Advancing Precision Medicine in Type 2 Diabetes through Machine Learning:

Treatment Comparisons and Risk Predictions

Department of Molecular and Clinical Medicine

Institute of Medicine

Sahlgrenska Academy, University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2024

Cover illustration: Moa Lugner

Moa.lugner@gu.se

ISBN 978-91-8069-781-1 (PRINT)
ISBN 978-91-8069-782-8 (PDF)

Printed in Borås, Sweden 2024
Printed by Stema Specialtryck AB

"I am in a charming state of confusion."

Ada Lovelace, 1815-1852

Advancing Precision Medicine in Type 2 Diabetes through Machine Learning:

Treatment Comparisons and Risk Predictions

Moa Lugner

Department of Molecular and Clinical Medicine, Institute of Medicine
Sahlgrenska Academy, University of Gothenburg
Gothenburg, Sweden

ABSTRACT

Precision medicine holds significant promise for improving the management of type 2 diabetes, yet its implementation faces several challenges. This thesis aims to address existing gaps in precision medicine for type 2 diabetes, focusing primarily on precision treatment, precision prognostic, and precision diagnosis. Each study within this thesis contributes to advancing the field in these key areas.

Study I compares the outcomes and safety of GLP-1 receptor agonists and SGLT-2 inhibitors for type 2 diabetes in a real-world setting, using propensity score matching and inverse probability treatment weighting to evaluate their practical effectiveness and side effects. Study II examines the validity and clinical utility of a proposed subclassification system for type 2 diabetes, using clustering analysis to explore potential subgroups and assess their ability to predict adverse events and diabetic complications. Study III identifies significant predictors for the development of type 2 diabetes, using an XGBoost classification model on a study population of 450,000 participants from the UK Biobank to project incidence and refine future risk scores. Study IV identifies principal predictors of cardiovascular complications and mortality in individuals with type 2 diabetes, using an XGBoost algorithm to analyze over 400 predictors and assess their impact on the risk of major adverse cardiovascular events.

Study I revealed similar effects of GLP-1 receptor agonists and SGLT-2 inhibitors on several cardiovascular outcomes, although GLP-1 receptor agonists were more effective in reducing stroke. Study II found distinct characteristics within potential subgroups of type 2 diabetes but limited predictive value in foreseeing adverse events. Study III identified HbA1c, BMI, waist circumference, and blood glucose levels as significant predictors of type 2 diabetes. The XGBoost model achieved high accuracy for 10-year type 2 diabetes prediction. Study IV highlighted the importance of variables such as age, Cystatin C, and pulse pressure in predicting cardiovascular events in individuals with type 2 diabetes. The model demonstrated high accuracy in predicting major adverse cardiovascular events.

In conclusion, this thesis contributes to advancing precision medicine in type 2 diabetes, providing valuable insights and paving the way for more personalized and effective approaches to treatment and management.

Keywords: Type 2 diabetes, epidemiology, observational, studies, Machine Learning, Cardiovascular disease, mortality, prediction

ISBN 978-91-8069-781-1 (PRINT)
ISBN 978-91-8069-782-8 (PDF)

SAMMANFATTNING PÅ SVENSKA

Bakgrund: Typ 2 diabetes är en heterogen sjukdom som uttrycker sig mycket varierat hos olika patienter. Precision inom diabetesvård syftar till att utforma behandlingsregimer efter individuella faktorer, men bristande kunskaper hindrar införandet av en mer personcentrerad behandlingsapproach. Det övergripande syftet med arbetet i denna avhandling är att bidra till utvecklingen inom precisionsmedicin för diabetes genom att öka kunskapen.

Metod: Vi har använt data från det svenska diabetesregistret samt UK Biobank för att besvara frågor rörande behandlingseffekter av SGLT-2-hämmare och GLP-1-analoger, subgrupperingar inom typ 2 diabetes och hur vi bäst kan förutsäga individers risk både för att utveckla typ 2 diabetes och för att drabbas av kardiovaskulära sjukdomar senare i sjukdomsförloppet. Vi har använt flera olika statistiska metoder, inklusive både traditionella och avancerade maskininlärningsmetoder, för att besvara frågeställningarna.

Resultat: Vi fann att behandling med SGLT-2-hämmare och GLP-1-analoger resulterade i liknande frekvenser av kardiovaskulär sjukdom och död, men att GLP-1-analoger i vissa grupper var effektivare som skydd mot stroke. Vi utvärderade en tidigare föreslagen subgruppering av typ 2 diabetes, men fann att den inte tillförde klinisk nytta genom att förutsäga framtida risk för kardiovaskulär sjukdom eller död. Vidare identifierades HbA1c, BMI och midjemått som de viktigaste prediktorerna för att förutsäga risk för diabetes, medan Cystatin C och ålder var mest effektiva för att förutsäga risk för framtida kardiovaskulär sjukdom hos personer med typ 2 diabetes.

Slutsats: Två av de nyare, mycket omtalade läkemedlen för typ 2 diabetes är jämförbara i sin förmåga att skydda mot kardiovaskulära händelser. Subgruppering av typ 2 diabetes har potential att bidra med precision, men det är ännu inte klarlagt hur en sådan subgruppering bör göras. Enkla, rutinmässigt insamlade kliniska variabler kan med god precision förutsäga risken för att utveckla diabetes, och Cystatin C utmärker sig som en stark prediktor för kardiovaskulära komplikationer hos personer med typ 2 diabetes.

LIST OF PAPERS

This thesis is based on the following studies, referred to in the text by their Roman numerals.

I. Lugner M, Sattar N, Miftaraj M, Ekelund J, Franzén S, Svensson A, Eliasson B. **Cardiorenal and other diabetes related outcomes with SGLT-2 inhibitors compared to GLP-1 receptor agonists in type 2 diabetes: nationwide observational study**
Cardiovascular Diabetology 2021; 20: 67.

II. Lugner M, Gudbjörnsdottir S, Sattar N, Svensson A, Miftaraj M, Eeg-Olofsson K, Eliasson B, Franzén S. **Comparison between data-driven clusters and models based on clinical features to predict outcomes in type 2 diabetes: nationwide observational study.** Diabetologia 2021; 64(9):1973-1981

III. Lugner M, Rawshani A, Eliasson B. **Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data.** Scientific Reports 2024; 14: 2102.

IV. Lugner M, Rawshani A, Eliasson B. **A Comprehensive Analysis of Predictive Factors for Cardiovascular Risk in Type 2 Diabetes: Insights from the UK Biobank.** Manuscript.

CONTENTS

ABBREVIATIONS

ADA	American Diabetes Association
AI	Artificial Intelligence
ASCVD	Atherosclerotic Cardiovascular Disease
BMI	Body Mass Index
CDR	The Cause of Death Register
CVD	Cardiovascular Disease
CVOT	Cardiovascular Outcome Trial
EASD	European Association for the Study of Diabetes
eGFR	Estimated glomerular Filtration Rate
GGT	Gamma-glutamyl Transferase
ESC	The European Society of Cardiology
FRM	Framingham Risk Score
HbA1c	Glycated Haemoglobin
IPTW	Inverse Probability Treatment Weighting
LADA	Latent Autoimmune Diabetes in Adults
ML	Machine Learning
MODY	Maturity Onset Diabetes of the Young
NPR	The National Patient Register
PDR	The Prescribed Drug Register
PMDI	The Precision Medicine in Diabetes Initiative

SNDR	The Swedish National Diabetes Register
UKPDS	UK Prospective Diabetes Study
WCSS	The Within-Cluster Sum of Squares
XGBoost	eXtreme Gradient Boosting

1 INTRODUCTION

Type 2 diabetes is a rapidly growing public health concern, affecting millions worldwide. Its rise is driven by factors such as aging populations, urbanization, and increasing prevalence of obesity and physical inactivity. The disease is not only chronic but also costly, primarily due to its numerous complications which can severely impact quality of life and increase mortality.

Precision medicine represents a transformative approach to healthcare, tailoring disease prevention, diagnosis, and treatment to the individual variations in genes, environment, and lifestyle. This approach shifts from a one-size-fits-all strategy, where treatments and prevention methods are standardized without considering individual differences, to more targeted interventions.

The main aim of this thesis is to bring precision medicine in type 2 diabetes one step closer to the clinic by attempting to fill some of the knowledge gaps that currently prevent its full utilization.

1.1 TYPE 2 DIABETES

Diabetes is a term that encompasses a group of chronic, metabolic diseases affecting the body's ability to regulate glucose levels. This dysregulation is caused by various mechanisms, which is why diabetes is divided into subcategories based on the cause for impairment. The major forms are Type 1 and Type 2 diabetes. Type 1 diabetes is an autoimmune condition in which the body's immune system targets and destroys the insulin-producing cells within the pancreas. There are also less common forms of diabetes, such as gestational diabetes, Maturity Onset Diabetes of the Young (MODY), and Latent Autoimmune Diabetes in Adults (LADA), among others. The WHO estimates that more than 422 million people worldwide are currently living with diabetes, with over 90% of all cases attributed to Type 2 diabetes[1].

PATHOPHYSIOLOGY: Type 2 diabetes is caused by a progressive decline in the body's ability to regulate glucose levels due to increasing insulin resistance and a varying degree of insulin deficiency. It is a multifactorial disease with risk factors that involve a complex interplay between genetics, environment, and lifestyle factors. Although predisposition for type 2 diabetes is influenced by non-modifiable risk factors such as ethnicity, increasing age and genetics, several modifiable factors significantly contribute to the development of the disease. The strongest connections between modifiable risk factors and Type 2 diabetes are obesity and physical inactivity[1].

COMPLICATIONS: Type 2 diabetes does not necessarily cause severe symptoms and can often go undiagnosed for several years[2]. However, the metabolic changes in the body, primarily hyperglycemia, promote the progression of complications typically associated with Type 2 diabetes[3]. The spectrum of complications associated with diabetes can be categorized into two main subtypes: macrovascular and microvascular. Microvascular complications encompass damage to the nervous system (neuropathy), kidneys (nephropathy), and eyes (retinopathy). Macrovascular complications, on the other hand, include cardiovascular diseases (CVD), stroke, and peripheral vascular disease. Overall, individuals with type 2 diabetes have a risk of death and cardiovascular events that is 2 to 4 times higher than that of the general population[4].

TREATMENT: The care of patients with Type 2 diabetes generally aims to improve overall health and prevent or delay diabetes-related complications. Most guidelines support the use of person-centered, holistic, and multifactorial

approaches to achieve this effectively[5]. Although blood glucose control is vital in type 2 diabetes with a major impact on microvascular complications[6], managing other risk factors such as hypertension and hyperlipidemia reduced macrovascular complication and overall mortality[7, 8]. There has been a shift in the treatment of Type 2 diabetes from a primarily glucose-centered approach to a comprehensive multi-risk strategy that targets several cardiovascular risk factors in addition to hyperglycemia[9].

