DP-next

It is well known that type 2 diabetes (T2D) in trial settings can be prevented in high-risk individuals through intensive lifestyle modifications. However, translating these findings into sustainable, real-world practices has proved difficult. The original interventions were too resource-intensive for large-scale use, while less intensive approaches only produced temporary weight loss without long-term diabetes prevention. Additionally, many of the highest at-risk groups are rarely identified in routine care, and participation in health promoting programs among these groups has declined. Despite these challenges, new opportunities for diabetes prevention have emerged, including advanced health data systems, machine learning, better understanding of behavior change, innovative intervention methods, real-time monitoring technologies, and widespread use of apps and smartphones. These resources remain underutilized but hold great potential to improve prevention strategies.

My PhD project is part of a larger project called Sustainable T2D Prevention for the 21st Century (DP-next). DP-Next is a project aimed at developing a sustainably, effective strategy for prevention of T2D in Denmark, Greenland and the Faroe Islands. The project has been funded by a Steno National Collaborative Grant from the Novo Nordisk Foundation. The DP-Next project consists of four work packages: WP1 (Management and Collaboration), WP2 (Risk Prediction), WP3 (Heterogeneity), WP4 (Intervention Development). My PhD project falls under WP2 of the DP-Next project. The aim of WP2 is to develop an operational risk model that calculates an individual T2D risk prediction for the entire population of Denmark, Greenland and the Faroe Islands. We will achieve this by designing a T2D risk prediction algorithm based exclusively on existing register-based data. Furthermore, my PhD project work is connected with WP4, where we will in collaboration derive subgroups based on heterogeneity in T2D risk.

Data

The T2D risk prediction model will be developed using synthetic data following the structure of the Danish register-data. All analyses will be conducted on a DST project database and transferred/validated in both real Danish register data, and in Greenlandic and Faroese data. Using register data available on DST, the type 2 diabetes outcome will be defined according to a previously validated classification (Isaksen et al., 2023). The analyses using Danish register data will be conducted on the DST project database at Steno Diabetes Center Aarhus, incorporating a wide range of data sources for model development from the following registries:

- CPR Register [1968-]
- LPR (Danish National Patient Registry) [1977-]

- LMDB (Danish National Prescription Registry) and the Hospital Medication Register [1995-]
- Register of Laboratory Results for Research [2011/2015Q4-]
- Cancer Register [1987-]
- Pathology Register [1997-]
- Diabase (RKKP) [2013-]
- RKKP (Danish Registry of Diabetes) [2022Q3-]
- Health Insurance [1990-]
- Medical Birth Register [1973-]
- DREAM database [1991-/2008-]
- HANDIC Register [2013-/2018-]
- FAIK/register of incomes [1990-]
- Vaccination Register [2013-/2016-]

My Project

My project aims to develop and adapt statistical methodology to create a super learner algorithm which produces diabetes risk prediction models based on Danish registry data. In collaboration with the research team of WP2, the algorithm will be applied to build models for the Danish population, and will be adapted to the Faroese and the Greenlandic populations. Specifically, I will apply the algorithm to build and validate a type 2 diabetes risk prediction model for the Danish population such that the model has the following features:

- Exclusively use register-based data such that the prediction model can be calculated for the entire population of Denmark, without the need to ask individuals for additional information.
- Is designed for yearly updates based on newly available / updated risk indicator data.
- Can incorporate longitudinal data (risk factor trajectories)
- Can incorporate time-dynamic risk indicators from family members

In collaboration with the research team of WP4 I will also develop statistical methods that can identify subgroups of the population with high type 2 diabetes risk. To achieve these overarching goals for the DP-next project, my PhD work is divided in the following three sub-projects.

Project 1: Joint inference of cause-specific risks (Constrained risk prediction)

When applying the type 2 diabetes (T2D) risk prediction model, it is essential to accurately report both the risk of T2D and the risk of death. This is necessary because, when investigating heterogeneity in T2D risk across the population, the competing risk of death must be taken into account, as it directly influences the estimated T2D risk. For example, if two individuals have similar covariate profiles but one has end-stage brain cancer, that individual will have a lower estimated risk of developing T2D solely due to an increased risk of death. Reporting both competing risks is also important for potential clinical applications: an individual with a high risk of death should not be targeted for preventive intervention based solely on an elevated T2D risk.

