Innovation Fund Denmark

# Template for Industrial PhD project description

This form **must** be filled in as it is. Delete the instructions in grey italics before you submit the application. You **must** use Calibri 11 pt. for the text, which must stay inside the boxes. The boxes and their borders may **not** be moved. If the borders are moved this can lead to an administrative rejection.

## Basic information

| | |
|---|---|
| Project title | Statistical Analyses of Cardiovascular Disease Preventive Interventions in Denmark |
| Industrial PhD candidate | Emilie Prang Nielsen |
| Company | The Danish Heart Foundation |
| University, centre/department | University of Copenhagen, Institute of Public Health |
| Any third parties | Bente Klarlund Pedersen, Center for Aktiv Sundhed |

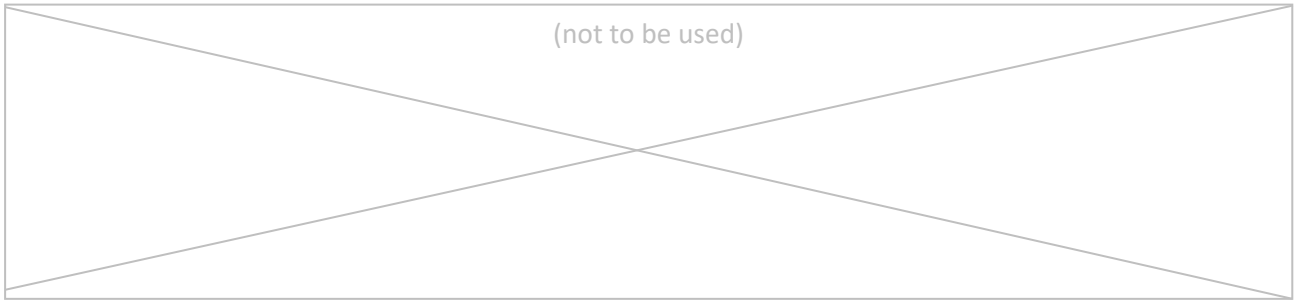
## A. Objectives and success criteria

- **The project's 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. The project's objective directly supports the Danish Heart Foundation's strategic goal of helping groups at high risk of cardiovascular disease (CVD) take care of their hearts, focusing on two preventive lifestyle domains, diet and physical activity. Evidence created through the project will support and guide the company's diet and physical exercise advice. Moreover, the results contribute to informing health policy decisions on cost-effective ways of preventing CVD.

- **The project's success criteria**

1) PhD-thesis containing at least three scientific articles. Results are presented at national and international conferences. 2) Semi-annual newsletter to update employees and members of the Danish Heart Foundation on preliminary research results. 3) Knowledge gained on methods to handle associations between cardiovascular health and grocery transaction/activity tracker data. 4) A report is completed on how the results on effective ways of preventing heart disease can be implemented in the company's patient activities, counselling offers and work to impact health policy decisions.



## B. Commercial significance and impact

- **The results' expected contribution to the company's business and/or earnings**

The Danish Heart Foundation wishes to contribute to the fight against heart disease significantly, and through time, extensive work has been done to support this strategy. This includes multiple aspects and perspectives within prevention, research, volunteer work, patient support, fundraising, communication and political work. Recently, there has been increased focus on collaborating broadly within the company, which this PhD project supports by ensuring collaboration especially between the research and prevention departments. This helps obtaining the most effective ways to "help high risk groups take care of their hearts", which is one of the newly established core subgoals in the strategy. Apart from contributing to new strategy goals and way of working in the organisation, this project will strengthen the quality of information and communication about how to live a heart healthy life. Furthermore, the project will assist the Danish Heart Foundation's political work by gaining knowledge on cost-effective CVD preventive interventions, which the company can use to influence health policy decisions. Lastly, the project will provide knowledge on how to establish and expand new products that increase earnings. A detailed plan for implementing results and realising these business contributions are given in below.