Medications for lowering glucose: Drugs available to treat hyperglycemia in type 2 diabetics include metformin, sulfonylureas, DPP-4 inhibitors, SGLT-2 inhibitors and GLP-1 receptor agonists.

Metformin is an oral diabetes medication that lowers HbA1c through several mechanisms, including inhibited gluconeogenesis, increased insulin sensitivity, and delayed intestinal glucose absorption. Due to its low risk of hypoglycemia, ability to maintain or reduce weight, favorable safety profile, and affordability, Metformin has long been the preferred initial treatment for type 2 diabetes[5, 10]. However, there is growing recognition of the effectiveness of alternative treatments.

Two new diabetes drugs that have become prominent in the treatment regimen for type 2 diabetes are SGLT-2 inhibitors and GLP-1 agonists. SGLT-2 inhibitors work by blocking the renal sodium-glucose cotransporter-2, which increases urinary glucose excretion and thereby reduces plasma glucose levels[11]. These inhibitors offer several cardiorenal protective qualities and reduce the risk of major adverse cardiovascular events (MACE), heart failure, and diabetic kidney disease[12]. GLP-1 receptor agonists enhance insulin secretion in response to glucose, suppress glucagon release, and slow gastric emptying, leading to reduced blood glucose levels and weight loss through increased satiety[13]. GLP-1 receptor agonists reduces risk of MACE in adults with established CVD or multiple cardiovascular risk factors[5]. Moreover, the fame of GLP-1 receptor agonists has surpassed beyond the medical field and become well-known amongst the general public under brand names such as Ozempic, Wegovy, and Saxenda due to their weight-reducing abilities**[14]**.

The advantages of SGLT-2 inhibitors and GLP-1 receptor agonists in improving cardiovascular and renal outcomes have been shown to occur regardless of whether metformin is used[15, 16]. Current guidelines therefore states that these medications should be considered for individuals with existing

or high risk of cardiovascular disease, heart failure, or chronic kidney disease, without necessarily depending on metformin[5].

1.2 PRECISION MEDICINE

Precision medicine tailors treatment to individual patient characteristics, taking into account genetic differences, environmental factors, and lifestyle choices. This approach diverges from the traditional 'one-size-fits-all' method in medicine, aiming to customize care, optimize treatment efficacy, and minimize adverse side effects. Strategies for achieving this include genetic mapping, the use of predictive tools, personalized care plans, and the integration of technologies such as bioinformatics, data management, and computing.

Precision medicine has advanced to varying degrees across different medical disciplines, with each field experiencing unique progress and facing distinct challenges. One of the fields most frequently highlighted in discussions of precision medicine is oncology. In this area, precision medicine has significantly transformed approaches to treatment, primarily through the use of targeted therapies and personalized treatment plans based on the genetic profiles of both the patient and the tumor[17].

For chronic, multifactorial diseases like type 2 diabetes, precision medicine presents greater challenges due to the disease's heterogeneity. While all diabetes patients share a common issue of disturbed glucose metabolism leading to hyperglycemia, individual cases can vary significantly. However, in certain types of monogenic diabetes, precision medicine has made more substantial progress, with targeted treatments based on identified specific gene defects[18].

Major ongoing projects are aiming to advance precision medicine in type 2 diabetes. The American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD) have collaborated on an initiative launched in 2018 known as the Precision Medicine in Diabetes Initiative (PMDI). This initiative encompasses various forms of diabetes, including type 1 and 2 diabetes, monogenic diabetes, and gestational diabetes.

The first Consensus Report from the PMDI, published in 2020, summarizes the current state of the field and addresses both the challenges remaining for

the implementation of precision medicine in type 2 diabetes and the advancements that have been made, as well as future prospects. It identifies five key pillars of precision diabetes: precision diagnostics, precision prevention, precision treatment, precision prognostics, and precision monitoring[18].

- **Precision diagnostics** At its most fundamental level, precision diagnostics in diabetes involves accurately categorizing patients into established subtypes such as Type 1, Type 2, MODY, LADA, etc. It also encompasses further stratification within these categories, particularly for type 2 diabetes, which is itself a relatively broad classification.
- **Precision treatment**: Precision treatment in diabetes focuses on selecting the most effective therapy based on an individual's unique factors to achieve the best possible outcome with minimal side effects. The goal is provide the right treatment to the right patient at the right time.
- **Precision prognosis**: Precision prognisus enhances the accuracy of predicting disease-related outcomes based on an individual's unique biological, environmental, and contextual information. This approach specifically aims to assess the risk and severity of diabetes complications, patient-centered outcomes, and potential early mortality.
- **Precision monitoring**: Precision monitoring in diabetes management involves detailed assessments using various tools and technologies. This includes continuous glucose monitoring systems, which are prominently used in type 1 diabetes, and applications that track physical activity, diet, sleep, and stress levels. These technologies provide crucial data that can be used to tailor individualized management plans, particularly in the proactive monitoring and treatment of diabetes.

The consensus report highlights advancements in precision medicine for diabetes, particularly emphasizing targeted therapies for monogenic diabetes and the application of biomarkers and genetic profiling to distinguish autoimmune diabetes from other types. The progress in precision medicine for managing type 2 diabetes appears more modest, largely due to the disease's complexity and diversity. This thesis will further detail some of the specific aspects of precision medicine in type 2 diabetes.

1.2.1 PRECISION DIAGNOSTICS IN TYPE 2 DIABETES

Diabetes is primarily classified into two types: type 1 and type 2 diabetes, with the latter encompassing over 90% of all diabetic patients[1]. This diagnosis of type 2 diabetes usually follows a process of exclusion, ruling out other potential causes of chronically elevated blood glucose levels. Type 2 diabetes is already a broad condition, affecting individuals who are young and old, thin and overweight, and varying widely in terms of overall health. It is no longer predominantly seen as an elderly person's disease but is increasingly diagnosed in younger individuals[19]. This younger demographic typically presents with a greater metabolic burden, marked by higher BMI, elevated blood pressure, and a more rapid deterioration of glucose regulation from the outset[20]. These factors contribute to more severe long-term health consequences for those who develop type 2 diabetes early in life[21].

This evolving demographic landscape challenges the traditional paradigms of diabetes management and demands a more nuanced approach to its classification. The recent consensus report highlights this need, pointing out that while genetic data provide valuable insights into the disease's etiology, they currently lack the predictive accuracy required for effective subgroup classification or treatment decisions. Clinical characteristics, although widely available and informative at diagnosis, can change over time, complicating long-term subgroup consistency[22]. It concludes that more research is needed on type 2 diabetes in order to define usefull subtypes.

1.2.2 PRECISION TREATMENT IN TYPE 2 DIABETES

Trials of therapeutic interventions traditionally don’t acknowledge the heterogeneity in type 2 diabetes in terms of clinical manifestation, disease progression, and pathogenesis. As we enhance our understanding of both genetic and phenotypic variants, we anticipate more tailored treatment options. Currently genetic mapping is not widely used to guide treatment decisions in type 2 diabetes, but phenotypical factors like patient characteristics and comorbidities is generally used to inform decision.

The characteristics of diabetes vary significantly among ethnic groups, which suggests that medication effects could also differ across populations. Although research into treatment responses based on ethnic grouping is limited, initial findings indicate variations. For example, African Americans have been shown to exhibit a greater glycemic response to metformin compared to European Americans[23]. However, the lack of diverse ethnic representation in clinical trials hampers our full understanding of these differential responses.

It also stresses the need for improved decision-support tools in clinical practice to facilitate the implementation of precision medicine.

The discussion extends to newer agents like GLP-1 receptor agonists and SGLT-2 inhibitors, underscoring the necessity for further research into their effects across different subpopulations. To enhance precision treatment in type 2 diabetes, studies on SGLT-2 inhibitors and GLP-1 1 receptor agonists have been reviewed to determine if demographic, clinical characteristics, and biomarkers at an individual level influenced the effects on glycemia, CVD, or renal failure. However, no substantial evidence was found to support the use of these individual factors in predicting treatment response[24].

Nonetheless, the advent of these medications has enabled a more personalized approach to managing type 2 diabetes, allowing clinicians to tailor treatments based on individual risk-benefit profiles. Even though we do not yet have the knowledge about how many individual factors such as ethnicity and specific biomarkers might affect treatment response, clinicians can use a broader view of the patient's specific risks to guide treatment options. For instance, SGLT-2 inhibitors are primarily recommended for patients with comorbid heart failure or kidney disease, while GLP-1 receptor agonists are favored for patients with obesity-related metabolic concerns[25].

1.2.3 PRECISION PROGNOSTICS IN TYPE 2 DIABETES

Risk stratification in medicine is crucial for effectively preventing adverse events. Today's healthcare heavily relies on various risk scores, which often inform treatment decisions. As these scores indirectly influence treatment options, it is essential for them to provide reliable risk assessments and produce outputs that are easily interpretable. Essential qualities include high predictive

ability and generalizability. Frequently, overestimating risks is preferable to underestimating them to avoid missing early treatment opportunities. The input variables should be simple to collect and cost-effective. Developing and evaluating risk assessment tools typically requires large datasets. Consequently, improvements have been seen in recent decades as data collection and storage have expanded.

Macrovascular complications such as myocardial infarction, peripheral vascular disease, and stroke are markedly increased in individuals with type 2 diabetes compared to the general population[4]. However, the risk varies significantly among those with type 2 diabetes[7]. Enhanced risk stratification for CVD and other major complications allows for intensified risk factor management in individuals at elevated risk.

Several risk scores are available to assess the risk of cardiovascular disease for both the general population and individuals with type 2 diabetes. Commonly used scores include the UK Prospective Diabetes Study (UKPDS) Risk Engine[26], the Framingham risk score (FRS)[27], SCORE2[28] and SCORE2-Diabetes[29], with the latter being recommended by the European Society of Cardiology (ESC)[30]. The input predictors in these models typically include gender, age, blood pressure, and lipid profiles.

However, accurately assessing cardiovascular disease risk in type 2 diabetes remains challenging. Many of the available models, including the previously mentioned SCORE2, UKPDS, and FRS, exhibit varying degrees of effectiveness in predicting outcomes when externally validated across different populations[31, 32, 33, 34, 35].

A systematic review of prognostic markers for CVD in type 2 diabetes identified NT-proBNP as the most effective predictor, a notable finding given that NT-proBNP is not included in any of the most commonly used risk models[26, 27, 28, 29]. It also concluded that risk scores generally perform better when deployed in cohorts similar to those from which they were derived. Therefore, population-specific risk models offer more potential than generic models, which perform modestly across various settings[36].

