Project description

1) Objectives

The overall goal of this project is to develop and apply statistical methods and machine learning to gain new and valuable knowledge about the effectiveness of cardiovascular health preventive interventions in Denmark. Most existing studies that link diet and cardiovascular health are questionnaire-based. Here, we present a novel approach by developing methods to analyse a large-scale time series of grocery shopping transaction data from various Danish supermarkets, which grants the opportunity to obtain a more detailed and unbiased picture of associations between actual dietary patterns and cardiovascular health. Furthermore, we will contribute with valuable knowledge on how to explore methods and algorithms for linking activity tracker data with national registries, which will allow us to assess interventions that aim to increase health through physical activity. These analyses will learn evidence based on a study of walking activities.

2) Success criteria

1) PhD-thesis containing at least three scientific articles. Results are presented at national and international conferences.
2) Knowledge gained on methods to handle associations between cardiovascular health and grocery transaction/activity tracker data.
3) A report is completed on how the results on effective ways of preventing heart disease can be implemented in the Danish Heart Foundations's patient activities, counselling offers and work to impact health policy decisions.

3) Introduction and state-of-the-art background

Diet

A healthy diet can help prevent cardiovascular disease (CVD) and decrease the risk of the disease worsening (Jensen et al., 2018). This is supported in the existing literature, where traditionally single nutrients have been linked to the risk of CVD (Cespedes et al., 2015), with trans fatty acids, sodium and omega 3 polyunsaturated fatty acids being the specific nutrients that are linked consistently. In recent years, focus has shifted towards the effects of overall diet, using a dietary pattern approach, as research has shown that single nutrients have effects of limited magnitude on CVD as compared to complex integrated dietary interventions, and it might be difficult to translate single nutrient-based recommendations into effective population wide interventions (Ravera et al., 2016).

Different diets have been investigated in multiple observational and intervention al studies, where the Mediterranean Diet (MED) and the Dietary Approach to Stop Hypertension (DASH) have been shown to reduce several aspects of CVD (Ravera et al., 2016). Consequently, current national nutritional guidelines focus on overall diet patterns including various parts of the diet. However, there is still a lack of evidence for such overall diet pattern recommendations (Nissen, 2016). A main weakness of previous questionnaire-based diet studies is that they do not give a very detailed picture of the actual dietary patterns (Bingham, 1991) and they cannot necessarily assess changes over time.

Analysing grocery transaction data grants the opportunity to investigate diet patterns, including the possibility of the effect of single nutrients or nutrient groups, considering different exposures based on a larger-scale objective data collection over time (Ransley et al., 2000). A Danish study from 2017 shows an association between unemployment and food purchase behaviour by analysing longitudinal data from a 5-year-period consisting of monthly grocery transaction data reported from households aligned with registry

data on unemployment (Smed et al., 2017). Another English study links childhood weight status and annual sales of unhealthy food (Wilsher et al., 2016).

A foundation for this project is the assumption that grocery transaction data from different su permarkets collected through time can give insight into the association between dietary patterns and CVD, which is an undiscovered field in the existing literature. How to approach this using existing and novel statistical and machine learning methods will be elaborated on in "Modelling approaches". The evidence learned from the analyses will be used, by adapting health economic theory to this framework (Komorowski, 2016), as a foundation to investigate what preventive diet interventions are cost-effective in helping high risk groups take care of their hearts and ensure a better quality of life for heart patients.

Physical activity

Walking groups and walking activities have well-documented health-promoting effects for both healthy persons and persons already living with CVD (Ried-Larsen, 2020). Thus, in concordance with the organisational strategy, a pilot phase of a walking project will be initiated in spring 2021, with data based on a pedometer on the wrist. The common practice to gather data on physical activity habits is still questionnaire based (Jensen et al., 2018), but an objective measure like a pedometer does not depend on memory or ability to answer, which gives more unbiased results (Harris et al., 2009; Matthiessen, 2016). Analyses of larger scale pedometer data have been found in other countries, such as Japan (Takamiya et al., 2019) and Finland (Hirvensalo et al., 2011), and in a Danish context, the national investigations of diet and physical exercise (DANSDA) also use trackers to collect physical activity data (Fagt et al., 2020).

The novelty in our study is that we aim to:

- Link tracker data to national registries containing health information.
- Analyse the impact on cardiovascular health of walking interventions, compared to mostly non-interventional studies in the literature.
- Use various machine learning techniques (see "Modelling approaches"), instead of descriptive statistics and logistic regression which are traditional methods for walking activity studies (Matthiesen et al., 2015).
- Develop health economic approaches to handle walking tracker data and analyse the cost-effectiveness concerning cardiovascular health on walking activities.