- **Plan for implementing results and realising business contributions**
  1) **Focus on broad collaborations to improve business processes:** This project is designed to assist the company's recent focus on broad collaborations and communication across departments and skill areas. Close collaborations with the prevention, public affairs and fundraising departments as well as communication through a semi-annual newsletter ensure a broad knowledge sharing. This helps getting constructive feedback during the PhD process to guarantee useful and high-quality evidence, which the company values highly.
  2) **Strengthening the quality of information and communication about how to live a heart healthy life:** This PhD project contributes to valuable knowledge on effective preventive interventions by analysing associations between cardiovascular health and diet patterns as well as physical activity, based on data directly from the target population. This will result in improving the company's communication in different channels, for example by more recipes on the website or assisting counselling by "Hjertelinjen", which will strengthen already existing and popular business products as well as contribute to positioning the company as a site of expertise driven by both internal and external research. The members value this highly, and the project is thus expected to attract more members and increase the earnings via fundraising.
  3) **Build up knowledge to increase political influence:** A goal for the Danish Heart Foundation is to be more visible and have more influence on the political scene. The results from this PhD project will grant the opportunity to affect the health area politically, by assisting with specific knowledge and suggestions on which structural interventions within diet and physical exercise are most effective. Examples could be to quantify a tax on specific foods with respect to the health gain or to suggest which walking activities are most cost-effective with respect to increased physical activity. These results will be discussed and used directly by the public affairs team as well as the strategic prevention team to ensure the highest political impact as possible.
  4) **Establishing new products that increase earnings:** The focus of this PhD project is to gather knowledge on how to obtain a heart healthy lifestyle effectively for groups at high risk of CVD. The results and the candidate's increasing knowledge will give the company the best possible foundation to develop new direct offers for the patients/members. The project is easy to profile and communicate, for example at the yearly TV-show "Hjertegalla", to create awareness of the ambition to unite patient offers with research and evidence, which is expected to increase funding. We will analyse data from the pilot phase of a walking initiative financed, now, by the Danish Heart Foundation and partly financed by Sundhedsstyrelsens Sundhedsfremmepulje. The health economic results of this PhD are expected to contribute positively to the possibility of getting more external funding, thus being able to expand to a larger national walking initiative. The goal is to expand from 10 municipalities in the pilot phase (spring 2021) to 85 municipalities during the next phase (spring 2022), for which more external funding is necessary.

## C. State-of-the-art and theoretical background

- **State-of-the-art and, if relevant, theoretical background for the Industrial PhD project's field of research**

**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 interventional 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 supermarkets 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" in section D. 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: 1) Link tracker data to national registries containing health information. 2) Analyse the impact on cardiovascular health of walking interventions, compared to mostly non-interventional studies in the literature. 3) Use various machine learning techniques (see "Modelling approaches", section D), instead of descriptive statistics and logistic regression which are traditional methods for walking activity studies (Matthiesen et al., 2015). 4) Develop health economic approaches to handle walking tracker data and analyse the cost-effectiveness concerning cardiovascular health on walking activities. This will contribute substantially to the company's knowledge about improving offers for members and expanding the project nationally.

- **State-of-the-art references**

Bingham, S. A. (1991). Limitations of the various methods for collecting dietary intake data. *Ann Nutr Metab*, *35*(3), 117-127. doi: 10.1159/000177635.

Cespedes, E. M., Hu, F. B. (2015, Apr). Dietary patterns: from nutritional epidemiologic analysis to national guidelines. *The American Journal of Clinical Nutrition*, *101*(5), 899-900. doi: 10.3945/ajcn.115.110213.

Fagt, S., Jensen, A., Sørensen, M., Trolle, E., Nordman, M., Christensen, T., Henriksen, K., Kørup, K., Ygil, K., Christensen, C., Matthiesen, J. (2020). Bag om danskernes kost og fysiske aktivitet i 2020-2021. *E-artikel fra TU fødevareinstitut*, 5. https://www.dansda.food.dtu.dk/resultater.

Harris, T. J., Owen, C. G., Victor, C. R., Ekelund, U., Adams, R., Cook, D. G. (2009). A comparison of questionnaire, accelerometer, and pedometer: measures in older people. *Medicine and Science in Sports and Exercise, 41(7)*, 1392-1402. doi: 10.1249/mss.0b013e31819b3533.

Hirvensalo M., Telama R., Schmidt M. D. (2011). Daily steps among Finnish adults: variation by age, sex, and socioeconomic position. *Scand J Public Health 2011, 39*, 669–77. doi: 10.1177/1403494811420324.

