression models, a positive association between obesity and the dietary pattern high in sweets and meat was observed. Additionally, a pattern high in processed dinner variables was associated with obesity, whereas a cake pattern and a porridge pattern were negatively associated with obesity. A direct effect of dietary patterns on obesity resulted in an indirect effect on the other risk factors for cardiovascular disease. After adjusting for the indirect effect of obesity, only direct effects of diet on triglycerides and HDL-cholesterol were observed.

Obesity is the fastest-growing risk factor for cardiovascular disease in both Norway and globally. It indirectly affects both health and quality of

life. Reversing the obesity trend is therefore important for public health. In particular, our analysis found consistent associations between obesity and patterns high in sweets, meat and processed dinner.

Sammendrag

Denne avhandlingen undersøker sammenhengen mellom kosthold og forekomsten av risikofaktorer for hjerte- og karsykdommer i data fra den sjuende Tromsøundersøkelsen (Tromsø7). Totalt deltok 21 083 personer i aldersgruppen 40 år og eldre i Tromsø7. Omtrent 10 000 av disse hadde svart på 90% eller mer på spørreskjemaet om kosthold. Disse svarene er grunnlag for våre analyser.

Vi brukte hierarkisk gruppering av matvariablene, faktoranalyse, treelettransformasjon og utforskende strukturelle ligningsmodeller til å finne kostholdsmønstre. Alle de fire metodene fant et kostholdsmønster med mye søtsaker og kjøtt. Dette kostholdet var vanligere blant de yngre deltagerne. Deltagere med høyt inntak av dette kostholdet var vanligvis mindre fysisk aktive. Et helsebevisst kostholdsmønster ble også funnet med alle metodene. Høyt inntak av dette kostholdet var vanligere for personer med høyere fysisk aktivitet og høyere utdanning.

Sammenhengen mellom kosthold og metabolsk syndrom, inkludert komponentene i definisjonen, ble undersøkt med logistisk regresjon. Komponentene i metabolsk syndrom er økt midjeomkrets, lavt HDL-kolesterolnivå, høyt triglyseridnivå, insulinresistens og høyt blodtrykk. Modeller med og uten kostholdsmønster ble sammenlignet for å undersøke hvor godt mønstrene predikerte metabolsk syndrom og de enkelte komponentene i definisjonen hos deltagerne. Når kostholdsmønsterene ble inkludert i modellene observerte vi at metabolsk syndrom og økt midjeomkrets ble merkbart bedre predikert. Kostholdsmønsteret med mye søtsaker og kjøtt var signifikant assosiert med metabolsk syndrom og økt midjeomkrets. Hos menn var et helsebevisst kostholdsmønster og et mer tradisjonelt kostholdsmønster assosiert med lavere triglyseridnivåer.

Vi brukte også utforskende strukturelle ligningsmodeller for å lage en helhetlig analyse av kostholdsmønster og kjente risikofaktorer for hjerte- og karsykdommer. Modellen hadde med analyse av kostholdsmønstre og en medieringsanalyse, hvor overvekt ble brukt som en mediator mellom kosthold og de andre risikofaktorene for hjerte- og karsykdom. I likhet med de logistiske regresjonsmodellene fant vi at et kosthold med mye søtsaker og kjøtt var signifikant assosiert med overvekt. I tillegg var kosthold som inneholdt mye prosessert middag assosiert med overvekt, mens kosthold med mye kake og grøt var assosiert med mindre overvekt. Når vi observerte en direkte effekt av kosthold på overvekt, observerte vi også at dette indirekte påvirket de andre risikofaktorene. Når overvekt var justert for, var det bare HDL-kolesterol og triglyserider som ble direkte påvirket av kosthold.

Overvekt er den risikofaktoren for hjerte- og karsykdommer som øker mest i både Norge og verden ellers, og som indirekte påvirker både helse og livskvalitet. Å snu denne trenden er viktig for folkehelsen. Våre analyser

viser at det er en entydig sammenheng mellom overvekt og kosthold med søtsaker, kjøtt og prosessert middag.

Takksigelser

Først og fremst vil jeg takke mine to nærmeste veiledere, Elinor og Sigrunn. Sammen har dere gitt meg utfyllende veiledning og god støtte i både store og små saker. Det har aldri vært et problem å komme til kontorene deres for å få svar på mine spørsmål. Veiledningen jeg har fått av dere har vært uvurderlig. Jeg ønsker også å takke Laila og Ola for veiledningen dere har gitt meg.

Jeg vil også takke Linn, Margaretha, Natalia, Rebekka og familien min for gode og interessante diskusjoner. Det har vært med på å gi arbeidet mitt en ekstra dybde. Jeg ønsker også å takke Sigrid som har laget illustrasjonene som er på forsiden av denne avhandlingen.

Til slutt ønsker jeg å takke Johann, mannen min, for støtten du har gitt meg i arbeidet mitt, spesielt innspurten. Og Robert, min lille sønn, for å gjøre hverdagen lysere.

Åse Mari Moe

Tromsø 2024

Contents

List of publications

Paper I
Moe, Å.M., Sørbye, S.H., Hopstock, L.A., Carlsen, M.H., Løvsletten, O. and Ytterstad, E. (2022), **Identifying dietary patterns across age, educational level and physical activity level in a cross-sectional study: the Tromsø Study 2015 - 2016**, *BMC Nutrition*, **8**, 102, https://doi.org/10.1186/s40795-022-00599-4.

Paper II
Moe, Å.M., Ytterstad, E., Hopstock, L.A., Løvsletten, O., Carlsen, M.H. and Sørbye, S.H. (2024), **Associations and predictive power of dietary patterns on metabolic syndrome and its components**, *Nutrition, Metabolism and Cardiovascular Diseases*, **34**, 3, 681-690, https://doi.org/10.1016/j.numecd.2023.10.029.

Paper III
Moe, Å.M. and Sørbye, S.H., **Analysis of dietary pattern effects on metabolic risk factors using structural equation modelling**, submitted.

CHAPTER 1

Introduction

Cardiovascular disease (CVD) and the rapid increase in obesity over the last decades represent a significant health burden worldwide. CVD is one of the main causes of disability in the world, and as obesity continues to rise, the burden of CVD is likely to increase (Haththotuwa et al., 2020).

Lifestyle is one of the most important risk factors for CVD and obesity. It encompasses modifiable factors such as unhealthy diets, physical inactivity, smoking and high alcohol consumption (Yusuf et al., 2004; O'Donnell et al., 2016). Among these, dietary habits are one of the most complex and challenging factors to collect and analyse effectively (Martínez et al., 1998). Diets are intricate, consisting of a wide variety of foods, dishes and beverages that are prepared and consumed with high daily variation (Satija et al., 2015; Jacobs, 2023). In this context, dietary pattern analysis is a promising method to understand the complex nature of dietary behaviour (Zhao et al., 2021).

This thesis examines dietary data collected using a food frequency questionnaire (FFQ) and employs several data reduction methods to analyse dietary patterns and their associations with risk factors for CVD. Figure 1.1 gives an overview of the methods and models used. A detailed presentation of the methods is provided in chapter 3. The food variables were either used to identify dietary patterns or used directly in the random forest algorithm. The analysis includes estimates of associations between dietary patterns and risk factors for CVD, using both standard logistic regression models and exploratory structural equation models (ESEM). ESEM is a method that incorporates causal assumptions and can be used for mediation analysis. In nutrition research, the method can be used to simultaneously identify dietary patterns and estimate the hypothesized diet-disease relationships. Additionally, the assessment of predictive power was included to evaluate model fit. This also includes comparison of the regression models with the random forest algorithm.

1.1 Metabolic syndrome

One of the condition investigated in this thesis is the Metabolic Syndrome (MetS). MetS is a condition where several risk factors for CVD are present

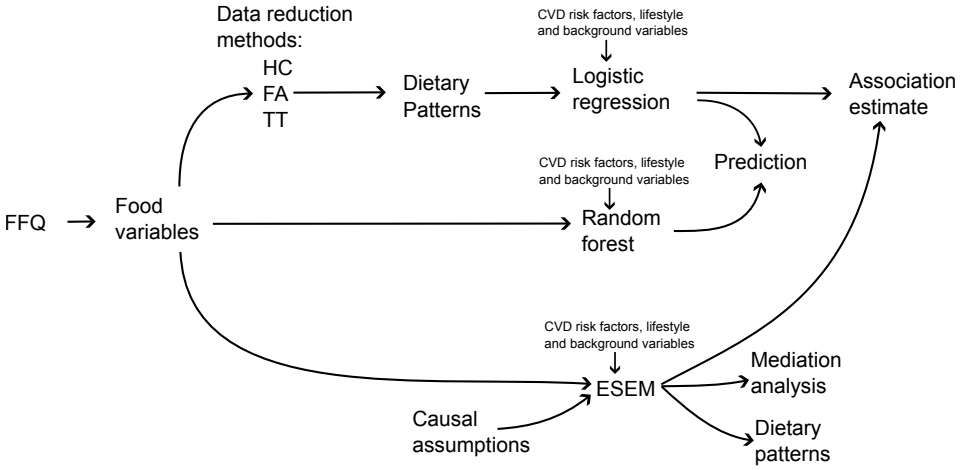


FIGURE 1.1. Overview of the methods and the analysis included in this thesis.
FFQ: food frequency questionnaire, HC: hierarchical clustering of food variables, FA: factor analysis, TT: treelet transform, CVD: cardiovascular disease, ESEM: exploratory structural equation model

in an individual. MetS has been found to double the risk of cardiovascular outcomes (Mottillo et al., 2010). The prevalence of MetS among adults is high in several populations. For example, in the United States, Iran, Brazil, Finland and Algeria, over-all prevalence indicate that every third adult live with MetS (Liang et al., 2021; Zabetian et al., 2007; de Siqueira Valadares et al., 2022; Haverinen et al., 2021; Ngwasiri et al., 2023). The prevalence of MetS typically increases with age (Kassi et al., 2011), and heterogeneity within regions and sex differences have been observed (de Siqueira Valadares et al., 2022).

The definition of MetS has evolved since the first attempt made by WHO in 1998, and today several definitions exist (Kassi et al., 2011). These definitions typically vary in the risk factors they include, the emphasized risk factor (if any), and the thresholds used for classifying individuals at elevated risk. There are also attempts to create a continuous measure of MetS based on cardiovascular risk factors, such as using principal component analysis (PCA) (Agarwal et al., 2012). Today there is a broad consensus on the types of measures that should be included, with most definitions incorporating measures of overweight, insulin resistance, hypertension and unfavourable lipid profiles (Grundy et al., 2005; Kassi et al., 2011). However, other measures closely related to MetS like pro-inflammation state or dysfunction

of high-density lipoprotein particles are not included in the definitions (Kassi et al., 2011; Onat and Hergenç, 2011).

One of the most used definitions is the one provided by the National Cholesterol Education Program Adult Treatment Panel III (ATPIII). This definition also includes a harmonised version where abdominal obesity is defined by population-specific thresholds (Kassi et al., 2011). The ATPIII includes five risk factors. Individuals are classified as having MetS if they have elevated risk of at least three of these risk factors. Other widely used definitions include those from the International Diabetes Federation (IDF) and the Joint Interim Statement (JIS). Similar prevalence of MetS has been observed using all of these definitions (Kassi et al., 2011; Zabetian et al., 2007; Haverinen et al., 2021).

There is some criticism of the definitions of MetS, particularly regarding the use of dichotomous risk factors, which can result in loss of information (Feldman et al., 2015). Additionally, there is uncertainty about whether the association between MetS and CVD risk provides additional insight beyond the sum of the individual components in the definition of MetS (Feldman et al., 2015). Therefore, the individual components of the MetS definition are also of interest and are examined in this thesis.

1.2 Aim of the thesis

The aim of this thesis is to derive and analyse dietary patterns using various statistical methods and to associate these with risk factors of CVD.

Specifically, the aim was divided into:

- Compare data-driven methods that result in both overlapping and non-overlapping dietary patterns (paper 1 and 2).
- Investigate the associations between dietary patterns and risk factors for CVD, such as MetS (paper 2 and 3).
- Use exploratory structural equation modelling to investigate the role of obesity as a potential mediator between diet and metabolic risk factors for CVD (paper 3).

1.3 Outline

In chapter 2, the data material is presented. This includes an explanation of the pre-processing of the FFQ responses and an overview of the study sample.

Chapter 3 first presents the methods used to identify data-driven dietary patterns, including examples from the Tromsø7 data. This gives a simplified overview, that can be useful for researchers wanting to analyse dietary patterns using data-driven methods. Then the background for analysing the association between dietary patterns and cardiovascular risk factors is explained, along with model evaluation by predictive power as investigated in paper 2. Next, causal assumptions are discussed, with a special focus on the role of body mass index (BMI), which introduces the ESEM model used in paper 3.

Chapter 4 gives a summary of each of the three papers included in this thesis. This chapter also includes tables of the estimated total effect of dietary patterns on CVD risk factors that was estimated in both paper 2 and 3.

Chapter 5 provides a discussion of the results and draws connections between the different methods. Strengths and limitations are also discussed.

CHAPTER 2

Data material

2.1 The Tromsø study

The Tromsø studies include seven population-based cross-sectional studies conducted between 1974 and 2016. The eight Tromsø study is currently being planned and is scheduled to take place between 2025 and 2026. The first study included only men, but both the study population and scope have expanded over the years.

In Tromsø7, all inhabitants in the municipality of Tromsø aged 40 years and above were invited. Biological samples, measurements and clinical examinations were conducted. The screening also included extensive questionnaires, and for the first time a detailed FFQ, which had been validated by Carlsen et al., 2010 and Carlsen et al., 2011. A detailed description of the measurements in the study can be found in Hopstock et al., 2022. The study had an attendance rate of 65%, including 21,083 participants.

FIGURE 2.1. Map of northern Europe, ©OpenStreetMap

2.2 FFQ data

The inclusion of an FFQ, offers a unique opportunity to investigate dietary patterns in the municipality of Tromsø, which represents a general Nordic population. The FFQ is one of the standard methods to measure dietary habits, alongside 24-hour recalls and food records. Among these methods, the FFQ is often preferred for large population studies due to its low cost and ease of distribution (Satija et al., 2015; Bailey, 2021). An FFQ is designed for a specific population and can vary in detail, ranging from around 10 to over 200 questions. It is used to assess the average intake of foods, beverages and dishes over a specified time period (Jacobs, 2023).

The FFQ used in Tromsø7 was paper-based and included 261 questions about dietary intake and supplements over the past year. Trained technicians reviewed the FFQs before scanning. The answers were further processed to calculate intake of grams per day (g/day) using the KBS AE14 food database and KBS software system at University of Oslo (KBS, version 7.3.). Additionally, the total energy intake (TEI, kilojoule/day) and the total water intake (TWI, g/day) were calculated. Missing values were assigned an intake of 0 g/day, like answers of "never/seldom" in the questionnaire.

The responses from the questionnaire were further aggregated into a smaller number of food variables, primarily based on the main food groups in the questionnaire. Paper 1 includes a complete list of items within each of the 33 food variables used in both paper 1 and paper 2. In paper 3, the "Meat Dinner", "Composite Dinner Dishes" and the "Fish Dinner" were separated into 3, 3 and 2 new variables, respectively. Paper 3 includes the list of items for the new food variables. No analysis in this thesis included data of supplement intake.

Energy adjustments are often used in diet-disease studies to account for differences in consumption due to varying energy needs and the potential relation between energy intake and disease (McCullough and Byrd, 2022; Willett et al., 1997). Energy adjustments have also been found to reduce error (Thompson et al., 2015). In this thesis, the food intake was adjusted by dividing it by TEI and then multiplying it by the population's mean energy intake, as shown in this equation:

$$\text{Food}^*_{ji} = \text{Food}_{ji}\frac{\overline{\text{TEI}}}{\text{TEI}_i}. \tag{2.1}$$

Food_{ji} represents the unadjusted food variable j for participant i, Food^*_{ji} is the adjusted food variable, $\overline{\text{TEI}}$ is the mean TEI for the entire population, and TEI_i is the TEI for participant i. This scaling adjusts the intake of each individual, representing their consumption as if everyone had TEI equal to

the population's mean intake, $\overline{\text{TEI}}$.

This approach was chosen to focus on the composition of food rather than the specific quantity consumed. Energy adjustment usually has a small impact on the resulting dietary patterns. For instance, the correlation between diet scores using either adjusted or unadjusted food variables is usually high (Edefonti et al., 2020). The importance of food variables is weighted differently based on dietary patterns. This weighting can be used to calculate a diet score for each individual.

2.3 CVD risk factors

In paper 2, the definition of MetS was based on the ATP III from 2001. However, the criterion for insulin resistance was modified from the original ATP III definition due to the use of non-fasting blood samples. Participants were classified as having MetS if they had three or more of the following risk factors:

- Insulin resistance: Self-reported diabetes and/or glycated haemoglobin (HbA1c) $\geq$6.1 %
- Low HDL-cholesterol level: HDL-C $\leq$1.0 mmol/L for men and HDL-C $\leq$1.3 mmol/L for women
- Elevated triglycerides: TG $\geq$1.7 mmol/L
- Hypertension: Blood pressure $\geq$130/85 mmHg and/or self-reported use of antihypertensive drugs
- Elevated waist circumference: WC $\geq$102 cm for men and WC $\geq$88 cm for women

In paper 3, the CVD risk factors were analysed as continuous measures. Additionally, C-reactive protein (CRP) was included as an extra risk factor. Given the highly skewed distribution of triglycerides, HDL-cholesterol and CRP, theses variables were log-transformed to normalize the data.

HbA1c was measured by high-performance liquid chromatography method using the Tosoh G8 instrument. For measurements of HDL-cholesterol and triglycerides, the enzymatic colorimetric method was used on the Cobas 8000 instrument. CRP was also analysed using the Cobas 8000 instrument by the immunturbidimetric method. Blood pressure was measured three times, and the average of the last two measurements was used in the analyses. Measurements were taken with the Dinamap ProCare 300 instrument. Waist circumference was measured at the umbilical level using a Seca measuring tape.

2.4 Lifestyle and background variables

In addition to dietary data, this thesis also included measures of the lifestyle factors physical activity level, smoking habits and alcohol consumption.

Physical activity level was collected using the Saltin-Grimby's activity level scale (Grimby et al., 2015). In Tromsø7, the questionnaire consisted of four categories: sedentary (mainly reading and watching TV), light (activities like walking or cycling at least 4 hours a week), moderate (vigorous sports and similar at least 4 hour a week) and vigorous activity (hard training multiple times during the week). As the vigorous activity group was small, it was decided to merge the moderate and vigorous activity group in all analysis. Daily smoking habits were answered using a question with the alternatives "never", "yes, now" and "yes, previously". In this thesis, the answers of "never" and "yes, previously" were aggregated into a group "non-smoker". Alcohol consumption was reported as part of the FFQ. We used an aggregated variable consisting of different types of alcohol consumption, such as beer and wine. Intake was represented as dl/day, using the simplified assumption that all beverage types have the same specific mass as water.

2.5 Study sample

The questionnaire was answered by 15,146 participants. Less than 90% of the questions were answered by 3,489 participants, and these participants were excluded from our analysis. As proposed in Carlsen et al., 2010, participants with extreme energy intake were also excluded from our analysis. We assessed extreme energy intake as the 1% with highest and the 1% with the lowest energy intake, adjusted for age, height, sex and physical activity level. Also, participants with extreme water intake were excluded using the same method. Figure 2.2 gives a detailed overview of the inclusion and exclusion criteria for the three papers included in this thesis. After the publication of paper 1, 13 participants in Tromsø7 had withdrawn their consent to participate in research.

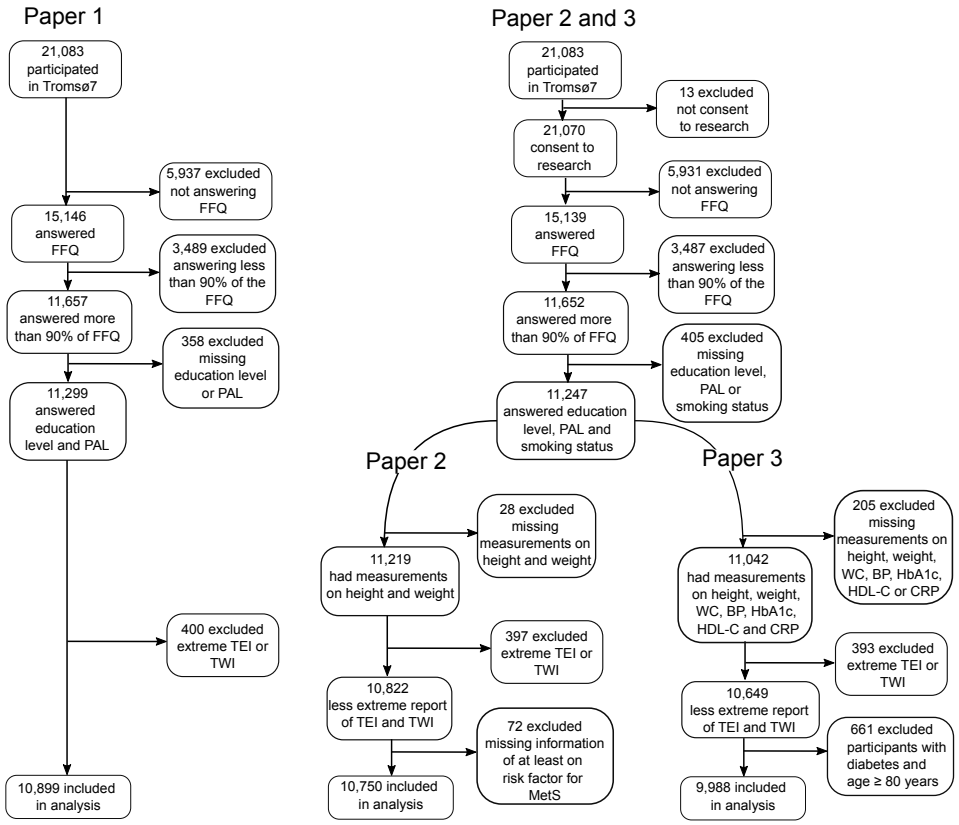


FIGURE 2.2. Flowchart of the study sample in paper 1 - 3. Tromsø7: The seventh Tromsø study, FFQ: Food frequency questionnaires, PAL: physical activity level, TEI: total energy intake, TWI: total water intake, MetS: Metabolic syndrome, WC: waist circumference, BP: blood pressure, HbA1c: hemoglobin A1c, HDL: HDL-cholesterol, CRP: C-reactive protein

CHAPTER 3

Methods and additional analysis

This chapter presents the data-driven methods used in this thesis to identify dietary patterns. These patterns were further included in a logistic regression model to investigate diet-disease associations. The chapter also presents the method used to assess the ability of dietary patterns to predict diet-disease outcomes. The role of BMI in this analysis is also discussed. This introduces the use of causal diagrams to explore causal assumptions and prior knowledge of causal relationships. ESEM, a method developed for causal inference, is introduced as an alternative to the regression models for investigating the diet-disease relationship.

3.1 Dietary patterns

The collected food intake data can be analysed in several ways, such as investigating nutrition intake and individual food variables. Since the beginning of the 21st century, research has shifted from mainly focusing on specific foods or nutrients to using dietary patterns that assess the entire diet (Hu, 2002). The idea behind this approach is that dietary patterns emphasize the complex interaction between nutrients and eating behaviours, providing a more extensive picture of food consumption (Cespedes and Hu, 2015). Dietary patterns also highlight that a healthy diet can be achieved through different choices and that dietary eating patterns can vary between countries and population groups. Several methods have been proposed to identify dietary patterns, including both data-driven approaches and evidence-based indices (Zhao et al., 2021).

In this thesis, three different data-driven methods where primarily investigated to identify dietary patterns using data from Tromsø7. These methods use the collected food intake to uncover dietary trends within the reported consumption. The resulting dietary patterns results in an overall summary of food consumption's in the specific population (Zhao et al., 2021; Cespedes and Hu, 2015).

In detail, the methods investigated include hierarchical clustering (HC) of food variables, factor analysis (FA) and treelet transform (TT). The main differences between these methods lie in how the food variables load on patterns. With HC, each food variable is assigned to only one pattern.

In contrast, FA and TT result in dietary patterns where food variables load into multiple patterns. However, TT allows for some of these loadings to be exactly zero, while FA does not. The number of dietary patterns a food variable loads into influences both the interpretation and the amount of information retained. The complexity of patterns identified by different methods have also been observed to affect their usefulness in analysing the association between diet and various health outcomes (Schoenaker et al., 2013).

HC, TT and FA all use the correlation between food variables to identify dietary patterns. A high positive correlation close to one between two food variables indicates that people who consume little or much of one of the foods similarly consume little or much of the other food variable. This suggests that the variables are consumed in a similar manner and are closely related.

3.1.1 Hierarchical clustering of food variables

Non-overlapping dietary patterns can be identified using HC of food variables. HC of food variables use the well-known hierarchical clustering algorithm, with the correlation matrix as the similarity measure between food variables. The clustering process begins by identifying the two most correlated variables and grouping them together. In the next step, the difference between the new cluster and the remaining variables or clusters is recalculated using a linkage method. The most common linkage methods are single, complete and average linkage, which correspond to the minimum, maximum or average distance between two clusters, respectively. This procedure is repeated, until all variables are clustered together.

It is important to note that this method differs from clustering participants, which is the most common hierarchical clustering method used to find dietary patterns in nutrition analysis (Gorst-Rasmussen et al., 2011). Clustering of participants usually base the similarity measure on the Euclidean distance between two individuals and not the correlation. Although HC of food variables is a well-established method, it has rarely been used to cluster food variables into dietary patterns.

The clustering procedure can be visualized using a dendrogram, such as the simplified version shown in figure 3.1. A dendrogram can also illustrate the similarity between variables or clusters, with lower height indicating greater similarity. To obtain the final clusters, the dendrogram is usually "cut" at a specific height. Fewer branches, and hence fewer clusters, are obtained by "cutting" the dendrogram closer to the top. All variables in the branch below a "cut" belong to the same cluster.

Recomputing the clustering procedure can be performed to assess the stability of the clusters. A stability matrix can be computed to determine

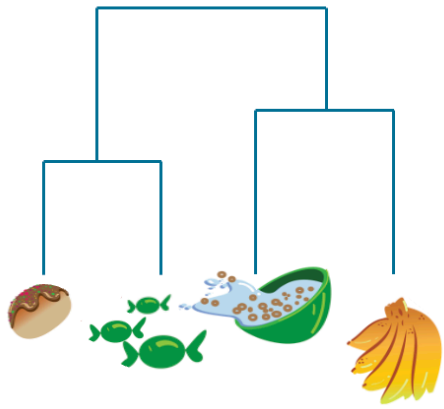
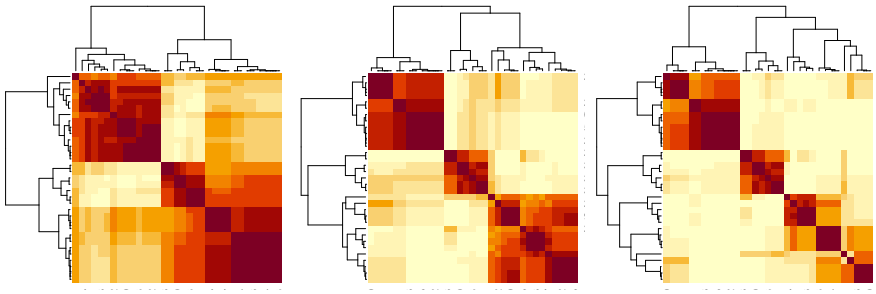


FIGURE 3.1. Illustration of a simplified dendrogram

how frequently pairs of variables are clustered together using a fixed number of clusters. The HC of food variables can be recomputed using methods such as bootstrapping. In bootstrapping, a bootstrap sample is drawn from the full sample with replacement, where the size of the bootstrap sample matches the original sample size. Figure 3.2 shows heatmaps of the stability matrix, varying the number of clusters among women in Tromsø7 using the dietary data and bootstrapping. This dataset is the same as the one used in paper 1. The food variables are sorted based on the dendrogram of the heatmaps, and may differ across the three figures.



(A) Two groups (B) Three groups (C) Four groups

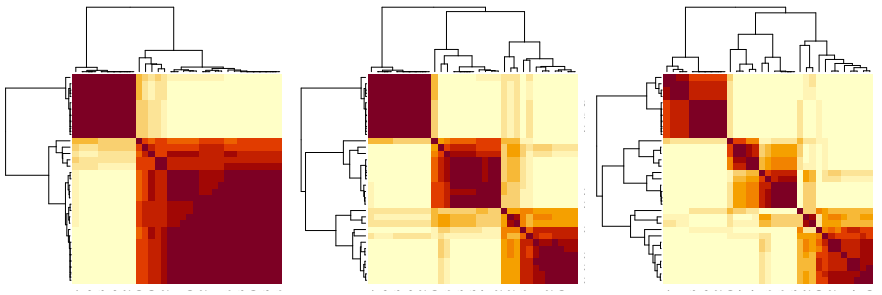
FIGURE 3.2. Stability matrices showing how often variables were clustered together using HC. The stability matrices were calculated based on a total of 100 bootstrap samples.

Using bootstrapping with two clusters results in one cluster containing sweets, snacks, meat, vegetables and fruit and a second cluster containing

bread, spreads and fish. Using three clusters resulted in one cluster for sweets, snacks and meat, another for vegetables and fruit and a third for bread, spreads and fish. Using four clusters, the bread and spread variables were separated from the fish and potato variables. The remaining clusters were similar to those found using three clusters.

In paper 1, the study sample was grouped into smaller cohorts based on similar background variables before calculation of the correlation matrix. It is well known that dietary intake data is influenced by various sources of errors, and it is recommended to analyse FFQ data at the group level (Slimani et al., 2015). We therefore used cohorts based on the background variables such as age, education level and physical activity level to reduce random errors.

According to the central limit theorem, using cohorts can make the distribution of the food variables more normal and reduce variance. Additionally, this approach makes differences in the background variables appear more clearly, as it minimizes noise and other sources of variation. Comparing the stability matrix found by cohorts and bootstrapping, we observed that cohorts gives slightly more stable patterns, see figure 3.2 and 3.3. This is most evident in the stability matrix using two groups, where few variables cluster together with variables from both groups. This indicate that the use of cohorts resulted in more stable results.