Emphasis is also placed on the need for well-designed studies that focus on the prognostic value of different markers for CVD in individuals with type 2 diabetes. There is a call for further research into the effectiveness of biomarkers beyond established risk factors, and the development of

population-specific cardiovascular risk prediction models, particularly in underrepresented populations. Additionally, there is support for research aimed at identifying and validating new biomarkers, as well as improving the predictive power and clinical utility of existing markers[36].

2 AIM

The overall aim of this thesis is to bridge existing gaps in the implementation of precision medicine for type 2 diabetes, mainly focusing on areas such as precision treatment, precision prognostic and precision diagnosis. The specific objectives of each study comprising this thesis are outlined as follows:

Study I aim to compare the outcomes and safety of two contemporary pharmacological treatments for type 2 diabetes in a real-world setting, thus providing insights into the practical effectiveness and side effects associated with each treatment.

Study II examines the validity and clinical utility of a proposed subclassification system for type 2 diabetes, assessing whether this detailed stratification can enhance therapeutic precision and improve patient outcomes.

Study III focuses on identifying the most significant predictors for the development of Type 2 diabetes. This effort aims not only to deepen our understanding of the disease's fundamental drivers but also to refine the development of future risk scores by elucidating the predictive value of various factors.

Study IV aims to identify the principal predictors of cardiovascular complications and mortality in individuals with type 2 diabetes. This research is intended to inform the development of future risk scores and assist in accurately identifying patients at elevated risk.

3 DATA SOURCES AND METHODS

3.1 DATA SOURCES

SWEDISH NATIONAL DIABETES REGISTER: Studies I and II main data source was the Swedish National Diabetes Register (SNDR), a national quality register that collects data on children and adults with diabetes. This data is gathered through both manual inputs and direct importation from healthcare journal systems and is updated annually. The SNDR encompasses approximately 85% of all adults with diabetes in Sweden, making it one of the largest diabetes registers in the world[37]. For adults, it has collected 70 different variables that provide information on clinical characteristics, treatments, risk factors, and comorbidities[38].

In Sweden, all residents are assigned a unique personal identity number. This number was used in Studies I and II to link data from SNDR with other national registries. For Study I, the linked registers included the Prescribed Drug Register (PDR), the Longitudinal Integration Database for Health Insurance and Labor Market Studies (LISA), the National Patient Register (NPR), and the Cause of Death Register (CDR). For Study II, the latter two registers were used. This linkage was carried out at the National Board of Health and Welfare through a pseudonymized process.

UK BIOBANK: Studies III and IV utilized data from the UK biobank (UKBB). The UKBB is a large-scale biomedical database and research resource containing in-depth genetic and health information from half a million UK volunteers. Initial enrollment took place between 2006 and 2010 at 22 different centers to ensure socioeconomic and ethnic heterogeneity. At baseline, individuals completed questionnaires on health, lifestyle, and nutrition, underwent measurements such as blood pressure and BMI, and provided biological samples like blood and urine[39].

Table 1: Details of the study designs used in the thesis studies

	Study I	Study II	Study III	Study IV
Study type	Observational, registered-based	Observational, registered-based	Observational, registered-based	Observational, registered-based
Data source	SNDR, PDR, LISA, NPR, CDR	SNDR, NPR, PDR	UKBB	UKBB
Study population	New users of SGLT2 or GLP1	Newly diagnosed type 2 diabetes	All individuals without diabetes at baseline	All individuals with type 2 diabetes at baseline
Study period	2013-2017	2002-2012	2006-2020	2006-2020
Analytic methods	Cox regression	K-means clustering	XGboost classification algorithm	XGboost classification algorithm
Variables included	NA	Age, Diabetes duration, HbA1c, BMI, SBP, DBP, LDL, HDL, Triglycerides, eGFR	N = 419	N = 437
Main outcome	mortality, MACE, fatal or non-fatal CVD, HF, renal disease, hyper- and hypoglycaemia, ketoacidosis, diabetic nephropathy, retinopathy	CVD	Type 2 diabetes	MACE

Abbreviations: UKBB: UK Biobank, NA: Not Applicable, HbA1c: Hemoglobin A1c, BMI: Body Mass Index, SBP: Systolic Blood Pressure, DBP: Diastolic Blood Pressure, LDL: Low-Density Lipoprotein, HDL: High-Density Lipoprotein, eGFR: Estimated Glomerular Filtration Rate, MACE: Major Adverse Cardiovascular Events, CVD: Cardiovascular Disease, HF: Heart Failure, PDR: Prescribed Drug Register, LISA: the Longitudinal Integration Database for Health Insurance and Labor Market Studies, NPR: the National Patient Register, CRD: the Cause of Death Register, SNRD: the Swedish National Diabetes Register

3.2 ETHICAL APPROVAL

All studies constituting this doctoral thesis has been approved by the The Regional Ethical Review Board of the University of Gothenburg, which conformed with the Helsinki Declaration of 1964, as revised in 2013, concerning human and animal rights.

3.3 STATISTICAL METHODS

3.3.1 OVERVIEW OF THE DIFFERENT STATISTICAL METHOD FOR EACH STUDY

The studies presented in this thesis utilized a diverse array of statistical methods, each selected to best suit the data structure of the dataset being analyzed and to address the specific research questions posed. Some of the perhaps less common and lesser-known methods will be explained in greater detail in the following sections than those that are more traditionally used. The first section provides a summary of the methods used in each study. For more detailed descriptions, readers are referred to the respective articles and any supplementary materials.

Study I: Participants were new users of GLP-1 receptor agonists or SGLT-2 inhibitors from 2013 to 2017. The primary analysis was conducted following the intention-to-treat principle. Outcomes assessed included MACE and other

diabetes-related complications. Propensity scores were estimated with a generalized boosted binomial regression model, and inverse probability treatment weighting (IPTW) was used to assess the average treatment effect. The study also employed Kaplan-Meier and Cox regression analyses to explore time-to-event data and compare treatment effects.

Study II: Participants in this study included individuals with newly diagnosed type 2 diabetes between 2002 and 2012. Outcomes measured were the first occurrences post-index of cardiovascular events. The study utilized K-means clustering to create up to 7 clusters based on the parameters age, BMI, HbA1c, systolic and diastolic blood pressure, HDL and LDL cholesterol, triglycerides, and estimated glomerular filtration rate (eGFR). These clusters were then evaluated as categorical predictors of outcomes. For comparison, Cox models were used, which considered all the aforementioned variables as independent factors, with concordance serving as the evaluation metric for model performance.

Study III: Participants were individuals without diabetes at baseline. They were followed for a minimum of 10 years, and the primary outcome was the incidence of type 2 diabetes, determined through hospital records, self-reports, and primary care records. An initial XGBoost model was employed, trained on 419 variables collected at baseline. A second, reduced model used only the 10 most important predictors. The performance of both models was evaluated using metrics such as ROC-AUC and accuracy. Additionally, SHAP values were used to assess the importance of each variable, providing insight into their influence on the model's predictions.

Study IV: Study IV involved individuals with type 2 diabetes, identified through self-reports, hospital and primary care records, and blood tests. The focus was on the occurrence of MACE within 10 years from baseline. Using an XGboost model trained on 437 baseline variables, the study assessed model performance with ROC-AUC and accuracy metrics. Similar to Study III, SHAP values helped elucidate the impact of each variable on model outcomes.

3.3.2 DETAILED METHOLOGY

PROPENSITY SCORE MATCHING: The primary aim of propensity score matching is to reduce bias in estimating treatment effects in observational studies, where random assignment to treatment groups is not feasible. This method is particularly effective in addressing bias-by-indication, which occurs when the choice of treatment is influenced by patient characteristics. By using

propensity score matching, confounding factors across treatment groups are balanced, creating a scenario where the primary difference between groups is the treatment itself, rather than the characteristics that might influence the choice of treatment. Propensity score matching approximates the conditions of a randomized controlled trial by ensuring that the likelihood of receiving a particular treatment is statistically independent of the observed characteristics. This simulated randomization helps in making causal inferences more reliable in studies where true randomization is not possible.

PROPENSITY SCORES AND IPTW IN STUDY I: In Study I, the propensity scores were estimated using a generalized boosted binomial regression model. These scores were then applied using the Inverse Probability Treatment Weighting (IPTW) method. IPTW adjusts the analysis by weighting each participant according to the inverse of their calculated propensity score for the treatment they actually received. This weighting method corrects for any imbalances in baseline characteristics between the different treatment groups, thereby enhancing the likelihood that observed differences in outcomes are attributable to the treatment itself, rather than to pre-existing differences among the participants.

SURVIVAL ANALYSIS: Survival analysis is a branch of statistics that deals with the analysis of time until the occurrence of an event of interest. It involves methods for modeling and analyzing the time it takes for an event such as death or disease recurrence, often considering the possibility that some observations may not experience the event (censorship). Survival analysis is pivotal in medical research to understand the effectiveness of treatments and the impact of risk factors on survival.

COX REGRESSION: The Cox regression or Cox proportional hazards model, examines the relationship between survival time and one or more predictors. It defines survival time as the duration from a specific starting point to an event or the study's end if the event doesn't occur. This model is crucial in medical research for exploring how different factors, such as genetic traits or behaviors like smoking, influence the time to events like death or heart attacks. The Cox model calculates hazard ratios, providing insights into how much a particular factor might increase the risk of an event during the study, and it effectively handles censored data where some participants' outcomes are unknown by the study's end.

KAPLAN-MEIER ESTIMATOR: Used to plot survival curves, the Kaplan-Meier estimator illustrates patient survival times following treatment. It's particularly valuable for showing the percentage of patients still alive at various time points after treatment initiation. This method can manage data censoring effectively, accommodating cases where patient follow-ups are incomplete. Survival curves generated by the Kaplan-Meier estimator are critical for visually comparing the effectiveness of different treatments over time.

3.3.3 MACHINE LEARNING

Machine learning (ML) and artificial intelligence (AI) are terms often used interchangeably, yet they describe different concepts. AI is a broad field focused on creating machines capable of mimicking human behaviors like reasoning, learning, and predicting. ML, a subset of AI, employs data-driven algorithms to learn from experience without explicit programming. It identifies patterns in data to make progressively better decisions. ML is categorized mainly into four types: supervised, unsupervised, semi-supervised, and reinforcement learning. 'Supervised' refers to algorithms trained using labeled data—data tagged with correct outputs to provide models with a baseline understanding of expected results, necessitating some human oversight. Conversely, unsupervised learning algorithms discern patterns in raw, unlabeled data, making them useful for discovering new patterns[40].