Jensen, H. A. R., Davidsen, M., Ekholm, O., & Christensen, A. I. (2018). *Danskernes sundhed - Den Nationale Sundhedsprofil 2017*. Sundhedsstyrelsen. https://www.sst.dk/da/udgivelser/2018/~/media/73EADC242CDB46BD8ABF9DE895A6132C.ashx.

Komorowski, M. (2016). *Markov Models and Cost Effectiveness Analysis: Applications in Medical Research*. Springer. doi: 10.1007/978-3-319-43742-2_24.

Matthiessen, J., Andersen, E., W., Raustorp, A., Knudsen, V. K., Sørensen, M. R. (2015). Reduction in pedometer-determined physical activity in the adult Danish population from 2007 to 2012. *Scandinavian Journal of Public Health*, *43 (5)*, 525-533. http://www.jstor.org/stable/45151003.

Matthiessen, J. (2016). Danske Kvinder er blevet mindre fysisk aktive. *E-artikel fra DTU fødevareinstitut*, *1*, 1-8. https://core.ac.uk/download/pdf/43256471.pdf.

Nissen, S. E. (2016, Apr). *U.S. Dietary Guidelines: An Evidence-Free Zone*. Annals of Internal Medicine. doi: 10.7326/m16-0035.

Ransley, J. K., Donnely, J. K., Khara, T. N., Botham, H., Arnot, H., Greenwood, D. C., Cade, J. E. (2000, Nov). The use of supermarket till receipts to determine the fat and energy intake in a UK population. *Public Health Nutrition*, *4*(6), 1279-1286. doi: 10.1079/PHN2001171.

Ravera, A., Carubelli, V., Sciatti, E., Bonadei, I., Gorga, E., Cani, D., Vizzardi, E., Metra, M., Lombardi, C. (2016, Jun). Nutrition and Cardiovascular Disease: Finding the Perfect Recipe for Cardiovascular Health. *Nutrients*, *8*(6). doi: 10.3390/nu8060363.

Ried-Larsen, M. (2020). *Notat vedr. følgeforskning på Hjerteforeningens (HF) "Gå-projekt"*. Center for Physical Activity Research. https://aktivsundhed.dk/da/.

Smed, S., Tetens, I., Lund, T. B., Holm, L., Nielsen, A. L. (2017, Nov). The consequences of unemployment on diet composition and purchase behaviour: a longitudinal study from Denmark. *Public Health Nutrition*, *21*(3), 580-592. doi: 10.1017/S136898001700266X.

Takamiya, T., Inoue, S. (2019, Sep). Trends in Step-determined Physical Activity among Japanese Adults from 1995 to 2016. *Med Sci Sports Exerc.*, *51*(9), 1852-1859. doi: 10.1249/MSS.0000000000001994.

Wilsher, S. H., Harrison, F., Yamoah, F., Fearne, A., Jones, A. (2016, Feb). The relationship between unhealthy food sales, socio-economic deprivation and childhood weight status: results of a cross-sectional study in England. *International Journal of Behavioral Nutrition and Physical Activity*, *13*(21). doi: 10.1186/s12966-016-0345-2.

## D. Project description

- **Hypotheses / research questions**

The existing evidence on preventing CVD through diet and physical activity, which the Danish Heart Foundation bases their recommendations on, is mostly learned from questionnaire-based studies. Furthermore, health economic methods concerning cost-effective CVD preventive interventions adapted to grocery shopping data and activity tracking data are somehow fragmented and need to be developed. Due to these unmet needs, this PhD project wishes to investigate the following:

1) Modelling electronic receipts from various Danish supermarkets to analyse associations between a time series of grocery shopping and cardiovascular health.
2) How can linkage between dietary patterns and cardiovascular health be used to assess and compare health economic preventive interventions?
3) How do we learn from physical activity tracking data to analyse the impact and cost-effectiveness of various walking interventions on cardiovascular health?

In the modelling process, different sociodemographic variables will be included to assess societal inequalities and identify groups at high risk of CVD based on their diet and physical activity patterns. In the following, a more detailed and comprehensive description of how the research questions will be investigated is found. Note that question 1) and 2) relate to the section "Diet" in below and question 3) relates to the section "Physical Activity".