The term risk of death can refer either to the risk of death without developing T2D or to the risk of death from any cause. The former is typically used in standard survival analysis, whereas the latter corresponds to the risk of death in the illness–death model. For this project, I will focus on the risk of death from any cause, as this measure also includes the risk of death occurring after T2D onset and therefore provides a more comprehensive view of competing risks.

When estimating cause-specific risks in practice, the standard approach is to estimate the risk of T2D and the risk of death separately. However, this is suboptimal when the goal is to report both risks simultaneously, since the risks are interdependent. When investigating heterogeneity in T2D risk across the population, the competing risk of death must be taken into account, as it directly influences the estimated T2D risk. For example, if two individuals have similar covariate profiles but one has end-stage brain cancer, that individual will have a lower estimated risk of developing T2D solely due to an increased risk of death. Reporting both competing risks is also important for potential clinical applications: an individual with a high risk of death should not be targeted for preventive intervention based solely on an elevated T2D risk.It is therefore necessary to estimate the cause-specific risks jointly, under appropriate constraints derived from the illness–death model. This issue has been examined previously, and a potential solution was proposed by (Li & Yang, 2016). However, that solution was developed only for the Cox proportional hazards model and has not been extended to other modeling frameworks.

The term risk of death can refer either to the risk of death without developing T2D or to the risk of death from any cause. The former is typically used in standard survival analysis, whereas the latter corresponds to the risk of death in the illness–death model. For this project, I will focus on the risk of death from any cause, as this measure also includes the risk of death occurring after T2D onset.

In this project, I aim to generalize this approach within the Super Learner algorithm (Polley & Van der Laan, 2010), ensuring that both the discrete and ensemble versions of the Super Learner satisfy the required constraint. The resulting Super Learner will produce the best-performing T2D risk prediction model by combining parametric, semi-parametric, and machine learning meth-

ods while maintaining coherence between the competing risks. Ultimately, the model will use Danish registry data to predict both the 5- and 10-year risks of developing T2D and the 5- and 10-year risks of death from any cause.

Project 2: Large-Language-Models in risk prediction

To build the most accurate risk prediction model for T2D, it is essential to incorporate all available data. However, this is not a straightforward task, as the Danish registry data used in this project are longitudinal, and the interpretation of missingness varies across registries. Two of the registries used in this project are the Danish National Prescription Registry (LMDB) and the Register of Laboratory Results for Research (RLRR). In the LMDB, a missing observation indicates that an individual was not prescribed any drugs. In contrast, in the RLRR, a missing observation does not imply that the individual lacks biomarker data - it merely indicates that no laboratory measurements were recorded during that period. Furthermore, the time points of measurements differ from person to person. For example, if we wish to examine the most recent measurement of a covariate X before time 0, the last measurement for person 1 might be $X(0-t)$, where $t = 1$ day, whereas for person 2 it could be $X(0-s)$, with $s \neq t$. This pattern is typical for most, if not all, covariates in the registries.

Such data structures are not straightforward to incorporate into traditional likelihood-based risk prediction frameworks, which typically rely on summarizing longitudinal information (e.g., using means or weighted averages). Consequently, new methodological approaches are required to fully exploit the richness and temporal dynamics of the Danish registry data.

To address this challenge, recent studies have proposed the use of machine learning methods, in particular Neural Networks. (Lee et al., 2019) build the Dynamic-DeepHit model that can incorporate longitudinal data measured at irregular timepoints to estimate the cause-specific cumulative incidence functions. However, it remains unclear how the model handles missing data or scales to large registry datasets. A similar approach was taken by (Wright et al., 2021), who employed recurrent neural networks (RNNs) to efficiently incorporate covariate histories into a risk prediction model, where the model was trained on danish registry data. A natural extension of these methods is to use Transformers instead of RNNs, which have the advantage of better handling large datasets and longitudinal data with long histories (Vaswani et al., 2017). Using Transformers will make the framework similar to that of large language models (LLMs), which recently have been used for disease risk prediction by (Shmatko et al., 2025), who proposed a modified GPT-based model to predict the risk of over 1,000 diseases using data from the UK Biobank.

In this project, I will develop a large language model for T2D risk prediction, following a similar approach to (Shmatko et al., 2025). Specifically, I will train a modified GPT-based model on Danish registry data to predict T2D risk exclusively. If this proves too challenging — given that this is a relatively unexplored area — I will instead employ methods similar to those proposed by (Lee et al., 2019) and (Wright et al., 2021).