(A) Two groups (B) Three groups (C) Four groups

FIGURE 3.3. Stability matrices showing how often variables were clustered together using HC. The stability matrices were calculated based on 100 different ways to make the cohorts.

Using cohorts with two groups resulted in one cluster containing sweets, snacks and meat and a second cluster with the remaining variables. Further, using three groups, food variables like vegetables, fruits, cheese and yoghurt were separated from the bread, fish, potato and dessert variables. The sweets, snacks and meat remained as a separate cluster. Using four groups,

bread, fish and potato were separated from the dessert variables. The remaining clusters were similar to those found with three groups. The dietary patterns found by FA and TT were more similar to the patterns identified using the bootstrap method than to those identified using cohorts.

Alternatively, the cohorts can be grouped based on other variables, such as health factors. This approach could reveal patterns more strongly associated with health outcomes. This approach is similar to how the reduced rank regression (RRR) method has been used in nutrition research to identify dietary patterns based on prior knowledge of disease risk (Weikert and Schulze, 2016). These methods heavily depend on the choice of response variables. In nutrition research using RRR, dietary patterns are commonly identified using nutrient intake or biomarkers (Weikert and Schulze, 2016). RRR was not included in the comparison of methods, as it is not a pure data-driven method. Whether grouping cohorts in other ways would result in stable patterns that are useful for understanding the diet-disease relationship remains to be investigated.

In figures 3.2 and 3.3, the dendrograms displayed with the stability matrix can be used to identify the most stable groups. Based on this, a stability score can be calculated by comparing the difference within and between groups. Based on the score in Bellec et al., 2010 and Rousseeuw, 1987, we calculated a stability score ranging from zero to one, where a score of one indicates that clusters have high internal stability and can be easily separated.

The stability matrix is first divided by the total number of recomputations of the clustering procedure to get a ratio of how often two variables are clustered together. For each variable, the mean stability of belonging to the most stable groups is subtracted by the mean stability of belonging to the second most stable group. The mean stability measure of all variables was used as the stability score.

The stability score using both bootstrapping and cohorts is displayed in figure 3.4. In comparison to the bootstrap analysis, we found that using cohorts resulted in higher stability scores for fewer than five groups. The difference between the methods decreased as the number of groups increased.

As an alternative, individuals can also be clustered. This was investigated in preliminary analyses. This divided the population into groups that were closely related to demographic variables. This was not surprising, as the dietary patterns were also associated with demographic factors. To simplify the analysis, only the clustering of food variables was included in the final analysis.

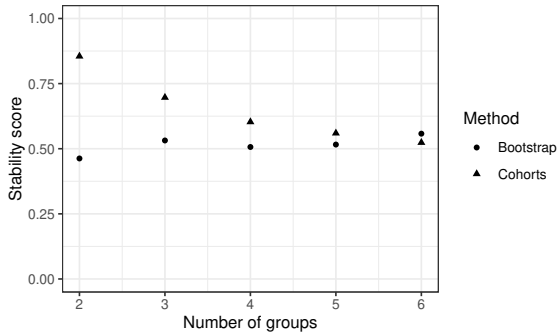


FIGURE 3.4. The stability score for two to six groups, using both bootstrapping and cohorts

3.1.2 Principal component analysis and factor analysis

PCA is a well-known data reduction method that aims to explain the structure of the covariance matrix of $\mathbf{X}$ by a reduced number of linear combinations of the original variables in $\mathbf{X}$ (Johnson and Wichern, 2014b). This linear transformation of p original variables, $\mathbf{X}$, can be expressed as:

$$\underset{(p\times 1)}{\boldsymbol{Y}} = \underset{(p\times p)}{\boldsymbol{A}^T}\ \underset{(p\times 1)}{\mathbf{X}} \tag{3.1}$$

The goal of PCA is to obtain transformations that result in uncorrelated $\boldsymbol{Y}_i$'s, which explain most of the variance in a dataset. This is achieved by using the eigenvectors of the covariance/correlation matrix of $\mathbf{X}$. For data reduction, a selection of the m eigenvectors with the highest eigenvalues is made, where $m < p$.

The method begins by computing the eigenvectors and corresponding eigenvalues of the estimated covariance/correlation matrix. The eigenvectors with the highest corresponding eigenvalues are then used to transform the data into a new set of orthogonal variables. The transformed data, known as principal component scores, can further be used as diet scores in the analysis of food patterns. Additionally, the results of PCA can be interpret by examining the eigenvectors (Johnson and Wichern, 2014b).

One way to determine the number of principal components to retain in PCA is by examine the scree plot (Johnson and Wichern, 2014b). The scree plot displays the eigenvalues of the estimated covariance/correlation matrix, which correspond to the sample variance of the principal components score. The scree plot helps identify the "elbow" point, where the eigenvalue starts to level off. At this point, additional components show a decrease in explained variation compared to the earlier ones. This indicate the appropriate number of components to retain. Figure 3.5 shows the scree plot for data from Tromsø7, using the same sample and food variables as analysed

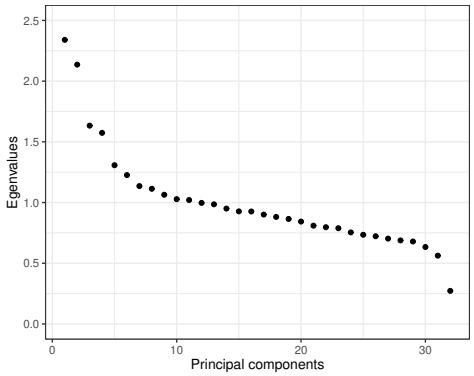


FIGURE 3.5. Scree plot using PCA on the dietary data in Tromsø7. An "elbow" can be observed around the first principal component.

in paper 2. In this plot, an "elbow" is observed around the fifth principal component.

From one sub-sample to another, the direction of an eigenvector can be reversed, meaning all its signs are opposite. Despite this change in direction, the eigenvector represents the same transformation. To ensure consistency in the direction of the eigenvectors across different sub-samples, the resulting eigenvector can be multiplied by -1 when necessary.

FA is a method that aims to capture the underlying structures in the covariance/correlation matrix using a few unobserved factors. The underlying structure does not necessarily lead to a transformation that explains the most of the variance in the original dataset, as is the case with PCA. The assumed underlying model for data in FA is (Johnson and Wichern, 2014a):

$$\underset{(p\times 1)}{\mathbf{X}} - \underset{(p\times 1)}{\boldsymbol{\mu}} = \underset{(p\times m)}{\mathbf{L}}\underset{(m\times 1)}{\mathbf{F}} + \underset{(p\times 1)}{\boldsymbol{\epsilon}} \tag{3.2}$$

The model also assumes that $\boldsymbol{\epsilon}$ and $\mathbf{F}$ are independent random vector, and that the expectation, and covariance of $\mathbf{F}$ and $\boldsymbol{\epsilon}$ are:

$$\begin{aligned}
\mathrm{E}[\mathbf{F}] &= \underset{(m\times 1)}{\mathbf{0}} \\
\mathrm{E}[\boldsymbol{\epsilon}] &= \underset{(p\times 1)}{\mathbf{0}} \\
\mathrm{Cov}[\mathbf{F}] &= \underset{(m\times m)}{\mathbf{I}} \\
\mathrm{Cov}[\boldsymbol{\epsilon}] &= \underset{(p\times p)}{\boldsymbol{\Psi}}
\end{aligned}$$

$\mathbf{X}$ is an observed random vector of p variables. The mean $\mathrm{E}[\mathbf{X}] = \boldsymbol{\mu}$, and the covariance matrix $\mathrm{Cov}[\mathbf{X}] = \boldsymbol{\Sigma}$. For standardised data, $\boldsymbol{\mu} = \mathbf{0}$, and the covariance matrix becomes equal to the correlation matrix. According to the model, the covariance matrix $\boldsymbol{\Sigma}$ can be explained by m underlying unobserved random factors $\mathbf{F}$, where $\mathbf{L}$ is the loading matrix. $\boldsymbol{\epsilon}$ is the vector of errors that accounts for additional source of variance, $\mathbf{I}$ is the identity matrix and $\boldsymbol{\Psi}$ is a diagonal matrix.

The model attempts to approximate the covariance matrix $\boldsymbol{\Sigma}$ as a combination of the explained covariances accounted for by the model ($\mathrm{Cov}[\mathbf{LF}]$) and the variance in the error ($\mathrm{Cov}[\boldsymbol{\epsilon}]$), as shown in:

$$\mathrm{Cov}[\mathbf{X}] = \boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi}$$

The loading matrix $\mathbf{L}$ is related to the factors and observed data by:

$$\mathrm{Cov}[\mathbf{X}, \mathbf{F}] = \mathbf{L}$$

The FA model is commonly estimated using either the principal component method or the maximum likelihood method (Johnson and Wichern, 2014a). First the covariance/correlation matrix is estimated. The principal component method starts like PCA, with computing pairs of eigenvalues $\hat{\lambda}$ and eigenvectors $\hat{\mathbf{e}}$ of the estimated covariance/correlation matrix. The estimated loading matrix is given by:

$$\hat{\mathbf{L}} = \left[\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1, \quad \sqrt{\hat{\lambda}_2}\hat{\mathbf{e}}_2, \quad \ldots, \quad \sqrt{\hat{\lambda}_m}\hat{\mathbf{e}}_m\right]$$

By including the assumption of normality of the factors $\mathbf{F}$ and the errors $\boldsymbol{\epsilon}$, the model can be estimated using the maximum likelihood method. The estimates of $\hat{\mathbf{L}}$ and $\hat{\boldsymbol{\Psi}}$ can be found numerically, for instance using the factanal() function in R.

The factors estimated from FA are interpreted as the underlying dietary patterns. Then the loading matrix is the covariance/correlations between the original observed food variables and the dietary patterns. The interpretation of the patterns can be hard. If more than one factor is estimated, the factor loadings can be rotated to obtain patterns with a simpler structure. This rotation does not affect the ability to reproduce the covariance matrix. The loading matrix still represents covariance/correlations, but for the rotated factors. A common rotation method used is the varimax rotation. When interpreting the dietary patterns, the variables with the highest absolute values in the loading matrix are emphasized. Variables with negative loading represent contrasts to the dietary patterns.

Similar to PCA, the scree plot can be explored to determine the number of factors to use in FA. The eigenvalues in FA do not directly correspond to the sample variance accounted for by the factors, but they still gives useful guidance in selecting the appropriate number of factors. In FA, using the

maximum likelihood method or rotating the factors result in lower explained variance. Based on the scree plot in figure 3.5, four patterns were identified in paper 2 using FA with the principal component method.

PCA and FA are the most common unsupervised methods in the literature for deriving dietary patterns (Zhao et al., 2021). However, it is rather common to refer to FA as PCA in the literature of dietary patterns, see for instance Hearty and Gibney, 2008; Varraso et al., 2012. Both methods aim to approximate the covariance or correlation matrix, but while FA assumes a statistical model, PCA does not (Joliffe and Morgan, 1992). Standardized data and the correlation matrix is usually used for food data, as it captures the relative consumption rather than the weight of the food items.

3.1.3 The treelet transform

TT is a data reduction method that combines concepts from HC and PCA. The method aims to identify the underlying structure within the data (Lee et al., 2008). This method allows for some loadings to be exactly zero, resulting in simpler structures compared to FA (Gorst-Rasmussen et al., 2011). A brief explanation of the TT methods, presented in full in Lee et al., 2008, is provided in the following paragraphs.

The first steps of TT shares many similarities with the first steps in HC. In HC, the two variables with the highest correlation are clustered, and their updated distances to the other variables are calculated using a linkage method. In contrast, TT transforms the two most highly correlated variables into two new uncorrelated variables using PCA. Of the two, the variable explaining most variation is used to represent the new group. The correlation matrix is updated based on the new variables.

The clustering process is repeated until all p original variables are grouped together. The transformation of the variables is recorded as a change of basis matrix $\mathbf{B}$, where each level l in the clustering procedure having its own basis $\mathbf{B}_l$. At level $l = 0$, the basis matrix is the identity matrix ($\mathbf{B}_0 = I$). Generally the transformed data $\boldsymbol{Y}_l$ at a given level l in the TT, can be found by:

$$\underset{(1\times p)}{\boldsymbol{Y}_l} = \underset{(p\times p)}{\boldsymbol{B}_l^T}\,\underset{(p\times 1)}{\mathbf{X}} \tag{3.3}$$

At the beginning, when $l = 0$, the transformed variables are the same as the original variables $\mathbf{X}$. Using the full matrix $\boldsymbol{B}_l$ with all the basis vectors does not result in data reduction. Instead, a selection of m vectors from $\boldsymbol{B}_l$, which explains the most variance of the total variance in the full original dataset, is used. This results in $\boldsymbol{Y}_l$ with dimensions $(1 \times m)$.

The clustering process is recorded in a clustering tree, as illustrated by figure 3.6. The complexity of the patterns identified by TT depends on the

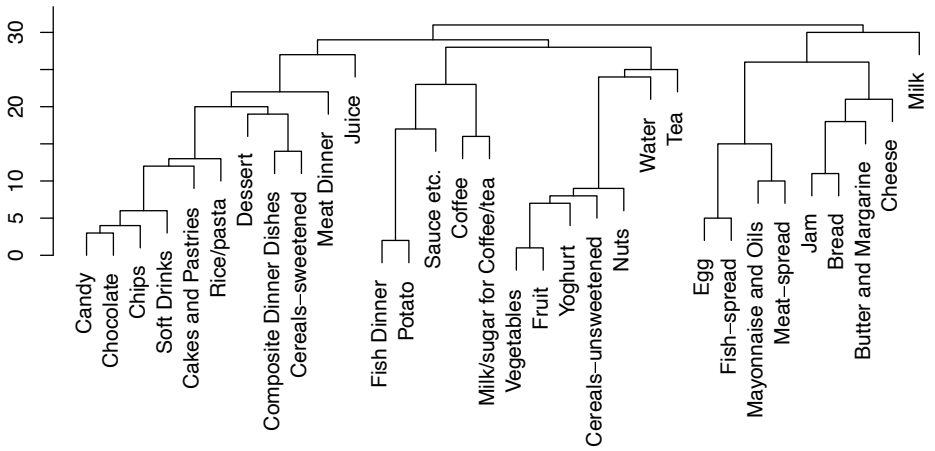


FIGURE 3.6. Clustering tree using TT and diet data from Tromsø7.

cut level of the clustering tree. At any level in the tree, you can "cut" the clustering tree and use the basis $\mathbf{B}_l$ from that level for further analysis.

One method to determine the optimal cut level is to select the level l where the basis matrix $\mathbf{B}_l$ results in a transformation that maximizes the explained variance. The number of basis vectors m from the different $\mathbf{B}_l$ is held constant at a fixed level (Lee et al., 2008).

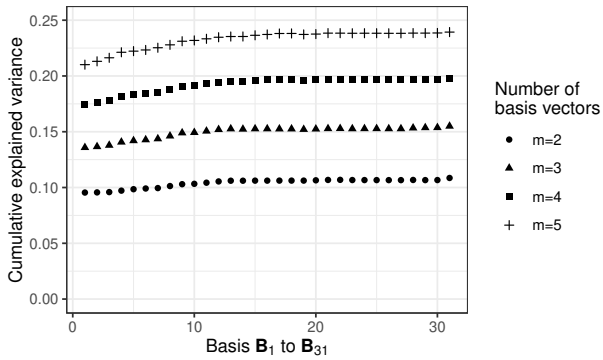


FIGURE 3.7. The mean of cross-validated cumulative explained variance using two to five basis vectors from $\mathbf{B}_l$ and the food data from Tromsø7. The basis vectors are varied using all the basis matrices found by different "cut" of the clustering tree.

Figure 3.7 shows the cumulative explained variance for a fixed number of basis vectors m using food data from Tromsø7. The cut level of the clustering three is varied from 1 to 31. At levels higher than 15-17, changes in the basis vectors due to the inclusion of more variables have little effect on the explained variance. The plot is based on cross-validation to avoid overfitting. However, in this example, it is evident that increasing the number of basis vectors m has a greater influence on the explained variance than cutting the tree at a higher point. This effect can be explained by a relatively low correlation between the variables in the dataset $\mathbf{X}$. A relatively low proportion of the variance in the full dataset is explained by the basis vectors.

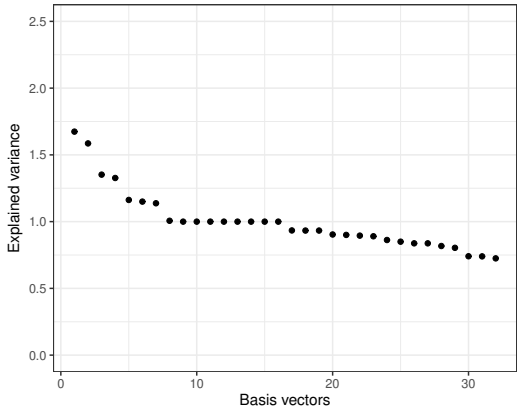


FIGURE 3.8. Scree plot of variance explained using the basis vectors of $\hat{\boldsymbol{B}}_{17}$.

Using the basis at a cut $l = 17$, the explained variance for each basis vector is shown in figure 3.8. As expected, comparing the scree plot for TT in figure 3.8 with the scree plot for PCA in figure 3.5 shows that the principal components found by PCA have a larger explained variance. In paper 2, four factors were chosen based on this scree plot and the interpretation of the clusters. Although the next three factors were considered, they were not selected because each contained only two variables and were not regarded as representing dietary patterns.

3.2 Regression analysis

In paper 2, the association between the identified dietary patterns and MetS, including its five components, was investigated using logistic regression. In all analyses, the dietary patterns were treated as continuous predictors for MetS and its components. This was done to avoid spurious interaction

effects, as well as loss of information and statistical power that can result from categorizing continuous variables (Thoresen, 2019).

3.2.1 Model evaluation by predictive power

The ability of dietary patterns to predict health-related outcomes was investigated by assessing the predictive power of the model. In paper 2, a hierarchy of models were investigated. First, the models were analysed including only age and education as predictors. In the next step, dietary patterns were included. Finally, the number of predictors were further extended by the lifestyle variables physical activity level, smoking and alcoholic consumption.

The estimated predictive power was based on randomly dividing the full dataset into a training and a test set. Since relying on a single split of test and training set can be unreliable (Bradley, 1997), the mean of 100 random splits of test and training sets were investigated in paper 2. All methods used the same 100 test and training sets. The training set was only used for model fitting, while the test set was used to evaluate the model's performance by comparing the predicted outcomes with the true outcomes. Overfitting was avoided by the use of test and training sets and resulted in a realistic measure of the model's ability to predict unseen data (Rousson and Goşoniu, 2007).

Depending on the response variable and the outcome of the model, the predictive power can be evaluated using different measures. In paper 2, all response variables were binary and modelled using logistic regression. In this case, the predicted outcome is a probability of belonging to a class, such as being classified with MetS. One approach to calculate the predictive power is to set a threshold on the predicted probability of belonging to a group and calculate the apparent error rate (APER). APER measures the proportion of misclassification (Tharwat, 2021). However, using a fixed threshold on the estimated probability does not fully utilize the probability estimates. An alternative is to use the receiver operating characteristics (ROC) curve, summarized by the area under the curve (AUC). This method takes into account the uncertainty in probability estimates by varying the threshold (Tharwat, 2021; Bradley, 1997). This has also been found to be the best way to evaluate methods using a single fit measure, due to its sensitivity and the fact that it does not rely on a specified threshold (Bradley, 1997).

3.2.2 Comparison with random forest

A known limitation of data-driven dietary patterns is that they do not necessarily predict health outcomes effectively, and a high proportion of the collected food data remains unused (Zhao et al., 2021; Michels and Schulze, 2005). For example, dietary patterns identified using FA often explain only

a relatively small proportion of the variance in the dataset (Michels and Schulze, 2005), with the total explained variance typically ranging from $20 - 30\%$ (Edefonti et al., 2020). Thus, there is potential to utilize more of the collected dietary data. To investigate the need of improvements in models, the predictive power can be useful (Shmueli, 2010).

In this context, random forest may offer evidence of improvements in the analysis of dietary patterns. The random forest algorithm is a powerful tool for prediction, capable of handling complex patterns (Biau and Scornet, 2016). Although random forest does not identify dietary patterns or estimate associations, it can be used to assess predictive power of including dietary data or not. A clear increase in predictive power using random forest compared to logistic regression could indicate limitations in the standard methods used to identify dietary patterns.

The random forest is constructed by making multiple decision trees. Every decision tree is built using a sub-sample of observations and variables, with splits put at different places to maximize prediction accuracy (Breiman, 2001). Growing the tree to the maximum depth is common, however research suggests that limiting the depth can achieve comparable accuracy (Nadi and Moradi, 2019). To predict a new outcome, random forest combines the predictions from all the individual trees, typically by averaging their outputs.

3.3 The role of BMI

In cross-sectional analysis, causal interpretation of associations can be problematic and must be done with care. However, discussing causality helps researchers understand the underlying assumptions and relationships. Causal diagrams, such as directed acyclic graphs (DAGs), offers a graphical presentation of the underlying causal knowledge or assumptions about how an exposure effect the outcomes. Typically, a variable can be connected to the predictor and response in three different ways, as illustrated in figure 3.9. The arrows in the diagram can be seen as indications of causal relationships, where an arrow points towards a measure further ahead in time. A collider is a variable that is caused by both the predictor and the response. A mediator is a variable that is caused by the predictor and effects the response. A confounder is an variable that has an effect both on both the predictor and the response.

Almost half of the publications in a recent meta-analysis by Fabiani et al., 2019 investigated association between data-driven dietary methods and MetS by adjusting for BMI. Some studies adjusted for BMI because it could be a potential confounder (Esmaillzadeh et al., 2007; Panagiotakos et al., 2007; Babio et al., 2009; Hong et al., 2012; Kang and Kim, 2016; Shakeri et al., 2019). On the other hand, BMI has also been considered as

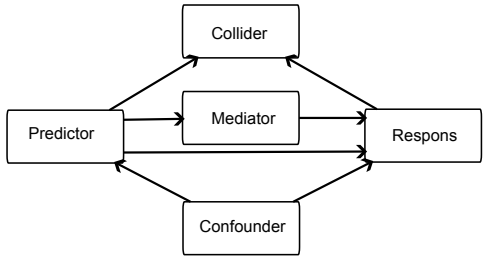


FIGURE 3.9. Causal diagram or DAG displaying collider, mediator and confounder variables

a mediator in certain analysis (Wagner et al., 2012). Overall, there was an insufficient explanation of BMI's role in these studies.

There is a broad consensus that obesity, commonly measured by BMI or waist circumference, impacts and disrupts various metabolic functions, including chronic inflammation, hypertension and altered lipid distribution (Lopez-Jimenez et al., 2022). Overweight is a risk factor for type 2 diabetes mellitus, and an important part of disease treatment is lifestyle changes (García-Molina et al., 2020). In this case dietary behaviour can indirectly act as a confounder. Also when dietary changes are made with the goal of losing weight, BMI may act as a confounder. However, this is based on the opposite assumption that diet affect weight. In this case, diet serves as a mediator. Some longitudinal studies suggest that diet affects overweight, showing for instance that consumption of ultra-processed foods (Canhada et al., 2020) and sugar-sweetened beverage (Santos et al., 2022) are associated with weight gain.

In principle, when using an FFQ, there is a temporal difference between the diet data collected and the biological samples and measurements taken during a study. An FFQ asks about the average intake over the past year rather than the current intake (Jacobs, 2023; Cade and Hutchinson, 2015). In this context, dietary behaviour is measured prior to the study, while BMI and other health outcomes are measured later as part of the main study. This suggests that BMI may act as a mediator. However, whether this temporal difference has a practical impact is another question. FFQs rely on long-term memory, and recent intake can have a large influence on recalled past intake (Cade and Hutchinson, 2015, p. 20). For instance, seasonal effects have been observed in FFQs distributed in summer versus winter (Shahar et al., 2001). Current intake can be captured using methods like the 24-hour recalls and food diaries.

If a variable acts as a mediator, it is possible to conduct a mediation analysis to investigate both the direct effect of the exposure on the outcome and the indirect effect through the mediator.

3.4 Exploratory structural equation model

Structural equation modelling (SEM) is a multivariate analysis technique that can model relationships between variables. SEM is developed for causal inference, where theory and course of events are important aspects in the construction of the model (Fairchild and McDaniel, 2017). This method facilitates mediation analysis of the effect of an predictor on an response. The mediation analysis includes test of direct effect without the mediator, indirect effect through the mediator and the total effect. The assumptions in the causal diagram are divided into several equations explaining the associations between variables. This is the structural part of the model.

SEM also includes a measurement part, where the observed variables are modelled as latent variables. The structure of the latent variables is predefined, including cross-loadings (Asparouhov and Muthén, 2009). ESEM is similar to SEM, but ESEM has an exploratory analysis in the measurement part. The measurement part in ESEM is like in an exploratory factor analysis, where the observed variables are modelled as latent variables without pre-defined structure. All observed variables are part of the latent variables. This gives more flexibility and is a recommended alternative to SEM when cross-loadings can't be ignored (Mai et al., 2018). It is also a good alternative when there is limited knowledge of the latent variables (Asparouhov and Muthén, 2009).

The most common method to fit models to continuous data is by using the maximum likelihood method. This method assumes a normal distribution. To correct test statistics and estimates for discrepancy of the normality assumption, bootstrapping and robust versions can be used. The model fit is commonly examined using the comparative fit index (CFI), the root mean square error of approximation (RMSEA) and the standardized root mean squared residual (SRMR). Cut points to consider the model as a good fit have been proposed like CFI > 0.90, RMSEA < 0.08 and SRMR < 0.08. However, there is ongoing debate over the use of such fit measures. For instance, the type of data can influence these measures, and they have been criticized for being too general (Marsh et al., 2004).

CHAPTER 4

Summary of papers

4.1 Paper 1: Identifying dietary patterns across age, educational level and physical activity level in a cross-sectional study: the Tromsø Study 2015 - 2016

Focus on overall dietary patterns has provided new insight into dietary behaviour by examining trends in intake rather than isolating specific food and nutritions, which only capture part of the total consumption (Cespedes and Hu, 2015). In paper 1, we focused on HC of food variables as a data-reduction method to analyse non-overlapping dietary patterns in Tromsø7. This approach provided clear and simple patterns by grouping variables based on the correlation matrix.

Additionally, the analysis was taken a step further to ensure stable patterns. First, individuals were grouped into cohorts based on similar background variables, such as age, education- and physical activity level. This approach reduced variance and resulted in more normally distributed food variables, which were then used to calculate the correlation matrix. To ensure the robustness of the clusters, a stability analysis was conducted by different grouping of these cohorts. This step ensured that unstable food variables were not forced into any specific cluster.

The analysis included a total of $10,899$ individuals aged 40 years and older, with 53.3% women. Dietary patterns were identified separately for women and men. The patterns identified were largely similar across sexes, with three dietary patterns found for both sexes: The Meat and Sweet pattern, the Traditional pattern and the Plant-based- and Tea pattern. Diet scores were calculated based on these patterns and used to investigate variations in dietary behaviour across different parts of the populations.

A linear model effectively explained much of the relationship between age and the diet scores for the Meat and Sweet diet and the Traditional diet. Specifically, the Meat and Sweet diet score decreased with age, while the Traditional diet score increased with age. When comparing the Plant-based- and Tea pattern across sexes, women had a higher score than men. Overall, individuals with lower education and lower physical activity had higher scores for the Meat and Sweet diet and the Traditional diet, whereas those with higher education and higher physical activity level had higher

score of the Plant-based- and Tea diet.

4.2 Paper 2: Associations and predictive power of dietary patterns on metabolic syndrome and its components

Paper 2 had a two-folded aim. The first aim was to use dietary patterns as predictors to estimate the association with MetS and its components elevated waist circumference, low HDL-cholesterol, elevated triglycerides, insulin resistance and hypertension. The second aim was to compare the predictive power of logistic regression models utilizing different data-driven methods.

Comparing various methods for deriving dietary patterns and analysing their association with health outcomes has been investigated (Schoenaker et al., 2013; Gorst-Rasmussen et al., 2011). However, few studies have assessed predictive power by splitting the data into test and training sets. This procedure helps evaluate model fit while avoiding overfitting.

In paper 2, the patterns identified in paper 1 were further explored in addition to dietary patterns identified by FA and TT. The effect on the predictive power of the inclusion of these patterns along with other lifestyle variables was investigated. Additionally, random forest was included for comparison. This method use the food variables directly and potentially include information that might be lost in the process of constructing dietary patterns.

The analysis included 10.750 participants aged 40 years or older from Tromsø7, with 53.3% women. Using patterns identified in paper 1, a positive association was observed between the Meat and Sweet pattern and MetS, as well as elevated waist circumference. The Plant-based- and Tea pattern was found to be negatively associated with hypertension among women and elevated triglycerides among men. Elevated triglycerides were also negatively associated with the Traditional pattern among men. Table 1 summarises the estimated odds ratios found in paper 2.

Expanding the analysis to include all three data-driven methods revealed that each method identified patterns like those found by HC of food variables. Specifically, the Meat and Sweet pattern and Plant-based- and Tea pattern, were also identified using FA and TT. Additionally, all four patterns identified by FA had a corresponding, but simplified and non-overlapping, version identified by TT.

Including dietary patterns in the logistic regression models resulted in a clear increase in predictive power for MetS and elevated waist circumference. For men, in the models that included dietary patterns, age and education as predictors, FA and random forest gave slightly better predictive power compared to HC and TT. However, adding other lifestyle variables resulted in only minor differences between the methods.

Note that table 5 in paper 2 contains a typo in the heading, where

	MetS	Elevated WC	Low HDL-C	Elevated TG	IR	Hyper-tension
Women:						
Meat and Sweets	1.11*	1.25*	1.10	1.06	1.11	1.04
Plant-based- and Tea	0.91	0.92	0.95	0.93	1.01	0.91
Traditional	0.94	1.07	1.00	0.91	0.93	1.02
Men:						
Meat and Sweets	1.16*	1.34*	1.08	1.03	1.01	1.04
Plant-based- and Tea	0.98	1.02	1.03	0.87*	1.04	0.93
Traditional	0.91	1.06	0.92	0.90*	0.95	0.98

TABLE 1. Total effect of dietary patterns on MetS and its components. MetS: Metabolic syndrome, WC: waist circumference, HDL-C: HDL-cholesterol, TG: triglycerides, IR: Insulin resistance. * P-value < 0.01.