- **Project execution**

### 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 their 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**

- Supporting the Danish Heart Foundation's strategy to help high risk groups take care of their hearts and ensure better quality of life for heart patients by identifying possible preventive diet interventions.
- Assisting health policy decisions concerning cost-effective diet interventions on both primary and secondary prevention of CVD.
- 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.

## 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**

- Informing health policy decisions about 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.
- Adapting various machine learning methods to handle physical activity tracking data.
- Investigating associations between walking activities and lifestyle (diet, smoking, alcohol).

## Other activities and project organisation

The methodological development, data mining and analyses will be carried out by the candidate, supervised by professor in biostatistics at University of Copenhagen, Thomas Gerds. In this process, relevant cardiological knowledge will be provided by professor in cardiology and senior consultant cardiologist Gunnar Gislason, as well as other physicians from the research department in the Danish Heart Foundation. Interpretation of results concerning diet patterns will be done in collaboration with dietitians and nutrition experts from the Danish Heart Foundation, both from the telephone counselling service (Hjertelinjen) to get the direct patient perspective, and the prevention department to get the strategic prevention perspective. Interpretation of results concerning physical activity will be done in collaboration with the walking project leaders in the company, co-supervisor Mads Lind as well as relevant subject matter experts in Center for Aktiv Sundhed led by third party supervisor Bente Klarlund Pedersen.

To obtain input concerning the possible political impact of the proposed cost-effective dietary/walking interventions on cardiovascular health, we will need support from the public affairs department and the prevention department, where co-supervisor Mads Lind can contribute with valuable inputs. During the PhD process, relevant project parties as well as the rest of the organisation will be updated (and able to give feedback) on preliminary results in semi-annual newsletter. Updates about the project will also be given continuously on the website open to the public to ensure transparency and a large degree of information from the research.

- **Project description references, if any**

Cryer, J. D. (2008). *Time Series Analysis: With Applications in R* (Second ed.). Springer.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning* (Second ed.) Springer.

Hernan, M. A., Robins, J. M. (2020). *Causal Inference: What If* (First ed.). Boca Raton: Chapman & Hall/CRC.

Hjerteforeningen (2020). Fakta om hjerte-kar sygdom i Danmark. https://hjerteforeningen.dk/alt-om-dit-hjerte/noegletal/.

Hoffmann, K. (2003). Application of a New Statistical Method to Derive Dietary Patterns in Nutritional Epidemiology. *American Journal of Epidemiology*, *159*(10). doi: 10.1093/aje/kwh134.

Jerez, J., Molina, I., Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine 50*(2), 105-115. doi: 10.1016/j.artmed.2010.05.002.

Kang, M., Rowe, D. A., Barreira, T. V., Robinson, T., Mahar, M. T. (2013). Individual-Centered Approach for Handling Physical Activity Missing Data. *Research Quarterly for Exercise and Sports*, *8*(2). 131-137. 10.1080/02701367.2009.10599546.

Komorowski, M. (2016). *Markov Models and Cost Effectiveness Analysis: Applications in Medical Research*. Springer.

Last, G., Brandt, A. (1995). *Marked Point Processes on the real line: the dynamical approach.* Springer.

Shen, P., Lai, U. (2020). Nonparametric analysis of recurrent gap time data. *Communications in Statistics - Theory and Methods*, (49), 3298-3312. doi: 10.1080/03610926.2019.1588322.

Tan, S. C., Lau, P. S., Yu, X. (2015, May). Finding Similar Time Series in Sales Transaction Data. *Current Approaches in Applied Artificial Intelligence*, 645-654. doi: 10.1007/978-3-319-19066-2_62.

Tan, P. N., Steinbach, M., Karpatne, A., Kumar, V. (2019). *Introduction to Data Mining*. Pearson.

Vliet, N., Suijkerbuijk, A., Blaeij, A., Wit, G., A., Gils, P., F., Polder, J. J. (2020). Ranking Preventive Interventions from Different Policy Domains: What Are the Most Cost-Effective Ways to Improve Public Health? *International Journal of Environmental Research and Public Health*, 17, 2160. doi: 10.3390/ijerph17062160.