Figure 1: Hierarchical Overview of Machine Learning Categories

CLUSTERING**:** In clustering, the task assigned to the ML model involves forming subgroups within a larger group. The goal is to maximize similarity within each subgroup (intra-similarity) and minimize similarity between different subgroups (inter-similarity). This essentially divides data based on common characteristics among data points. Several algorithms with different strategies exist to achieve this, one of the most common being K-means clustering.

K-MEANS CLUSTERING: K-means is a popular unsupervised machine learning algorithm used to partition a dataset into k clusters. In K-means clustering, the human operator decides the number of clusters, 'k', to be created. For example, if k = 5, the model will create five clusters aiming to maximize intra-cluster similarity. It operates iteratively to group data points into clusters based on feature similarity. The algorithm begins by randomly selecting 'k' initial centroids from the data points, which are the central points of the initial clusters. Each data point is assigned to the nearest centroid, with proximity typically measured using Euclidean distance. Each point joins the cluster with the closest centroid. After all points are assigned, the centroid of each cluster is recalculated by taking the mean of all points assigned to that cluster, effectively centering the centroid. These assignment and update steps are repeated. Data points may shift between clusters based on newly calculated

centroids. The algorithm continues this process until the centroids stabilize and changes are negligible. This indicates optimal clustering with high intra-cluster homogeneity and distinct inter-cluster separation[41].

ELBOW METHOD: When using k-means clustering, especially in fields like medical research, the selection of the number of clusters, k, is pivotal. This is because the primary goal of clustering in this context is usually to uncover underlying structures within the dataset that are not immediately apparent. The aim is to find groups that accurately reflect different stages or types of diseases, hence improving precision diagnostic.

The choice of k critically influences how finely the data is divided into clusters. Too few clusters may oversimplify the variations within the data, while too many clusters might interpret random noise as significant differences, which can lead to model overfitting. The optimal value of k captures the precise level of detail that aligns with significant medical distinctions. Choosing k involves balancing intra-cluster similarity against inter-cluster separation. Naturally, intra-cluster similarity increases as more clusters are formed because each cluster then tends to encompass more homogeneous data points. Logically, if each data point (or patient) represented its own cluster, intra-cluster dissimilarity would be nonexistent. However, such an extreme division would be clinically irrelevant as it provides no meaningful insight for medical analysis or decision-making.

Introduction to k-Means Clustering Using a School Gym Analogy

Imagine a class of children heading to their school gymnastics class, where they are going to play a game that emphasizes similarity among team members in terms of height and weight. The children, who represent data points in k-means clustering, are arranged on the gym floor based on their height and weight: the tallest on one side and the shortest on the other, similarly for weight from heaviest to lightest.

Step 1: Choosing Initial Team Leaders (Centroids). Initially, four team leaders, akin to the centroids in k-means, randomly position themselves somewhere on the floor among the children. This random placement of team leaders is crucial as it sets the initial conditions for the clustering process, much like how initial centroids are chosen in the k-means algorithm.

Step 2: Assigning Teams Based on Proximity. In this step, the distance between each child and the various team leaders is measured. Each child then joins the team whose leader is closest to them. This mirrors the assignment phase in k-means where each data point is assigned to the nearest cluster based on a distance metric (usually Euclidean distance), which serves as a measure of similarity.

Step 3: Updating Team Leaders' Positions. Once all children are assigned to teams, each team leader moves to the center of their team members, ensuring they are perfectly centered. This is analogous to recalculating the centroids in k-means, where each new centroid is computed as the mean of all points in the cluster.

Iteration of Assignment and Updating. The process of measuring distances (Step 2) and updating team leaders' positions (Step 3) is repeated. After each round, if a different team leader is now closer, a child may switch teams. These steps are repeated iteratively until the teams no longer change, which indicates that the clusters have been optimized in terms of internal similarity and external dissimilarity.

Convergence. The division is completed when no children change teams between rounds, meaning the best possible grouping has been achieved based on the initial setup. This condition of no change signifies convergence in the k-means algorithm, often indicating that further iterations will not significantly improve the partitioning.

To guide the human operator in choosing k, several methods are available, with the elbow method being one of the most popular. This method effectively demonstrates the trade-offs between inter-cluster separation and intra-cluster compactness in a simple, visual manner. In this approach, k-means clustering is performed multiple times with different values of k. The Within-Cluster Sum of Squares (WCSS) is calculated each time and plotted against the number of clusters, k. Typically, the graph shows a sharp bend or 'elbow' where the reduction in WCSS becomes less pronounced as k increases, indicating that the quality of the clusters does not markedly improve with further divisions.

3.3.4 CLASSIFICATION TASK AND XGBOOST

To solve the tasks in studies III and IV, a completely different branch of ML was required. The data provided was labeled, meaning it informed the model of the expected output (type 2 diabetes in study III and MACE in study IV). This allowed the model to train on data where the outcomes for the respective participants were already known.

CLASSIFICATION TASKS: The basics of classification tasks are quite simple. The goal is to correctly label outputs based on various inputs and thereby categorize the output into predetermined categories. For example, a task might require the model to distinguish between humans and dogs. The input in this case could be questions like 'Does it have fur: yes/no?' or 'Does it walk on two legs: yes/no?' The output is binary; it can either be a dog or a human. Initially, the model does not have the human understanding of what a human or a dog is, so it needs to be trained on data where the answers are provided. It will then, through iterative learning, begin to recognize the connection between having fur and more than two legs with being a dog. This part of the process, where we allow the model to learn about the relationship between input and output, is called model training.

After the training is complete, we provide the model with a new set of data; this time, we do not tell it whether it is a dog or a human, we just provide information about any fur and number of legs. We can then assess how often the model is correct in its labeling. This part of the process is called model evaluation.

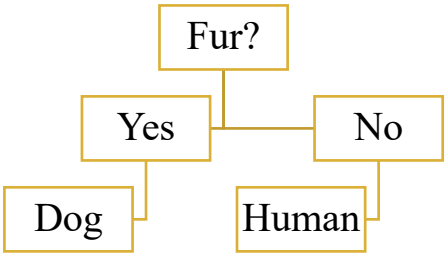
APPLICATION OF CLASSIFICATION TASKS IN MEDICINE: In medicine, classification tasks are often used for prediction, such as estimating the likelihood of contracting a specific disease within a certain timeframe. This

information is then used to determine the necessity of initiating preventive measures to avoid the disease in question.

Neither in study III nor in study IV was our ultimate goal to develop prediction models for disease risk. Instead, we aimed to utilize the knowledge that the trained model had gained about the relationships between different inputs and outcomes. In the simplistic example of distinguishing between dogs and humans, the task does not require the sophisticated capacity of ML algorithms, as human brains can easily discern the association between having fur and being a dog. However, as the complexity of the input data increases, the relationship between inputs and outputs becomes less clear, and eventually too complex for the human mind to recognize and comprehend. This complexity is exacerbated by the fact that input variables often have non-linear relationships with the outputs and interact with each other. Given that there were more than 400 input variables in both study III and IV, ML classification tools, particularly Xgboost, with its inherent capability to capture complex relationships, were deemed suitable.

XGBOOST (EXTREME GRADIENT BOOSTING): At its core, XGBoost utilizes decision trees as its base learners. A decision tree functions like a flowchart, dividing data based on specific conditions. For instance, using the dog/human analogy, one criterion might be 'does it have fur?' If yes, the data point moves in one direction; if no, it moves in another (figure 2). In Xgboost, several week learnings (simple decision trees) are combined to form one strong learner. As the name implies, gradiant boosting is involved. Gradient boosting is a concept where small models are build in a sequenced process, where each new model focuses on improving in errors the previous one has made. Essentially, each tree attempts to correct the mistakes of the ones before it by asking: "Where did the last prediction go wrong, and by how much?". Then it tries to produce new values (adjustments) that when added to the previous predictions, lead to more accurate results. This process gradually nudges the predictions closer to the true values. The series of corrections continues until the model reaches a predetermined number of trees (set by the human operator). By predicting the errors rather than the actual output, each subsequent model in the sequence needs only to improve incrementally upon what has already been achieved, making the learning process more manageable and the model more accurate. The final prediction is a weighted sum of all of the tree predictions, where more accurate trees has a bigger influense.

Figure 2: Example of decision tree



PRECTICAL EXAMPLE: Imagine you are teaching a child to hit a target. The first throw is far to the left of the target. Instead of focusing on hitting the target directly again, you advise adjusting the throw slightly to the right. The child's next throw is closer but still slightly off. You make another small suggestion for adjustment. Each piece of advice (like each new tree in XGBoost) is aimed at correcting errors from the previous attempt rather than achieving perfection in one go

HYPERPARAMETERS AND TUNING: A hyperparameter is a setting in a ML algorithm that is established before the algorithm begins learning from the data. Unlike parameters, which the algorithm learns during training, hyperparameters are set by the human operator. Hyperparameters act like a set of rules that the model must adhere to during training, controlling factors such as the number of trees to build and how complex each tree is allowed to be.

Initially, it's often unclear which values for the hyperparameters will produce the best model performance. Therefore, a process known as hyperparameter tuning is used to identify the most effective settings. One common method used in hyperparameter tuning is the creation of a hyperparameter grid, which involves defining a grid of hyperparameter values to systematically explore their effects. Since most hyperparameters have a range within which the optimal values are likely to fall, this grid allows the search for the best hyperparameters to be strategically focused within this range, streamlining the process and improving efficiency in finding the most effective model settings.

XGBoost has become one of the most popular algorithms for classification and regression tasks due to several favorable qualities. These include its speed, ability to handle complex relationships in data, regularization techniques that prevent overfitting, and a clever method for handling missing data[42].