FA and HC have been swapped. However, this typo does not affect the main conclusion regarding the predictive power analysis. Generally, the differences of the predictive power using the different data-reduction methods were small. This is consistent with findings in Gorst-Rasmussen et al., 2011, which reported minimal differences in Akaike's Information Criterion when comparing FA and TT in Cox regression models for myocardial infarction. Additionally, the logistic regression models gave comparable predictive power as using random forest.

4.3 Paper 3: Analysis of dietary pattern effects on metabolic risk factors using structural equation modelling

The literature on dietary patterns and association with MetS and its components reveals some uncertainty regarding the role of BMI. To address this, paper 3 utilizes the exploratory structural equation model (ESEM) to delve deeper into the complex interactions between dietary patterns, overweight, and metabolic risk factors for CVD.

BMI and waist circumference, which measure total and central obesity respectively, were used to assess overweight in the population. ESEM integrates FA, mediation analysis, and regression into one comprehensive model. This approach allowed for a thorough examination of the role of overweight as a potential mediator and how obesity influence estimated effects between dietary patterns and metabolic risk factors. A simplified illustration of the model is shown in figure 4.1.

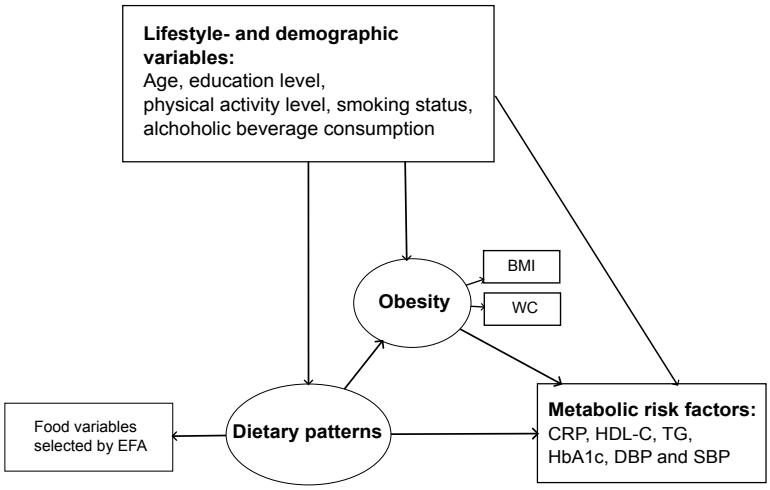


FIGURE 4.1. Simplified illustration of the ESEM model. EFA: exploratory factor analysis, BMI: body mass index, WC: waist circumference, CRP: C-reactive protein, HDL-C: HDL-cholesterol, TG: triglycerides, HbA1c: glycated hemoglobin, DBP: diastolic blood pressure, SBP: systolic blood pressure

The analysis included 9,988 participants aged 40 to 79 years without known diabetes from Tromsø7, of whom 53.7% were women. Our analysis revealed three common dietary patterns among women and men: a Snacks and Meat, a Health-conscious, and a Processed Dinner pattern. Additionally, a Porridge pattern was identified for women only and a Cake pattern

was identified among men.

Direct effects of dietary patterns were mainly observed on HDL-cholesterol and triglycerides. This was also observed in additional analyses included in the appendix of paper 3. Specifically, in the main analysis and for both sexes, the health-conscious pattern showed a direct favourable effect on HDL-cholesterol, and among women, it also had a favourable effect on triglycerides. Among men, the Snacks and Meat pattern had an unfavourable direct effect on triglycerides, while the Cake pattern showed a favourable effect.

Most dietary patterns identified had a direct effect on obesity. This included an unfavourable effect of the Snacks and Meat pattern, the Health-conscious pattern (men only) and the Processed Dinner pattern, whereas the Porridge pattern and Cake pattern had favourable negative effects on obesity. It was also seen that obesity was significantly associated with all other investigated CVD risk factors. This is expected, as the definition of MetS is based on the observed clustering of CVD risk factors (Grundy et al., 2005). Our assumption that obesity acts as a mediator, combined with the association between obesity and the other risk factors, resulted in a significant indirect effect of dietary patterns on all risk factors when these patterns were associated with obesity.

	Obesity	CRP	HDL-C	TG	HbA1c	SBP	DBP
Women:							
Snacks and Meat	0.08*	0.06*	−0.02*	0.00	−0.01	0.02	0.01
Health-conscious	0.02	−0.03	0.01	−0.03*	−0.01	−0.05	0.00
Processed Dinner	0.20*	0.11*	−0.02*	0.05*	0.01	0.09*	0.02
Porridge	−0.12*	−0.05*	0.00	−0.02	0.00	−0.06	−0.04
Men:							
Snacks and Meat	0.11*	0.08*	−0.03*	0.06*	0.00	0.07	0.05*
Health-conscious	0.07*	0.04*	0.01	−0.01	−0.02	0.02	0.01
Processed Dinner	0.22*	0.06*	−0.02*	0.02	0.01	0.06	0.05
Cake	−0.10*	−0.06*	0.02*	−0.04*	−0.02	0.01	−0.02

TABLE 2. Total effect of dietary patterns on CVD risk factors. HDL-C: HDL-cholesterol, TG: triglycerides, SBP: systolic blood pressure, DBP: diastolic blood pressure. * P-value < 0.01.

The indirect effect, combined with the direct effect, could lead to a significant total effect of a dietary pattern on the CVD risk factors. For both sexes, a total unfavourable effect of the Snacks and Meat pattern and the Processed Dinner pattern were observed on CRP and HDL-cholesterol.

Additionally, among men, the Snacks and Meat pattern showed an unfavourable positive effect on triglycerides and diastolic blood pressure. In women, an unfavourable effect of the Possessed Dinner pattern was observed on triglycerides and Systolic blood pressure. The health-conscious pattern had a favourable total effect on triglycerides among women. Among men, an unfavorable total effect of the health-conscious pattern on CRP was observed. Both the cake pattern and the Porridge pattern had favourable effect on CRP, the Cake pattern also demonstrated favourable effect on HDL-cholesterol and triglycerides.

CHAPTER 5

Discussion

In this thesis, dietary patterns in Tromsø7 were investigated using several data-driven methods. The patterns were further associated with MetS, and its components using logistic regression. This analysis also included an investigation of the predictive power when dietary patterns were included in the models as predictors. The analysis was enhanced further by the use of ESEM that incorporate causal assumptions. The ESEM can simultaneously identify dietary patterns, conduct mediation analysis and estimate associations between dietary patterns and CVD risk factors.

5.1 The identified dietary patterns

The HC, FA, TT and ESEM all identified a pattern high in sweets and meat and a health-conscious pattern. In addition, six other patterns were identified. The HC identified a traditional pattern with food variables like fish, potato, bread and cakes. FA and TT identified both a fish dinner pattern and a spread pattern. ESEM identified a pattern high on processed meat and fish, as well as a porridge pattern (women) or a cake pattern (men).

Age had a strong influence on the identification of dietary patterns, regardless of the method used. A significant trend with age was observed for the patterns high in sweets and meat, where the younger participants had a higher intake of these patterns. A significant negative trend with age was also observed for the health-conscious patterns found by ESEM, but not for the pattern identified by HC. The age trend was positive for both the traditional and processed dinner patterns, with younger participants having a lower intake of these patterns. The results are similar with other populations, where it has been observed that there is a higher preference for dietary patterns with sweets and meat among the younger populations, and that food commonly seen as traditional in a population is more consumed at higher age (Karageorgou et al., 2019; Dinu et al., 2021). The hypothesis of a global transition from traditional to westernized dietary patterns has been proposed (Popkin and Gordon-Larsen, 2004). Different countries are at various stages in this transition (Popkin, 2021). For example, Norway

is far along in this transition, with a study finding that one in three purchases were a sweet and ultra-processed product (Solberg et al., 2016). On the other hand, other trends have been observed that slow down the transition (Popkin, 2021). For instance changes in behaviour have been observed (Popkin and Gordon-Larsen, 2004). This may also be the case in Tromsø7, where the health-conscious patterns identified by ESEM were associated with younger ages. The data were measured at a single time point, and it is also possible that eating behaviour changes with age.

The most consistent finding for the health-conscious patterns was that they were more common among participants with higher education. On the other hand, patterns including sweets and meat were often less consumed by individuals with higher education. This is commonly observed in population studies, where healthier patterns often are associated with higher education level and other indicators of socioeconomic status like income (Mayén et al., 2014). One proposed explanation for these differences is the variation in prices between healthy foods and energy-dense foods (Darmon and Drewnowski, 2015; Schoenberg et al., 2013). People with lower socioeconomic status may also have less time and space for cooking, which can lead to higher use of prefabricated food. The environment and social influences can also have an effect on dietary patterns (Schoenberg et al., 2013).

The patterns high in sweets and meat, often referred to as westernized dietary patterns in the literature, were associated with lower physical activity level. Also, the processed dinner and the traditional patterns were associated with smoking and lower activity level. As the name implies, the health-conscious patterns were associated with other health-conscious choices like less smoking and higher activity level. This clustering of lifestyle behaviours is well known (Kris-Etherton et al., 2022). Evidence supports interactions between lifestyle behaviours, where improvement in one lifestyle behaviour can influence others (Kris-Etherton et al., 2022). In our analysis, we also observed that the porridge pattern and the cake pattern were associated with non-smoking and lower alcohol consumption, both of which are part of healthier lifestyle.

5.2 Dietary patterns and their association with CVD risk factors

The association between obesity and dietary patterns with high loadings on sweets and meat was consistently positively associated with obesity in both the logistic regression models and ESEM. Also the processed dinner pattern identified by ESEM was positively associated with obesity for both sexes. There is growing evidence that diets with a high intake of processed food, meats and sweets are risk factors for obesity and CVD (Mu et al., 2017; Min et al., 2017; Fabiani et al., 2019). The observed effects on obesity of

the dietary pattern with high loadings on sweets and meat and the pattern with high loadings on processed dinner variables are consistent with this observation. The dietary patterns high in sweets and meat and the processed dinner pattern also had a total unfavourable effect on HDL-cholesterol for both sexes in the ESEM model, but not in the logistic regression models.

Dietary patterns rich in vegetables and fruit, such as the well-studied Mediterranean diet, have been associated with improvements in several risk factors. For instance, reductions in blood pressure and BMI, along with improvements in the lipid profile, have been observed (Ndanuko et al., 2016; Castro-Barquero et al., 2020). The health-conscious pattern shares similarities with the Mediterranean diet, with high loading on vegetables and fruit. The health-conscious pattern had a total favourable effect on triglycerides in the logistic regression models and ESEM, but the results were inconsistent across sexes and models (see table 1 and 2 in chapter 4). On the other hand, the health-conscious pattern was positively associated with obesity among men in the ESEM analysis. However, the model also found a favourable direct effect of the health-conscious pattern on HDL-cholesterol. Obesity seems to act as a competitive mediator in this case, where no significant total effect was found. Among women, a direct favourable effect of the health-conscious patterns on triglycerides was observed.

Surprisingly, negative associations were observed between obesity and the porridge pattern and the cake pattern. Cake and cookies have been associated with lower risk of chronic diseases in von Ruesten et al., 2013. One proposed explanation is under-reporting of unhealthy food choices (von Ruesten et al., 2013), an explanation that is supported by additional analysis. For instance, models that initially showed a negative association between cakes and obesity, the association disappeared or changed direction when the model was adjusted for energy misreporting or under-reporting (Gottschald et al., 2016; Mendez et al., 2011). Under-reporting of energy intake also includes under eating (Poslusna et al., 2009). It is also possible that overweight participants have been motivated to manage their wight by reducing the consumption of unhealthy foods and lowering their daily energy intake.

In the ESEM analysis, obesity was included as a mediator between the dietary patterns and the other CVD risk factors. When a direct effect of a dietary pattern on obesity was significant, the indirect effect of the pattern through obesity was also significant. The indirect effect could strongly influence on the estimated total effect. This was for instance the case for CRP, where no significant direct effect was observed for any of the dietary patterns, but a total effect. Zhang et al., 2023 also observed that obesity had a significant mediation role in the effect of dietary patterns on CRP. Only a direct effect of diet on HDL-cholesterol and triglycerides was observed after

adjusting for the indirect effect of obesity.

Obesity is one of the risk factors for CVD that has increased the most worldwide (Castro-Barquero et al., 2020). This increase counteracts the beneficial reductions in other risk factors, such as hypertension, elevated triglycerides and low HDL-cholesterol, smoking and inactivity observed in several high-income countries (Lopez and Adair, 2019). The global increase in obesity is becoming a major health challenge, and preventing and treating obesity is important to further reduce the CVD burden (Lopez and Adair, 2019). In this context, a shift from the Westernized dietary pattern to healthier dietary patterns has been highlighted as one of the strategies to reduce risk factors for CVD (Castro-Barquero et al., 2020). However, lower CVD risk are often achieved by targeting several lifestyle behaviours (Kris-Etherton et al., 2022; Tsai et al., 2020).

5.3 Comparison of methods deriving dietary patterns

FA is one of the most common methods used to identify dietary patterns. The patterns identified by FA include all food variables. This complexity has been highlighted as a disadvantage of FA, as it makes the patterns harder to interpret (Zhao et al., 2021). The TT has been proposed as an alternative to obtain simpler patterns (Gorst-Rasmussen et al., 2011). In this thesis, also HC was included as a method to identify simpler patterns. In the literature of dietary patterns this is a novel approach. Clustering of food variables using HC is uncommon, despite being a well-known and relatively easy method to implement. To get more stable patterns in HC, cohorts of smaller groups with similar backgrounds variables were used. Among the three main data-driven methods investigated, TT produced the simplest patterns in terms of the number of variables included, closely followed by HC. Neither HC nor TT resulted in non-overlapping patterns in our analysis.

The interpretation of the dietary patterns identified by FA and TT was quite similar, when the focus was on the food variables with high loading. However, the simplified patterns by TT did not capture information about food variables with negative loading. When the goal is to identify the overall dietary patterns, the complexity in FA can also be seen as a flexibility and a strength (Imamura and Jacques, 2011). Generally, FA and TT identified similar patterns in our analysis. Also HC using bootstrapping resulted in patterns similar to those from FA and TT. High consistency between FA and TT has also been reported in other studies as well (Cunha et al., 2010; Gorst-Rasmussen et al., 2011; Schoenaker et al., 2013).

The patterns found by the different data reduction-methods were further included as predictors in logistic regression models, and the predictive performance of the models was investigated. In some cases FA gave slightly better predictive performance. However, the simpler patterns identified by

HC and TT had, in most of the cases, comparable predictive performance. We concluded that the difference in predictive performance between the different data reduction-methods was minimal. It can be noted that for most risk factors the inclusion of dietary patterns generally had little effect on the predictive power. The exception was the models for MetS and waist circumference. Waist circumference is a commonly used measure of obesity.

The reason for this low increase in predictive power by the inclusion of dietary patterns can be due to loss of information in the construction of dietary patterns. A relatively small percentage of the variation in the food variables is explained by the dietary patterns (McCann et al., 2001). In a meta-analysis of Fabiani et al., 2019 the typical percentages of explained variance in the included papers were around $20 - 30\%$. As a consequence, data-driven methods do not necessary identify patterns that reflect diet-disease relationships (Zhao et al., 2021). To investigate this, also random forest was included in the predictive performance comparison. This method can use the food variables directly, without the need to construct dietary patterns. In our analysis, the predictive performance of the random forest algorithm was comparable to that of the logistic regression models. This suggests that the patterns identified using data-driven methods were as effective as random forest in detecting associations between diet and CVD risk factors.

Already in the aggregation of the food variables, loss of information occur. In McCann et al., 2001 it was observed that similar patterns were identified by different level of aggregation. However, more detailed food variables resulted in stronger estimated risk. It was therefore suggested that more detailed information on food intake may be necessary to differentiate between individuals with and without disease (McCann et al., 2001). Therefore, more detailed food variables were used in the ESEM analysis compared to the previous analysis using logistic regression.

Regarding the choice of method, the ESEM was found to be a suitable approach for analysing diet-disease risk. The dietary patterns, the estimated effect of these pattens on CVD risk factors, and the effect of the lifestyle and demographic variables could be derived simultaneously. The method also support mediation analysis. This gives an overall model, as opposed to the logistic regression models, where each CVD risk factor had its own separate model. The use of ESEM or the construction of a directed acyclic graph (DAG) can also lead to a better understanding of the interplay between dietary patterns and health outcomes. A DAG is a graphical representation of the causal assumption. In both cases, the analyst must consider the underlying causal assumptions when constructing the model or creating a DAG. This is particularly the case if the role of a variable is uncertain. Obesity is one such variable, as its role in the relationship

between dietary patterns and risk factors for CVD appears to be uncertain. This is also useful in regression models, as these models are also constructed to investigate causal assumptions (Kevin Kelloway, 1995). Even though a model is constructed based on causal assumptions, the data itself can limit conclusions about causal relationships (Kevin Kelloway, 1995). This is the case with cross-sectional data. When using cross-sectional data, the course of events can´t guide the causal assumptions because the data is collected at a single time point.

Generally, the estimated diet-disease relationships were observed both across sexes and statistical approaches. We observed similar associations using logistic regression models and ESEM. Additionally, similar patterns occurred across all methods investigated, and the predictive performance was similar across the data reduction methods HC, FA and TT. The similarities across the different analyses strengthen the reliability of the results.

5.4 Strengths and limitations in dataset

An important strength of the data material is the high number of participants and a relatively high participation rate in Tromsø7. The large number of participants facilitated separate analyses for women and men. The large number of participants also facilitated a thorough analysis of predictive power, where the dataset could be split into test and training sets.

The data was collected at one time point using a cross-sectional design. This limits the causal interpretation of the results. Any causal conclusions based on a cross-sectional design must be done with care, as observed associations may have other explanations.

The use of self-reported data collected using FFQ has several limitations. It is well known that the traditional methods of capturing food consumption using methods like FFQs, 24-hour recalls and food records have shortcomings and are prone to both under- and over-reporting of food variables (Subar et al., 2015).

Missing values is often a problem in FFQs, as they consist of a long list of predefined food variables. Typically, the relevance of questions varies across the population and not all questions are relevant for everyone. Additionally, long FFQs can be very time-consuming to answer, and result in a relatively high participant burden. The handling of the missing values is therefore an important aspect. In the food data from Tromsø7, this problem is to some degree ignored. Missing values are coded as “seldom or not consumed” in Tromsø7. For foods that are rarely consumed, this is often a satisfactory assumption (Fraser et al., 2009). However, for frequently consumed foods, this is usually not the case (Fraser et al., 2009). The automatic coding of missing answers as zero intakes limits the ability to handle missing data and is often incorrect (Fraser et al., 2009). To reduce this problem, the analysis

in this thesis only includes participants answering more than 90% of the questionnaire.

There is an ongoing effort to improve the collection of food data, both for the researcher and for the participants (Das et al., 2022). Alternatives and modifications of the traditional methods to collect dietary data have been explored. The modifications include digitalization of traditional methods, for instance including assistance during the filling in of the questionnaire, and the processing of answers has been automated (Das et al., 2022; Tanweer et al., 2022). This work has contributed to making 24-hour recalls and food record alternatives to FFQs in large studies by reducing costs and making the methods more feasible for researchers (Subar et al., 2015). 24-hour recalls and food record are better at capturing current diet and do not rely on long-term memory or a restricted list of food variables (Satija et al., 2015).

The use of mobile phones has also been proposed. For instance, there have been attempts to use images taken on smartphones to classify foods and estimate energy and nutrient intake (Archundia Herrera and Chan, 2018). However, these methods are also prone to errors and under-reporting. Examples of under-reporting and errors include missing images or images that are difficult to analyse due to poor quality (Tanweer et al., 2022). Whether the new methods are more effective, less burdensome, and result in fewer error is still unclear. Methods relying on mobile phones also have potential for improvement (Tanweer et al., 2022).

CHAPTER 6

Conclusion

Dietary patterns were robust to the data-reduction methods used. For all investigated data-driven methods, two main dietary patterns were consistently identified. One pattern had a high loading on sweets and meat and the other was a health-conscious pattern. Additionally, a traditional pattern, a processed meat dinner pattern, and four other more narrowly focused patterns were identified.

The dietary patterns varied significantly by demographic variables. The younger part of the population had a higher intake of the sweets and meat patterns and a lower intake of the traditional and the processed dinner patterns. Healthier dietary intake, including high intake of the health-conscious patterns and low intake of the sweet and meat patterns and processed dinner patterns, were generally more common among individuals with higher education and higher physical activity level.

The data-reduction methods also gave robust results in measures of predictive performance. The predictive performance was investigated by including dietary patterns as predictors in logistic models of CVD risk factors. The models for MetS and obesity resulted in the clearest increase of predictive power with the inclusion of dietary patterns. Using random forest with individual food variables had comparable predictive performance to the logistic regression models.

Throughout this thesis, obesity was the risk factor most strongly associated with dietary patterns. Specifically, patterns with high loadings on food variables such as sweets, meat and processed dinners were consistently associated with obesity. For the other risk factors, the results were more inconsistent. The health-conscious patterns were associated with lower triglyceride levels in some models. We also observed that the indirect effect through obesity strongly influenced the total effect of the dietary patterns on the other CVD risk factors. Only HDL-cholesterol and triglycerides were directly affected by diet after adjusting for obesity.

The consistent association between obesity and more unhealthy patterns highlights how diet, an important part of lifestyle, most likely contributes to the ongoing rise in obesity. This finding is consistent with other research (Min et al., 2017).

Bibliography

Agarwal, S., D. R. Jacobs Jr., D. Vaidya, C. T. Sibley, N. W. Jorgensen, J. I. Rotter, Y.-D. I. Chen, Y. Liu, J. S. Andrews, S. Kritchevsky, B. Goodpaster, A. Kanaya, A. B. Newman, E. M. Simonsick and D. M. Herrington (2012). Metabolic Syndrome Derived from Principal Component Analysis and Incident Cardiovascular Events: The Multi Ethnic Study of Atherosclerosis (MESA) and Health, Aging, and Body Composition (Health ABC). *Cardiology Research and Practice*, 2012(1), 919425. DOI: https://doi.org/10.1155/2012/919425.

Archundia Herrera, M. C. and C. B. Chan (2018). Narrative Review of New Methods for Assessing Food and Energy Intake. *Nutrients*, 10(8). ISSN: 2072-6643. DOI: https://doi.org/10.3390/nu10081064.

Asparouhov, T. and B. Muthén (2009). Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3), 397–438. DOI: https://doi.org/10.1080/10705510903008204.

Babio, N., M. Bulló, J. Basora, M. Martínez-González, J. Fernández-Ballart, F. Márquez-Sandoval, C. Molina and J. Salas-Salvadó (2009). Adherence to the Mediterranean diet and risk of metabolic syndrome and its components. *Nutrition, Metabolism and Cardiovascular Diseases*, 19(8), 563–570. DOI: https://doi.org/10.1016/j.numecd.2008.10.007.

Bailey, R. L. (2021). Overview of dietary assessment methods for measuring intakes of foods, beverages, and dietary supplements in research studies. *Current Opinion in Biotechnology*, 70, 91–96. ISSN: 0958-1669. DOI: https://doi.org/10.1016/j.copbio.2021.02.007.

Bellec, P., P. Rosa-Neto, O. C. Lyttelton, H. Benali and A. C. Evans (2010). Multi-level bootstrap analysis of stable clusters in resting-state fMRI. *NeuroImage*, 51(3), 1126–1139. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2010.02.082.

Biau, G. and E. Scornet (2016). A random forest guided tour. *TEST*, 25(2), 197–227. DOI: 10.1007/s11749-016-0481-7.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), 1145–1159. ISSN: 0031-3203. DOI: https://doi.org/10.1016/S0031-3203(96)00142-2.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. DOI: https://doi.org/10.1023/A:1010933404324.

Cade, J. and J. Hutchinson (2015). Study Design: Population-Based Studies. *Nutrition Research Methodologies*. John Wiley & Sons, Ltd. Chap. 4, pp. 48–70. ISBN: 9781119180425. DOI: https://doi.org/10.1002/9781119180425.ch2.

Canhada, S. L., V. C. Luft, L. Giatti, B. B. Duncan, D. Chor, M. d. J. M. d. Fonseca, S. M. A. Matos, M. d. C. B. Molina, S. M. Barreto and R. B. Levy (2020). Ultra-processed foods, incident overweight and obesity, and longitudinal changes in weight and waist circumference: the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil). *Public Health Nutrition*, 23(6), 1076–1086. DOI: https://doi.org/10.1017/S1368980019002854.

Carlsen, M. H., A. Karlsen, I. T. L. Lillegaard, J. M. Gran, C. A. Drevon, R. Blomhoff and L. F. Andersen (2011). Relative validity of fruit and vegetable intake estimated from an FFQ, using carotenoid and flavonoid biomarkers and the method of triads. *British Journal of Nutrition*, 105(10), 1530–1538. DOI: https://doi.org/10.1017/S0007114510005246.

Carlsen, M. H., I. T. Lillegaard, A. Karlsen, R. Blomhoff, C. A. Drevon and L. F. Andersen (2010). Evaluation of energy and dietary intake estimates from a food frequency questionnaire using independent energy expenditure measurement and weighed food records. *Nutrition Journal*, 9(1), 37. ISSN: 1475-2891. DOI: https://doi.org/10.1186/1475-2891-9-37.

Castro-Barquero, S., A. M. Ruiz-León, M. Sierra-Pérez, R. Estruch and R. Casas (2020). Dietary Strategies for Metabolic Syndrome: A Comprehensive Review. *Nutrients*, 12(10). ISSN: 2072-6643. DOI: https://doi.org/10.3390/nu12102983.

Cespedes, E. M. and F. B. Hu (2015). Dietary patterns: from nutritional epidemiologic analysis to national guidelines2. *The American Journal of Clinical Nutrition*, 101(5), 899–900. ISSN: 0002-9165. DOI: https://doi.org/10.3945/ajcn.115.110213.

Cunha, D. B., R. M. V. R. d. Almeida and R. A. Pereira (2010). A comparison of three statistical methods applied in the identification of eating patterns. *Cadernos de Saúde Pública*, 26(11), 2138–2148. ISSN: 0102-311X. DOI: https://doi.org/10.1590/S0102-311X2010001100015.

Darmon, N. and A. Drewnowski (2015). Contribution of food prices and diet cost to socioeconomic disparities in diet quality and health: a systematic review and analysis. *Nutrition Reviews*, 73(10), 643–660. ISSN: 0029-6643. DOI: https://doi.org/10.1093/nutrit/nuv027.

Das, S. K., A. J. Miki, C. M. Blanchard, E. Sazonov, C. H. Gilhooly, S. Dey, C. B. Wolk, C. S. H. Khoo, J. O. Hill and R. P. Shook (2022). Perspective: Opportunities and Challenges of Technology Tools in Dietary

and Activity Assessment: Bridging Stakeholder Viewpoints. *Advances in Nutrition*, 13(1), 1–15. ISSN: 2161-8313. DOI: https://doi.org/10.1093/advances/nmab103.

De Siqueira Valadares, L. T., L. S. B. de Souza, V. A. Salgado Júnior, L. de Freitas Bonomo, L. R. de Macedo and M. Silva (2022). Prevalence of metabolic syndrome in Brazilian adults in the last 10 years: a systematic review and meta-analysis. *BMC Public Health*, 22(1), 327. DOI: https://doi.org/10.1186/s12889-022-12753-5.

Dinu, M., G. Pagliai, I. Giangrandi, B. Colombini, L. Toniolo, G. Gensini and F. Sofi (2021). Adherence to the Mediterranean diet among Italian adults: results from the web-based Medi-Lite questionnaire. *International Journal of Food Sciences and Nutrition*, 72(2), 271–279. DOI: https://doi.org/10.1080/09637486.2020.1793306.

Edefonti, V., R. D. Vito, M. Dalmartello, L. Patel, A. Salvatori and M. Ferraroni (2020). Reproducibility and Validity of A Posteriori Dietary Patterns: A Systematic Review. *Advances in Nutrition*, 11(2), 293–326. DOI: https://doi.org/10.1093/advances/nmz097.

Esmaillzadeh, A., M. Kimiagar, Y. Mehrabi, L. Azadbakht, F. B. Hu and W. C. Willett (2007). Dietary patterns, insulin resistance, and prevalence of the metabolic syndrome in women2. *The American Journal of Clinical Nutrition*, 85(3), 910–918. ISSN: 0002-9165. DOI: https://doi.org/10.1093/ajcn/85.3.910.

Fabiani, R., G. Naldini and M. Chiavarini (2019). Dietary Patterns and Metabolic Syndrome in Adult Subjects: A Systematic Review and Meta-Analysis. *Nutrients*, 11(9). ISSN: 2072-6643. DOI: https://doi.org/10.3390/nu11092056.

Fairchild, A. J. and H. L. McDaniel (2017). Best (but oft-forgotten) practices: mediation analysis 1,2. *The American Journal of Clinical Nutrition*, 105(6), 1259–1271. ISSN: 0002-9165. DOI: https://doi.org/10.3945/ajcn.117.152546.

Feldman, R. D., T. J. Anderson and R. M. Touyz (2015). Metabolic Syndrome Sinkholes: What to Do When Occam's Razor Gets Blunted. *Canadian Journal of Cardiology*, 31(5), 601–604. ISSN: 0828-282X. DOI: https://doi.org/10.1016/j.cjca.2014.12.035.

Fraser, G. E., R. Yan, T. L. Butler, K. Jaceldo-Siegl, W. L. Beeson and J. Chan (2009). Missing Data in a Long Food Frequency Questionnaire: Are Imputed Zeroes Correct? *Epidemiology*, 20(2). DOI: https://doi.org/10.1097/EDE.0b013e31819642c4.

García-Molina, L., A.-M. Lewis-Mikhael, B. Riquelme-Gallego, N. Cano-Ibáñez, M.-J. Oliveras-López and A. Bueno-Cavanillas (2020). Improving type 2 diabetes mellitus glycaemic control through lifestyle modification implementing diet intervention: a systematic review and meta-analysis. *European Journal of Nutrition*, 59(4), 1313–1328. DOI: https://doi.org/10.1007/s00394-019-02147-6.