4) Project execution

To fulfill the project objectives and answer the research questions, the project will be executed as described in the following.

Diet

The Storebox app and SMIL study

To investigate research question 1) and 2), we will use longitudinal transaction data from the SMIL database, which contains electronic receipts from the Storebox app. Storebox is an app, where the user can gather receipts from various Danish supermarkets, including Netto, Føtex, MENY, Bilka and REMA 1000. Credit card information is typed in the app, and each time the credit card is used for purchase, the user receives an electronic receipt. Currently, the app has over 1 million users in Denmark.

After the app is downloaded, users can choose to become a part of the SMIL study (run by Aalborg University Hospital), thereby giving permission to share their receipts with researchers in a protected environment where their identity is encrypted. Consenting users will provide t heir Civil Registration Number which is collected with a Storebox identification code and transferred to Statistics Denmark. The SMIL study was initiated in 2018 and there are currently (Apr 2021) around 10,000 participants, with 40 million single transactions. The participants can choose to leave the study at any time.

To sum up, the facilities of Statistics Denmark combined with Storebox make it possible to combine a time series of grocery shopping patterns with different relevant health data, ensuring a high level of individual integrity.

Registries

We will use the following registries provided by Statistics Denmark:

- The Danish Civil Registration System: sex, date of birth, vital status, civil status, immigration/emigration status, country of origin and area of residence.
- The Danish National Patient Registry: hospital contacts with attached discharge diagnosis codes and operative procedures coded by the ICD-10 system since 1994.
- The Danish National Prescription Registry: claimed prescriptions with ATC-codes.
- The Population's Education Registry: attained educational level.
- The Danish income register: income and household information.

Categorisation of transactions

The large-scale transaction data from Storebox contains a huge number of unique foods from different categories. For example, many different specific fruit juices can be found in various supermarkets, which leads to categorisation being necessary. The food institute at DTU (the Danish Technical University) maintains the Frida Food database (https://frida.fooddata.dk/), where the majority of food sources sold in Denmark can be categorized into approximately 1,100 Frida food names, such as "cabbage, red, raw" or "ymer, low fat", then again into approximately 40 higher level Frida food groups (such as "root and tuber vegetables" or "biscuits and cookies"), and then again into 16 higher level categories, such as "fish and fish products" or "egg and egg products". By using available categories combined with manual sorting, the products from the Storebox receipts have been categorized at DTU to more general food groups based on the Frida database. This categorisation is not complete, but is an ongoing process that needs to be maintained.

Modelling approaches

The structure of the purchase data is quite complex: each participant has multiple observations of food transactions over time, and the transactions are quite irregular, meaning that the frequency of transactions differs over the weeks, months and from person to person. Furthermore, we will have new participants entering the study and participants dropping out, or even participants dropping out and entering again. A good starting point to handle the longitudinal data is to consider the econometric approach to time series analysis (Cryer, 2008). A central part of our work is to adapt these methods to handle another framework, namely transactions of groceries over time. The approach to handle data as recurrent event data with time gaps will also be considered (Shen et al., 2020). Other tools which we plan to use to handle this complex time structure are unsupervised clustering methods, for example k-means clustering, where small clusters of similar time series can be created (Tan et al., 2015).

Each receipt contains information about the purchased products: product name, the amount that was bought, the price of the product, discounts (if any), purchase date and time and the price of the entire order. Based on existing literature for supermarket transaction data, it will be assessed how to model these variables sufficiently, by investigating whether to use absolute price for an item, relative price as compared to the total price, how to incorporate the discounts, whether to consider participants only or entire households and so on.

Structures and patterns in the purchases will be assessed using various data mining techniques, for example the unsupervised machine learning technique called "association rule mining", where data will be considered as market basket transactions. Here, the basic idea is to find frequent itemsets in the transactions and form association rules that represent the relationship between these itemsets (Tan et al., 2019; Hastie et al., 2009). A central part of our work will be to adapt the framework of marked point

processes (Last et al., 1995) to use in this association analysis, which has not been seen in the existing literature, and this work will thus lie in the field in between unsupervised and supervised learning. Furthermore, extensions on how to associate these grocery shopping patterns to the risk of CVD do not exist and need to be developed. Here, the causality between food purchase patterns and CVD risk will be considered (Hernan et al., 2020). In the modelling process, different sociodemographic variables from the national registries such as gender, age, income and education will be taken into account. In this way, we can assess societal inequalities in CVD, such as the fact that people in Denmark with no higher than basic education have two to three times higher risk of dying of CVD as compared to their peers with a higher education (Hjerteforeningen, 2020).