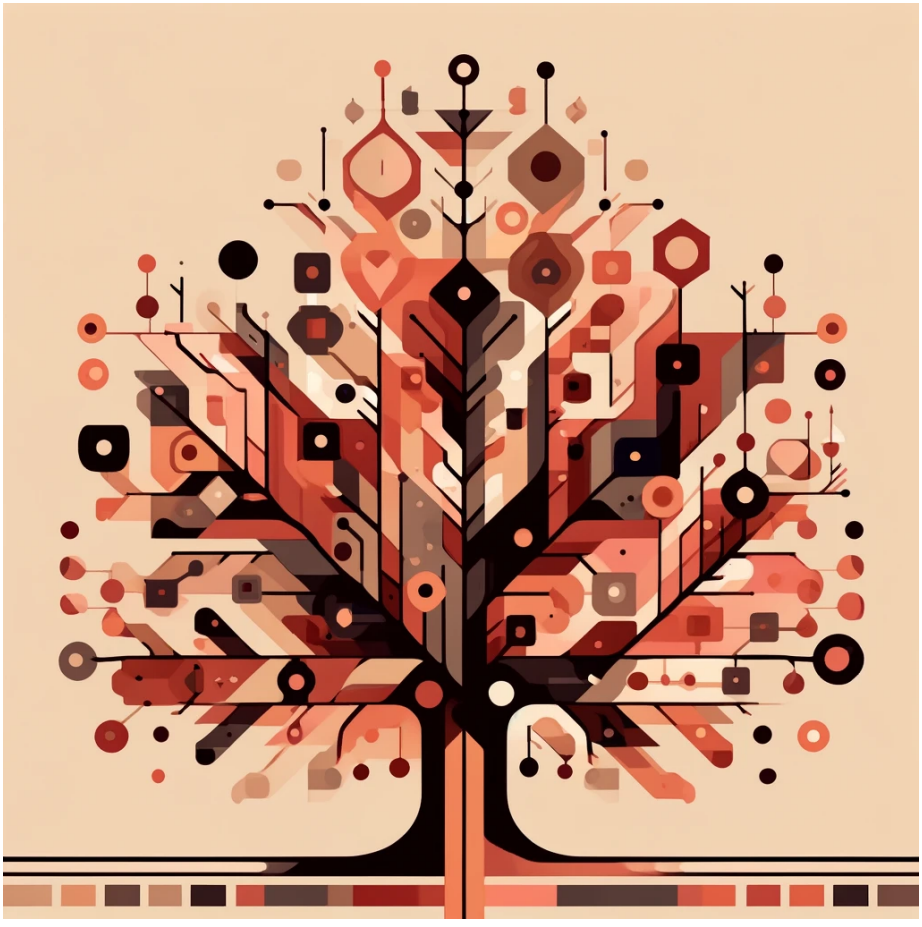
SHAP (SHAPLEY ADDITIVE EXPLANATIONS): Understanding why a model makes a certain prediction is often as important as the prediction itself. In Studies III and IV, elucidating the reasoning behind the model's decisions was, in fact, the primary purpose of training the models. ML algorithms are frequently criticized for their lack of interpretability and are often described as “black boxes”—systems where complex calculations occur unseen, and predictions appear to emerge from nowhere. As a result, simpler models, such as linear models, are sometimes favored for their straightforward interpretability, even though they might sacrifice some accuracy compared to more complex models. To address this issue, various methods have been developed to provide insight into the "reasoning" behind the model's predictions. Among the most utilized are SHAP values, which help to elucidate these contributions in a comprehensible manner.

Shap originates from cooperative game theory where it was developed as a tool to fairly distribute the profits. It enabled a valuation of how important each player was to an overall cooperation, and what share that player therefore could expect to get from the overall profit. In a prediction modell, this is instead used to figure out how importante each feature is for the final prediction. For each individual prediction, SHAP calculates how much each feature in the model shifts the prediction from baseline. Each feature are in this way assigned its own SHAP value, that can be both positive or negative[43].

Example of SHAP Values in a Model Predicting Income: Suppose a person in a dataset has a predicted salary of $30K per month, based on features such as age, education, and sex. To understand the influence of specific features on this predicted salary, we use SHAP values:

SHAP Value for Sex: If the SHAP value for the feature 'sex=male' is +0.05, it indicates that being male contributes positively to increasing the predicted salary.

SHAP Value for Education Level: Suppose the SHAP value for 'Education Level = Bachelor's Degree' is +0.03. This lower SHAP value compared to 'sex=male' suggests that having a Bachelor’s degree contributes positively but has a slightly less impact on increasing the predicted salary than the sex being male.



RESULTS

This section provides a summary of the main findings from the studies included in this thesis. Detailed results are available in the respective articles.

Study I: The study included 9,648 participants treated with GLP-1 receptor agonists and 12,097 with SGLT-2 inhibitors, with median follow-up times of 1.7 and 1.1 years respectively. The most commonly used GLP-1 receptor agonists were liraglutide, dulaglutide, and exenatide once weekly. In the SGLT-2 inhibitors group, empagliflozin and dapagliflozin were predominantly used.

Clinical characteristics between the groups showed minimal differences according to the propensity score-adjusted analysis. In terms of outcomes, there were no significant differences in mortality, heart failure, and stroke between treatments. The reduction in HbA1c was 10.05 mmol/mol for GLP-1 receptor agonists users and 9.15 mmol/mol for SGLT-2 inhibitors users. Body weight reduction was 3.46 kg for SGLT-2 inhibitors users compared to 2.49 kg for GLP-1 receptor agonists.

Patients treated with GLP-1 receptor agonists had a significantly lower risk of stroke in the group without established CVD compared to those treated with SGLT-2 inhibitors, with a hazard ratio (HR) of 1.73 (95% CI 1.11-2.68), and in those without pre-existing cardiorenal disease, HR 2.08 (95% CI 1.21-3.58). The risk of MACE in those without established CVD was also lower for GLP-1 receptor agonists, with an HR of 1.27 (95% CI 1-1.62), though this result was underpowered (p=0.053).

Study II: The study involved 114,231 individuals from the SNDR with a mean age of 63 years, where 43% were women. The mean BMI was 30.5 kg/m2, and the average HbA1c was 54 mmol/mol. The study's median follow-up was 5.2 years. Clustering analysis, using the K-means algorithm, explored the optimal number of clusters for the population, with the elbow plot showing no clear cut-off points, suggesting difficulty in defining distinct groups. The silhouette method suggested two clusters as optimal, although silhouette scores were low, indicating weak clustering. Gap statistics slightly favored five clusters.

The four-cluster model showed distinct characteristics: Cluster 1 had a low age at diagnosis, high HbA1c, and relatively high BMI and LDL-cholesterol. Cluster 2 consisted of older individuals with lower HbA1c levels and high blood pressure. Cluster 3 included the oldest participants with the lowest HbA1c, BMI, and lipid levels, but the highest eGFR. Cluster 4 had the youngest individuals with the highest BMI and relatively high HbA1c. The clusters' characteristics aligned with known subtypes of diabetes, like MOD, MARD, and SIDD, based on various clinical parameters.

The ability of the clusters to predict mortality and cardiovascular disease events was compared using Cox regression models. The highest concordance achieved using cluster-based models was 0.66 for mortality and 0.68 for CVD. The concordance for mortality predictions was highest with the Cox model using splines for all variables (0.78), closely followed by the ordinary Cox model (0.77). Similarly, for cardiovascular events, the splines model showed the highest concordance (0.78), indicating superior predictive ability over the cluster models.

Study III: The study enrolled 448,277 individuals, with a median follow-up of 12.16 years, during which 12,148 participants developed type 2 diabetes. Those who developed diabetes were typically older by an average of 2.5 years and had higher BMI (31.5 kg/m^2 compared to 27.0 kg/m^2) at recruitment than those who did not develop the disease. The most robust predictors of type 2 diabetes identified were HbA1c, BMI, waist circumference, blood glucose levels, the number of first-degree relatives with diabetes, Gamma-Glutamyl Transferase (GGT), waist-hip ratio, HDL cholesterol, age, and serum urate levels.

The main model had an ROC-AUC of 0.90 and an accuracy of 0.92. In comparison, the reduced model with only 10 features had slightly lower performance metrics with an ROC-AUC of 0.88 and similarly high accuracy of 0.92. Both models exhibited high specificity (0.93), but the main model had better sensitivity (0.62) compared to the reduced model (0.57).

Study IV: The study followed 26,301 individuals diagnosed with type 2 diabetes, over a ten-year period, during which 2,367 participants experienced MACE, including 604 strokes, 1,103 myocardial infarctions, and 1,118 cardiovascular deaths. The average age of participants experiencing MACE was slightly higher at 61.71 years compared to 59.11 years for those without MACE. Both groups had a similar mean BMI of 31.70 kg/m^2.

Smoking status varied significantly between groups; 47.0% of those in the MACE group were current smokers compared to 41.6% in the non-MACE group. The MACE group also had higher systolic blood pressure averages (144.11 mmHg) than their non-MACE counterparts (141.42 mmHg), and slightly higher triglyceride levels. The duration of diabetes prior to the study was longer in the MACE group, averaging 13.04 years, and their HbA1c levels were also higher at 56.80 mmol/mol.

Regarding treatment, insulin was more commonly used in the MACE group (24.8%) compared to the non-MACE group (16.3%), while other common glucose-lowering treatments like Biguanides and Sulfonylureas were similarly used across both groups.

In terms of model performance, the accuracy was high at 0.90 and the ROC-AUC was 0.74.

SUMMARY OF RESULTS: These studies collectively highlight the clinical outcomes and safety profiles of contemporary diabetes treatments, the challenges in identifying distinct subgroups within type 2 diabetes, and the importance of key predictors for diabetes development and cardiovascular complications. The findings emphasize the need to incorporate diverse factors, including clinical characteristics, biomarkers, and lifestyle variables, into risk scores to more accurately predict an individual's actual risk of developing diabetes and its complications.

4 DISCUSSION

Currently, precision medicine in type 2 diabetes represents a promising concept rather than a standard practice. The studies presented in this thesis collectively strive to bridge existing knowledge gaps, advancing the field of type 2 diabetes towards a more precise and personalized approach to medicine.

SGLT-2 INHIBITORS AND GLP-1 RECEPTOR AGONISTS: In recent years, the range of drugs that are both effective and tolerable for managing type 2 diabetes has expanded substantially. GLP-1 analogues and SGLT-2 inhibitors have quickly become integral components of the diabetes treatment arsenal, driven by increasing evidence of their safety, glucose-regulatory effectiveness, and protective effects against serious diabetes-related complications [44]. Despite numerous studies comparing these drugs to placebos in large cardiovascular outcome trials (CVOTs), there remains a scarcity of direct head-to-head studies that compare these therapies. Consequently, Study I aimed to facilitate as direct a comparison as possible, enabling the assessment of differential treatment efficacy and safety profiles.

We utilized real-world data to compare GLP-1 receptor agonists and SGLT-2 inhibitors, matching the groups with propensity score and IPTW. The findings indicated that both treatments had similar effects on HbA1c levels and body weight reduction, as well as comparable rates of several cardiorenal outcomes. In subgroup analyses of participants without established CVD at baseline, GLP-1 receptor agonists proved more effective in preventing strokes.

Due to the continuing absence of direct head-to-head studies comparing GLP-1 receptor agonists with SGLT-2 inhibitors, most of our knowledge about the differences in effect between these two drug classes is based on indirect comparisons of results from CVOTs. These results show similar reductions in all cause mortality, cardiovascular mortality and myocardial infarction between the two drug classes[44]. The differences noted relate to heart failure, where SGLT2 inhibitors provide better protection than GLP-1 receptor agonists, and stroke, where the inverse relationship holds, with GLP-1 receptor agonists more effectively reducing risk[44].