## E. Expected publications

| Proposed title and date of publication | Proposals for one or more acknowledged research journals as desired place of publication |
|---|---|
| Modelling Dietary Patterns by Analysing Associations between Grocery Shopping Items Based on a Time Series of Electronic Receipts from Various Danish Supermarkets (spring 2022) | Journal of the Royal Statistical Society (Series A) |
| A Health Economic Approach to Analyse and Compare different Cardiovascular Disease Preventive Interventions (spring 2023) | Statistics in Medicine |
| Analysing the Impact of Various Walking Activity Interventions on Health using a Time Series of Data from a Physical Activity Tracker (fall 2023) | American Journal of Epidemiology |
| PhD thesis (spring 2024) | |

**Non-academic publications:**

- Newsletter published semi-annually (start fall 2021) on the Danish Heart Foundation's internal website, public website and the members' magazine "Hjertenyt".
- A report on how the project results can be implemented in the Danish Heart Foundation's activities, counselling and as a tool to impact health policy decisions.

## F. Courses, conferences og stays abroad

**Courses at the PhD school of Faculty of Health and Medical Sciences:**

- Mandatory: Responsible Conduct of Research 1: An Introduction (2 ETCS).
- Mandatory: Responsible Conduct of Research 2: Getting Ready for Submission of Manuscript and Thesis (0,6 ETCS).
- Elective generic: Academic Writing – How to Create good texts (1,5 ETCS).
- Elective generic: Communication & presentation in the academic context (1,5 ETCS).
- Elective generic: Advanced Topics in Data Analysis (5 ETCS).
- Elective specialist: Epidemiology 2: Analysing continuous and dichtomonous data from observational studies (6 ETCS).

**Course at CBS:** Mandatory industrial PhD course (5 ETCS).
Furthermore, the PhD student will follow relevant statistical methodical and subject matter courses at University of Copenhagen as well as abroad, so that the total number of courses will be around 30 ETCS.

**Conferences, seminars**
Apart from relevant seminars and conferences in Denmark, the PhD student plans to participate in ICEPH (International Conference on Epidemiology and Public Health) in Sydney planned February 2022 and ISCB (International Society for Clinical Biostatistics) planned in August 2023.

**Stays abroad**
A three-month stay is planned at for example Stanford University, Berkeley University or University of Auckland, which all have distinct departments, courses and researchers within statistics and epidemiology. The intention is to gain and learn from highly qualified partners in these fields.

## G. Dissemination plan

| Dissemination at the company | | |
|---|---|---|
| **Dissemination type** | **When** | **Time in hours** |
| Report on implementation of results in the Danish Heart Foundation | Spring 2024 | 120 |
| Semi-annual newsletter published on the public website, internal website and the members' magazine Hjertenyt | Start fall 2021 | 100 |
| Other presentations at the Danish Heart Foundation | Continuously | 50 |
| **Dissemination at the university** | | |
| **Dissemination type** | **When** | **Time in hours** |
| Presentation at webinars/conferences in Denmark and abroad | Continuously | 100 |
| Internal seminar/webinar series at section of biostatistics | Continuously | 70 |
| Scientific articles | Start spring 2022 | 400 |
| PhD thesis | | - |
| **Total dissemination 840 hours** | | |

(not to be used)

## H. Structure and time schedule

During the project, it is expected that the PhD student on average is at the company 2-3 days a week and at the university 2-3 days a week. A newsletter on preliminary research results on the website and in Hjertenyt is a milestone for each semester during the entire project period.