Gorst-Rasmussen, A., C. C. Dahm, C. Dethlefsen, T. Scheike and K. Overvad (2011). Exploring Dietary Patterns By Using the Treelet Transform. *American Journal of Epidemiology*, 173(10), 1097–1104. ISSN: 0002-9262. DOI: https://doi.org/10.1093/aje/kwr060.

Gottschald, M., S. Knüppel, H. Boeing and B. Buijsse (2016). The influence of adjustment for energy misreporting on relations of cake and cookie intake with cardiometabolic disease risk factors. *European Journal of Clinical Nutrition*, 70(11), 1318–1324. DOI: https://doi.org/10.1038/ejcn.2016.131.

Grimby, G., M. Börjesson, I. H. Jonsdottir, P. Schnohr, D. S. Thelle and B. Saltin (2015). The "Saltin–Grimby Physical Activity Level Scale" and its application to health research. *Scandinavian Journal of Medicine & Science in Sports*, 25(S4), 119–125. DOI: https://doi.org/10.1111/sms.12611.

Grundy, S. M., J. I. Cleeman, S. R. Daniels, K. A. Donato, R. H. Eckel, B. A. Franklin, D. J. Gordon, R. M. Krauss, P. J. Savage, S. C. Smith, J. A. Spertus and F. Costa (2005). Diagnosis and Management of the Metabolic Syndrome. *Circulation*, 112(17), 2735–2752. DOI: https://doi.org/10.1161/CIRCULATIONAHA.105.169404.

Haththotuwa, R. N., C. N. Wijeyaratne and U. Senarath (2020). Chapter 1 - Worldwide epidemic of obesity. *Obesity and Obstetrics (Second Edition)*. Ed. by T. A. Mahmood, S. Arulkumaran and F. A. Chervenak. Second Edition. Elsevier, pp. 3–8. ISBN: 978-0-12-817921-5. DOI: https://doi.org/10.1016/B978-0-12-817921-5.00001-1.

Haverinen, E., L. Paalanen, L. Palmieri, A. Padron-Monedero, I. Noguer-Zambrano, R. Sarmiento Suárez and H. Tolonen (2021). Comparison of metabolic syndrome prevalence using four different definitions –a population-based study in Finland. *Archives of Public Health*, 79(1), 231. DOI: https://doi.org/10.1186/s13690-021-00749-3.

Hearty, Á. P. and M. J. Gibney (2008). Comparison of cluster and principal component analysis techniques to derive dietary patterns in Irish adults. *British Journal of Nutrition*, 101(4), 598–608. DOI: https://doi.org10.1017/S0007114508014128.

Hong, S., Y. Song, K. H. Lee, H. S. Lee, M. Lee, S. H. Jee and H. Joung (2012). A fruit and dairy dietary pattern is associated with a reduced

risk of metabolic syndrome. *Metabolism*, 61(6), 883–890. ISSN: 0026-0495. DOI: https://doi.org/10.1016/j.metabol.2011.10.018.

Hopstock, L. A., S. Grimsgaard, H. Johansen, K. Kanstad, T. Wilsgaard and A. E. Eggen (2022). The seventh survey of the Tromsø Study (Tromsø7) 2015–2016: study design, data collection, attendance, and prevalence of risk factors and disease in a multipurpose population-based health survey. *Scandinavian Journal of Public Health*, 50(7), 919–929. DOI: https://doi.org/10.1177/14034948221092294.

Hu, F. B. (2002). Dietary pattern analysis: a new direction in nutritional epidemiology. *Current Opinion in Lipidology*, 13(1).

Imamura, F. and P. F. Jacques (2011). Invited Commentary: Dietary Pattern Analysis. *American Journal of Epidemiology*, 173(10), 1105–1108. ISSN: 0002-9262. DOI: https://doi.org/10.1093/aje/kwr063.

Jacobs, D. R. (2023). Challenges in Research in Nutritional Epidemiology. *Nutritional Health: Strategies for Disease Prevention*. Ed. by N. J. Temple, T. Wilson, D. R. Jacobs Jr. and G. A. Bray. Cham: Springer International Publishing, pp. 21–31. ISBN: 978-3-031-24663-0. DOI: https://doi.org/10.1007/978-3-031-24663-0_2.

Johnson, R. and D. Wichern (2014a). Factor Analysis and Inference for Structured Covariance Matrices. *Applied Multivariate Statistical Analysis*. 6th ed. Harlow: Pearson. Chap. 9, pp. 21–31. ISBN: 978-1-292-02494-3.

— (2014b). Principal Components. *Applied Multivariate Statistical Analysis*. 6th ed. Harlow: Pearson. Chap. 8, pp. 21–31. ISBN: 978-1-292-02494-3.

Joliffe, I. and B. Morgan (1992). Principal component analysis and exploratory factor analysis. *Statistical Methods in Medical Research*, 1(1), 69–95. DOI: https://doi.org/10.1177/096228029200100105.

Kang, Y. and J. Kim (2016). Gender difference on the association between dietary patterns and metabolic syndrome in Korean population. *European Journal of Nutrition*, 55(7), 2321–2330. DOI: https://doi.org/10.1007/s00394-015-1127-3.

Karageorgou, D., E. Magriplis, A. Mitsopoulou, I. Dimakopoulos, I. Bakogianni, R. Micha, G. Michas, M. Chourdakis, T. Ntouroupi, S. Tsaniklidou, K. Argyri, D. Panagiotakos, A. Zampelas, E. Fappa, E.-M. Theodoraki, E. Trichia, T.-E. Sialvera, A. Varytimiadi, E. Spyreli, A. Koutelidakis, G. Karlis, S. Zacharia, A. Papageorgiou, G. Chrousos, G. Dedoussis, G. Dimitriadis, I. Manios and E. Roma (2019). Dietary patterns and lifestyle characteristics in adults: results from the Hellenic National Nutrition and Health Survey (HNNHS). *Public Health*, 171, 76–88. ISSN: 0033-3506. DOI: https://doi.org/10.1016/j.puhe.2019.03.013.

Kassi, E., P. Pervanidou, G. Kaltsas and G. Chrousos (2011). Metabolic syndrome: definitions and controversies. *BMC Medicine*, 9(1), 48. DOI: https://doi.org/10.1186/1741-7015-9-48.

Kevin Kelloway, E. (1995). Structural Equation Modelling in Perspective. *Journal of Organizational Behavior*, 16(3), 215–224. ISSN: 08943796, 10991379.

Kris-Etherton, P. M., P. A. Sapp, T. M. Riley, K. M. Davis, T. Hart and O. Lawler (2022). The Dynamic Interplay of Healthy Lifestyle Behaviors for Cardiovascular Health. *Current Atherosclerosis Reports*, 24(12), 969–980. DOI: https://doi.org/10.1007/s11883-022-01068-w.

Lee, A. B., B. Nadler and L. Wasserman (2008). Treelets: An Adaptive Multi-Scale Basis for Sparse Unordered Data. *The Annals of Applied Statistics*, 2(2), 435–471. DOI: https://doi.org/10.1214/07-AOAS137.

Liang, X. P., C. Y. Or, M. F. Tsoi, C. L. Cheung and B. M. Y. Cheung (2021). Prevalence of metabolic syndrome in the United States National Health and Nutrition Examination Survey (nhanes) 2011–2018. *European Heart Journal*, 42(Supplement_1), ehab724.2420. ISSN: 0195-668X. DOI: https://doi.org/10.1093/eurheartj/ehab724.2420.

Lopez, A. D. and T. Adair (2019). Is the long-term decline in cardiovascular-disease mortality in high-income countries over? Evidence from national vital statistics. *International Journal of Epidemiology*, 48(6), 1815–1823. ISSN: 0300-5771. DOI: https://doi.org/10.1093/ije/dyz143.

Lopez-Jimenez, F., W. Almahmeed, H. Bays, A. Cuevas, E. Di Angelantonio, C. W. le Roux, N. Sattar, M. C. Sun, G. Wittert, F. J. Pinto and J. P. H. Wilding (2022). Obesity and cardiovascular disease: mechanistic insights and management strategies. A joint position paper by the World Heart Federation and World Obesity Federation. *European Journal of Preventive Cardiology*, 29(17), 2218–2237. ISSN: 2047-4873. DOI: https://doi.org/10.1093/eurjpc/zwac187.

Mai, Y., Z. Zhang and Z. Wen (2018). Comparing Exploratory Structural Equation Modeling and Existing Approaches for Multiple Regression with Latent Variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(5), 737–749. DOI: https://doi.org/10.1080/10705511.2018.1444993.

Marsh, H. W., K.-T. Hau and Z. Wen (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal*, 11(3), 320–341. DOI: https://doi.org/10.1207/s15328007sem1103_2.

Martínez, M. E., J. R. Marshall and L. Sechrest (1998). Invited Commentary: Factor Analysis and the Search for Objectivity. *American Journal*

of Epidemiology, 148(1), 17–19. ISSN: 0002-9262. DOI: https://doi.org/10.1093/oxfordjournals.aje.a009552.

Mayén, A.-L., P. Marques-Vidal, F. Paccaud, P. Bovet and S. Stringhini (2014). Socioeconomic determinants of dietary patterns in low- and middle-income countries: a systematic review. *The American Journal of Clinical Nutrition*, 100(6), 1520–1531. ISSN: 0002-9165. DOI: https://doi.org/10.3945/ajcn.114.089029.

McCann, S. E., J. R. Marshall, J. R. Brasure, S. Graham and J. L. Freudenheim (2001). Analysis of patterns of food intake in nutritional epidemiology: food classification in principal components analysis and the subsequent impact on estimates for endometrial cancer. *Public Health Nutrition*, 4(5), 989–997. DOI: https://doi.org/10.1079/PHN2001168.

McCullough, L. E. and D. A. Byrd (2022). Total Energy Intake: Implications for Epidemiologic Analyses. *American Journal of Epidemiology*, 192(11), 1801–1805. ISSN: 0002-9262. DOI: https://doi.org/10.1093/aje/kwac071.

Mendez, M. A., B. M. Popkin, G. Buckland, H. Schroder, P. Amiano, A. Barricarte, J.-M. Huerta, J. R. Quirós, M.-J. Sánchez and C. A. González (2011). Alternative Methods of Accounting for Underreporting and Overreporting When Measuring Dietary Intake-Obesity Relations. *American Journal of Epidemiology*, 173(4), 448–458. ISSN: 0002-9262. DOI: https://doi.org/10.1093/aje/kwq380.

Michels, K. B. and M. B. Schulze (2005). Can dietary patterns help us detect diet-–disease associations? *Nutrition Research Reviews*, 18(2), 241–248. DOI: https://doi.org/10.1079/NRR2005107.

Min, M., X. Li-Fa, H. Dong, W. Jing and B. Ming-Jie (2017). Dietary patterns and overweight/obesity: a review article. *Iranian journal of public health*, 46(7), 869.

Mottillo, S., K. B. Filion, J. Genest, L. Joseph, L. Pilote, P. Poirier, S. Rinfret, E. L. Schiffrin and M. J. Eisenberg (2010). The Metabolic Syndrome and Cardiovascular Risk. *Journal of the American College of Cardiology*, 56(14), 1113–1132. DOI: https://doi.org/10.1016/j.jacc.2010.05.034.

Mu, M., L.-F. Xu, D. Hu, J. Wu and M.-J. Bai (2017). Dietary patterns and overweight/obesity: a review article. *Iranian journal of public health*, 46(7), 869.

Nadi, A. and H. Moradi (2019). Increasing the views and reducing the depth in random forest. *Expert Systems with Applications*, 138, 112801. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2019.07.018.

Ndanuko, R. N., L. C. Tapsell, K. E. Charlton, E. P. Neale and M. J. Batterham (2016). Dietary Patterns and Blood Pressure in Adults: A Systematic Review and Meta-Analysis of Randomized Controlled Trials.

Advances in Nutrition, 7(1), 76–89. ISSN: 2161-8313. DOI: https://doi.org/10.3945/an.115.009753.

Ngwasiri, C., M. Kinoré, S. Samadoulougou and F. Kirakoya-Samadoulougou (2023). Sex-specific-evaluation of metabolic syndrome prevalence in Algeria: insights from the 2016–2017 non-communicable diseases risk factors survey. *Scientific Reports*, 13(1), 18908. DOI: https://doi.org/10.1038/s41598-023-45625-y.

O'Donnell, M. J., S. L. Chin, S. Rangarajan, D. Xavier, L. Liu, H. Zhang, P. Rao-Melacini, X. Zhang, P. Pais, S. Agapay, P. Lopez-Jaramillo, A. Damasceno, P. Langhorne, M. J. McQueen, A. Rosengren, M. Dehghan, G. J. Hankey, A. L. Dans, A. Elsayed, A. Avezum, C. Mondo, H.-C. Diener, D. Ryglewicz, A. Czlonkowska, N. Pogosova, C. Weimar, R. Iqbal, R. Diaz, K. Yusoff, A. Yusufali, A. Oguz, X. Wang, E. Penaherrera, F. Lanas, O. S. Ogah, A. Ogunniyi, H. K. Iversen, G. Malaga, Z. Rumboldt, S. Oveisgharan, F. Al Hussain, D. Magazi, Y. Nilanont, J. Ferguson, G. Pare and S. Yusuf (2016). Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. *The Lancet*, 388(10046), 761–775. DOI: https://doi.org/10.1016/S0140-6736(16)30506-2.

Onat, A. and G. Hergenç (2011). Low-grade inflammation, and dysfunction of high-density lipoprotein and its apolipoproteins as a major driver of cardiometabolic risk. *Metabolism*, 60(4), 499–512. ISSN: 0026-0495. DOI: https://doi.org/10.1016/j.metabol.2010.04.018.

Panagiotakos, D. B., C. Pitsavos, Y. Skoumas and C. Stefanadis (2007). The Association between Food Patterns and the Metabolic Syndrome Using Principal Components Analysis: The ATTICA Study. *Journal of the American Dietetic Association*, 107(6), 979–987. ISSN: 0002-8223. DOI: https://doi.org/10.1016/j.jada.2007.03.006.

Popkin, B. M. and P. Gordon-Larsen (2004). The nutrition transition: worldwide obesity dynamics and their determinants. *International Journal of Obesity*, 28(3), S2–S9. DOI: https://doi.org/10.1038/sj.ijo.0802804.

Popkin, B. M. (2021). Measuring the nutrition transition and its dynamics. *Public Health Nutrition*, 24(2), 318–320. DOI: https://doi.org/10.1017/S136898002000470X.

Poslusna, K., J. Ruprich, J. H. M. de Vries, M. Jakubikova and P. van't Veer (2009). Misreporting of energy and micronutrient intake estimated by food records and 24-hour recalls, control and adjustment methods in practice. *British Journal of Nutrition*, 101(S2), S73–S85. DOI: 10.1017/S0007114509990602.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied*

Mathematics, 20, 53–65. ISSN: 0377-0427. DOI: https://doi.org/10.1016/0377-0427(87)90125-7.

Rousson, V. and N. F. Goşoniu (2007). An R-square coefficient based on final prediction error. *Statistical Methodology*, 4(3), 331–340. ISSN: 1572-3127. DOI: https://doi.org/10.1016/j.stamet.2006.11.004.

Santos, L. P., D. P. Gigante, F. M. Delpino, A. P. Maciel and R. M. Bielemann (2022). Sugar sweetened beverages intake and risk of obesity and cardiometabolic diseases in longitudinal studies: A systematic review and meta-analysis with 1.5 million individuals. *Clinical Nutrition ESPEN*, 51, 128–142. ISSN: 2405-4577. DOI: https://doi.org/10.1016/j.clnesp.2022.08.021.

Satija, A., E. Yu, W. C. Willett and F. B. Hu (2015). Understanding Nutritional Epidemiology and Its Role in Policy. *Advances in Nutrition*, 6(1), 5–18. ISSN: 2161-8313. DOI: https://doi.org/10.3945/an.114.007492.

Schoenaker, D. A., A. J. Dobson, S. S. Soedamah-Muthu and G. D. Mishra (2013). Factor Analysis Is More Appropriate to Identify Overall Dietary Patterns Associated with Diabetes When Compared with Treelet Transform Analysis. *The Journal of Nutrition*, 143(3), 392–398. DOI: https://doi.org/10.3945/jn.112.169011.

Schoenberg, N. E., B. M. Howell, M. Swanson, C. Grosh and S. Bardach (2013). Perspectives on Healthy Eating Among Appalachian Residents. *The Journal of Rural Health*, 29(s1), s25–s34. DOI: https://doi.org/10.1111/jrh.12009.

Shahar, D., N. Yerushalmi, F. Lubin, P. Froom, A. Shahar and E. Kristal-Boneh (2001). Seasonal variations in dietary intake affect the consistency of dietary assessment. *European Journal of Epidemiology*, 17(2), 129–133. DOI: https://doi.org/10.1023/A:1017542928978.

Shakeri, Z., P. Mirmiran, S. Khalili-Moghadam, F. Hosseini-Esfahani, A. Ataie-Jafari and F. Azizi (2019). Empirical dietary inflammatory pattern and risk of metabolic syndrome and its components: Tehran Lipid and Glucose Study. *Diabetology & Metabolic Syndrome*, 11(1), 16. DOI: https://doi.org/10.1186/s13098-019-0411-4.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, 25(3), 289–310. DOI: https://doi.org/10.1214/10-STS330.

Slimani, N., H. Freisling, A.-K. Illner and I. Huybrechts (2015). Methods to Determine Dietary Intake. *Nutrition Research Methodologies*. John Wiley & Sons, Ltd. Chap. 4, pp. 48–70. ISBN: 9781119180425. DOI: https://doi.org/10.1002/9781119180425.ch4.

Solberg, S. L., L. Terragni and S. I. Granheim (2016). Ultra-processed food purchases in Norway: a quantitative study on a representative sample of

food retailers. *Public Health Nutrition*, 19(11), 1990–2001. DOI: https://doi.org/10.1017/S1368980015003523.

Subar, A. F., L. S. Freedman, J. A. Tooze, S. I. Kirkpatrick, C. Boushey, M. L. Neuhouser, F. E. Thompson, N. Potischman, P. M. Guenther, V. Tarasuk, J. Reedy and S. M. Krebs-Smith (2015). Addressing Current Criticism Regarding the Value of Self-Report Dietary Data. *The Journal of Nutrition*, 145(12), 2639–2645. ISSN: 0022-3166. DOI: https://doi.org/10.3945/jn.115.219634.

Tanweer, A., S. Khan, F. N. Mustafa, S. Imran, A. Humayun and Z.-n. Hussain (2022). Improving dietary data collection tools for better nutritional assessment – A systematic review. *Computer Methods and Programs in Biomedicine Update*, 2, 100067. ISSN: 2666-9900. DOI: https://doi.org/10.1016/j.cmpbup.2022.100067.

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, 17(1), 168–192. DOI: https://doi.org/10.1016/j.aci.2018.08.003.

Thompson, F. E., S. I. Kirkpatrick, A. F. Subar, J. Reedy, T. E. Schap, M. M. Wilson and S. M. Krebs-Smith (2015). The National Cancer Institute's Dietary Assessment Primer: A Resource for Diet Research. *Journal of the Academy of Nutrition and Dietetics*, 115(12), 1986–1995. ISSN: 2212-2672. DOI: https://doi.org/10.1016/j.jand.2015.08.016.

Thoresen, M. (2019). Spurious interaction as a result of categorization. *BMC Medical Research Methodology*, 19(1), 28. DOI: https://doi.org/10.1186/s12874-019-0667-2.

Tsai, M.-C., C.-C. Lee, S.-C. Liu, P.-J. Tseng and K.-L. Chien (2020). Combined healthy lifestyle factors are more beneficial in reducing cardiovascular disease in younger adults: a meta-analysis of prospective cohort studies. *Scientific Reports*, 10(1), 18165. DOI: https://doi.org/10.1038/s41598-020-75314-z.

Varraso, R., J. Garcia-Aymerich, F. Monier, N. Le Moual, J. De Batlle, G. Miranda, C. Pison, I. Romieu, F. Kauffmann and J. Maccario (2012). Assessment of dietary patterns in nutritional epidemiology: principal component analysis compared with confirmatory factor analysis123. *The American Journal of Clinical Nutrition*, 96(5), 1079–1092. ISSN: 0002-9165. DOI: https://doi.org/10.3945/ajcn.112.038109.

Von Ruesten, A., S. Feller, M. M. Bergmann and H. Boeing (2013). Diet and risk of chronic diseases: results from the first 8 years of follow-up in the EPIC-Potsdam study. *European Journal of Clinical Nutrition*, 67(4), 412–419. DOI: https://doi.org/10.1038/ejcn.2013.7.

Wagner, A., J. Dallongeville, B. Haas, J. Ruidavets, P. Amouyel, J. Ferrières, C. Simon and D. Arveiler (2012). Sedentary behaviour, physical activity

and dietary patterns are independently associated with the metabolic syndrome. *Diabetes & Metabolism*, 38(5), 428–435. DOI: https://doi.org/10.1016/j.diabet.2012.04.005.

Weikert, C. and M. B. Schulze (2016). Evaluating dietary patterns: the role of reduced rank regression. *Current Opinion in Clinical Nutrition & Metabolic Care*, 19(5). DOI: https://doi.org/10.1097/MCO.0000000000000308.

Willett, W., G. Howe and L. Kushi (1997). Adjustment for total energy intake in epidemiologic studies. *The American Journal of Clinical Nutrition*, 65(4), 1220S–1228S. ISSN: 0002-9165. DOI: https://doi.org/10.1093/ajcn/65.4.1220S.

Yusuf, S., S. Hawken, S. Ôunpuu, T. Dans, A. Avezum, F. Lanas, M. McQueen, A. Budaj, P. Pais, J. Varigos and L. Lisheng (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *The Lancet*, 364(9438), 937–952. DOI: https://doi.org/10.1016/S0140-6736(04)17018-9.

Zabetian, A., F. Hadaegh and F. Azizi (2007). Prevalence of metabolic syndrome in Iranian adult population, concordance between the IDF with the ATPIII and the WHO definitions. *Diabetes Research and Clinical Practice*, 77(2), 251–257. ISSN: 0168-8227. DOI: https://doi.org/10.1016/j.diabres.2006.12.001.

Zhang, S., X. Yang, L. E, X. Zhang, H. Chen and X. Jiang (2023). The Mediating Effect of Central Obesity on the Association between Dietary Quality, Dietary Inflammation Level and Low-Grade Inflammation-Related Serum Inflammatory Markers in Adults. *International Journal of Environmental Research and Public Health*, 20(5). ISSN: 1660-4601. DOI: https://doi.org/10.3390/ijerph20053781.

Zhao, J., Z. Li, Q. Gao, H. Zhao, S. Chen, L. Huang, W. Wang and T. Wang (2021). A review of statistical methods for dietary pattern analysis. *Nutrition Journal*, 20(1), 37. DOI: https://doi.org/10.1186/s12937-021-00692-7.

Paper I

Identifying dietary patterns across age, educational level and physical activity level in a cross-sectional study: the Tromsø Study 2015 - 2016

BMC Nutrition, **8**, 102, 2022.

Moe *et al. BMC Nutrition* (2022) 8:102
https://doi.org/10.1186/s40795-022-00599-4

BMC Nutrition

RESEARCH Open Access

Identifying dietary patterns across age, educational level and physical activity level in a cross-sectional study: the Tromsø Study 2015 - 2016

Åse Mari Moe[1*], Sigrunn H. Sørbye[1], Laila A. Hopstock[2], Monica H. Carlsen[3], Ola Løvsletten[2] and Elinor Ytterstad[1]

Abstract

Background: A healthy diet can decrease the risk of several lifestyle diseases. From studying the health effects of single foods, research now focuses on examining complete diets and dietary patterns reflecting the combined intake of different foods. The main goals of the current study were to identify dietary patterns and then investigate how these differ in terms of sex, age, educational level and physical activity level (PAL) in a general Nordic population.

Methods: We used data from the seventh survey of the population-based Tromsø Study in Norway, conducted in 2015-2016. The study included 21,083 participants aged 40 — 99 years, of which 72% completed a comprehensive food frequency questionnaire (FFQ). After exclusion, the study sample included 10,899 participants with valid FFQ data. First, to cluster food variables, the participants were partitioned in homogeneous cohorts according to sex, age, educational level and PAL. Non-overlapping diet groups were then identified using repeated hierarchical cluster analysis on the food variables. Second, average standardized diet intake scores were calculated for all individuals for each diet group. The individual diet (intake) scores were then modelled in terms of age, education and PAL using regression models. Differences in diet scores according to education and PAL were investigated by pairwise hypothesis tests, controlling the nominal significance level using Tukey's method.

Results: The cluster analysis revealed three dietary patterns, here named the Meat and Sweets diet, the Traditional diet, and the Plant-based- and Tea diet. Women had a lower intake of the Traditional diet and a higher preference for the Plant-based- and Tea diet compared to men. Preference for the Meat and Sweets diet and Traditional diet showed significant negative and positive trends as function of age, respectively. Adjusting for age, the group having high education and high PAL compared favourably with the group having low education and low PAL, having a significant lower intake of the Meat and Sweets and the Traditional diets and a significant higher intake of the Plant-based- and Tea diet.

Conclusions: Three dietary patterns (Meat and Sweets, Traditional, and Plant-based- and Tea) were found by repeated clustering of randomly sampled homogeneous cohorts of individuals. Diet preferences depended

*Correspondence: ase.m.moe@uit.no

[1] Department of Mathematics and Statistics, UiT The Arctic University of Norway, Tromsø, Norway
Full list of author information is available at the end of the article

significantly on sex, age, education and PAL, showing a more unhealthy dietary pattern with lower age, low education and low PAL.

Keywords: Diet groups, Dietary patterns, FFQ, Hierarchical cluster analysis, Population studies

Background

Nutrition plays a critical role in the well-being and development of all human beings. An unhealthy diet can have severe consequences and contribute to lifestyle related conditions such as obesity, cardiovascular diseases, type 2 diabetes and cancer [1–3]. This requires a strong focus on promoting healthy diets to decrease individual risk of lifestyle diseases [4].

During the last few decades, research focus has switched from studying the health effects related to intake of single foods and nutrients to studying dietary patterns which reflect the combined intake of different foods and nutrients. A plant-based dietary pattern might reduce the risk of certain chronic diseases, while a dietary pattern high in red meet and added sugar, can potentially increase this risk [2, 5, 6]. A focus on the overall diet, rather than single nutrients, have the practical benefit of giving individuals flexibility in adapting to a healthier diet, avoiding strict nutrition advice [7].

Eating patterns vary across age, culture, lifestyle and socioeconomic status. Preferences for a healthier diet have been seen to increase with age [8–10] while diets high in meat and sweets are most prevalent among the younger population [11, 12]. A healthier diet is typically positively associated with lifestyle factors like physical activity level and educational level [11]. Specifically, higher education have been observed to give higher compliance with the national recommendations [13]. Differences in eating habits between different socioeconomic groups can contribute to inequalities in health. Investigation of dietary patterns across age and population groups is therefore important to understand and target preventive health measures.

Data used to study dietary preferences are commonly collected retrospectively by food frequency questionnaires (FFQ), in which individual intake values for a wide range of food items, dishes and beverages can be aggregated into groups of food variables. Such data can be analysed using a variety of statistical methods, see the recent review by Zhao et al. [14]. These methods include investigator-based a priori approaches, which are based on predefined diet quality scores for different food or nutrition items [15]. A posteriori approaches include classical data-driven methods for dimension reduction like principal component analysis, factor analysis and clustering [16–18].

The first aim of this study was to identify and describe dietary patterns in the seventh survey of the Tromsø Study. This is a comprehensive community based health survey situated in the north of Norway, and should reflect a general Nordic population. Dietary data was collected using a validated FFQ and robust dietary pattern analyses were conducted. A second important objective was to investigate whether there were any associations between the intakes of different diet patterns and age, sex, educational level and PAL in this community survey.

Methods

The Tromsø Study is a population-based study conducted in the municipality of Tromsø, Norway. Seven surveys have been conducted between 1974 and 2016 (Tromsø1-Tromsø7) to which total birth cohorts and random population samples have been invited [19].

Study sample

The present study used data from the seventh survey (Tromsø7), conducted in 2015-2016. All inhabitants of Tromsø municipality aged 40 years or above were invited and 21,083 women and men attended (65% attendance). All of the participants received an FFQ to complete on paper and to be returned by postal mail in a pre-paid envelope. The study sample comprised those who answered the FFQ ($n = 15,146$). We excluded those who completed less than 90% of the FFQ ($n = 3,489$), those with missing values on educational level or PAL ($n = 358$) and those with extreme values of total energy intake or total water intake ($n = 400$), see details in Additional file 1. The final study sample then comprised 10,899 participants (Fig. 1).

Measurements

Self-reported educational level was dichotomized as Low (including primary/partly secondary with up to 9 years of schooling, and upper secondary with 10-12 years of schooling) or High (short tertiary with < 4 years college/university and long tertiary with ≥ 4 years college/university). Self-reported leisure-time PAL was dichotomized as Low or High according to the validated Saltin-Grimby questionnaire [20]. Specifically, the category of low PAL here includes both sedentary and light exercise, ranging from those who are almost completely inactive to those who do light physical activity at least 4 hours a week (Level 1 and 2 of the original Saltin-Grimby scale). High PAL ranges from regular and moderate training at least 4 hours a week to vigorous hard training for

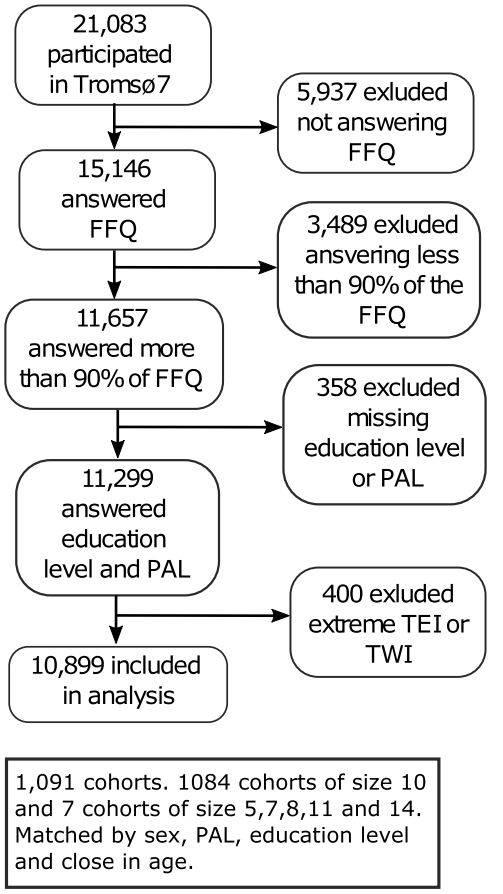


Fig. 1 Flowchart of the study sample. FFQ: food frequency questionnaire

competitions (level 3 and 4 of the original Saltin-Grimby scale).