We hypothesise that detecting patterns in the grocery transactions like mentioned will be useful to suggest optimal and effective CVD diet interventions. To assess cost-effectiveness of suggested interventions, we will include the effect measures QALY (incremental cost per quality-adjusted life-year) and ICER (incremental cost-effectiveness ratio) (Komorowski, 2016).

Another method that will be explored is reduced rank regression, which can be used efficiently in nutritional epidemiology by choosing a disease-specific response variable and determining combinations of food transactions that explain as much response variation as possible (Hoffmann, 2003). This method is not new in epidemiological literature, however, ways of adapting this to a large-scale transaction data need to be developed. Further modelling approaches will be considered as part of the PhD project.

Perspectives

- Presenting novel methods to investigate actual grocery habits/dietary patterns based on transactions instead of questionnaires.
- Analysing grocery transactions as a time series (seasonality), which can contribute to knowing precisely when the interventions will be most effective.
- Addressing association between grocery habits and CVD for different exposures (linking various Danish registries with the transaction data).
- Addressing dietary patterns using data driven machine learning techniques, which can assist in discovering effective interventions concerning cardiovascular health.
- Assisting health policy decisions concerning cost-effective diet interventions on both primary and secondary prevention of CVD.

Physical Activity

Data sources

We will use data from a prospective cohort study with one year follow-up organised by the Danish Heart Foundation, which will be initiated in the spring 2021. Currently, it is expected that the cohort will consist of around 25% of the 4,400 participants already signed up for the expansion of the walking project. Data consisting of physical activity will be gathered continuously over a year with activity trackers (Garmin Vivofit 4) placed on the wrist. Data is uploaded directly to an online database, through the online system Easytrial (https://www.easytrial.net/), where electronic case report forms can be created and the physical activity data can be stored.

Information on the following will be collected through the activity tracker:

- Total volume, intensity and frequency of physical activity.
- Step counts (by walking).
- Overall sleeping patterns / quality.

Information on the following will be collected through electronic questionnaires three times a year:

- Lifestyle factors (smoking, alcohol, diet, weight). It is planned to use a modified version of the questionnaire used in the national health profile (http://www.danskernessundhed.dk/).
- Quality of life (general health, emotional health, physical health, social activities). It is planned to use the SF36 or SF12 questionnaire.

Data will be linked to relevant administrative and health registries concerning education, income, visits to the general practitioners, diagnoses, prescriptions and hospitalisations. In this way we can address societal inequalities and identify groups at high risk of CVD based on their physical activity patterns.

Modelling approaches

Data relies on participants wearing the tracker for a year, which will likely create missing time gaps and a quite unstable time series, as some participants might use the tracker more frequently than others. This will be taken into account in the modelling process, for example by a nonparametric analysis of recurrent gap time data (Shen et al., 2020) or using individual and group information centred methods to handle missing tracker data (Kang et al., 2013). A novel approach of adapting different machine learning methods, such as k-nearest neighbour self-organisation maps, to handle missing activity tracker data will be considered (Jerez et al., 2010).

Adherence to the planned walking activities and change in the number of steps walked after the intervention will be assessed, including the variation between participants using a mixed-effects model with a random effect on subject. To take societal inequalities into account, multiple sociodemographic factors (sex, age, education, income) as well as smoking, alcohol and dietary patterns will be included in the model. Apart from this more traditional approach, we will also adapt machine learning methods to our framework, such as support vector machines, to classify the degree of adherence to and impact of the different walking interventions (Hastie et al., 2009).

Part of the work will explore how to link the data on physical activity to both the dietary patterns derived from the transaction data described in the "Diet" section and cardiovascular health. The effect of the walking activities on health economic costs (for example quantified by number of hospitalisations) will also be explored. Here, we will adapt existing health economic theory to the framework of walking interventions based on activity tracking data to be able to rank and compare interventions by cost-effectiveness (Vliet, 2020; Komorowski, 2016). Further modelling approaches will be considered as part of the PhD project.

Perspectives

- Adapting various machine learning methods to handle physical activity tracking data.
- Investigating associations between walking activities and lifestyle (diet, smoking, alcohol).
- Informing health policy decisions about the impact of walking activities on health benefits through physical activity based on data directly reflecting the target population.
- Determining the level of adherence to the walking activities, which can support decisions about how to make subgroup specific effective interventions.