However, CVOTs are conducted in highly controlled environments with frequent monitoring and intervention, which may not accurately represent real-world adherence, compliance, and lifestyle variations. The strict inclusion and exclusion criteria of CVOTs result in homogeneous study populations that do

not reflect the diverse characteristics of patients encountered in everyday clinical practice. Consequently, the findings of CVOTs may not be generalizable to broader populations. In contrast, real-world studies offer greater generalizability, practical relevance, inclusion of diverse and high-risk patient groups, and the ability to monitor real-world treatment patterns and outcomes. Therefore, they serve as a necessary complement to the knowledge gained from CVOTs.

In addition to Study I, there have been a number of real-world studies on cardiovascular outcomes and side effect profiles for SGLT-2 inhibitors and GLP-1 receptor agonists in different populations. A meta-analysis of real-world studies published in 2022 compared the results of eight different studies, including Study I[45]. The combined result showed no significant difference in composite cardiovascular outcomes between treatments with SGLT-2 inhibitors and GLP-1 receptor agonists. As seen in randomized controlled trials, SGLT-2 inhibitors appear to be superior to GLP-1 receptor agonists in reducing the risk of hospitalization due to heart failure[45]. We observed the same trend in Study I, although the difference was not statistically significant. In the compilation of real-world studies, Study I stands out regarding stroke[45]. We observed a clear trend where GLP-1 receptor agonists were more effective than SGLT-2 inhibitors in preventing stroke, whereas the other studies found equivalent efficacy. The credibility of our result is strengthened by the fact that this difference has also been noted in CVOTs[44]. There are several potential reasons why this difference was observed in our study but not in others. The studies were conducted in different populations, with variations in the types of individuals included. For instance, the mean age in the included studies varies from 50 years[46] to 71[47]. A common limitation across almost all studies, including Study I, is the relatively short follow-up period, with one exception that achieved a median follow-up of 4.3 years[48]. This Danish study found no significant differences in standardized 3-year absolute risk for any major cardiovascular endpoints[48].

In summary, while GLP-1 receptor agonists and SGLT-2 inhibitors have both demonstrated significant benefits in the management of type 2 diabetes, the nuances of their effectiveness and safety profiles require further elucidation through direct comparisons. Study I contributes valuable real-world evidence, highlighting the comparable impacts of these treatments on glycemic control and cardiovascular outcomes. It also underscores the superior efficacy of GLP-1 receptor agonists in stroke prevention within certain subgroups. However, the limitations inherent in real-world studies, including variations in study

populations and relatively short follow-up periods, must be considered when interpreting these findings. Ultimately, the complementary insights gained from both CVOTs and real-world studies are crucial for optimizing therapeutic strategies and improving patient outcomes in diverse clinical settings.

TO CLUSTER OR NOT TO CLUSTER**:** Debates about precision diagnostics in type 2 diabetes often highlight the need for a more refined classification of the broad and diverse group currently identified as type 2 diabetes. However, the optimal approach to achieving such subgrouping remains elusive. Numerous efforts have been made to define subgroups based on phenotypic[49] and genetic data[50, 51] but the real challenge lies in establishing a stable, clinically beneficial subgrouping. Such a classification would ideally predict complication risks or treatment responses more accurately, enhancing therapeutic outcomes.

In 2018, a pivotal article using cluster analysis identified five potential subgroups of diabetes based on six clinical variables, igniting widespread interest due to its potential implications[49]. The clusters consists of Severe Autoimmune Diabetes (SAID), characterized by autoantibodies; Severe Insulin-Deficient Diabetes (SIDD) with insulin deficiency; Severe Insulin-Resistant Diabetes (SIRD) marked by insulin resistance; Mild Obesity-Related Diabetes (MOD) associated with obesity; and Mild Age-Related Diabetes (MARD), typically seen in older adults. In Study II, we aimed to validate these subgroups in a different population and assess their clinical utility in terms of predictive value. Our findings indicated that while similar groupings could be replicated within the SNDR population, they did not offer predictive advantages over simple risk calculations for forecasting future risks of CVD and mortality. Furthermore, we observed that the quality of clustering did not significantly vary with the number of clusters, suggesting either a lack of distinct, underlying groups within type 2 diabetes or that the clinical variables used for clustering were not ideally suited for this purpose.

While Study II focuses on data-driven subclassification, the predominant approach for subgrouping type 2 diabetes follows a markedly different strategy[52]. This approach is expert-driven, utilizing existing knowledge of the disease and predefined cut-offs for common variables such as HbA1c, age, and body composition. One example of this approach is the proposed recommendation to set individual HbA1c targets according to factors such as the duration of diabetes and the presence of diabetes complications[53]. This individualized approach aims to achieve a more appropriate risk-to-benefit

ratio for all patients by considering the unique characteristics of each person. It highlights the importance of phenotypic traits that are easy to collect and factor in, such as age, body weight, complications, and duration of diabetes[53]. Other proposed systems aim to stratify patients based on disease severity by incorporating key risk factors such as diabetes duration, HbA1c levels, microvascular complications, comorbidities, and prescribed treatments[54]. These systems are closely linked to previously discussed risk scores for diabetes complications[27, 29]. The advantages of these expert-driven strategies are that the relationship between the input variables and the outcomes is often well understood, and they are typically based on variables that are routinely collected. Consequently, the outputs often make sense to both researchers and clinicians since they are grounded in existing knowledge about the disease drivers. However, a systematic review published in 2023, which examined 51 such different simple subclassifications, found that none had been replicated and most lacked association with clinical outcomes[55].

The appeal of data-driven strategies lies in their ability to generate hypotheses and reveal new insights into diseases that might have previously been overlooked. However, these strategies often yield results that are more complex and harder to interpret, necessitating a greater degree of evidence and careful verification. This complexity means that clinicians may require more substantial proof before basing treatment decisions on these findings. Consequently, the path towards clinical implementation of such data-driven insights can be longer, as the results may not be as intuitive as those derived from expert-driven methods, demanding more rigorous evaluation to ensure they enhance precision in clinical use.

Among the data-driven subgroupings, the clusters proposed by Ahlqvist et al. [49] remain the most widely replicated and validated[55, 56]. The aforementioned systematic review reported that these clusters have been replicated across 22 different studies[55]. It has been repeatedly demonstrated that employing similar clustering techniques with the same or closely related clinical variables across different populations typically yields clusters akin to those proposed by Ahlqvist et al[57, 58, 59, 60]. Additionally, individuals in different diabetes subgroups have been shown to exhibit varying risks for specific complications[60, 61, 62, 63]. However, several challenges remain before the subgrouping proposed here can be implemented in clinical practice.

Individuals within different clusters often overlap, many situated on the edges and displaying nearly equal likelihoods of fitting into adjacent clusters[57]. In

Study II, we found that the number of clusters used to categorize patients with type 2 diabetes seems arbitrary, which challenges the idea that these clusters reflect distinct disease groups with unique pathophysiological traits. Additionally, cluster boundaries are not fixed; individuals frequently shift between clusters over a five-year span after initial classification[60]. To demonstrate clinical utility, it is crucial to prove that clustering can precisely predict the risk of future complications. Although research has consistently shown variation in the complication spectrum across different clusters[60, 61, 62, 63], Study II and other studies have confirmed that predicting risks based on individual-level data is more effective than using cluster-based approaches[22].

Thus, despite its appealing concept and proven reproducibility, along with observed differences in complication spectra between clusters, the results of Study II and other studies show that this clustering system is neither robust enough nor does it bring clinical benefit in terms of risk prediction. In its current state, these weaknesses argue against this subgrouping system providing precision for patients with type 2 diabetes.

In conclusion, while machine learning and other data-driven techniques offer promising insights into diabetes subtypes, they face challenges in validation, interpretability, and clinical integration[55]. Conversely, expert-driven strategies provide a practical framework for managing Type 2 diabetes, although they may lack the precision of data-driven approaches. Moving forward, integrating insights from both methodologies could enhance precision diagnostics in diabetes care, aiming to develop a subclassification system that accurately predicts health outcomes and informs tailored interventions for individual patients. This endeavor will require ongoing research, interdisciplinary collaboration, and constant model refinement based on clinical data to ensure that subclassifications are scientifically valid and practically feasible.

PREDICTORS FOR TYPE 2 DIABETES**:** Understanding which factors can identify individuals at high risk of developing type 2 diabetes serves multiple purposes. This knowledge deepens our understanding of the disease processes driving its development and supports practical applications such as risk estimation and the identification of high-risk individuals. One of the challenges of type 2 diabetes is that its initial stages often remain asymptomatic, allowing the adverse effects of chronic hyperglycemia to progress undetected for years.

Recent estimates indicate that over 34% of all US adults meet the criteria for prediabetes. Consequently, it is recommended to screen everyone aged 35 to 70 who is overweight or obese for type 2 diabetes[64]. Outside the US, discussions continue regarding which populations and risk groups should be screened for type 2 diabetes[65]. By identifying high-risk individuals early, monitoring and preventive measures can be implemented, thereby reducing the overall disease burden.

In Study III, our objective was to impartially sort and rank various factors to determine which are most predictive of developing type 2 diabetes. We included as much information about the participants as possible in this analysis to conduct the most comprehensive search feasible. For this reason, we utilized data from the UK Biobank, whose impressive size allowed us to consider a multitude of factors relating to physical composition, comorbidities, biomarkers, and lifestyle, among others.

In Study III's comprehensive analysis of 419 factors, HbA1c, BMI, waist circumference, and blood glucose levels were identified as the most significant predictors of type 2 diabetes. Together with the number of first-degree relatives with diabetes, GGT, waist-hip ratio, HDL cholesterol, age, and serum urate levels, these factors comprise the top ten strongest predictors. Initially, our model, which included all 419 factors, demonstrated a strong predictive capability for assessing the risk of developing type 2 diabetes over a 10-year period. To assess the impact of a more focused approach, we developed a secondary model that included only these top ten predictors. Although this reduced model also performed commendably, it was slightly less effective according to some evaluation metrics.

The strength, and also the challenge, of this study is the enormous amount of data used in the analysis. We analyzed data from 450,000 individuals, each with roughly 400 variables. This provided an almost unique opportunity to paint a high-resolution picture of various participants, allowing us to impartially evaluate a wide range of lifestyle, health, and socio-demographic factors to identify the strongest predictors for the development of type 2 diabetes. Because the ranking was largely data-driven and unmanipulated, each factor was given a chance to distinguish itself in its ability to predict disease.