The FFQ, validated by Carlsen et al. [21], included questions about frequency and amount of intake of 244 food items, dishes and beverages. Calculation of total energy- (TEI), total water- (TWI), food- and nutrient intakes in kilojoule (kJ), and grams (g) per day, respectively, was performed using the food composition database and nutrient calculation system KBS at the University of Oslo, database AE14 (based on the Norwegian food composition tables 2014 and 2015), software version 7.3 [22].

We aggregated the intakes from the original 244 food items into 33 new variables plus one variable not used in this study. The aggregation was done in a supervised manner, each variable representing the total intake for groups of solid food and beverages (see Table S1, Additional file 2). This type of aggregation is in coherence with literature, in which questions of FFQs are typically summarized by 30-60 variables, see e.g. Karageorgou et al. [17].

Preprocessing: Scaled intake values and partition of the study sample in cohorts

To identify similarities in dietary intake and the composition of food preferences, each food variable was scaled according to the individual energy intake. The scaled intake values were calculated by

$$\text{Food}^*_{ji} = \text{Food}_{ji}\frac{\overline{\text{TEI}}}{\text{TEI}_i}$$

where Food^*_{ji} and Food_{ji} represent the scaled and unscaled intake of food variable j for individual i, respectively. The scaling factor divides the total mean intake for all individuals ($\overline{\text{TEI}}$), with the total energy intake for individual i (TEI_i).

Although we excluded participants with extreme values of intake, some participants still had remarkably high scaled intake values for some of the food variables. Due to inherent uncertainty of FFQ data on the individual level, it has been recommended that FFQ data are used on group level [23]. For clustering food variables, we therefore chose to divide the study sample in cohorts of approximately 10 participants, resulting in a total of 1,091 cohorts. The partition in cohorts was performed by random sampling of the participants, under the constraint that the cohorts should be homogeneous with respect to background variables. This implied that the participants assigned to each cohort were matched by sex, similar age, educational level (low/high) and PAL (low/high). The intake values for the different food variables were then averaged within each of the cohorts, reducing the effect of extreme and possible erroneous values.

Statistical analysis

An hierarchical agglomerative clustering method was applied to group similar food variables in non-overlapping clusters. The pairwise distances between the 33 aggregated food variables were measured by Spearman's rank correlation. The food variables were then merged according to the complete linkage method into clusters. The cluster analysis was repeated for 100 different random samplings of a total of 1091 cohorts. The final diet groups were based on these 100 repetitions, where each food variable was assigned to the diet group in which it occurred most often. This was done to assess the variability of the cluster results and increase the validity of the resulting diet groups [24]. Also, the random sampling of

cohorts was used to decide on the number of clusters giving the most stable classification result.

After establishing dietary patterns by cluster analysis, we calculated individual diet intake scores. The score was found by first standardizing the individual scaled intake values ($Food^*_{ji}$) for each food variable *j*. This standardization was done to weight the intake of all food variables equally. For each individual, we then calculated the average of the standardized food variables within each diet group. This made the final intake scores comparable across diets. Note that these individual diet intake scores reflect diet preferences in the sense that larger values imply larger intake of the specific diet. Also, the sum over all individual scores within each diets is 0.

The scores for each diet were then modelled as the dependent variable in regression models on age, including educational level (low/high) and PAL (low/high) as categorical covariates. The association between the diet scores and age was modelled either using a linear or non-linear trend, in which the non-linear trend was modelled using natural cubic splines. Different models were evaluated using the adjusted coefficient of determination (R^2_{adj}). Subsequent analyses included analysis of covariance and post hoc pairwise testing of diet scores between groups having different combinations of the two categorical variables. Specifically, this included comparison of groups characterized by having low education and low PAL ($L_{edu}L_{PAL}$), low education and high PAL ($L_{edu}H_{PAL}$), high education and low PAL ($H_{edu}L_{PAL}$), and high education and high PAL ($H_{edu}H_{PAL}$). The analysis was performed separately for men and women and p-values were adjusted using Tukey's method, ensuring a global significance level of 0.05.

Results

Characteristics of the study sample and food intake

The study sample consisted of 5807 women and 5092 men (Table 1). Slightly more than half (53.1%) had high education. This percentage was significantly different for the age groups below and above 60 years, for both women and men ($p < 0.001$ in both). In total 28.3% reported high PAL. The percentage of high PAL was lower in women (22.3%) compared to men (35.1%, $p < 0.001$). Also, the percentage of high PAL decreased with age for both sexes ($p < 0.001$ in both).

The 33 aggregated food variables comprised 7 beverage variables and 26 variables representing solid food items (see Table S2, Additional file 2). Most of the beverage variables were recorded with a maximum intake of more than 3000 g/day, and the largest mean and median scaled intake values were reported for Coffee, Water and Milk. Among the solid food variables, the highest mean and median values of intake were seen for Vegetables, Fruit, Bread, Meat Dinner and Potato, followed by Fish Dinner and Composite Dinner Dishes (in decreasing order). Almost all of the respondents (98-99%) reported an intake for these variables, while about 50% - 60% reported an intake of food variables like Mayonnaise and Plant-based Oils, Chips, Candy and Breakfast Cereals.

Table 1 Descriptive statistics for age, educational level and physical activity level (PAL), by sex

			All	Women	Men
N			10899	5807	5092
Age (sd)			57.1 (10.7)	56.5 (10.5)	57.9 (10.9)
Educational level, n(%):					
	Low (L_{edu})		5114 (46.9%)	2649 (45.6%)	2465 (48.4%)
	High (H_{edu})		5785 (53.1%)	3158 (54.4%)	2627 (51.6%)
		40 — 59 years old		2302 (64.9%)	1600 (56.8%)
		60 years old and above		856 (37.9%)	1027 (41.5%)
PAL, n(%):					
	Low (L_{PAL})		7818 (71.7%)	4511 (77.7%)	3307 (64.9%)
	High (H_{PAL})		3081 (28.3%)	1296 (22.3%)	1785 (35.1%)
		40 — 59 years old		923 (26.0%)	1120 (39.8%)
		60 years old and above		373 (16.5%)	665 (29.2%)
Education and PAL groups, n (%):					
	$L_{edu}L_{PAL}$		3971 (36.4%)	2234 (38.5%)	1737 (34.1%)
	$L_{edu}H_{PAL}$		1143 (10.5%)	415 (7.1%)	728 (14.3%)
	$H_{edu}L_{PAL}$		3847 (35.3%)	2277 (39.2%)	1570 (30.8%)
	$H_{edu}H_{PAL}$		1938 (17.8%)	881 (15.2%)	1057 (20.8%)

Identification of diet groups by cluster analysis

The cluster analyses identified three main diet groups (Table 2). These dietary patterns are referred to as: The Meat and Sweets diet, including Composite Dinner Dishes, Meat-spread, Meat Dinner, and a variety of sweets; The Traditional diet, including Bread, Fish-spread, Fish Dinner, Milk and Potato; The Plant-based- and Tea diet, including Cereals, Fruit, Nuts, Tea, Vegetables and dairy products.

The given dietary patterns were based on observing the number of times each food variable was clustered to each of the three diet groups, choosing the diet group having the maximum frequency of the 100 random samplings (see Table S3, Additional file 2). A total of five food variables (Beverages with Alcohol, Butter and Margarine, Egg, Juice and Water) were not clearly classified to one of the given three diet groups as these variables frequently switched between clusters in repeated analyses. All but four of the food variables (Beverages with Alcohol, Butter and Margarine, Milk/sugar for Coffee/tea and Water) were clustered to the same diet group for men and women.

Identifying dietary patterns across age, level of education and PAL

To make the diet scores comparable for men and women in subsequent analysis, these were calculated using only the food variables classified to the same diet group for both sexes. The average scores for the Meat and Sweets diet were approximately zero for both women (sd $= 0.42$) and men (sd $= 0.39$). For the Traditional diet, the average score for women was equal to -0.05 (sd $= 0.33$). The corresponding score for men was 0.05 (sd $= 0.34$), giving a significant difference between the sexes ($p < 0.001$). Women had a higher intake of the Plant-based- and Tea diet with an average score of 0.13 (sd $= 0.48$), while the corresponding score was equal to -0.22 (sd $= 0.37$) for men ($p < 0.001$). The correlation between the scores for the Meat and Sweets diet and the Traditional diet was -0.36. The corresponding correlations in scores for the Plant-based- and Tea- diet, compared with the other two diets, were both -0.27.

For both women and men, the diet scores for the Meat and Sweets diet showed a significantly decreasing linear trend as function of age ($p < 0.001$), see Fig. 2. In contrast, the scores for the Traditional diet were observed to have a significant positive linear trend as function of age ($p < 0.001$). For the Plant-based- and Tea diet, we chose to model the association between the diet scores and age non-linearly, as this gave a clear relative increase in the adjusted coefficient of determination for women (see Table S4, Additional file 2). For women, the estimated non-linear functions were seen to increase until the approximate age of 60, and then decrease for higher ages. For men, the average score for the Plant-based- and Tea diet increased slightly with age.

The given results were calculated without including interaction terms in the regression models as these did not increase the adjusted coefficient of determination for the different models (see Table S4, Additional File 2). Specifically, interaction between education and PAL was non-significant in all cases. Thus, the estimated linear and non-linear curves of diet scores as function of age were parallel for the four groups $L_{edu}L_{PAL}$, $L_{edu}H_{PAL}$, $H_{edu}L_{PAL}$ and $H_{edu}H_{PAL}$. However, the mean levels of these curves were shifted vertically, indicating potential significant differences between these groups in terms of the intakes of the different diets.

To further investigate differences in dietary preferences between groups, we performed four pairwise hypotheses tests for differences in the mean diet score. This was repeated for all three diets and both sexes, resulting in a total of 24 tests (Table 3). The group with lower education (L_{edu}) showed a significantly higher intake of the Meat and Sweets and the Traditional diets than those with high education (H_{edu}), and also

Table 2 Diet groups from cluster analysis of food variables

	Meat and Sweets	Traditional	Plant-based- and Tea	Inconclusive*
All	Candy, Chips, Chocolate, Soft Drinks, Composite Dinner Dishes, Meat-spread, Mayonnaise and Plant-based Oils, Meat Dinner, Rice/pasta, Sauce etc.	Bread, Cakes and Pastries, Breakfast Cereals (Sweetened), Coffee, Dessert, Fish Dinner, Fish-spread , Milk, Potato, Jam	Cheese, Breakfast Cereals and Porridge (Unsweetened), Fruit, Nuts, Tea, Vegetables, Yoghurt	Egg, Juice
Only men	Water	Milk/sugar for Coffee/tea		Beverages with Alcohol, Butter and Margarine
Only women		Butter and Margarine	Beverages with Alcohol, Milk/sugar for Coffee/tea	Water

* Inconclusive include variables not clearly defined to the either of the three main diet groups

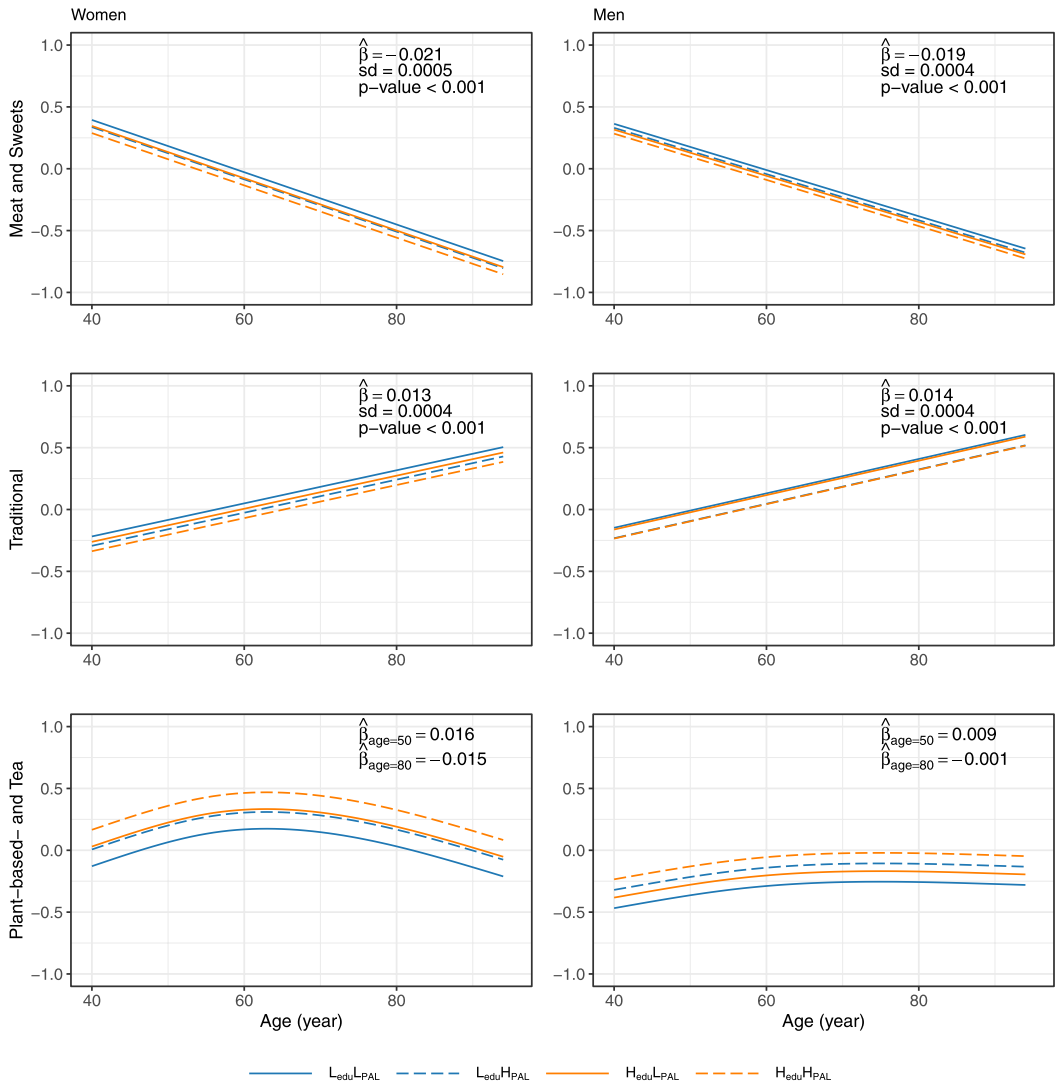


Fig. 2 Diet scores for each of the three dietary patterns from the cluster analysis, by age. Estimated trends of diet scores as a function of age, based on data from 5807 women (left panels) and 5092 men (right panels). Four lines are included, where educational level is low(blue)/high (orange) and PAL is low(unbroken lines)/high (dashed lines). For the linear models, the figures include the estimated slopes ($\hat{\beta}$), standard deviations (sd) and the p-value evaluating whether the slope is significantly different from 0. For the non-linear models, the figures include the slope estimates at the age of 50 and 80

a significantly lower intake of the Plant-based- and Tea diet for both sexes. Similar results were found in comparing the groups with low (L_{PAL}) versus high PAL (H_{PAL}), except that no significant difference was found for the Traditional diet among men. Naturally, these differences in diet preferences are most clear in comparing the groups $L_{edu}L_{PAL}$ with $H_{edu}H_{PAL}$. Among men, the group $L_{edu}H_{PAL}$ had a significant higher intake of the Traditional diet and a significant lower intake of the Plant-based- and Tea diet, than $H_{edu}L_{PAL}$.

Table 3 Comparison of differences (sd) in diet score between education (low/high) and PAL groups (low/high)

	Meat and sweets		Traditional		Plant-based- and Tea	
Comparisons	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**
$L_{edu} - H_{edu}$	0.06∗ (0.01)	0.03∗ (0.01)	0.08∗ (0.01)	0.08∗ (0.01)	−0.14∗ (0.01)	−0.15∗ (0.01)
$L_{PAL} - H_{PAL}$	0.05∗ (0.01)	0.05∗ (0.01)	0.04∗ (0.01)	0.01 (0.01)	−0.16∗ (0.01)	−0.09∗ (0.01)
$L_{edu}L_{PAL} - H_{edu}H_{PAL}$	0.11∗ (0.01)	0.08∗ (0.01)	0.12∗ (0.01)	0.09∗ (0.01)	−0.29∗ (0.02)	−0.23∗ (0.01)
$L_{edu}H_{PAL} - H_{edu}L_{PAL}$	0.01 (0.02)	−0.01(0.01)	0.03 (0.01)	0.07∗ (0.01)	0.02 (0.02)	−0.06∗ (0.02)

* $p < 0.05$ by Tukey's method

Discussion

The given cross-sectional study in a general Nordic population identified three diet groups, referred to as the Meat and Sweets diet, the Traditional diet, and the Plant-based- and Tea diet. In both men and women, the diet score reflecting the intake of the Meat and Sweets diet showed a significant negative trend with increasing age. The opposite trend was seen for the Traditional diet, in which the scaled intake increased as a function of age. Women had a lower consumption of the Traditional diet and a higher consumption of the Plant-based- and Tea diet than men. Diet preferences were significantly dependent on education level and PAL. Participants with high education had lower diet scores for the Meet and Sweets and the Traditional diets and higher scores for the Plant-based- and Tea diet compared to participants with low education. Participants with high PAL had lower diet scores of the Meat and Sweets diet and the Traditional diet (only women), and higher diet scores of the Plant-based- and Tea diet, compared to participants with low PAL.

The three diet groups identified in this population have similarities to the western, traditional and prudent diets [17, 18, 25]. The Meat and Sweets diet group is similar to the western dietary pattern, characterized by high intakes of red and processed meat (here represented by Composite Dinner Dishes, Meat-spread and Meat Dinner), high intake of sugar drinks (here Soft Drinks) and sweets like Candy and Chocolate. The derived Traditional diet was similar to traditional patterns found in Scandinavia and contained food variables like Bread, Fish-spread, Fish Dinner, Milk and Potato [11, 12, 25]. The Plant-based- and Tea diet group contained some of the food variables being characteristic of a prudent or healthy dietary pattern, including Fruit, Nuts and Vegetables. The similarities in the clustering results for men and women are in accordance with other studies [12, 17].

The age-related differences in diet preferences are in accordance with previous findings where a western diet has shown to be more common among younger ages [17, 26, 27], while more traditional- and prudent diet scores either increase with age or no significant association is found [11, 17, 26–28]. Unless younger individuals gradually change their diet to consuming more traditional foods as they grow older, these findings imply a transition from the Traditional to the Meat and Sweets dietary pattern in the population. However, in this study, we only have data for the population above 40 years of age. Unfavourable developments of diet habits have been observed in the Nordic population between 2011 and 2014, with a decrease in the consumption of fish and whole grains [29]. On the other hand, changes to more healthy dietary patterns by increasing age have been observed for groups with higher education [8–10]. The nutrition transition hypothesis of global dietary convergence to a western diet has been tested and rejected in a study by Azzam [30] over the period 1993-2013, where some countries changed from having a low Western Diet Index (traditional diet) to a high Western Diet Index, and countries previously characterized as having a high Western Diet Index had declined towards a traditional diet during the same period. National health policies on speeding up a transition to healthier dietary patterns in countries with high Western Diet Index, may explain this phenomenon.

The observed dependence between dietary preference and educational level is in accordance with the findings in Nilsen et al. [31]. In this population, a previous study found that participants with higher education had higher odds of nutrition intake in accordance with the Nordic nutrition recommendations [31]. Similar trends have been observed in other populations. For instance, in a study of dietary patterns in Australians aged 55-65 years, high consumption of red- and processed meet and refined grains were associated with low education and low PAL [32]. On the other hand, for different ethnic groups in Netherlands, lower education was not necessarily associated with a significant poorer diet [33].

In studies on food prices and socioeconomic status (SES), Rydén and Hagfors [34] found that healthy diets are

more expensive than unhealthy diets in Sweden. Further, children of parents with low education and manual low-skill occupations had the cheapest and most unhealthy diets. It has also been observed that tax and subsidies on unhealthy and healthy food, respectively, improves healthy eating in lower SES [35]. In a Norwegian study on vegetables and fish [36], persons with lower education reported less knowledge in how to prepare these food items compared with persons with higher education. Also, persons with higher education ate significantly more vegetables and fish than persons with lower education. Other findings of the role of education on diet choices are reviewed by Pampel et al. [37]. They summarize research on how problem-solving skills from schooling help making healthier choices and other related questions.

In a German study [38], adults with lower education reported that they consumed energy-dense foods more frequently then adults with higher education. Higher levels of physical work activity among adults with lower education may partly explain why they consume more energy-dense foods. This can also be the case for persons who are physically active in their leisure time. On the other hand, these individuals can be more health-conscious, a factor that also affects food choice. Joo et al. [39] found indication of exercise as a motivation for healthier diets among young adults. Dependent on type of exercise, it was observed that exercise reduced the preference for the western diet and increased the preference for the prudent pattern.

Major strengths of the given analyses include the high Tromsø7 study attendance, and the use of an extensive FFQ to capture the full habitual diet. Further, the diet preferences between groups are investigated using age as a continuous variable. Analyses of diet preferences are often performed by stratifying participants in terms of age groups. This might result in loss of information as the influence of age is assumed constant for participants within a certain age span, e.g. 10 years [40].

Collection of self-reported food intake by FFQs is influenced by several sources of error, including missing data. In the given study, missing data were originally registered as zero, making it impossible to separate zero-intake values from missing data. This limitation of the data material was to a large degree reduced by excluding participants answering less than 90% of the FFQ. The most extreme intake values were excluded in accordance with Lundblad et al. [22], taking into account that intake depend on sex, PAL and body composition.

Conclusions

The given cluster analysis identified three dietary patterns, referred to as the Meat and Sweets diet, the Traditional diet, and the Plant-based- and Tea diet. Preference for the Meat and Sweets diet and the Traditional diet showed significant negative and positive trends as a function of age, respectively. Adjusting for age, a more unhealthy dietary pattern was associated with low educational level and low PAL.

Abbreviations

Tromsø7: The seventh survey of the Tromsø Study; FFQ: Food frequency questionnaire;; TEI: Total energy intake; TWI: Total water intake; PAL: Physical activity level; $L_{edu}L_{PAL}$: low education and low PAL; $L_{edu}H_{PAL}$: low education and high PAL; $H_{edu}L_{PAL}$: high education and low PAL; $H_{edu}H_{PAL}$: high education and high PAL; KJ: Kilojoule; g: gram; KBS: Kostberegningssystemet (food composition database and nutrient calculation system); SES: Socioeconomic status.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40795-022-00599-4.

Additional file 1: Appendix.

Additional file 2: Table S1-S4.

Acknowledgements

We would like to thank all Tromsø Study participants for their patience.

Authors' Contributions

ÅMM performed the statistical analysis, contributed to the study design and interpretation of the results. EY and SHS contributed to the study design and interpretation of the results. LAH and MHC contributed to the collection and preprocessing of data, interpretation of the results and provided knowledge on the data material. OL contributed to the study design. All authors have contributed to the manuscript, and read and approved the final version.

Funding

Open access funding provided by UiT The Arctic University of Norway (incl University Hospital of North Norway). The main funding sources for the data collection in Tromsø7 were UiT The Arctic University of Norway, North Norwegian Regional Health Authority, Norwegian Ministry of Health and Care Services, University Hospital of North Norway, and Troms County.

Availability of data and materials

The dataset analysed in the current study are obtained from a third party (the Tromsø Study) and are not publicly available. There are legal restrictions set on data availability in order to control for data sharing, including publication of datasets with the potential of reverse identification of de-identified sensitive study participant information. The data can, however, be made available from the Tromsø Study for bona fide researchers upon application to the Tromsø Study Data and Publication Committee. Contact information: The Tromsø Study, Department of Community Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway; e-mail: tromsous@uit.no. Detailed instructions on how to apply are given at the webpage of the Tromsø Study https://uit.no/research/tromsostudy.

Declarations

Ethics approval and consent to participate

The seventh survey of the Tromsø Study was approved by The Regional Committee of Medical and Health Research Ethics (REC) North (reference 2014/940) and the Norwegian Data Protection Authority (reference 14/01463-4/CGN). All procedures performed were in accordance with the 1964 Declaration of Helsinki and its later amendments. All participants gave written informed consent. This study was approved by REK Nord (reference 2019/1137) and the Norwegian Data Protection Authority.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Mathematics and Statistics, UiT The Arctic University of Norway, Tromsø, Norway. [2]Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway. [3]Division of Nutritional Epidemiology, University of Oslo, Oslo, Norway.

Received: 25 March 2022 Accepted: 6 September 2022
Published online: 15 September 2022

References

1. Seifu CN, Fahey PP, Hailemariam TG, Frost SA, Atlantis E. Dietary patterns associated with obesity outcomes in adults: an umbrella review of systematic reviews. Public Health Nutr. 2021;24(18):6390–414. https://doi.org/10.1017/S1368980021000823.
2. Medina-Remón A, Kirwan R, Lamuela-Raventós RM, Estruch R. Dietary patterns and the risk of obesity, type 2 diabetes mellitus, cardiovascular diseases, asthma, and neurodegenerative diseases. Crit Rev Food Sci Nutr. 2018;58(2):262–96. https://doi.org/10.1080/10408398.2016.1158690.
3. Tabung FK, Brown LS, Fung TT. Dietary Patterns and Colorectal Cancer Risk: a Review of 17 Years of Evidence (2000–2016). Curr Colorectal Cancer Rep. 2017;13(6):440–54. https://doi.org/10.1007/s11888-017-0390-5.
4. Thow AM, Downs SM, Mayes C, Trevena H, Waqanivalu T, Cawley J. Fiscal policy to improve diets and prevent noncommunicable diseases: from recommendations to action. Bull World Health Organ. 2018;96(3):201–10. https://doi.org/10.2471/blt.17.195982.
5. Heidemann C, Schulze MB, Franco OH, Van Dam RM, Mantzoros CS, Hu FB. Dietary Patterns and Risk of Mortality From Cardiovascular Disease, Cancer, and All Causes in a Prospective Cohort of Women. Circulation. 2008;118(3):230–7. https://doi.org/10.1161/circulationaha.108.771881.
6. Drake I, Sonestedt E, Ericson U, Wallström P, Orho-Melander M. A Western dietary pattern is prospectively associated with cardio-metabolic traits and incidence of the metabolic syndrome. Br J Nutr. 2018;119(10):1168–76. https://doi.org/10.1017/S000711451800079X.
7. Mozaffarian D. Dietary and Policy Priorities for Cardiovascular Disease, Diabetes, and Obesity. Circulation. 2016;133(2):187–225. https://doi.org/10.1161/circulationaha.115.018585.
8. Harrington JM, Dahly DL, Fitzgerald AP, Gilthorpe MS, Perry IJ. Capturing changes in dietary patterns among older adults: a latent class analysis of an ageing Irish cohort. Public Health Nutr. 2014;17(12):2674–86. https://doi.org/10.1017/S1368980014000111.
9. Thorpe MG, Milte CM, Crawford D, McNaughton SA. Education and lifestyle predict change in dietary patterns and diet quality of adults 55 years and over. Nutr J. 2019;18(1):67. https://doi.org/10.1186/s12937-019-0495-6.
10. Kristal AR, Hedderson MM, Patterson RE, Neuhauser ML. Predictors of Self-initiated, Healthful Dietary Change. J Am Diet Assoc. 2001;101(7):762–6. https://doi.org/10.1016/S0002-8223(01)00191-2.
11. Knudsen VK, Matthiessen J, Biltoft-Jensen A, Sørensen MR, Groth MV, Trolle E, et al. Identifying dietary patterns and associated health-related lifestyle factors in the adult Danish population. Eur J Clin Nutr. 2014;68(6):736–40. https://doi.org/10.1038/ejcn.2014.38.
12. Petrenya N, Rylander C, Brustad M. Dietary patterns of adults and their associations with Sami ethnicity, sociodemographic factors, and lifestyle factors in a rural multiethnic population of northern Norway - the SAMINOR 2 clinicalsurvey. BMC Public Health. 2019;19(1632):72–83. https://doi.org/10.1186/s12889-019-7776-z.
13. Darmon N, Drewnowski A. Does social class predict diet quality? Am J Clin Nutr. 2008;87(5):1107–17. https://doi.org/10.1093/ajcn/87.5.1107. https://academic.oup.com/ajcn/article-pdf/87/5/1107/23918937/znu00508001107.pdf.
14. Zhao J, Li Z, Gao Q, Zhao H, Chen S, Huang L, et al. A review of statistical methods for dietary pattern analysis. Nutr J. 2021;20(1). https://doi.org/10.1186/s12937-021-00692-7.
15. Fransen HP, Ocké MC. Indices of diet quality. Curr Opin Clin Nutr Metab Care. 2008;11(5):559–65. https://doi.org/10.1097/MCO.0b013e32830a49db.
16. Newby PK, Tucker KL. Empirically Derived Eating Patterns Using Factor or Cluster Analysis: A Review. Nutr Rev. 2004;62(5):177–203. https://doi.org/10.1111/j.1753-4887.2004.tb00040.x.
17. Karageorgou D, Magriplis E, Mitsopoulou AV, Dimakopoulos I, Bakogianni I, Micha R, et al. Dietary patterns and lifestyle characteristics in adults: results from the Hellenic National Nutrition and Health Survey (HNNHS). Public Health. 2019;171:76–88. https://doi.org/10.1016/j.puhe.2019.03.013.
18. Mosalmanzadeh N, Jandari S, Soleimani D, Shadmand Foumani Moghadam MR, Khorramrouz F, Araste A, et al. Major dietary patterns and food groups in relation to rheumatoid arthritis in newly diagnosed patients. Food Sci Nutr. 2020;8(12):6477–86. https://doi.org/10.1002/fsn3.1938.
19. Jacobsen BK, Eggen AE, Mathiesen EB, Wilsgaard T, Njølstad I. Cohort profile: The Tromsø Study. Int J Epidemiol. 2012;41(4):961–7. https://doi.org/10.1093/ije/dyr049.
20. Grimby G, Börjesson M, Jonsdottir IH, Schnohr P, Thelle DS, Saltin B. The Saltin-Grimby Physical Activity Level Scale and its application to health research. Scand J Med Sci Sports. 2015;25(S4):119–25. https://doi.org/10.1111/sms.12611.
21. Carlsen MH, Lillegaard ITL, 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):37. https://doi.org/10.1186/1475-2891-9-37.
22. Lundblad MW, Andersen LF, Jacobsen BK, Carlsen MH, Hjartåker A, Grimsgaard S, et al. Energy and nutrient intakes in relation to National Nutrition Recommendations in a Norwegian population-based sample: the Tromsø Study 2015–16. Food Nutr Res. 2019;63(0). https://doi.org/10.29219/fnr.v63.3616.
23. Slimani N, Freisling H, Illner AK, Huybrechts I. Methods to Determine Dietary Intake. In: Nutrition Research Methodologies. John Wiley & Sons, Ltd; 2015. p. 48–70. https://doi.org/10.1002/9781119180425.ch4.
24. Funtikova AN, Benítez-Arciniega AA, Fitó M, Schröder H. Modest validity and fair reproducibility of dietary patterns derived by cluster analysis. Nutr Res. 2015;35(3):265–8. https://doi.org/10.1016/j.nutres.2014.12.011.
25. Engeset D, Alsaker E, Ciampi A, Lund E. Dietary patterns and lifestyle factors in the Norwegian EPIC cohort: The Norwegian Women and Cancer (NOWAC) study. Eur J Clin Nutr. 2005;59(5):675–84. https://doi.org/10.1038/sj.ejcn.1602129.
26. Gazan R, Béchaux C, Crépet A, Sirot V, Drouillet-Pinard P, Dubuisson C, et al. Dietary patterns in the French adult population: a study from the second French national cross-sectional dietary survey (INCA2) (2006–2007). Br J Nutr. 2016;116(2):300–15. https://doi.org/10.1017/S0007114516001549.
27. Krieger JP, Pestoni G, Cabaset S, Brombach C, Sych J, Schader C, et al. Dietary Patterns and Their Sociodemographic and Lifestyle Determinants in Switzerland: Results from the National Nutrition Survey menuCH. Nutrients. 2018;11(1):62. https://doi.org/10.3390/nu11010062.
28. Kant AK. Dietary patterns and health outcomes. J Am Diet Assoc. 2004;104(4):615–35. https://doi.org/10.1016/j.jada.2004.01.010.
29. Matthiessen J, Andersen LF, Barbieri HE, Borodulin K, Knudsen VK, Kørup K, et al. The Nordic Monitoring System 2011-2014: status and development of diet, physical activity, smoking, alcohol and overweight. Copenhagen: Nordic Council of Ministers; 2016. https://doi.org/10.6027/TN2016-561.
30. Azzam A. Is the world converging to a 'Western diet'? Public Health Nutr. 2021;24(2):309–17. https://doi.org/10.1017/s136898002000350x.
31. Nilsen L, Hopstock LA, Skeie G, Grimsgaard S, Lundblad MW. The Educational Gradient in Intake of Energy and Macronutrients in the General Adult and Elderly Population: The Tromsø Study 2015–2016. Nutrients. 2021;13(2). https://doi.org/10.3390/nu13020405.
32. Thorpe MG, Milte CM, Crawford D, Mcnaughton SA. A comparison of the dietary patterns derived by principal component analysis and cluster analysis in older Australians. Int J Behav Nutr Phys Act. 2016;13(1):30. https://doi.org/10.1186/s12966-016-0353-2.
33. Yau A, Adams J, White M, Nicolaou M. Differences in diet quality and socioeconomic patterning of diet quality across ethnic groups:

cross-sectional data from the HELIUS Dietary Patterns study. Eur J Clin Nutr. 2020;74(3):387–96. https://doi.org/10.1038/s41430-019-0463-4.
34. Rydén PJ, Hagfors L. Diet cost, diet quality and socio-economic position: how are they related and what contributes to differences in diet costs? Public Health Nutr. 2011;14(9):1680–92. https://doi.org/10.1017/s1368980010003642.
35. Mcgill R, Anwar E, Orton L, Bromley H, Lloyd-Williams F, O'Flaherty M, et al. Are interventions to promote healthy eating equally effective for all? Systematic review of socioeconomic inequalities in impact. BMC Public Health. 2015;15(1). https://doi.org/10.1186/s12889-015-1781-7.
36. Skuland SE. Healthy Eating and Barriers Related to Social Class. The case of vegetable and fish consumption in Norway. Appetite. 2015;92:217–26. https://doi.org/10.1016/j.appet.2015.05.008.
37. Pampel FC, Krueger PM, Denney JT. Socioeconomic Disparities in Health Behaviors. Annu Rev Sociol. 2010;36(1):349–70. https://doi.org/10.1146/annurev.soc.012809.102529.
38. Finger JD, Tylleskär T, Lampert T, Mensink GBM. Dietary Behaviour and Socioeconomic Position: The Role of Physical Activity Patterns. PLoS ONE. 2013;8(11): e78390. https://doi.org/10.1371/journal.pone.0078390.
39. Joo J, Williamson SA, Vazquez AI, Fernandez JR, Bray MS. The influence of 15-week exercise training on dietary patterns among young adults. Int J Obes. 2019;43(9):1681–90. https://doi.org/10.1038/s41366-018-0299-3.
40. Van Walraven C, Hart RG. Leave 'em Alone - Why Continuous Variables Should Be Analyzed as Such. Neuroepidemiology. 2008;30(3):138–9.

Publisher's Note

Appendix

As over- or underreporting of total food intake is common [1], we chose to exclude data showing extreme intake values. Several methods have been suggested to identify such extreme intakes [4, 2, 3]. Here, we fitted regression models to identify extreme absolute values of residuals. Specifically, we used the total intake values, TEI and TWI, as response variables in separate regression models, as these typically reflect extreme values of the individual food and beverage intake, respectively. These variables were log-transformed to give approximate Gaussian distributions.

The variables TEI and TWI have been seen to depend on body size and composition, PAL, sex and age. We therefore modelled TEI and TWI using height, weight, PAL and age as explanatory variables in the regression models, in addition to sex. Missing values of height and weight were imputed using their respective sex-specific average values. The regression models were fitted using ordinary least squares, providing estimates for the residuals. For each of the regression models, we excluded all cases corresponding to the two percent highest absolute values of the residuals, resulting in 400 participants having an unrealistic food intake.

References

[1] M.H. Carlsen, I.T.L. Lillegaard, A. Karlsen, R. Blomhoff, C.A. Drevon, and L.F. Andersen. "Evaluation of energy and dietary intake estimates from a food frequency questionnaire using independent energy expenditure measurement and weighed food records". In: *Nutrition Journal* 9.1 (2010), p. 37. DOI: 10.1186/1475-2891-9-37.

[2] G.R. Goldberg, A.E. Black, S.A. Jebb, T.J. Cole, P.R. Murgatroyd, W.A. Coward, and A.M. Prentice. "Critical evaluation of energy intake data using fundamental principles of energy physiology: 1. Derivation of cut-off limits to identify under-recording". In: *European journal of clinical nutrition* 45.12 (1991), pp. 569–581.

[3] Marie W. Lundblad, Lene Frost Andersen, Bjarne K. Jacobsen, Monica Hauger Carlsen, Anette Hjartåker, Sameline Grimsgaard, and Laila A. Hopstock. "Energy and nutrient intakes in relation to National Nutrition Recommendations in a Norwegian population-based sample: the Tromsø Study 2015–16". In: *Food & Nutrition Research* 63.0 (2019). ISSN: 1654-661X. DOI: 10.29219/fnr.v63.3616.

[4] J.J. Rhee, L. Sampson, E. Cho, M.D. Hughes, F.B. Hu, and W.C. Willett. "Comparison of Methods to Account for Implausible Reporting of Energy Intake in Epidemiologic Studies". In: *American Journal of Epidemiology* 181.4 (2014), pp. 225–233. DOI: 10.1093/aje/kwu308.

Table S1-S4

Table S1: The 244 food items included in the FFQ, aggregated into 33 new food variables plus one variable not used in this study. The food variables represent groups of solid food and beverages.

Groups	Variables
Bread (nr. 1)	FOOD_BREAD_WHITE_T7, FOOD_BREAD_WHOLEGRAIN50_T7, FOOD_BREAD_WHOLEGRAIN100_T7, FOOD_CRISPBREAD_WHITE_T7, FOOD_CRISPBREAD_WHOLEGRAIN_T7
Butter and Margarine (nr. 2)	FOOD_BUTTER_B_T7, FOOD_MARGARINE_BREMYKT_B_T7, FOOD_MARGARINE_BRELETT_B_T7, FOOD_MARGARINE_SOFT_SFSE_B_T7, FOOD_MARGARINE_VITA_B_T7, FOOD_MARGARINE_SFV_ LIGHT_B_T7, FOOD_MARGARINE_MELANGE_B_T7, FOOD_MARGARINE_OTHER_B_T7
Mayonnaise and Plant-based Oils (nr. 3)	FOOD_OIL_OLIVE_OTHER_B_T7, FOOD_MAYONNAISE_B_T7, FOOD_S_MAYONNAISE_SALAD_T7, FOOD_S_MAYONNAISE_SAL_LIGHT_T7
Cheese (nr. 4)	FOOD_S_CHEESE_WHEY_T7, FOOD_S_CHEESE_WHEY_LIGHT_T7, FOOD_S_CHEESE_WHITE_T7, FOOD_S_CHEESE_WHITE_LIGHT_T7, FOOD_S_CHEESE_BLUE_DESSERT_T7, FOOD_S_CHEESE_SOFT_T7, FOOD_S_CHEESE_SOFT_LIGHT_T7, FOOD_S_COTTAGECHEESE_T7
Meat-spread (nr. 5)	FOOD_S_LIVERPASTE_T7, FOOD_S_LIVERPASTE_LIGHT_T7, FOOD_S_SERVELAT_T7, FOOD_S_HAM_BOILED_T7, FOOD_S_SALAMI_T7
Fish-spread (nr. 6)	FOOD_S_CAVIARSPREAD_T7, FOOD_S_CAVIARSPREAD_SVOLVAR_T7, FOOD_S_MACKEREL_TOMATOSAUCE_T7, FOOD_S_SALMON_TROUT_SMOKED_T7, FOOD_S_SARDINES_HERRING_T7, FOOD_S_TUNA_T7, FOOD_S_SCHRIMP_CRAB_T7
Egg (nr. 7)	FOOD_S_EGG_T7
Jam (nr. 8)	FOOD_S_JAM_MARMELADE_T7, FOOD_S_JAM_LIGHT_T7, FOOD_S_PEANUTBUTTER_T7, FOOD_S_CHOCOLATE_NUT_SPREAD_T7, FOOD_S_SWEET_SPREAD_T7

Groups	Variables
Breakfast Cereals and Porridge (Unsweetened) (nr. 9)	FOOD_PORRIDGE_OATMEAL_T7, FOOD_OATMEAL_4GRAIN_T7, FOOD_CEREAL_UNSWEETENED_T7, FOOD_CEREAL_ALLBRAN_T7
Breakfast Cereals - Sweetened (nr. 10)	FOOD_CEREAL_SWEETENED_T7, FOOD_CORNFLAKES_T7, FOOD_CEREAL_HONEY_T7, FOOD_CEREAL_PUFFED_RICE_OAT_T7, FOOD_JAM_CEREAL_T7, FOOD_SUGAR_CEREAL_T7
Milk (nr. 11)	FOOD_MILK_WHOLE_T7, FOOD_MILK_SEMISKIMMED_T7, FOOD_MILK_EXTRASEMISKIMMED_T7, FOOD_MILK_SKIMMED_T7, FOOD_MILK_ BIOLA_CULTURA_NA_T7, FOOD_MILK_ BIOLA_CULTURA_FL_T7, FOOD_MILK_FL_T7, FOOD_HOTCHOCOLATE_T7
Yoghurt (nr. 12)	FOOD_YOGHURT_DRINK_T7, FOOD_YOGHURT_NATURAL_T7, FOOD_YOGHURT_FRUIT_T7, FOOD_YOGHURT_GOMORGEN_MUSLI_T7, FOOD_YOGHURT_LIGHT_FRUIT_T7, FOOD_YOGHURT_LIGHT_MUSLI_T7
Water (nr. 13)	FOOD_WATER_TAP_T7, FOOD_WATER_BOTTLE_T7
Juice (nr. 14)	FOOD_JUICE_ORANGE_T7, FOOD_JUICE_APPLE_OTHER_T7, FOOD_NECTAR_APPLE_OTHER_T7
Soft Drinks (nr. 15)	FOOD_SAFT_SUGAR_T7, FOOD_SAFT_ARTIFICIAL_T7, FOOD_SOFTDRINK_SUGAR_T7, FOOD_SOFTDRINK_ARTIFICIAL_T7, FOOD_ICETEA_SUGAR_T7, FOOD_ICETEA_ARTIFICIAL_T7, FOOD_BEER_NONALCOHOLIC_T7
Beverages with Alcohol (nr. 16)	FOOD_BEER_STRONG_PILS_T7, FOOD_BEER_LIGHT_T7, FOOD_CIDER_ALCOPOPS_T7, FOOD_WINE_RED_T7, FOOD_WINE_WHITE_T7, FOOD_WINE_FORTIFIED_T7, FOOD_LIQUOR_T7, FOOD_COCKTAIL_T7
Coffee (nr. 17)	FOOD_COFFEE_BOILED_T7, FOOD_COFFEE_FILTERED_T7, FOOD_COFFEE_INSTANT_T7, FOOD_COFFEE_ESPRESSO_T7, FOOD_COFFEE_LATTE_T7, FOOD_COFFEE_CAPPUCINO_T7
Tea (nr. 18)	FOOD_TEA_BLACK_T7, FOOD_TEA_GREEN_T7, FOOD_TEA_HERBS_T7

Groups	Variables
Meat Dinner (nr. 19)	FOOD_SAUSAGE_REDMEAT_T7, FOOD_SAUSAGE_REDMEAT_LIGHT_T7, FOOD_SAUSAGE_CHICKEN_TURKEY_T7, FOOD_SAUSAGE_HOTDOG_PORK_T7, FOOD_SAUSAGE_HOTDOG_CHICKEN_T7, FOOD_HAMBURGER_WITH_BUN_T7, FOOD_KARBONADEBURGER_T7, FOOD_MEATBALL_MEATLOAF_T7, FOOD_STEW_MINCED_MEAT_T7, FOOD_STEAK_PORK_BEEF_LAMB_T7, FOOD_CHOPS_PORK_BEEF_LAMB_T7, FOOD_ROAST_PORK_BEEF_LAMB_T7, FOOD_ROAST_GAMEMEAT_T7, FOOD_STEW_MEAT_T7, FOOD_STEW_MEAT_LAPSKAUS_T7, FOOD_BACON_T7, FOOD_CHICKEN_GRILLED_T7, FOOD_CHICKEN_FILLET_T7, FOOD_WOK_T7, FOOD_STEW_CHICKEN_T7
Composite Dinner Dishes (nr. 20)	FOOD_TACO_SHELLS_MEAT_SALAD_T7, FOOD_WRAP_TORTILLA_T7, FOOD_KEBAB_T7, FOOD_LASAGNA_MOUSAKKA_T7, FOOD_PIZZA_T7, FOOD_CALZONA_T7, FOOD_PIE_QUICHE_T7, FOOD_SPRINGROLLS_T7, FOOD_PORRIDGE_SOURCREAM_T7, FOOD_PORRIDGE_RICE_MILK_T7, FOOD_PANCAKES_T7, FOOD_SOUP_VEGETABLE_T7, FOOD_DISH_VEGETARIAN_T7, FOOD_NUDLES_INSTANT_T7, FOOD_OMELETTE_T7
Fish Dinner (nr. 21)	FOOD_FISH_BURGER_PUDDING_T7, FOOD_FISH_BALLS_T7, FOOD_FISH_LEAN_BOILED_T7, FOOD_FISH_LEAN_FRIED_T7, FOOD_FISH_STICKS_T7, FOOD_HERRING_FRESH_SMOKED_T7, FOOD_MACKEREL_FRESH_SMOKED_T7, FOOD_SALMON_TROUT_T7, FOOD_STEW_SOUP_FISH_T7, FOOD_FISH_BAKED_GRATIN_T7, FOOD_WOK_SEAFOOD_VEGETABLES_T7, FOOD_SCHRIMP_CRAB_T7
Potato (nr. 22)	FOOD_POTATOES_BOILED_BAKED_T7, FOOD_POTATOES_MASHED_T7, FOOD_POTATOSALAD_MAJONNAISE_T7, FOOD_POTATO_GRATIN_CREAM_T7, FOOD_POTATOES_FRIED_T7, FOOD_FRENCHFRIES_DEEPFRIED_T7, FOOD_FRENCHFRIES_OVENBAKED_T7
Rice/pasta (nr. 23)	FOOD_RICE_T7, FOOD_PASTA_T7, FOOD_HOTDOGBUN_POTATOWRAP_T7

Groups	Variables
Vegetables (nr. 24)	FOOD_CARROT_T7, FOOD_CABBAGE_T7, FOOD_RUTABAGA_T7, FOOD_CAULIFLOWER_T7, FOOD_BROCCOLI_T7, FOOD_BRUSSELSSPROUT_T7, FOOD_ONION_T7, FOOD_LETTUCE_T7, FOOD_BELLPEPPER_T7, FOOD_AVOCADO_T7, FOOD_TOMATO_T7, FOOD_CORN_T7, FOOD_VEGETABLES_MIX_FROZEN_T7, FOOD_SALAD_MIX_T7, FOOD_BEANS_LENTILS_T7, FOOD_S_VEGETABLES_BREAD_T7
Sauce etc. (nr. 25)	FOOD_SAUCE_BROWN_WHITE_T7, FOOD_SAUCE_BEARNAISE_T7, FOOD_BUTTER_MARGARINE_MELT_T7, FOOD_BUTTER_HERB_T7, FOOD_MAYONNAISE_REMOULADE_T7, FOOD_MAYONNAISE_LIGHT_T7, FOOD_SOURCREAM_T7, FOOD_SOURCREAM_LIGHT_T7, FOOD_SOURCREAM_EXTRALIGHT_T7, FOOD_SALADDRESSING_T7, FOOD_SALADDRESSING_LIGHT_T7, FOOD_SALADDRESSING_OIL_T7, FOOD_SOYSAUCE_T7, FOOD_PESTO_T7, FOOD_SALSA_TOMATOSAUCE_T7, FOOD_KETCHUP_T7, FOOD_MUSTARD_T7
Fruit (nr. 26)	FOOD_APPLE_T7, FOOD_PEAR_T7, FOOD_BANANA_T7, FOOD_ORANGE_T7, FOOD_CLEMENTINE_T7, FOOD_GRAPEFRUIT_T7, FOOD_PEACH_NECTARINE_T7, FOOD_KIWI_T7, FOOD_GRAPE_T7, FOOD_MELON_T7, FOOD_STRAWBERRY_T7, FOOD_RASPBERRY_T7, FOOD_BLUEBERRY_T7, FOOD_CLOUDBERRY_T7, FOOD_RAISIN_T7, FOOD_FRUIT_DRIED_T7, FOOD_FRUIT_B_T7, FOOD_FRUIT_HERMETIC_T7, FOOD_FRUITSALAD_T7
Dessert (nr. 27)	FOOD_ICECREAM_T7, FOOD_ICELOLLY_SORBET_T7, FOOD_PUDDING_T7, FOOD_SAUCE_VANILLA_T7, FOOD_CREAM_WHIPPED_T7
Cakes and Pastries (nr. 28)	FOOD_SWEET_BUN_PRETZEL_T7, FOOD_SWEET_ROLL_CUSTARD_T7, FOOD_PASTRY_DANISH_T7, FOOD_MUFFIN_CAKE_NOICING_T7, FOOD_WAFFLE_T7, FOOD_LEFSE_T7, FOOD_CAKE_CHOCOLATE_BROWNIE_T7, FOOD_CAKE_SPONGE_CREAM_T7, FOOD_BISCUIT_SWEET_T7, FOOD_TREAT_SNOWBALL_T7
Chocolate (nr. 29)	FOOD_CHOCOLATE_T7, FOOD_CHOCOLATE_DARK_T7, FOOD_CHOCOLATE_CONFECTIONS_T7
Candy (nr. 30)	FOOD_PASTILLES_SUGARFREE_T7, FOOD_CANDY_LICORICE_OTHER_T7, FOOD_CANDY_MIX_T7
Chips (nr. 31)	FOOD_CHIPS_POTATOE_T7, FOOD_SNACS_SALTY_T7

Groups	Variables
Nuts (nr. 32)	FOOD_PEANUT_CASHEW_T7, FOOD_ALMOND_HAZELNUT_WALNUT_T7, FOOD_FRUIT_NUT_MIX_T7
Supplements (nr. 33) Not included in this study	FOOD_SUPPL_CODLIVEROIL_T7, FOOD_SUPPL_CODLIVEROIL_C_T7, FOOD_SUPPL_FISHOIL_OMEGA3_T7, FOOD_SUPPL_SEALOIL_CAPSULA_T7, FOOD_SUPPL_SANASOL_T7, FOOD_SUPPL_BIOVIT_T7, FOOD_SUPPL_MULTIVIT_MINERAL_T7, FOOD_SUPPL_MULTIVIT_TAB_T7, FOOD_SUPPL_IRON_SULFATE_T7, FOOD_SUPPL_IRON_HEME_T7, FOOD_SUPPL_IRON_FERROCHEL_T7, FOOD_SUPPL_IRON_FLORADIX_T7, FOOD_SUPPL_VIT_B_T7, FOOD_SUPPL_VIT_C_T7, FOOD_SUPPL_VIT_D_T7, FOOD_SUPPL_VIT_E_T7, FOOD_SUPPL_FOLATE_T7
Milk/sugar for Coffee/tea (nr. 34)	FOOD_SUGAR_COFFEE_T7, FOOD_SUGAR_TEA_T7, FOOD_SWEETENERS_COFFEE_TEA_T7, FOOD_MILK_CREAM_COFFEE_TEA_T7

Table S2: Summary of individual food intake (g/day) for the final study sample (n = 10899), adjusted for the individual energy intake (J/day).

Food variables	Mean	Median	25% percentile	75% percentile	Max	Consumers (%)
Bread	148	144	105	186	487	99.2
Butter and Margarine	12	8	0	19	214	72.5
Mayonnaise and Plant-based Oils	5	1	0	7	172	51.2
Cheese	36	30	16	49	290	97.2
Meat-spread	16	13	7	21	133	92.9
Fish-spread	23	17	7	32	277	88.7
Egg	18	14	8	24	390	86.6
Jam	15	10	2	20	283	75.5
Breakfast Cereals and Porridge (unsweetened)	25	8	0	33	636	62.6
Breakfast Cereals (Sweetened)	6	3	1	7	110	81.4
Milk	366	309	132	538	3,201	83.8
Yoghurt	41	18	0	57	1,063	74.5
Water	919	709	430	1,078	10,691	98.5
Juice	86	26	0	108	2,481	62.1
Soft Drinks	149	44	0	143	10,018	70.1
Beverages with Alcohol	166	103	36	219	3,172	89.5
Coffee	934	738	424	1,137	12,845	93.8
Tea	176	36	0	208	9,512	62.1
Meat Dinner	125	118	84	159	553	99.4
Composite Dinner Dishes	89	81	50	119	476	98.6
Fish Dinner	96	87	57	125	429	99.2
Potato	100	90	54	134	485	98.8
Rice/pasta	38	27	12	50	780	93.0
Vegetables	221	188	116	288	1,753	99.9
Sauce etc.	28	24	15	36	190	99.4
Fruit	206	174	96	278	1,517	99.3
Dessert	11	7	2	14	324	81.7
Cakes and Pastries	22	17	8	30	208	91.9
Chocolate	8	5	2	10	177	84.4
Candy	6	2	0	7	160	62.1
Chips	4	2	0	6	124	60.9
Nuts	15	8	2	20	178	82.3
Milk/sugar for Coffee/tea	16	0	0	4	1,616	38.6

Table S3: The number of times each of the 33 food variables is categorized to the three diet groups using cluster analysis on 100 random samplings of the cohorts.

Food variables	Women			Men		
	Meat and Sweets	Traditional	Plant-based-and Tea	Meat and Sweets	Traditional	Plant-based-and Tea
Candy	100	0	0	100	0	0
Chips	100	0	0	100	0	0
Chocolate	100	0	0	99	0	1
Soft Drinks'	100	0	0	100	0	0
Composite Dinner Dishes	100	0	0	100	0	0
Rice/pasta	100	0	0	100	0	0
Mayonnaise and Plant-based Oils'	90	8	2	88	2	10
Meat-spread	96	2	2	90	0	10
Meat Dinner	96	2	2	92	1	7
Sauce etc.	96	2	2	92	1	7
Water	17	40	43	75	7	18
Cakes and Pastries	1	94	5	0	90	10
Dessert	1	94	5	0	90	10
Bread	0	100	0	5	90	5
Fish spread	0	90	10	0	69	31
Jam	0	100	0	0	99	1
Coffee	0	100	0	0	98	2
Fish Dinner	0	76	24	0	99	1
Milk	2	93	5	5	92	3
Potato	0	100	0	0	99	1
Breakfast Cereals (sweetened)	2	93	5	1	80	19
Butter and Margarine	19	76	5	52	38	10
Cheese	2	9	89	1	11	88
Breakfast Cereals and Porridge (unsweetened)	4	1	95	2	7	91
Fruit	0	6	94	0	21	79
Nuts	5	1	94	7	4	89
Tea	7	1	92	0	11	89
Vegetables	0	6	94	0	12	88
Yoghurt	2	21	77	0	23	77
Beverages with Alcohol	8	1	91	51	3	46
Milk/sugar for Coffee/tea	14	10	76	0	75	25
Egg	13	39	48	29	43	28
Juice	28	36	36	48	6	46

Table S4: The adjusted coefficient of determination (R^2_{adj}) for different regression models of diet scores. The diet scores were modeled by age, PAL and education, considering both a linear and non-linear association with age and also inclusion of interaction terms between age, PAL and education.

Models	Women			Men		
	Meat and Sweets	Traditional	Plant-based-and Tea	Meat and Sweets	Traditional	Plant-based-and Tea
Linear	0.259	0.237	0.043	0.262	0.232	0.071
Non-linear	0.272	0.238	0.064	0.271	0.232	0.075
Linear with interaction	0.259	0.241	0.050	0.263	0.233	0.071
Non-linear with interaction	0.273	0.243	0.064	0.271	0.234	0.075

Paper II

Associations and predictive power of dietary patterns on metabolic syndrome and its components

Nutrition, Metabolism and Cardiovascular Diseases, **34**, 3, 681-690, 2024

Nutrition, Metabolism & Cardiovascular Diseases (2024) **34**, 681–690

Available online at www.sciencedirect.com

Nutrition, Metabolism & Cardiovascular Diseases

journal homepage: www.elsevier.com/locate/nmcd

Associations and predictive power of dietary patterns on metabolic syndrome and its components

Åse Mari Moe [a], Elinor Ytterstad [a,*], Laila A. Hopstock [b], Ola Løvsletten [c], Monica H. Carlsen [d], Sigrunn H. Sørbye [a]

[a] Department of Mathematics and Statistics, UiT The Arctic University of Norway, Tromsø, Norway
[b] Department of Health and Care Sciences, UiT The Arctic University of Norway, Tromsø, Norway
[c] Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway
[d] Division of Nutritional Epidemiology, University of Oslo, Oslo, Norway

Received 30 January 2023; received in revised form 11 October 2023; accepted 23 October 2023
Handling Editor: A. Siani
Available online 31 October 2023

KEYWORDS
Dietary patterns;
Dimensionality reduction techniques;
Food frequency questionnaire;
Metabolic syndrome;
Predictive power;
The Tromsø study

Abstract *Background and aims:* Metabolic syndrome (MetS) defines important risk factors in the development of cardiovascular diseases and other serious health conditions. This study aims to investigate the influence of different dietary patterns on MetS and its components, examining both associations and predictive performance.
Methods and results: The study sample included 10,750 participants from the seventh survey of the cross-sectional, population-based Tromsø Study in Norway. Diet intake scores were used as covariates in logistic regression models, controlling for age, educational level and other lifestyle variables, with MetS and its components as response variables. A diet high in meat and sweets was positively associated with increased odds of MetS and elevated waist circumference, while a plant-based diet was associated with decreased odds of hypertension in women and elevated levels of triglycerides in men. The predictive power of dietary patterns derived by different dimensionality reduction techniques was investigated by randomly partitioning the study sample into training and test sets. On average, the diet score variables demonstrated the highest predictive power in predicting MetS and elevated waist circumference. The predictive power was robust to the dimensionality reduction technique used and comparable to using a data-driven prediction method on individual food variables.
Conclusions: The strongest associations and highest predictive power of dietary patterns were observed for MetS and its single component, elevated waist circumference.

1. Introduction

Metabolic syndrome (MetS) is a phenotype, commonly defined by elevated blood glucose, elevated blood pressure, elevated waist circumference, elevated triglycerides and low high-density lipoprotein (HDL) cholesterol [1,2]. According to the large multicohort study by Scuteri et al. [3], about one in four of the adult European population has MetS and the prevalence increases by age. This poses serious health challenges as individuals with MetS have an

* Corresponding author. Department of Mathematics and Statistics, UiT The Arctic University of Norway, 9037 Tromsø, Norway.
E-mail address: elinor.ytterstad@uit.no (E. Ytterstad).

https://doi.org/10.1016/j.numecd.2023.10.029

elevated risk of developing cardiometabolic disease such as diabetes and cardiovascular disease [4].