The model developed in this project will be build on Danish registry data exclusively, and will predict the 5- and 10-year risk of developing type 2 diabetes and predict the 5- and 10-year risk of dying from any cause. This model will then be compared with the Super Learner Algorithm devolped in Project 1, and the best performing T2D risk prediction model will be used for the DP-Next project.

Project 3: Identifying subgroups with high diabetes risk

WP4 of the DP-Next project aims to develop targeted intervention strategies for the prevention of type 2 diabetes (T2D) among subgroups of the Danish population identified as being at high risk. These high-risk subgroups will be identified in the present project using the best performing T2D risk prediction model developed in Projects 1 and 2. The subgroups will be defined based on covariate patterns - for example, one subgroup might consist of males under 40 years of age with a body mass index (BMI) of 25 or higher. The overall goal is to identify specific covariate patterns that define the segments of the Danish population with the highest risk of developing T2D.

Because the aim of WP4 is to recruit high-risk individuals from the Danish population into newly developed intervention strategies, the identified subgroups should be both recruitable and suitable for intervention. Therefore, I will aim to identify multiple high-risk subgroups with diverse covariate patterns, allowing the WP4 team to select those most appropriate for their recruitment and intervention objectives.

The subgroup analysis carried out in this project will be entirely data driven. Data driven subgroup analysis methods are already well devolped, with methods such as (Atzmueller & Puppe, 2006) who has developed the SD-Map algorithm, which relies on the Frequent Patern Growth (FP-growth) algorithm, for exhaustive and efficient subgroup discovery. Other methods such as regression trees or cluster analysis are also frequently used for subgroup discovery, with examples such as (Su et al., 2009) or (Parikh et al., 2021). These methods are however notoriously unstable, especially for large data (Wani, 2024), and will need carefull considerations to identify the true high riks subgroups of the population.

In this project, I will apply the best-performing T2D risk prediction model from Projects 1 and 2 to the entire Danish population, thereby obtaining an estimated T2D risk for each individual. I will then use the SD-Map or cluster analysis to identify the subgroups with the highest average T2D risk. To make the search more stable and the subgroups more interpretable, the subgroups will be defined based on 2-4 covariates. All subgroups with an average risk exceeding some threshold ϵ will be reported to the WP4 team, who can then select the subgroups most suitable for their intervention study.

Bibliography

Atzmueller, M., & Puppe, F. (2006). Sd-map–a fast algorithm for exhaustive

subgroup discovery. *European Conference on Principles of Data Mining and Knowledge Discovery*, 6–17.

Isaksen, A. A., Sandbæk, A., & Bjerg, L. (2023). Validation of register-based diabetes classifiers in danish data. *Clinical Epidemiology*, 569–581.

Lee, C., Yoon, J., & Van Der Schaar, M. (2019). Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *Ieee Transactions on Biomedical Engineering*, *67*(1), 122–133.

Li, G., & Yang, Q. (2016). Joint inference for competing risks survival data. *Journal of the American Statistical Association*, *111*(515), 1289–1300.

Parikh, R. B., Linn, K. A., Yan, J., Maciejewski, M. L., Rosland, A.-M., Volpp, K. G., Groeneveld, P. W., & Navathe, A. S. (2021). A machine learning approach to identify distinct subgroups of veterans at risk for hospitalization or death using administrative and electronic health record data. *Plos One*, *16*(2), e0247203.

Polley, E. C., & Van der Laan, M. J. (2010). *Super learner in prediction.*

Shmatko, A., Jung, A. W., Gaurav, K., Brunak, S., Mortensen, L. H., Birney, E., Fitzgerald, T., & Gerstung, M. (2025). Learning the natural history of human disease with generative transformers. *Nature*, 1–9.

Su, X., Tsai, C.-L., Wang, H., Li, B., & others. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, *10*(2).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wani, A. A. (2024). Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions. *Peerj Computer Science*, *10*, e2286.

Wright, M. N., Kusumastuti, S., Mortensen, L. H., Westendorp, R. G., & Gerds, T. A. (2021). Personalised need of care in an ageing society: The making of a prediction tool based on register data. *Journal of the Royal Statistical Society Series a: Statistics in Society*, *184*(4), 1199–1219.