Our results reinforce several well-known and widely used predictors such as HbA1c, glucose, and anthropometric measures, but also highlight some less established ones such as GGT and urate.

Timely diagnosis and initiation of treatment are crucial in type 2 diabetes, since the disease can be postponed or even avoided completely following early interventions such as lifestyle modifications or pharmacotherapy[66, 67]. As a result, numerous researchers have focused on developing predictive models for type 2 diabetes. Initially, these models employed traditional statistical learning methods such as linear regression. Other commonly used methodologies include Cox proportional hazards models and logistic regressions[68]. More recently, a wide variety of machine learning techniques have been incorporated into this field of study. Popular choices include Support Vector Machines, Random Forests, Gradient Boosting Trees, and Deep Neural Networks[69].

Common predictors used to predict the onset of type 2 diabetes include BMI, HbA1c, age, lipid status, family history of diabetes, hypertension, and waist circumference[68, 70]. The number of predictors included in models for forecasting type 2 diabetes varies widely, and there is currently no consensus on which factors should be included to maximize prediction accuracy[69].

The findings from Study III reinforce the efficacy of traditional risk factors—such as HbA1c, BMI, waist circumference, plasma glucose, family history of diabetes, waist-hip ratio, HDL cholesterol, and age—in predicting the onset of type 2 diabetes. These predictors are notably advantageous due to their ease of collection, often as part of routine clinical practice. Additionally, we identified two less common predictors: serum urate levels and GGT. Although these are not typically used in predictive modeling for type 2 diabetes, both have shown robust associations with the disease, reinforcing the validity of our findings[71, 72].

In study III, HbA1c, was identified as the top predictor. It plays a dual role in the context of type 2 diabetes as it is both a critical predictor and an integral component of the disease's pathophysiology. Unlike other measures of glucose metabolism such as fasting plasma glucose, HbA1c offers the advantage of stability over time. It is less affected by daily fluctuations and external factors such as stress and illness, making it a reliable indicator[73]. The robustness of HbA1c as a predictor for type 2 diabetes is clear, generating similar results as established risk scores combining several different clinical features[74].

Although we did not directly analyze the optimal number of predictors, our results are revealing. The model that utilized only the top 10 predictors performed nearly as well as the primary model that incorporated all 419 factors, indicating a trade-off between the loss of information and dataset

conciseness. This nearly maintained accuracy suggests that larger datasets do not necessarily lead to better predictions. This aligns well with other research, which found that decreasing the number of predictors can actually improve model accuracy[69]. Our findings imply that selecting the most relevant predictors may be more beneficial than using a large number of less significant factors.

To conclude, Study III effectively demonstrates the predictive strength of select biomarkers and clinical metrics in forecasting the onset of type 2 diabetes. By methodically analyzing a vast array of factors from the UK Biobank, the study underscores the profound significance of HbA1c, BMI, waist circumference, and plasma glucose levels as top predictors. These findings advocate for a targeted approach in diabetes risk assessment, emphasizing the utility of a condensed set of well-chosen predictors over broader datasets.

PREDICTORS FOR CARDIOVASCULAR DISEASE IN TYPE 2 DIABETES

In Study IV, we aimed to further refine the method developed during our work on Study III. To achieve this, we opted for a similar setup but shifted our focus to predicting the risk of MACE in individuals with type 2 diabetes. Cardiovascular disease is the most common serious complication of type 2 diabetes. While risk assessments are crucial for identifying high-risk individuals, existing models still perform suboptimally[75].

Among the 437 factors analyzed in our study, Cystatin C, age, and established CVD emerged as the most significant predictors, with only one diabetes-specific factor, HbA1c, ranking among the top ten most important. This highlights that although type 2 diabetes significantly increases the risk of CVD, the primary risk factors resemble those in the general population. Given that diabetes-specific risk scores—incorporating factors such as glucose levels, diabetes duration, and diabetes treatments—often perform comparably or sometimes less effectively than general population scores[75], our findings suggest that diabetes-specific models might not need unique diabetes-specific factors but could instead simply adjust the overall risk estimate upwards, reflecting a broader risk profile.

That Cystatin C stands out as the most important predictor is an interesting and promising finding from several aspects. Cystatin C usually serves as a sensitive and stable marker of kidney function[76]. Elevated levels are strongly associated with several cardiovascular outcomes, including stroke, myocardial infarction, and cardiovascular mortality[77, 78, 79]. It is also easy and relatively inexpensive to collect. Combining these facts with the results from Study IV highlights the promise of Cystatin C as a predictor of CVD, yet none of the established risk scores incorporate it in their models[75].

MACHINE LEARNING IN MEDICAL RESEACH: The potential for ML in medicine is enormous, largely due to its unparalleled ability to discern patterns within complex datasets that far exceed human capabilities. The availability of large quantities of data is also steadily increasing, driven by a recognition of the value of data preservation and enhanced electronic storage capacities. However, significant challenges are associated with the application of ML in medicine. An overwhelming number of studies and models are being produced, making it difficult to discern truly valuable contributions[80]. Furthermore, there has been a persistent lack of clear guidelines on how the development of ML models should be reported, which is crucial for their validation in subsequent stages and eventual integration into clinical practice[80]. Notably, initiatives like TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) and TRIPOD-AI+ aim to address these issues by promoting transparency and thoroughness in reporting[81]. They provide easy-to-use checklists to facilitate adherence to these standards.

Efforts to synthesize the current state of research on prediction models for type 2 diabetes have revealed several such shortcomings[69, 82]. The variability in models, driven by differing algorithms and the selection and number of predictors included, poses significant challenges in integrating the results into a coherent whole. Furthermore, the diversity in the evaluation metrics used to assess these models adds another layer of complexity. Compounding these issues is a lack of transparency in many studies regarding model training and feature selection, with reports often limited to selected evaluation measures.

Studies III and IV deliberately do not seek to compete with the plethora of existing predictive models, which often vary widely in terms of complexity and effectiveness. Instead, we aim to provide clear and unbiased insights into which predictors should be prioritized in the development of future models, thereby aiding in the refinement of predictive strategies and tools. By

providing comprehensive details about the predictors included, the methodologies employed in model development, and a thorough evaluation of the model's performance, we enhance the reproducibility and scrutiny of our work. This level of transparency not only facilitates a deeper understanding of our model's capabilities and limitations but also sets a standard for reporting that can be emulated in future studies. Both Study III and Study IV classify as so-called predictor finding studies, and therefore do not belong to the type of studies that TRIPOD-AI is aimed at. Despite that, we have tried to follow these established guidelines as closely as possible to contribute to increased quality and transparent reporting in the field.

In Summary: The realization of ML's full potential in medicine is currently constrained by factors such as a lack of transparency and a focus on innovation at the expense of validation. However, efforts by organizations like TRIPOD are working to standardize and harmonize practices within the field, fostering more reliable and clinically applicable advancements. In our studies, we have conscientiously aligned our methodologies with these proposed guidelines, thereby contributing to the necessary effort to improve rigor and clarity in this somewhat disordered field.

4.1 LIMITATIONS

The studies included in this thesis present several limitations, some inherent to their observational, registry-based nature. We have endeavored to mitigate these limitations through the use of carefully selected statistical techniques. However, issues such as residual confounding and missing data must be carefully considered when interpreting the results.

In Studies I and II, we utilized data from the SNDR, which encompasses virtually all individuals diagnosed with diabetes in Sweden, thus providing a representative sample of this population. However, the generalizability of our findings to other populations with different ethnic compositions and genetic backgrounds remains uncertain. Study I included exclusion criteria that could further limit generalizability, particularly to individuals with severe obesity or kidney failure. A significant challenge specific to this study was minimizing the risk of confounding by indication - the possibility that patient characteristics influencing treatment decisions could also affect outcomes. To address this, we employed propensity score matching combined with IPTW, a

widely recognized method for achieving pseudo-randomization in observational studies[83]. Nevertheless, we cannot definitively rule out the impact of residual confounding. The relatively short follow-up period (average 1.7 years for GLP-1 receptor analogues and 1.1 years for SGLT-2 inhibitors) also limited our ability to assess long-term treatment effects.

A major limitation of Study II was the unavailability of certain variables used in the original study by Ahlqvist et al., specifically two estimates of beta-cell function and insulin resistance[49]. Instead, we included other clinically relevant factors related to type 2 diabetes, though the impact of this substitution on our findings is difficult to determine.

Studies III and IV utilized data from the UK Biobank, which comprises voluntary participants and thus may exhibit a healthy volunteer bias, as these participants are generally healthier than the broader UK population[84]. Both studies also faced the challenge of developing predictive models with unbalanced data, a common issue when the outcome of interest, such as disease contraction, affects a small proportion of the study population. We addressed this by downsampling the majority class to balance the groups more evenly. Although this approach is effective, some evaluation metrics suggest that model performance was suboptimal in identifying the minority group.

Despite the UK Biobank being one of the world's most comprehensive databases, replete with extensive participant information, it still lacks certain predictors that could potentially prove useful in risk prediction. For instance, NT-proBNP has consistently been shown to be a significant predictor of CVD in individuals with diabetes[36]. It would have been insightful to assess its performance relative to factors we identified as having the greatest predictive ability, such as Cystatin C.

The need for research on diverse ethnic groups to enhance the generalizability of findings to various populations has been repeatedly emphasized throughout this thesis. It is a fundamental component of precision medicine. Unfortunately, all our studies primarily involved participants of White European descent, representing a significant limitation of this work and an area that should be prioritized in future research endeavors.

Another limitation that should be mentioned is the lack of genetic data in our analyses. Genetics plays a significant role in precision medicine, although in type 2 diabetes, it is still primarily theoretical at this stage. Our limitation to

phenotypic data may therefore reduce the comprehensiveness of our findings. By not including genetic factors, we might miss potential insights into individual susceptibilities and the underlying mechanisms of type 2 diabetes and cardiovascular disease. Future studies incorporating genetic data could provide a more holistic view and potentially improve the predictive power and personalized treatment strategies for these conditions. Despite this, our study's focus on readily available phenotypic data ensures practical applicability and relevance in current clinical settings.

4.2 CONCLUSION

This thesis has endeavored to advance the field of precision medicine in type 2 diabetes by addressing significant gaps through a series of focused studies. Our investigations have broadened the understanding of treatment effects, risk factor identification, and subgroup classification within this diverse patient population, laying the groundwork for more personalized and effective therapeutic strategies.