Lifestyle risk factors, such as having an unhealthy diet, play an important role in the development of MetS and several noncommunicable diseases [1,5]. Previous meta-analyses and reviews have shown that consumption of healthy diets, like the well-established Mediterranean diet, is associated with a lower prevalence of MetS [6] while consumption of diets high in meat and sweets can increase the risk of MetS [7,8].

In general, the interplay between dietary patterns and health can be complex and estimated associations might depend on both the study population, data collection and the methods used. Methods to derive dietary patterns are commonly divided into a priori and a posteriori approaches, see the recent review by Zhao et al. [9] for an overview. A priori methods give predefined dietary patterns based on indices and scientific evidence of diet quality, including examples like the Healthy Eating Index [10]. Such dietary patterns are useful in assessing how closely individual diets align with known healthy eating patterns. Here, we focus on a posteriori methods in which novel dietary patterns are derived using different dimensionality reduction techniques on the observed dietary intake. This provides insight into the actual dietary habits of the given study population. Also, scores summarizing the resulting dietary patterns are very useful as these can be used directly as covariates in statistical models explaining health outcomes [9].

The study aim of this paper was twofold. Making use of the data-driven dietary patterns derived in a previous study by Moe et al. [11], we first assessed significant associations between these and MetS and its defining components. Second, we derived new dietary patterns using different dimensionality reduction techniques and investigated the predictive ability of these. Specifically, prediction here implied that the study sample was divided randomly in training and test sets. The models were fitted to individuals in the training sets and then applied to predict MetS or its components for individuals in the corresponding test sets. This gives a realistic evaluation of the predictive ability of the models and how well they generalize to new observations.

2. Methods

2.1. Study population

The Tromsø Study is a population-based health study consisting of 7 consecutive cross-sectional studies (Tromsø1 - Tromsø7) conducted between 1974 and 2016 in Tromsø, Norway. Data collection included questionnaires and interviews, biological sampling, and clinical examinations.

2.2. Study sample

We used data from the seventh survey of the Tromsø Study (Tromsø7 2015–2016), described by Hopstock et al. [12], inviting all inhabitants of Tromsø municipality 40 years and older (N = 32,591). A total of 21,083 (65 %) women and men aged 40–99 years participated, and of these 15,139(72 %) returned a food frequency questionnaire (FFQ). Participants who withdraw their consent to research were excluded (n = 13). In accordance with Lundblad et al. [13], we excluded participants who answered less than 90 % of the questions in the FFQ (n = 3487). Further, we excluded participants with missing information of height or weight, or on educational level, physical activity level and/or smoking status (n = 433). We also excluded participants with unrealistic total energy intakes (kJ/day) or water intakes (g/day) (n = 397). Unrealistic intake values were defined in terms of the 1 % upper and lower quantile values, adjusting intake values for height, weight, physical activity level and age [11]. Finally, we excluded participants with missing information on at least one of the MetS components (n = 72). Thus, the final study sample included 10,750 participants (Fig. S1, Supplementary).

2.3. MetS and its components

MetS was defined by the Adult Treatment Panel III (ATP III) criteria from 2001 [2] as having three or more of the following five risk factors [14,2]:

- Insulin resistance: Self-reported diabetes and/or glycated hemoglobin (HbA1c) $\geq$6.1 %
- Hypertension: Blood pressure $\geq$130/85 mmHg and/or self-reported use of antihypertensive drugs
- Elevated waist circumference: WC $\geq$ 102 cm for men and WC $\geq$ 88 cm for women
- Elevated triglycerides: TG $\geq$ 1.7 mmol/L
- Low HDL cholesterol level: HDL-C $\leq$ 1.0 mmol/L for men and HDL-C $\leq$ 1.3 mmol/L for women

Due to non-fasting blood samples, the original ATP III criterion of insulin resistance based on a fasting serum glucose level was replaced by HbA1c. HbA1c was analysed by high-performance liquid chromatography with Tosoh G8 (Tosoh Bioscience, San Francisco, USA). Blood pressure was measured three times after 2 min seated rest using an automatic oscillometric digital device (Dinamap ProCare 300 monitor, GE Healthcare, Norway). The average of the last two measurements was used in the analysis. Waist circumference was measured with a Seca measuring tape at the umbilical level. HDL cholesterol and triglycerides were analysed by enzymatic colorimetric methods with Cobas 8000 c702 (Roche Diagnostics, Mannheim, Germany).

2.4. Dietary data

The paper FFQ used in Tromsø7 has previously been evaluated by Carlsen et al. [15] and described in detail by Lundblad et al. [13]. The questionnaire includes 261 questions on habitual food and beverage consumption (frequency and amount). Answers were checked manually by trained technicians before scanning. Food intakes in grams/day (g/day) were calculated using the KBS AE14 food database and KBS software system at University of

Oslo (KBS, version 7.3.) based on the Norwegian food composition tables 2014–2015. Missing values on frequency of intake in the FFQ were automatically coded to never/seldom by the KBS system. The questions were aggregated in a supervised manner, providing 33 food variables including alcohol consumption, as described by Moe et al. [11]. To account for variations in total energy intake, the intake values for each food variable and individual were multiplied by the total mean intake for all individuals divided by the total energy intake for each individual, see Ref. [11] for further explanation.

2.5. Covariates

Participants registered age was given in terms of years. Self-reported educational level included four categories: Primary/partly secondary education (up to 10 years of schooling), upper secondary education (more than 10 years of schooling not including college/university), short tertiary education (less than 4 years at college/university), or long tertiary education (4 years or more at college/university). Self-reported leisure-time physical activity level was collected using the Saltin-Grimby's activity level scale [16] and included four categories of physical activity: sedentary (mainly reading/watching TV), light (walking/biking more than 4 h/week), moderate (exercise more than 4 h/week) and vigorous (hard exercise/competitive sports more than 4 h/week). Due to few individuals in the vigorous physical activity group, the levels of moderate and vigorous physical activity were combined as one category named high activity level. Self-reported smoking status was current daily smoker, previous smoker or never-smoker, dichotomized into current or non-smoker. Those who smoked occasionally (i.e., not daily) were categorized as non-smokers. Frequency and the amount of alcoholic beverage intake were collected as part of the FFQ.

2.6. Estimation of associations between diet scores and MetS and its components using logistic regression

In a first step, we investigated the associations between dietary patterns and MetS and its individual components. This was performed using the dietary patterns already derived in the study by Moe et al. [11], who analysed the FFQ of Tromsø7 using hierarchical clustering. This method identified three diet groups named the Meat and Sweets diet (Candy, Chips, Chocolate, Composite dinner dishes, Meat-spread, Mayonnaise and oils, Meat dinner, Rice/pasta, Sauce, Soft drinks), the Plant-based- and Tea diet (Cheese, Cereal (unsweetened), Fruit, Nuts, Tea, Vegetables, Yoghurt) and the Traditional diet (Bread, Cakes and pastries, Cereal (sweetened), Coffee, Dessert, Fish dinner, Fish-spread, Jam, Milk, Potato). Five of the 33 aggregated food variables were not included in any of the diet groups as the classification of these switched between groups [11].

In the current study, we used the individual intake scores for each of the three diets previously computed by Moe et al. [11], as covariates in logistic regression models. The response variables of these models included the binary outcomes of MetS and its five defining risk factors according to ATP III. These were modelled as linear functions of the diet intake scores, adjusted for linear effects of age, educational level, physical activity, smoking and alcohol intake. The resulting models are described generically by Eq. (1) in which we used a logit link function. This represents our main model of interest. In addition, we have investigated a model that also adjusts for the linear effect of body mass index (Section S1.5.1, Supplementary)

$$\text{Response} \sim \text{Age} + \text{Education} + \text{Diet scores} + \text{Physical activity} + \text{Smoking} + \text{Alcohol}, \quad (1)$$

To account for multiple testing, the confidence level was set equal to 99 % and all models were fitted separately for women and men.

2.7. Deriving new dietary patterns by factor analysis and the treelet transform

The hierarchical clustering method groups food variables into non-overlapping dietary patterns. To explore the use of different dimensionality reduction methods on the 33 food variables, we applied exploratory factor analysis and the treelet transform which are both commonly applied to derive dietary patterns [9]. All of these three methods perform grouping based on the correlation between the individual intake values of food variables. However, the derived patterns typically differ in terms of the number of diet groups found and the weights given to each food variable.

Factor analysis models the relationship between variables in terms of linear combinations of unobserved common factors, here representing different dietary patterns. The loadings of each food variable on each factor were calculated using the principal component method. Further, we applied varimax rotation to simplify interpretations of the loadings. The number of patterns was determined using a scree plot, reflecting the size of the eigenvalues of the correlation matrix. We have not used a specific threshold value for the loadings, but interpreted the factors in terms of groups of food variables having the most positive versus most negative loadings. Mostly this includes food variables with absolute values of loadings larger than 0.20.

The treelet transform estimates loadings on each food variable by combining hierarchical clustering and principal component analysis as described by Lee et al. [17]. The method works by subsequently finding the two variables with the highest correlation and then rotate these variables locally using principal component analysis. The procedure is summarized by a hierarchical tree. To get the final dietary patterns, the hierarchical tree is cut, and a chosen number of factors explaining most of the variation is extracted [17]. The cut level was here determined using a measure of the explained variance of the chosen number of factors (varied from 2 to 10) and 10-fold cross-validation. The number of factors was chosen based on the explained variance using the final cut level.

Diet scores for each factor were calculated by summing the products of intakes of food groups weighted by loading coefficients. Using hierarchical clustering, all food variables within a diet are given equal weight.

For the given study population, we have previously observed only minimal differences in the derived dietary patterns for men and women [11]. Therefore, the dietary patterns using all three data reduction methods were found for men and women simultaneously. Also, the individual diet scores were scaled for men and women simultaneously having zero mean and a standard deviation of one. This was done to simplify comparison between dietary patterns across methods.

2.8. Evaluation of predictive performance

In a second step, we evaluated the predictive performance using the derived dietary patterns in predicting MetS and its five components. To evaluate predictive performance, the individuals in the study sample was allocated randomly to training sets (70 %) and test sets (30 %). We then fitted models to each of the training sets, providing predictions for the individuals in the corresponding test sets.

The predictive performance was investigated using logistic regression models of increasing complexity. The simplest model (M1) is described by Eq. (2) and includes linear effects of age and educational level. This serves as an important background model as both age and educational level are known to be important predictors of disease. The second model (M2) is given by Eq. (3) and was used to investigate the predictive power of the dietary patterns, adjusted for age and educational level. Finally, we fitted the full model (M3) defined by Eq. (1), to assess the additional predictive power of the lifestyle variables physical activity, smoking and alcohol consumption. The predictive power including body mass index as a covariate is reported in Section S1.5.2, Supplementary.

$$\text{Response} \sim \text{Age} + \text{Education} \quad (2)$$

$$\text{Response} \sim \text{Age} + \text{Education} + \text{Diet scores} \quad (3)$$

A general concern using dimensionality reduction techniques is that the summarized intake scores for specific dietary patterns might not reveal important predictors among the original food variables [18,9]. We therefore applied the random forest algorithm [19], performing predictions based on the individual food variables. The algorithm fits several decision trees, where each tree is constructed using a subsample of the variables [19]. The random forest algorithm is data-driven and does not assume a specific model construction like in the logistic regression case. This implies that we do not have to worry about violating model assumptions, possible interaction effects or collinearity among predictor variables. We implemented the random forest algorithm using 500 decision trees, in which each tree used 63.2 % of the training sample. Both the number of variables used in each tree and the depth of the trees were optimised to give a best possible prediction.

The prediction of MetS and its components corresponds to a classification problem with two classes. To evaluate the predictive power of the different methods, we used the measure AUC (Area under the curve) giving values between 0.5 (no discrimination) to 1 (perfect classification). The predictions were repeated for 100 random partitions of training and test sets to get average measures of predictive power.

3. Results

In the included 5715 women and 5035 men, 21.6 % and 26.8 % were classified with MetS, respectively (Table 1). Slightly more than half of the participants (53.1 %), men and women combined, reported short or long tertiary education. More than half of the participants (58.9 %) reported a light physical activity level of at least 4 h a week, while more than a quarter of the participants (28.3 %) reported a high physical activity level. The prevalence of smoking was 13.2 % in women and 11.1 % in men. Adjusted for energy intake, women had a median alcoholic beverage consumption of 0.65 dl/day while the corresponding consumption among men was 1.36 dl/day.

3.1. Estimation of associations by logistic regression models

Odds ratios for MetS using Eq. (1) are displayed in Table 2. A one unit increase of the Meat and Sweets diet score was

Table 1 Overview of the study sample.

	Women		Men	
Total number, *n*(%)	5715	(53.2 %)	5035	(46.8 %)
MetS, *n*(%)	1235	(21.6 %)	1349	(26.8 %)
Components of MetS, *n*(%)				
Insulin resistance	551	(9.6 %)	662	(13.1 %)
Hypertension	2536	(44.4 %)	3115	(61.9 %)
Elevated waist circumference	3118	(54.6 %)	2033	(40.4 %)
Elevated triglycerides	1324	(23.2 %)	2004	(39.8 %)
Low HDL cholesterol	1137	(19.9 %)	903	(17.9 %)
Age (year), mean (sd)	56.4	(10.5)	57.9	(10.9)
Educational level, *n*(%)				
Primary/partly secondary	1151	(20.1 %)	977	(19.4 %)
Upper secondary	1450	(25.4 %)	1460	(29.0 %)
Short tertiary	1066	(18.7 %)	1161	(23.1 %)
Long tertiary	2048	(35.8 %)	1437	(28.5 %)
Physical activity level, *n*(%)				
Sedentary	713	(12.5 %)	661	(13.1 %)
Light	3726	(65.2 %)	2606	(51.8 %)
High	1276	(22.3 %)	1768	(35.1 %)
Smoking status, *n*(%)				
Non-smoker	4959	(86.8 %)	4475	(88.9 %)
Current smoker	756	(13.2 %)	560	(11.1 %)
Alcohol (dl/day), median (sd) (Adjusted for energy intake)	0.65	(1.41)	1.36	(2.50)

Table 2 Odds ratios (OR) with confidence intervals (CI) in predicting MetS for women and men.

Covariates	Women		Men	
	OR	99 % CI	OR	99 % CI
Dietary patterns				
Meat and Sweets	1.11	(1.01, 1.22)	1.16	(1.05, 1.29)
Plant-based- and Tea	0.91	(0.83, 1.00)	0.98	(0.86, 1.12)
Traditional	0.94	(0.84, 1.03)	0.91	(0.82, 1.01)
Age	1.03	(1.02, 1.04)	1.02	(1.01, 1.03)
Educational level				
Primary/partly secondary	1		1	
Upper secondary	0.89	(0.70, 1.13)	0.87	(0.69, 1.10)
Short tertiary	0.60	(0.45, 0.79)	0.78	(0.60, 1.00)
Long tertiary	0.51	(0.40, 0.67)	0.48	(0.37, 0.63)
Physical activity level				
Sedentary	1		1	
Light	0.58	(0.46, 0.73)	0.65	(0.51, 0.82)
High	0.30	(0.22, 0.41)	0.35	(0.27, 0.46)
Smoking status				
Non-smoker	1		1	
Current smoker	1.16	(0.91, 1.48)	0.82	(0.62, 1.07)
Alcohol	0.90	(0.84, 0.97)	1.00	(0.97, 1.04)

associated with an 11 % and 16 % increased odds of MetS for women and men respectively. The odds of MetS was positively associated with higher age. Further, the odds of MetS was significantly lower by higher educational level, higher physical activity level and higher alcohol consumption (women only).

The odds ratios of the diet score variables using the five separate components of MetS as response variables in Eq. (1), are shown in Table 3. A higher intake of the Meat and Sweets diet was positively associated with the odds of elevated waist circumference. A higher intake of the Plant-based- and Tea diet was associated with decreased odds of hypertension (women only) and elevated triglycerides (men only).

The odds ratios for all covariates in Eq. (1), using the five components of MetS as response variables, are given in the Supplementary, Tables S1–S5. The odds of all components were higher by higher age, except for the case of elevated triglycerides in men, and low HDL-C. Compared with the reference group, individuals with long tertiary education had significantly lower odds of all components except for elevated triglycerides in men. The odds mostly decreased with higher activity level for all components, except for the case of hypertension in men. Female smokers had significantly higher odds of insulin resistance, elevated triglycerides and low HDL-C, and significantly lower odds of hypertension and elevated waist circumference, compared to non-smokers. Alcohol consumption was associated with decreased odds of insulin resistance (women only) and low HDL cholesterol. In men, alcohol consumption was also associated with increased odds of elevated waist circumference and hypertension.

3.2. Dietary patterns found by factor analysis and the treelet transform

Based on the screeplot, factor analysis (FA) gave four factors, all corresponding to eigenvalues above 1.5. The estimated loadings for the four factors are given in Table 4, the factors being named FA1, FA2, FA3 and FA4. FA1 showed high loadings on various sweets, chips, soft drinks, rice/pasta, meat dinner and sauce, having the most negative loadings for fish dinner, potato, bread, milk and jam. FA2 was characterized by high loadings on unsweetened cereal, yoghurt and plant-based food variables like nuts, vegetables and fruit, and negative loadings for potato, bread and some spreads. FA3 showed positive loadings on vegetables, fruit, bread and spreads, also including egg and cheese, in contrast to negative loadings on sweetened cereals, dessert, cakes and pastries, milk and composite dinner dishes. The factor FA4 showed high loadings for potato, fish dinner and sauce.

Applying the treelet transform (TT), the explained variance started to decrease around a cut level of 17 which was therefore used in the final analysis. This resulted in four factors, referred to as TT1, TT2, TT3 and TT4. The factors are similar to the patterns found by factor analysis, but in a simplified version as each factor only included 3–6 food variables with non-zero loadings. Specifically, TT1 included candy, chips, chocolate, soft drinks, cakes and pastries, and rice/pasta. TT2 included nuts, vegetables, yoghurt, unsweetened cereal and fruits. Further, TT3

Table 3 Odds ratios (OR) with confidence intervals (CI) for the five components of MetS, based on intake scores for the Meat and Sweets diet (HC1), the Plant-based- and Tea diet (HC2) and the Traditional diet (HC3). Abbreviations: Waist circumference (WC), Triglycerides (TG), HDL cholesterol (HDL-C).

Response	Sex	HC1		HC2		HC3	
		OR	99 % CI	OR	99 % CI	OR	99 % CI
Insulin resistance	Women	1.11	(0.97, 1.27)	1.01	(0.89, 1.14)	0.93	(0.81, 1.07)
	Men	1.01	(0.87, 1.18)	1.04	(0.88, 1.22)	0.95	(0.84, 1.09)
Hypertension	Women	1.04	(0.95, 1.13)	0.91	(0.84, 0.99)	1.02	(0.93, 1.12)
	Men	1.04	(0.94, 1.15)	0.93	(0.82, 1.05)	0.98	(0.89, 1.08)
Elevated WC	Women	1.25	(1.15, 1.36)	0.92	(0.86, 1.00)	1.07	(0.99, 1.17)
	Men	1.34	(1.21, 1.48)	1.02	(0.90, 1.15)	1.06	(0.96, 1.16)
Elevated TG	Women	1.06	(0.97, 1.17)	0.93	(0.85, 1.01)	0.91	(0.82, 1.00)
	Men	1.03	(0.94, 1.14)	0.87	(0.77, 0.98)	0.90	(0.82, 0.99)
Low HDL-C	Women	1.10	(1.00, 1.21)	0.95	(0.87, 1.04)	1.00	(0.90, 1.11)
	Men	1.08	(0.95, 1.21)	1.03	(0.89, 1.19)	0.92	(0.82, 1.04)

Table 4 Loadings for factors found by factor analysis (FA1 - FA4) where loadings with absolute value smaller than 0.10 are marked with -. Using the treelet transform (TT1 - TT4), missing loadings are equal to 0.

	FA1	FA2	FA3	FA4	TT1	TT2	TT3	TT4
Dessert	–	–	−0.34	0.18				
Cakes and pastries	0.35	–	−0.26	−0.13	0.30			
Candy	0.52	–	−0.11	–	0.50			
Chips	0.51	–	–	–	0.45			
Chocolate	0.32	–	−0.13	−0.15	0.50			
Soft drinks	0.42	−0.12	–	–	0.36			
Rice/pasta	0.39	–	–	–	0.30			
Meat dinner	0.22	–	0.12	0.23				
Composite dinner dishes	0.17	0.16	−0.36	–				
Sauce etc.	0.34	−0.10	–	0.54				0.41
Fish dinner	−0.24	0.12	–	0.62				0.64
Potato	−0.23	−0.34	−0.18	0.65				0.64
Bread	−0.31	−0.39	0.29	−0.25				
Butter	–	−0.42	–	−0.20				
Cheese	–	0.12	0.29	−0.29				
Egg	–	–	0.43	–			0.52	
Mayonnaise and oils	0.15	−0.32	0.29	–			0.48	
Fish-spread	−0.16	–	0.56	0.15			0.52	
Meat-spread	0.17	−0.26	0.49	−0.15			0.48	
Jam	−0.29	−0.19	−0.15	−0.14				
Milk	−0.32	−0.17	−0.28	−0.13				
Cereal (sweetened)	−0.18	–	−0.52	−0.14				
Cereal (unsweetened)	−0.19	0.54	−0.18	−0.23		0.39		
Fruit	–	0.33	0.22	0.19		0.52		
Nuts	–	0.56	–	−0.19		0.38		
Vegetables	–	0.54	0.22	0.25		0.52		
Yoghurt	–	0.36	–	–		0.40		
Juice	–	–	–	−0.14				
Coffee	−0.13	–	–	0.31				
Tea	–	–	–	–				
Milk/sugar for coffee/tea	−0.14	–	–	–				
Water	–	–	–	–				
Explained variance (%)	7.3	6.7	5.1	4.9	5.2	5.0	4.2	4.1
Cumulative (%)	7.3	14.0	19.1	24.0	5.2	10.2	14.4	18.6

included food variables related to bread meals like mayonnaise, egg, meat-spread and fish-spread. The pattern given by TT4 included fish, potato and sauce. Using the treelet transform with the given tree cut level, fifteen of the original food variables were discarded (Table 4).

Using factor analysis and the treelet transform, the first factor identified a dietary pattern with high loadings on sweets and also on meat (only FA1). These patterns were clearly correlated with the Meat and Sweets diet found by hierarchical clustering, the correlation being 0.87 for FA1 and 0.77 with TT1. Also, a plant-based dietary pattern was found using all three dimensionality reduction methods. The correlations of the Plant-based- and Tea diet versus FA2 and TT2 were 0.81 and 0.89, respectively. The last two dietary patterns using factor analysis and the treelet transform represented a mix of the dietary patterns found by hierarchical clustering, but were similar to each other. The third factor using both methods had high loadings on spread (FA3 and TT3), the correlation between these patterns being 0.72. The fourth factors represented a fish dinner dietary pattern, the correlation between FA4 and TT4 being 0.87 (Table S6, Supplementary).

3.3. Predictive power of different dietary patterns

Average measures of AUC predicting MetS and its components for 100 test sets are displayed in Table 5, using the defined models of different complexity (M1 - M3) and the three dimensionality reduction techniques. For comparison, we also report the corresponding AUC values for the random forest algorithm (RF) using the covariates as given by M2 and M3.

These AUC values ranged from 0.580 to 0.767, having standard deviations between 0.009 and 0.019. The largest absolute increase in performance including diet scores was seen in predicting elevated waist circumference among men, in which AUC increased from 0.58 (M1) to 0.68 using M2 and the random forest algorithm. The corresponding AUC-values among women were 0.61–0.65. Inclusion of diet scores also increased the AUC-value in predicting MetS for men, from a value of 0.59–0.64. Inclusion of lifestyle variables (M3) improved the predictive performance for all components and methods, with an average increase in the AUC-value of 0.03 compared with M2. The largest improvement using M3 versus M2 was seen in

Table 5 AUC values predicting MetS and its components including the covariates age and educational level (M1), adding diet scores (M2) and adding lifestyle variables (M3). The methods include hierarchical clustering (HC), factor analysis (FA), the treelet transform (TT) and random forest (RF).

Response	Sex	AUC		AUC			
		M1	M2/M3	HC	FA	TT	RF
MetS	Women	0.639	M2	0.655	0.649	0.646	0.653
			M3	0.683	0.678	0.678	0.672
	Men	0.587	M2	0.620	0.600	0.600	0.638
			M3	0.652	0.636	0.637	0.662
Insulin resistance	Women	0.704	M2	0.710	0.705	0.707	0.682
			M3	0.726	0.720	0.724	0.691
	Men	0.689	M2	0.698	0.687	0.692	0.690
			M3	0.711	0.701	0.706	0.699
Hypertension	Women	0.762	M2	0.764	0.763	0.762	0.751
			M3	0.767	0.767	0.766	0.753
	Men	0.686	M2	0.689	0.687	0.686	0.682
			M3	0.693	0.690	0.690	0.684
Elevated waist circumference	Women	0.609	M2	0.640	0.631	0.632	0.648
			M3	0.665	0.655	0.658	0.668
	Men	0.582	M2	0.641	0.606	0.613	0.676
			M3	0.684	0.661	0.668	0.700
Elevated triglycerides	Women	0.600	M2	0.616	0.608	0.607	0.611
			M3	0.642	0.637	0.638	0.626
	Men	0.586	M2	0.599	0.599	0.593	0.600
			M3	0.619	0.619	0.616	0.616
Low HDL-C	Women	0.603	M2	0.608	0.610	0.604	0.598
			M3	0.655	0.655	0.654	0.629
	Men	0.582	M2	0.583	0.582	0.580	0.584
			M3	0.643	0.643	0.642	0.628

predicting low HDL-C, giving a relative increase of 8.3 % for women and 10.7 % for men using treelet transform.

Overall, the best power was seen in M3, predicting hypertension in women, in which AUC was equal to 0.77. Prediction of insulin resistance in women gave an AUC value equal to 0.70 using M1. This increased to an AUC value of 0.73 using M3 and hierarchical clustering.

The given analysis shows that the predictive performance was robust to the dimensionality reduction method used to derive dietary patterns. This is also the case using diet scores categorized according to quartiles, giving very similar results (Table S7, Supplementary). Also, the predictive performance was similar to using the random forest algorithm on individual food variables. Averaged over the six response variables, the AUC values using different methods ranged from 0.658 (FA) for men to 0.690 (HC) for women using M3. The predictive performance was slightly lower for men, in which AUC ranged from 0.616 (TT,RF) to 0.767 (HC,FA).

4. Discussion

4.1. Associations between dietary patterns and MetS and its components

In our study based on data from a general population in Norway, we found an association between dietary patterns and cardiometabolic health variables by applying logistic regression analysis. The odds of MetS increased with the intake of the Meat and Sweets diet in both sexes, and we estimated a decrease ($0.01 < p < 0.05$) in the odds of MetS with intake of the Plant-based- and Tea diet in women. This is in accordance with previous research, summarized in two meta-analyses [7,8], where a meat-based or so-called Western diet was associated with increased odds of MetS. They also both found that a healthy diet decreased the odds of MetS, giving a larger relative decrease for women compared with men, which is comparable to our finding for the Plant-based- and Tea diet. A similar association by sex was found in Grosso et al. [20], between a Mediterranean diet, which is considered as a healthy diet, and MetS. However, in a small study ($n = 808$) by Babio et al. [21], this association was significant in men only. According to a recent meta-analysis [22], individual studies examining the associations between the consumption of a Mediterranean diet and cardiovascular disease (CVD) have yielded conflicting results. Some studies have reported a beneficial effect of the Mediterranean diet on CVD in both sexes [23], while others have reported a benefit only in men [24]. The meta-analysis conducted by [22] concluded that a Mediterranean diet is equally beneficial for both sexes.

Considering the five separate components of MetS, the most interesting finding was the association between the Meat and Sweets diet and waist circumference. A similar, but non-significant effect on waist circumference of an unhealthy diet was seen in a meta-analysis by Rezagholizadeh et al. [25]. The Meat and Sweets dietary pattern is primarily an energy-dense diet, characterised by high levels of total fat, saturated fat and sugar compared to the other two dietary patterns identified through hierarchical clustering. Diets high in saturated fat and sugar have been shown to be associated with higher odds of obesity [26].

Another two components of MetS, hypertension and elevated TG, were found to be related to the Plant-based- and Tea diet, where increased intake of the diet was associated with decreased odds of hypertension (women only), and decreased odds of elevated TG (men only). A meta-analysis by Godos et al. [6] found a similar effect of a Mediterranean diet on hypertension, and one of the included studies [21] demonstrated a significant association between a Mediterranean diet and elevated TG. However, the overall result from the meta-analysis was non-significant.

Our study did not show any associations between dietary patterns and insulin resistance. In a Norwegian study of type 2 diabetes [27], it was found that approximately one-third of the patients controlled their diabetes through diet. This indicates that some of our participants with insulin resistance may have had a different diet earlier in adulthood, which could also explain the lack of an association between the current registered diet and insulin resistance.

4.2. Dietary patterns and predictive performance

To study predictive power measured by AUC, three models with increasing complexity were considered and fitted separately to women and men. The simplest model included only age and educational level as covariates (M1). Further, this model was expanded with diet score variables (M2) and lifestyle variables (M3). The predictive power was generally slightly lower for men compared to women. However, for both sexes, the largest absolute increase in predictive power due to the inclusion of diet score variables (comparing M2 with M1) was found in predicting MetS and elevated waist circumference. This is in accordance with the increased odds of MetS and elevated waist circumference by higher scores of Meat and Sweets, a diet known to be connected to weight gain. To our knowledge, few studies have investigated the influence of diets on predictive performance. Remyaa et al. [28] used different machine-learning tools on dietary data to predict body weight, reporting improved prediction on a test set.

The differences in predictive performance using dietary patterns found by hierarchical clustering, factor analysis and the treelet transform were relatively small. All methods identified a diet high in sweets and a plant-based diet, having high inter-related correlations across the different methods. Both factor analysis and treelet transform yielded four similar dietary patterns. The high correlation observed between factors found by these two methods is in accordance with previous studies [29–31].

Our results indicate that the analysis is robust to differences in dietary patterns, where hierarchical clustering on average gave slightly higher predictive power. Using factor analysis, none of the estimated loadings will be exactly zero and thereby none of the food variables are discarded. The treelet transform can also give overlapping groups but typically with more sparse patterns as many of the loadings are exactly zero [9,29]. This facilitates interpretation of each factor [29] but comes at the expense of lost information due to discarded variables as observed in a study of dietary patterns and diabetes [30].