| | Spring | Fall |
|---|---|---|
| 2021 | | **Focus**<br>• Recent developments in machine learning methods for longitudinal data.<br>• Grouping and ranking grocery items according to heart disease.<br>• Grocery shopping transaction data preparation using learnings from course "Advanced Topics in Data Analysis".<br>**Milestones**<br>• Detailed project plan.<br>• Methodical knowledge on how to handle transaction data.<br>• Data ready for analysis. |
| 2022 | **Focus**<br>• Grocery shopping transaction data analysis using national registries, developed methods and learnings from course "Analysing continuous and dichtomonous data from observational studies".<br>• Participation in International Conference on Epidemiology and Public Health (ICEPH).<br>**Milestones**<br>• Manuscript for submission (section E).<br>• Presentation of project at ICEPH. | **Focus**<br>• Statistical method development using health economic theory.<br>• Transaction data implementation.<br>**Milestones**<br>• New models for health economic CVD preventive interventions.<br>• Presentation and discussion with relevant stakeholders in the company on how to implement results on health economic CVD preventive interventions within diet. |
| 2023 | **Focus**<br>• Transaction data implementation of health economic methods using national registries.<br>• Methodical development: how to handle a time series of physical activity tracking data in a health economic framework.<br>• Study abroad (3 months).<br>**Milestones**<br>• Manuscript for submission (section E).<br>• Methodical knowledge on health economic approaches to handle tracking data. | **Focus**<br>• Tracking data implementation.<br>• Participation in conference International Society for Clinical Biostatistics (ISCB).<br>**Milestones**<br>• Analysis on health economic CVD preventive interventions within physical activity.<br>• Presentation and discussion with relevant stakeholders in the company on how to implement results.<br>• Manuscript for submission (section E).<br>• Presentation of project at ISCB. |
| 2024 | **Completing the project**<br>• Thesis writing.<br>**Milestones**<br>• PhD thesis.<br>• Complete report on how the Danish Heart Foundation can implement the results. | |

## I. Time allocation

| Allocation of the Industrial PhD candidate's time | in months | in % of project time |
|---|---|---|
| In Danish division of host company | 18 months | 50 % |
| In non-Danish divisions of host company | 0 months | 0 % |
| At other companies or organisations | 0 months | 0 % |
| At the host university | 15 months | 42 % |
| At other universities and research institutions | 3 months | 8 % |

## J. Company

- **The company and its activities**

**History:** The Danish Heart Foundation was founded by a group of doctors in 1962. At that time, the heart-lung machine revolutionised the treatment of heart disease, however, a large number of people still died of cardiovascular disease, which was a threat to the Danish welfare state. Therefore, there was good reason to form a foundation that could carry out research and gather important knowledge and information in the interest of patients. In the beginning, the Danish Heart Foundation focused mostly on research in treatment of patients, but during the 1970s, the organisation started preventive work by communicating the importance of physical exercise, diet and the harmful effects of smoking. Recently, the preventive work has been targeted on children and people at high risk of CVD.
**Patient support:** Since then, the Danish Heart Foundation's role as the advocate of the heart patients has strengthened, and rehabilitation, counselling and support for patients and their relatives have become activities beside research and prevention. Some concrete interventions to support patients are the campaign "Ta' cyklen Danmark", the walking paths "Hjertestier" and many heart healthy recipes on the website. Another place that patients can seek various advice, for example concerning diet and physical activity, is through the free telephonic "Hjertlinje", where both nurses, physicians, dietitians and psychologists are situated to provide professional counselling. Recently, the "Børnehjertelinjen" has been established, supporting families with advice on how to handle life with a heartsick child. Furthermore, the Children's Heart Foundation has been founded, which raises money for heartsick children and their families.
**Economy and size:** Today, the Danish Heart Foundation is the second largest patient organisation with currently around 130,000 members, more than 10,000 volunteers (in 93 local associations across the country) and an annual revenue of 182 million kr. in 2020. The revenues mainly come from heritage, subscriptions, individuals, companies, donations and sale of services. The Danish Heart Foundation also has the TV-show "Hjertegalla" each year, where millions are raised to support the cause. The organisation is housed in Vognmagergade 7, 3rd floor, 1120 Copenhagen, where there are currently around 130 employees, led by CEO Anne Kaltoft.
**Research:** The Danish Heart Foundation funds scientific research in cardiovascular epidemiology to improve prevention and treatment. Annually, around 50 million kr. is donated to support research in cardiovascular disease, and the Danish Heart Foundation therefore contributes significantly to research in the area in Denmark. The foundation publishes HjerteTal.dk which is an online encyclopedia of occurrence, mortality and treatment of cardiovascular disease in Denmark from 2006 and onwards. The Research Department of the company was established in 2014 and is doing independent research.
**Political work:** The focus of the Danish Heart Foundation is mainly research, patient support and prevention. However, the Danish Heart Foundation is also working for political influence to improve cardiovascular patients' terms and is doing international disease-fighting work through membership of the European Heart Network. Another goal for the company is to be more influential on the political scene in Denmark, and the health economic perspective in this project is thus a novel and needed perspective in the research within the Danish Heart Foundation.

- **The candidate's placement in the company**

The industrial PhD candidate will have office at the Danish Heart Foundation in Copenhagen. The candidate will be located in the company's research department, which is led by the company supervisor, professor in cardiology and senior consultant cardiologist Gunnar Gislason. Therefore, it will be possible for daily support and inputs. Currently, the research department consists of 6 PhD students, 3 postdocs, 2 senior researchers and professor in biostatistics Thomas Gerds. The candidate will thus be placed in a research environment with a wide professional range, so there are many opportunities for support on specific medical practices and register-based research in cardiology. The candidate will be the first PhD student within biostatistics.
The candidate will collaborate closely with the prevention department, especially the co-superviser PhD in sociology, Mads Lind, who is involved closely in the walking initiative and specialises in strategic prevention and political decisions. Relevant dietitians and nutrition experts from the counselling service (Hjertelinjen) will support the direct patient perspective of the project, and a dietitian from the prevention department will support the strategic prevention perspective. The public affairs department will support the health political perspective, to obtain the largest possible impact based on the project results.

- **Any exit strategy**

The Danish Heart Foundation is one of the largest disease-fighting organisations in Denmark with a distinct research section which can support the PhD candidate. However, in the unlikely case where the funding for the project breaks down, the section of biostatistics at Copenhagen University will finance the rest of the project.

## K. University

- **Description of the university and centre / department**

The section of Biostatistics is part of the Department of Public Health and the biostatistical resource for all parts of the Faculty of Health Sciences, University of Copenhagen. The Department of Public Health was established on 1 January 1997 at the Faculty of Health Sciences, University of Copenhagen. The department carries out research, teaching and acts as a consultant within the area of public health. The aim of public health research is to create a scientific foundation for improving the health of the population. The research investigates the health status of the population and what efforts to make to improve the health of the population and reduce morbidity and mortality.
The section of Biostatistics conducts general biostatistical research, methodological development in biostatistics, participates in medical research projects and undertakes teaching of students at various educations, undergraduate as well as postgraduate. The section offers biostatistical advice to PhD students and staff in all disciplines of the Faculty of Health Sciences. The biostatistical research focuses on causal inference, latent variable models, time series analysis, survival analysis, diagnostic and predictive models, statistical computing and epidemiological methods. The applied statistical collaboration covers all areas of medical research and ranges from short consultations to collaborative projects lasting several years. The section has an academic staff of approximately 50. Thus, there will be a great opportunity for relevant support for the PhD student within the biostatistical field.
The supervisor at the Section of Biostatistics will be Professor Thomas A. Gerds. His research has a focus on the statistical analysis of binary, longitudinal and time-to-event data, development of statistical strategies for building risk prediction models and tools in bioinformatics and machine learning with applications in high dimensional data. He has published several scientific articles in the field of biostatistics. Thomas has helped the research unit of the Danish Heart Foundation with their biostatistical work of their projects and has since August 2016 been permanently assigned to the research unit two days a week. Thomas' extensive experience and network within biostatistics will provide the PhD student relevant support at a high level.

## L. Third parties

- **Description of any third parties**

Center for Aktiv Sundhed (CFAS) is a translational research center at Rigshospitalet that aims, through research, to develop new training forms that can be used as ideal treatment within a broad variety of chronic diseases, including cardiovascular disease. The aim is that new research discoveries regarding exercise as medicine should be implemented and anchored in society, leading to change of praxis in a relatively short time, not least to the benefit of the growing number of patients with chronic diseases, who may prosper from individualised physical training. CFAS is led by Prof., DMSc. Bente Klarlund Pedersen, who has done extensive research within physical activity and prevention of chronic diseases. One of Prof. Pedersen's strengths is her ability to convey this line of research to other professionals, a broader part of the Danish population and politicians/decision makers. Both Prof. Pedersen's extended knowledge on physical activity, disease and health, and her communicative skills will be important to this PhD project, as these features are not covered by any other supervisor. Therefore, Prof. Pedersen will be able to guide, support and provide relevant knowledge in these areas, which will benefit the project.