4.3. Random forest and predictive performance

The random forest algorithm utilizes all food variables separately as opposed to the dietary patterns from the data reduction techniques, where important predictors among the original food variables may be lost [32]. However in our study, dietary patterns from hierarchical clustering, factor analysis and the treelet transform were comparable in predictive performance with the random forest algorithm. This indicates that the dietary patterns captured important food variables as predictors for MetS and its components, and that the model assumptions for logistic regression were acceptable.

4.4. Implications

Our study uncovers associations between a diet rich in meat and sweets, as well as a plant-based diet, and cardiovascular disease risk factors. However, further research is recommended to investigate causality, particularly through longitudinal studies.

The healthy plant-based diet comprises fruit, nuts, vegetables, unsweetened breakfast cereals and porridge, cheese, yoghurt, and tea. Some of these food groups may already be subsidized, but more can be done for consumers to make healthier food choices.

The predictive ability of how well our models generalize to new observations, was found to be similar for different dimensionality reduction techniques. This is important when statistical models are used to assess the risk of health outcomes for individuals not included in the original study sample.

4.5. Strengths and limitations

A major strength of this study is the use of a large-size population-based sample of a general adult community-dwelling population. Further, the data collection includes the use of validated questionnaires, analysis of blood samples performed at a ISO-standardized laboratory, and clinical examinations such as blood pressure and height- and weight measurements performed by trained personnel using calibrated equipment and standardised methods. Another strength is that the predictive performance was investigated by using several dimensionality reduction methods in addition to the random forest algotithm, by randomly and repeatedly splitting the dataset into a test- and a training set. The random forest algorithm also contributed to an evaluation of the regression model assumptions as well as on whether important features in the food variables were captured in the dietary patterns.

An important limitation of the given study is that a cross-sectional study design cannot establish causal relationships between dietary patterns and MetS or its components, only associations. Another limitation is that the logistic regression models assumed linear effects of the diet score variables. Additionally, the chosen number of dietary groups using the different dimensionality reduction methods could potentially impact the estimated models and predictive performance.

5. Conclusions

Based on a cross-sectional study of a general adult population, dietary patterns were found to be associated with cardiometabolic health, particularly abdominal obesity. A diet high in meat and sweets was positively associated with the odds of MetS and its individual component elevated waist circumference. Similar conclusions were drawn when assessing the predictive power of different dietary patterns. The derived dietary patterns were robust to the dimensionality reduction method used and the predictive performance using logistic regression models was comparable to using a data-driven method on food variables.

Author contribution statement

ÅMM performed the statistical analysis, contributed to the study design and interpretation of the results. EY and SHS contributed to the study design and interpretation of the results. LAH and MHC contributed to the collection and preprocessing of data, interpretation of the results and provided knowledge on the data material. OL contributed to the study design. All authors have contributed to the manuscript, and read and approved the final version.

Declaration of competing interest

None declared.

Acknowledgements

We would like to thank all Tromsø Study participants for their patience.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.numecd.2023.10.029.

References

[1] Peters R, Ee N, Peters J, Beckett N, Booth A, Rockwood K, et al. Common risk factors for major noncommunicable disease, a systematic overview of reviews and commentary: the implied potential for targeted risk reduction. Ther Adv Chronic Dis 2019;10:204062231988039. https://doi.org/10.1177/2040622319880392.

[2] Grundy SM, Cleeman JI, Daniels SR, Donato KA, Eckel RH, Franklin BA, et al. Diagnosis and management of the metabolic syndrome. Circulation 2005;112(17):2735–52. https://doi.org/10.1161/circulationaha.105.169404.

[3] Scuteri A, Laurent S, Cucca F, Cockcroft J, Cunha PG, Mañas LR, et al. Metabolic syndrome across Europe: different clusters of risk factors. Eur J Prev Cardiol 2020;22(4):486–91. https://doi.org/10.1177/2047487314525529.

[4] Ford ES. Risks for all-cause mortality, cardiovascular disease, and diabetes associated with the metabolic syndrome: a summary of the evidence. Diabetes Care 2005;28(7):1769–78. https://doi.org/10.2337/diacare.28.7.1769.

[5] Nilsson PM, Tuomilehto J, Rydén L. The metabolic syndrome – what is it and how should it be managed? Eur J Prev Cariol 2019;26(2 suppl):33–46. https://doi.org/10.1177/2047487319886404.

[6] Godos J, Zappalà G, Bernardini S, Giambini I, Bes-Rastrollo M, Martinez-Gonzalez M. Adherence to the Mediterranean diet is inversely associated with metabolic syndrome occurrence: a meta-analysis of observational studies. Int J Food Sci Nutr 2017;68(2):138–48. https://doi.org/10.1080/09637486.2016.1221900. pMID: 27557591.

[7] Rodríguez-Monforte M, Sánchez E, Barrio F, Costa B, Flores-Mateo G. Metabolic syndrome and dietary patterns: a systematic review and meta-analysis of observational studies. Eur J Nutr 2017;56(3):925–47. https://doi.org/10.1007/s00394-016-1305-y.

[8] Fabiani R, Naldini G, Chiavarini M. Dietary patterns and metabolic syndrome in adult subjects: a systematic review and meta-analysis. Nutrients 2019;11(9). https://doi.org/10.3390/nu11092056.

[9] Zhao J, Li Z, Gao Q, Zhao H, Chen S, Huang L, et al. A review of statistical methods for dietary pattern analysis. Nutr J 2021;20(1):37. https://doi.org/10.1186/s12937-021-00692-7.

[10] Kennedy ET, Ohls J, Carlson S, K F. The healthy eating index: design and applications. J Am Diet Assoc 1995;95(10):1103–8. https://doi.org/10.1016/S0002-8223(95)00300-2.

[11] Moe ÅM, Sørbye SH, Hopstock LA, Carlsen MH, Løvsletten O, Ytterstad E. Identifying dietary patterns across age, educational level and physical activity level in a cross-sectional study: the Tromsø Study 2015 - 2016. BMC Nutr 2022;8(1):102. https://doi.org/10.1186/s40795-022-00599-4.

[12] Hopstock LA, Grimsgaard S, Johansen H, Kanstad K, Wilsgaard T, Eggen AE. The seventh survey of the Tromsø Study (Tromsø7) 2015–2016: study design, data collection, attendance, and prevalence of risk factors and disease in a multipurpose population-based health survey. Scand J Public Health 2022;50(7):919–29. https://doi.org/10.1177/14034948221092294.

[13] Lundblad MW, Andersen LF, Jacobsen BK, Carlsen MH, Hjartåker A, Grimsgaard S, et al. Energy and nutrient intakes in relation to national nutrition recommendations in a Norwegian population-based sample: the Tromsø study 2015–16. Food Nutr Res 2019;63. https://doi.org/10.29219/fnr.v63.3616.

[14] Herder M, Arntzen K, Johnsen SH, Mathiesen EB. The metabolic syndrome and progression of carotid atherosclerosis over 13 years. The Tromsø study. Cardiovasc Diabetol 2012;11:77. https://doi.org/10.1186/1475-2840-11-77.

[15] Carlsen MH, Lillegaard ITL, 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. https://doi.org/10.1186/1475-2891-9-37.

[16] Grimby G, Börjesson M, Jonsdottir IH, Schnohr P, Thelle DS, Saltin B. The "saltin–grimby physical activity level scale" and its application to health research. Scand J Med Sci Sports 2015;25(S4):119–25. https://doi.org/10.1111/sms.12611.

[17] Lee AB, Nadler B, Wasserman L. Treelets-An adaptive multi-scale basis for sparse unordered data. Ann Appl Stat 2008;2(2):435–71. https://doi.org/10.1214/07-AOAS137.

[18] Hoffmann K, Schulze MB, Schienkiewitz A, Nöthlings U, Boeing H. Application of a new statistical method to derive dietary patterns in nutritional epidemiology. Am J Epidemiol 2004;159(10):935–44. https://doi.org/10.1093/aje/kwh134.

[19] Breiman L. Random forests. Mach Learn 2001;45:5–32. https://doi.org/10.1023/a:1010933404324.

[20] Grosso G, Stepaniak U, Micek A, Topor-Mądry R, Stefler D, Szafraniec K, et al. A Mediterranean-type diet is associated with better metabolic profile in urban Polish adults: results from the HAPIEE study. Metab Clin Exp 2015;64(6):738–46. https://doi.org/10.1016/j.metabol.2015.02.007.

[21] Babio N, Bulló M, Basora J, Martínez-González MA, Fernández-Ballart J, Márquez-Sandoval F, et al. Adherence to the Mediterranean diet and risk of metabolic syndrome and its components. Nutr Metabol Cardiovasc Dis 2009;19(8):563–70. https://doi.org/10.1016/j.numecd.2008.10.007. https://www.sciencedirect.com/science/article/pii/S0939475308002226.

[22] Pant A, Gribbin S, McIntyre D, Trivedi R, Marschner S, Laranjo L, et al. Primary prevention of cardiovascular disease in women with a Mediterranean diet: systematic review and meta-analysis. Heart 2023;109:1208–15. arXiv, https://heart.bmj.com/content/early/2023/02/14/heartjnl-2022-321930.full.pdf, doi:10.1136/heartjnl-2022-321930. https://heart.bmj.com/content/early/2023/02/14/heartjnl-2022-321930.

[23] Neelakantan N, Koh WP, Yuan JM, van Dam RM. Diet-quality indexes are associated with a lower risk of cardiovascular, respiratory, and all-cause mortality among Chinese adults. J Nutr 2018;148(8):1323–32. https://doi.org/10.1093/jn/nxy094.

[24] Strengers JG, den Ruijter HM, Boer JMA, Asselbergs FW, Verschuren WMM, van der Schouw YT, et al. The association of the Mediterranean diet with heart failure risk in a Dutch population. Nutr Metabol Cardiovasc Dis 2021;31(1):60–6. https://doi.org/10.1016/j.numecd.2020.08.003.

[25] Rezagholizadeh F, Djafarian K, Khosravi S, Shab-Bidar S. A posteriori healthy dietary patterns may decrease the risk of central obesity: findings from a systematic review and meta-analysis. Nutr Res 2017;41:1–13. https://doi.org/10.1016/j.nutres.2017.01.006.

[26] Livingstone KM, Sexton-Dhamu MJ, Pendergast FJ, Worsley A, Brayner B, McNaughton SA. Energy-dense dietary patterns high in free sugars and saturated fat and associations with obesity in young adults. Eur J Nutr 2022;61(3):1595–607. https://doi.org/10.1007/s00394-021-02758-y.

[27] Bakke Å, Cooper JG, Thue G, Skeie S, Carlsen S, Dalen I, et al. Type 2 diabetes in general practice in Norway 2005-2014: moderate improvements in risk factor control but still major gaps in complication screening. BMJ Open Diabetes Res Care 2017;5(1): e000459. https://doi.org/10.1136/bmjdrc-2017-000459.

[28] Ramyaa R, Hosseini O, Krishnan GP, Krishnan S. Phenotyping women based on dietary macronutrients, physical activity, and body weight using machine learning tools. Nutrients 2019;11(7): 1681. https://doi.org/10.3390/nu11071681.

[29] Gorst-Rasmussen A, Dahm CC, Dethlefsen C, Scheike T, Overvad K. Exploring dietary patterns by using the treelet transform. Am J Epidemiol 2011;173(10):1097–104. https://doi.org/10.1093/aje/kwr060.

[30] Schoenaker DAJM, Dobson AJ, Soedamah-Muthu SS, Mishra GD. Factor analysis is more appropriate to identify overall dietary patterns associated with diabetes when compared with treelet transform analysis. J Nutr 2013;143(3):392–8. https://doi.org/10.3945/jn.112.169011.

[31] Assi N, Moskal A, Slimani N, Viallon V, Chajes V, Freisling H, et al. A treelet transform analysis to relate nutrient patterns to the risk of hormonal receptor-defined breast cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC). Publ Health Nutr 2016;19(2):242–54. https://doi.org/10.1017/S1368980015000294.

[32] Schulze MB, Hoffmann K, Kroke A, Boeing H. Risk of hypertension among women in the EPIC-potsdam study: comparison of relative risk estimates for exploratory and hypothesis-oriented dietary patterns. Am J Epidemiol 2003;158(4):365–73. https://doi.org/10.1093/aje/kwg156.

S1 Supplementary material

S1.1 Flowchart for all participants

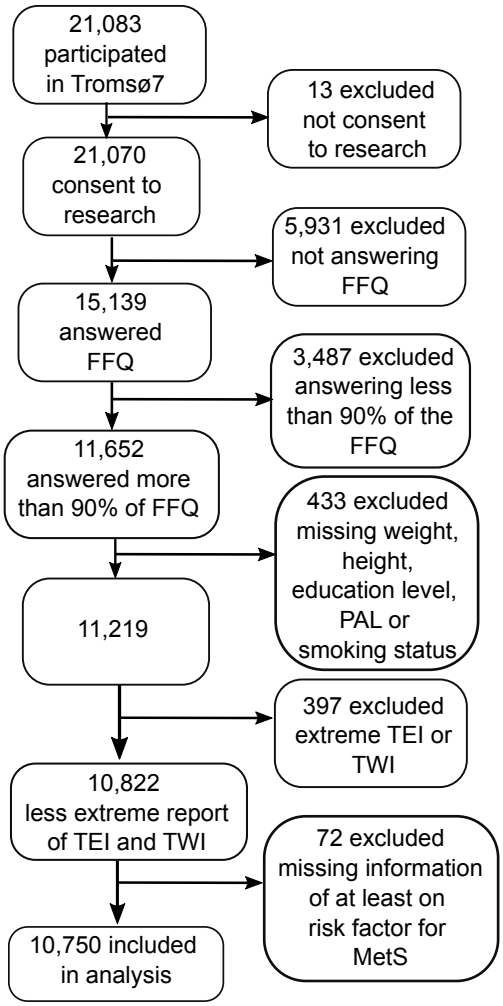


Figure S1: Flowchart illustrating exclusion of participants due to not giving consent to research, not answering the food frequency questionnaire (FFQ), having unrealistic total energy intake (TEI) or total water intake (TWI). Also, participants not answering questions on weight, height, educational level, physical activity level (PAL), smoking status or at least of of the components of Metabolic syndrome (MetS) were excluded.

S1.2 Estimating associations between diet scores and the components of MetS

This section presents the full results fitting the main logistic regression model (M3), assessing associations between the dietary patterns found by hierarchical clustering and the components of MetS. Each response variable is modelled as a linear function of diet scores, adjusted for the linear effects of age, educational level, physical activity level, smoking status and alcohol consumption:

$$\text{Response} \quad \sim \quad \text{Age} + \text{Education} + \text{Diet scores} + \text{Physical activity} + \text{Smoking} + \text{Alcohol}$$

The response variables include insulin resistance, hypertension, elevated waist circumference, elevated levels of triglycerides and low HDL-cholesterol (Table S1 - S5).

Table S1: Odds ratios (OR) with confidence intervals (CI) in explaining insulin resistance for women and men.

	Women		Men	
Covariates	OR	99% CI	OR	99% CI
Dietary patterns:				
Meat and Sweets	1.11	(0.97, 1.27)	1.01	(0.87, 1.18)
Plant-based- and Tea	1.01	(0.89, 1.14)	1.04	(0.88, 1.22)
Traditional	0.93	(0.81, 1.07)	0.95	(0.84, 1.09)
Age	1.06	(1.05, 1.08)	1.06	(1.05, 1.07)
Educational level:				
Primary/partly secondary	1		1	
Upper secondary	0.76	(0.55, 1.04)	0.91	(0.68, 1.23)
Short tertiary	0.66	(0.45, 0.97)	0.85	(0.61, 1.17)
Long tertiary	0.55	(0.39, 0.79)	0.59	(0.41, 0.83)
Physical activity level:				
Sedentary	1		1	
Light	0.72	(0.53, 1.00)	0.59	(0.44, 0.80)
High	0.46	(0.30, 0.72)	0.41	(0.29, 0.58)
Smoking status:				
Non-smoker	1		1	
Current smoker	1.42	(1.02, 1.96)	1.31	(0.93, 1.82)
Alcohol	0.82	(0.73, 0.92)	0.96	(0.90, 1.01)

Table S2: Odds ratios (OR) with confidence intervals (CI) explaining hypertension for women and men.

	Women		Men	
Covariates	OR	99% CI	OR	99% CI
Dietary patterns:				
Meat and Sweets	1.04	(0.95, 1.13)	1.04	(0.94, 1.15)
Plant-based- and Tea	0.91	(0.84, 0.99)	0.93	(0.82, 1.05)
Traditional	1.02	(0.93, 1.12)	0.98	(0.89, 1.08)
Age	1.10	(1.09, 1.11)	1.07	(1.06, 1.08)
Educational level:				
Primary/partly secondary	1		1	
Upper secondary	0.89	(0.70, 1.13)	0.90	(0.70, 1.15)
Short tertiary	0.68	(0.52, 0.88)	0.81	(0.63, 1.05)
Long tertiary	0.62	(0.49, 0.78)	0.69	(0.53, 0.88)
Physical activity level:				
Sedentary	1		1	
Light	0.66	(0.51, 0.84)	0.98	(0.76, 1.25)
High	0.51	(0.38, 0.68)	0.90	(0.69, 1.17)
Smoking status:				
Non-smoker	1		1	
Current smoker	0.72	(0.57, 0.90)	0.74	(0.57, 0.95)
Alcohol	1.04	(0.98, 1.09)	1.06	(1.02, 1.10)

Table S3: Odds ratios (OR) with confidence intervals (CI) explaining elevated waist circumference for women and men.

	Women		Men	
Covariates	OR	99% CI	OR	99% CI
Dietary patterns:				
Meat and Sweets	1.25	(1.15, 1.36)	1.34	(1.21, 1.48)
Plant-based- and Tea	0.92	(0.86, 1.00)	1.02	(0.90, 1.15)
Traditional	1.07	(0.99, 1.17)	1.06	(0.96, 1.16)
Age	1.03	(1.02, 1.03)	1.02	(1.01, 1.03)
Educational level:				
Primary/partly secondary	1		1	
Upper secondary	1.06	(0.85, 1.33)	0.95	(0.76, 1.19)
Short tertiary	0.79	(0.62, 1.01)	0.91	(0.71, 1.15)
Long tertiary	0.62	(0.50, 0.77)	0.55	(0.43, 0.70)
Physical activity level:				
Sedentary	1		1	
Light	0.54	(0.43, 0.69)	0.45	(0.35, 0.57)
High	0.30	(0.23, 0.39)	0.25	(0.19, 0.33)
Smoking status:				
Non-smoker	1		1	
Current smoker	0.79	(0.64, 0.99)	0.66	(0.51, 0.85)
Alcohol	0.99	(0.94, 1.04)	1.06	(1.03, 1.10)

Table S4: Odds ratios (OR) with confidence intervals (CI) explaining elevated triglycerides for women and men.

	Women		Men	
Covariates	OR	99% CI	OR	99% CI
Dietary patterns:				
Meat and Sweets	1.06	(0.97, 1.17)	1.03	(0.94, 1.14)
Plant-based- and Tea	0.93	(0.85, 1.01)	0.87	(0.77, 0.98)
Traditional	0.91	(0.82, 1.00)	0.90	(0.82, 0.99)
Age	1.02	(1.01, 1.03)	0.98	(0.97, 0.98)
Educational level:				
Primary/partly secondary	1		1	
Upper secondary	0.94	(0.75, 1.20)	1.04	(0.83, 1.31)
Short tertiary	0.71	(0.54, 0.93)	1.08	(0.85, 1.37)
Long tertiary	0.58	(0.45, 0.74)	0.82	(0.65, 1.04)
Physical activity level:				
Sedentary	1		1	
Light	0.70	(0.55, 0.89)	0.81	(0.64, 1.02)
High	0.40	(0.29, 0.54)	0.49	(0.38, 0.64)
Smoking status:				
Non-smoker	1		1	
Current smoker	1.50	(1.19, 1.88)	1.08	(0.85, 1.38)
Alcohol	0.94	(0.88, 1.00)	1.01	(0.98, 1.05)

Table S5: Odds ratios (OR) with confidence intervals (CI) explaining low HDL-cholesterol for women and men.

	Women		Men	
Covariates	OR	99% CI	OR	99% CI
Dietary patterns:				
Meat and Sweets	1.10	(1.00, 1.21)	1.08	(0.95, 1.21)
Plant-based- and Tea	0.95	(0.87, 1.04)	1.03	(0.88, 1.19)
Traditional	1.00	(0.90, 1.11)	0.92	(0.82, 1.04)
Age	0.98	(0.97, 0.99)	0.98	(0.97, 0.99)
Educational level:				
Primary/partly secondary	1		1	
Upper secondary	0.83	(0.64, 1.07)	1.07	(0.81, 1.41)
Short tertiary	0.55	(0.41, 0.74)	1.00	(0.74, 1.35)
Long tertiary	0.55	(0.42, 0.72)	0.65	(0.48, 0.89)
Physical activity level:				
Sedentary	1		1	
Light	0.61	(0.48, 0.78)	0.63	(0.48, 0.82)
High	0.33	(0.24, 0.45)	0.35	(0.26, 0.48)
Smoking status:				
Non-smoker	1		1	
Current smoker	1.42	(1.12, 1.81)	1.21	(0.90, 1.61)
Alcohol	0.81	(0.74, 0.88)	0.87	(0.82, 0.92)

S1.3 Correlations between dietary patterns found by different methods

Table S6: Correlation matrix between dietary pattern found by hierarchical clustering (HC1 - HC3), factor analysis (FA1 - FA4) and the treelet transform (TT1 - TT4). The patterns HC1 - HC3 denote the previously found diet groups in [1], representing the Meat and Sweets diet, the Plant-based- and Tea diet and the Traditional diet.

	HC1	HC2	HC3	PCA1	PCA2	PCA3	PCA4	TT1	TT2	TT3	TT4
HC1	1	-0.15	-0.31	0.87	-0.12	-0.05	0.01	0.77	-0.13	0.17	-0.06
HC2		1	-0.25	-0.07	0.81	0.21	-0.05	-0.11	0.89	-0.06	-0.11
HC3			1	-0.49	-0.35	-0.17	0.37	-0.26	-0.23	0.06	0.44
PCA1				1	0.06	-0.13	-0.11	0.87	-0.04	-0.01	-0.22
PCA2					1	-0.03	0.01	-0.03	0.85	-0.19	-0.13
PCA3						1	0.05	-0.24	0.17	0.72	-0.06
PCA4							1	-0.23	0.08	0.07	0.87
TT1								1	-0.09	-0.13	-0.23
TT2									1	-0.07	-0.04
TT3										1	-0.04
TT4											1

S1.4 Predictive power using categorised diet scores

Table S7: AUC values predicting MetS and its components including the covariates age, educational level and categorised (quartiles) diet scores (M2) and adding lifestyle variables (M3). The methods include hierarchical clustering (HC), factor analysis (FA) and the treelet transform (TT).

			AUC		
Response	Sex	M2/M3	HC	FA	TT
MetS	Women	M2:	0.650	0.654	0.649
		M3:	0.678	0.682	0.680
	Men	M2:	0.593	0.612	0.596
		M3:	0.635	0.645	0.635
Insulin resistance	Women	M2:	0.705	0.709	0.710
		M3:	0.718	0.723	0.724
	Men	M2:	0.686	0.692	0.690
		M3:	0.701	0.706	0.704
Hypertension	Women	M2:	0.763	0.763	0.762
		M3:	0.766	0.766	0.765
	Men	M2:	0.686	0.688	0.684
		M3:	0.688	0.689	0.686
Elevated waist circumference	Women	M2:	0.625	0.634	0.625
		M3:	0.652	0.661	0.655
	Men	M2:	0.601	0.631	0.605
		M3:	0.657	0.675	0.662
Elevated triglycerides	Women	M2:	0.602	0.607	0.602
		M3:	0.632	0.634	0.632
	Men	M2:	0.594	0.596	0.590
		M3:	0.620	0.620	0.617
Low HDL cholesterol	Women	M2:	0.611	0.609	0.598
		M3:	0.650	0.651	0.646
	Men	M2:	0.576	0.574	0.573
		M3:	0.643	0.639	0.639

S1.5 Results adjusting for body mass index (BMI)

According to a recent review [2], about half of the studies on diet and MetS included BMI as a confounder in the analysis. One problem is that BMI is highly correlated with elevated waist circumference which is one of the defining components of MetS, here used as a response variable. In our material this correlation is about 0.9. Due to this similarity with one of the response variables, we have chosen to report results adjusting for BMI separately in this supplement.

This section includes estimation of associations and predictive power when the models are adjusted for BMI. This implies that we run logistic regression models expressed by

$$\text{Response} \quad \sim \quad \text{Age} + \text{Education} + \text{Diet scores} + \text{Physical activity} + \text{Smoking} + \text{Alcohol} + \text{BMI}$$

where the response denotes MetS or one of its five components. Similarily, we have included BMI as an explanatory variable using the random forest algorithm.

S1.5.1 Estimation of associations including BMI

Adjusting for BMI, the positive association between the Meat and Sweets diet score and MetS was no longer significant, giving an odds ratio of 0.97 for women and 1.05 for men (Table S8, see Table 3 for comparison). The positive association between the Meat and Sweets diet and elevated waist circumference remained significant for men, but reducing the odds ratio from 1.34 to 1.25. For women, the odds ratio decreased from 1.25 to 1.05, giving a non-significant association. A possible interpretation of these results is that BMI acts as a mediator, fully or partially mediating the effects of diet scores on MetS and its components.

Adjusting for BMI, a higher intake of the Plant-based- and Tea diet was still associated with decreased odds of elevated triglycerides in men. In addition, a higher intake of the Traditional diet was associated with decreased odds of MetS for both women (OR = 0.88) and men (OR = 0.84).

Table S8: Odds ratios (OR) with confidence intervals (CI) for MetS and its five components, based on intake scores for the Meat and Sweets diet (HC1), the Plant-based- and Tea diet (HC2) and the Traditional diet (HC3). Abbreviations: Waist circumference (WC), Triglycerides (TG), HDL cholesterol (HDL-C).

		HC1		HC2		HC3	
Response	Sex	OR	99% CI	OR	99% CI	OR	99% CI
MetS	Women	0.97	(0.87, 1.08)	0.92	(0.82, 1.02)	0.88	(0.79, 0.98)
	Men	1.05	(0.93, 1.18)	0.98	(0.84, 1.13)	0.84	(0.75, 0.94)
Insulin	Women	1.03	(0.89, 1.19)	1.03	(0.90, 1.17)	0.91	(0.79, 1.05)
resistance	Men	0.95	(0.82, 1.11)	1.05	(0.88, 1.24)	0.93	(0.81, 1.06)
Hypertension	Women	0.97	(0.88, 1.06)	0.92	(0.84, 1.00)	1.00	(0.91, 1.10)
	Men	0.97	(0.87, 1.08)	0.92	(0.82, 1.05)	0.96	(0.87, 1.06)
Elevated	Women	1.05	(0.93, 1.19)	0.88	(0.79, 0.99)	0.96	(0.85, 1.09)
WC	Men	1.25	(1.07, 1.45)	1.04	(0.87, 1.24)	1.02	(0.88, 1.18)
Elevated	Women	0.98	(0.88, 1.08)	0.94	(0.85, 1.03)	0.87	(0.79, 0.97)
TG	Men	0.97	(0.88, 1.07)	0.86	(0.76, 0.97)	0.88	(0.79, 0.97)
Low	Women	1.00	(0.91, 1.11)	0.96	(0.87, 1.06)	0.97	(0.87, 1.08)
HDL-C	Men	1.01	(0.89, 1.14)	1.03	(0.88, 1.20)	0.89	(0.79, 1.00)

S1.5.2 Predictive power including BMI

To investigate the predictive power of the models adjusted for BMI, these were fitted and run for the same 100 training and test sets as before. Inclusion of BMI as a confounder naturally increased the predictive power in all cases (Table S9, see Table 5 for comparison). Specifically, the overall average AUC value across methods, response variables and gender, increased from 0.67 using M3, to 0.78 when BMI was included. Unsurprisingly, the highest power was seen in predicting elevated waist circumference in which the overall average AUC value increased from about 0.67 to 0.93, implying almost perfect classification. Inclusion of BMI was also seen to clearly increase the overall average AUC value in predicting MetS, giving a power of approximately 0.82. For the other components, the average absolute increase in AUC including BMI was approximately 0.05 (insulin resistance), 0.03 (hypertension), and 0.08 (elevated triglycerides and low HDL-cholesterol).

Table S9: AUC values predicting MetS and its components including the covariates age, educational level, diet scores, lifestyle variables and BMI. The methods include hierarchical clustering (HC), factor analysis (FA), the treelet transform (TT) and random forest (RF).

		AUC			
Response	Sex	HC	FA	TT	RF
MetS	Women	0.818	0.819	0.817	0.813
	Men	0.826	0.826	0.825	0.817
Insulin resistance	Women	0.775	0.774	0.774	0.741
	Men	0.751	0.748	0.750	0.734
Hypertension	Women	0.789	0.789	0.788	0.772
	Men	0.732	0.731	0.731	0.716
Elevated waist circumference	Women	0.931	0.930	0.930	0.921
	Men	0.941	0.941	0.941	0.928
Elevated triglycerides	Women	0.724	0.725	0.724	0.718
	Men	0.694	0.692	0.692	0.684
Low HDL cholesterol	Women	0.738	0.738	0.738	0.718
	Men	0.713	0.715	0.714	0.693

References

[1] Å. M. Moe, S. H. Sørbye, L. A. Hopstock, M. H. Carlsen, O. Løvsletten, E. Ytterstad, Identifying dietary patterns across age, educational level and physical activity level in a cross-sectional study: the Tromsø Study 2015 - 2016, BMC Nutr 8 (1) (2022) 102. doi:10.1186/s40795-022-00599-4.

[2] R. Fabiani, G. Naldini, M. Chiavarini, Dietary Patterns and Metabolic Syndrome in Adult Subjects: A Systematic Review and Meta-Analysis, Nutrients 11 (9) (2019). doi:10.3390/nu11092056.

Paper III

Analysis of dietary pattern effects on metabolic risk factors using structural equation modelling

Submitted.


Chips
Pizza