Our results have provided deeper insights into observed real-world treatment effects with GLP-1 receptor agonists and SGLT-2 inhibitors, offering guidance to clinicians in choosing between these therapies in the absence of direct head-to-head trials. Our findings underscore the importance of both traditional and novel biomarkers in predicting diabetes-related outcomes, and demonstrate that type 2 diabetes can be forecasted with only a few routinely collected variables. We have also highlighted the complexity in subgrouping type 2 diabetes and the remaining challenges in demonstrating the clinical utility of proposed subgroups.

Ultimately, the advancement of precision medicine in type 2 diabetes will depend on ongoing research, interdisciplinary collaboration, and the continual refinement of models based on emerging data.

5 FUTURE PERSPECTIVES

Precision medicine in type 2 diabetes faces numerous challenges before it can be effectively implemented in clinical practice.

While GLP-1 receptor agonists and SGLT-2 inhibitors are generally associated with positive outcomes, the expected effects at the individual level remain uncertain[24]. Investigating the impact of various diabetes medications in relation to patient characteristics—such as ethnic background, phenotypic traits, and genetic variations—could significantly improve the personalization and effectiveness of treatments. Accurately predicting which individuals will be poor responders to specific treatments would allow us to avoid unnecessary treatments and potential side effects, thereby optimizing therapeutic outcomes. As these treatments become more prevalent and users have longer exposure, there are valuable opportunities to conduct more extensive real-world comparisons. These studies could beneficially include subgroup analyses to better understand how treatments affect different individuals with type 2 diabetes.

There are still no established subgroups within type 2 diabetes, despite the potential clinical utility of such classifications. It may be that clinical variables such as age, HbA1c levels, and blood pressure alone are not sufficient to define clear clusters[22]. Future research could explore whether more complex models that integrate a broad array of phenotypic variables with genetic data might be more successful in identifying stable group divisions.

Regarding risk scores for CVD in type 2 diabetes, there is significant potential not only to develop new, effective models but also to enhance those that are already established and validated. For instance, it would be intriguing to explore whether widely accepted models such as SCORE2-diabetes could be improved by incorporating Cystatin C, which emerged as the most significant predictor in our study. Research suggests that adding a measure of kidney function (creatinine-based eGFR) could indeed improve predictions for the standard SCORE2 model [48]. Whether this holds true for the diabetes-adjusted model and how Cystatin C-based measures perform compared to creatinine remains unknown.

ACKNOWLEDGEMENT

I would like to extend my heartfelt gratitude to all my colleagues, friends, and family who supported me throughout the development of this thesis.

First and foremost, I want to thank my main supervisor, Björn Eliasson, for his patient guidance, inspiring conversations, and, above all, for instilling in me the joy of research. Although I have no other main supervisor to compare him to, I find it hard to imagine being in better hands. The team we formed with my co-supervisor, Araz Rawshani, who brought innovative ideas and exceptional technical expertise, has been invaluable to me.

I would also like to thank my co-authors for their excellent cooperation. In particular, I want to mention Stefan Franzén, who first helped me appreciate the allure of statistics.

I am fortunate to be surrounded by smart and driven individuals at home who are always eager to discuss and provide interesting ideas and new perspectives. Both my brother, Jesper, and my husband, Joel, have significantly contributed to the breadth of the content in this thesis.

As I have the opportunity to express my gratitude, I want to thank my wonderful colleagues at Omtanken Grimmered and my friends at Vagnhallen Crossfit for making my days so enjoyable and interesting.

Finally, I would like to acknowledge some of the most important people in my life: my best friends Emma, Bea, and Sofia, and, of course, my family. Thank you for being there for me through both the best and worst days. None of this would have been possible without you.

A special thank you to my husband, Joel, and our girl, Cena. Love you!

REFERENCES

1. Galicia-Garcia, U., et al. Pathophysiology of type 2 diabetes mellitus. Int J Mol Sci. 2020;21(17).
2. Saeedi, P., et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas, 9(th) edition. Diabetes Res Clin Pract. 2019;157:107843.
3. DeFronzo, R.A., et al. Type 2 diabetes mellitus. Nat Rev Dis Primers. 2015;1:15019.
4. Rawshani, A., et al. Mortality and cardiovascular disease in type 1 and type 2 diabetes. N Engl J Med. 2017;376(15):1407-18.
5. Davies, M.J., et al. Management of hyperglycaemia in type 2 diabetes, 2022. A consensus report by the american diabetes association (ada) and the european association for the study of diabetes (easd). Diabetologia. 2022;65(12):1925-66.
6. Stratton, I.M., et al. Additive effects of glycaemia and blood pressure exposure on risk of complications in type 2 diabetes: A prospective observational study (ukpds 75). Diabetologia. 2006;49(8):1761-9.
7. Sattar, N., et al. Twenty years of cardiovascular complications and risk factors in patients with type 2 diabetes: A nationwide swedish cohort study. Circulation. 2023;147(25):1872-86.
8. Gaede, P., et al. Multifactorial intervention and cardiovascular disease in patients with type 2 diabetes. N Engl J Med. 2003;348(5):383-93.
9. Association, A.D. 10. Cardiovascular disease and risk management: Standards of medical care in diabetes—2021. Diabetes Care. 2020;44(Supplement_1):S125-S50.
10. Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes (ukpds 34). Uk prospective diabetes study (ukpds) group. Lancet. 1998;352(9131):854-65.
11. Dai, Z.C., et al. Role and mechanisms of sglt-2 inhibitors in the treatment of diabetic kidney disease. Front Immunol. 2023;14:1213473.
12. Tsapas, A., et al. Comparative effectiveness of glucose-lowering drugs for type 2 diabetes: A systematic review and network meta-analysis. Ann Intern Med. 2020;173(4):278-86.
13. Muller, T.D., et al. Glucagon-like peptide 1 (glp-1). Mol Metab. 2019;30:72-130.
14. Liu, Y., et al. The weight-loss effect of glp-1ras glucagon-like peptide-1 receptor agonists in non-diabetic individuals with overweight or

obesity: A systematic review with meta-analysis and trial sequential analysis of randomized controlled trials. Am J Clin Nutr. 2023;118(3):614-26.
15. Crowley, M.J., et al. Effects of liraglutide on cardiovascular outcomes in type 2 diabetes patients with and without baseline metformin use: Post hoc analyses of the leader trial. Diabetes Care. 2020;43(9):e108-e10.
16. Neuen, B.L., et al. Sodium-glucose co-transporter-2 inhibitors with and without metformin: A meta-analysis of cardiovascular, kidney and mortality outcomes. Diabetes Obes Metab. 2021;23(2):382-90.
17. Naithani, N., et al. Precision medicine: Uses and challenges. Med J Armed Forces India. 2021;77(3):258-65.
18. Chung, W.K., et al. Precision medicine in diabetes: A consensus report from the american diabetes association (ada) and the european association for the study of diabetes (easd). Diabetes Care. 2020;43(7):1617-35.
19. Geiss, L.S., et al. Prevalence and incidence trends for diagnosed diabetes among adults aged 20 to 79 years, united states, 1980-2012. Jama. 2014;312(12):1218-26.
20. Steinarsson, A.O., et al. Short-term progression of cardiometabolic risk factors in relation to age at type 2 diabetes diagnosis: A longitudinal observational study of 100,606 individuals from the swedish national diabetes register. Diabetologia. 2018;61(3):599-606.
21. Hillier, T.A., Pedula, K.L. Complications in young adults with early-onset type 2 diabetes: Losing the relative protection of youth. Diabetes Care. 2003;26(11):2999-3005.
22. Dennis, J.M., Shields, B.M., Henley, W.E., Jones, A.G., Hattersley, A.T. Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared with models based on simple clinical features: An analysis using clinical trial data. Lancet Diabetes Endocrinol. 2019;7(6):442-51.
23. Williams, L.K., et al. Differing effects of metformin on glycemic control by race-ethnicity. J Clin Endocrinol Metab. 2014;99(9):3160-8.
24. Young, K.G., et al. Treatment effect heterogeneity following type 2 diabetes treatment with glp1-receptor agonists and sglt2-inhibitors: A systematic review. Commun Med (Lond). 2023;3(1):131.
25. Liu, W., et al. Machine-learning versus traditional approaches for atherosclerotic cardiovascular risk prognostication in primary prevention cohorts: A systematic review and meta-analysis. Eur Heart J Qual Care Clin Outcomes. 2023;9(4):310-22.
26. Kavaric, N., Klisic, A., Ninic, A. Cardiovascular risk estimated by ukpds risk engine algorithm in diabetes. Open Med (Wars). 2018;13:610-7.

27. Jahangiry, L., Farhangi, M.A., Rezaei, F. Framingham risk score for estimation of 10-years of cardiovascular diseases risk in patients with metabolic syndrome. J Health Popul Nutr. 2017;36(1):36.
28. group, S.w., collaboration, E.S.C.C.r. Score2 risk prediction algorithms: New models to estimate 10-year risk of cardiovascular disease in europe. Eur Heart J. 2021;42(25):2439-54.
29. Group, S.C.-D.W., the, E.S.C.C.R.C. Score2-diabetes: 10-year cardiovascular risk estimation in type 2 diabetes in europe. Eur Heart J. 2023;44(28):2544-56.
30. Marx, N., et al. 2023 esc guidelines for the management of cardiovascular disease in patients with diabetes: Developed by the task force on the management of cardiovascular disease in patients with diabetes of the european society of cardiology (esc). European Heart Journal. 2023;44(39):4043-140.
31. van Dieren, S., et al. External validation of the uk prospective diabetes study (ukpds) risk engine in patients with type 2 diabetes. Diabetologia. 2011;54(2):264-70.
32. Szymonifka, J., et al. Cardiovascular disease risk prediction for people with type 2 diabetes in a population-based cohort and in electronic health record data. JAMIA Open. 2020;3(4):583-92.
33. van Trier, T.J., et al. Validation of systematic coronary risk evaluation 2 (score2) and score2-older persons in the epic-norfolk prospective population cohort. Eur J Prev Cardiol. 2024;31(2):182-9.
34. Kasim, S.S., et al. Validation of the general framingham risk score (frs), score2, revised pce and who cvd risk scores in an asian population. Lancet Reg Health West Pac. 2023;35:100742.
35. Kist, J.M., et al. Score2 cardiovascular risk prediction models in an ethnic and socioeconomic diverse population in the netherlands: An external validation study. EClinicalMedicine. 2023;57:101862.
36. Ahmad, A., et al. Precision prognostics for cardiovascular disease in type 2 diabetes: A systematic review and meta-analysis. medRxiv. 2023.
37. Bak, J.C.G., Serné, E.H., Kramer, M.H.H., Nieuwdorp, M., Verheugt, C.L. National diabetes registries: Do they make a difference? Acta Diabetol. 2021;58(3):267-78.
38. Gudbjornsdottir, S., Cederholm, J., Nilsson, P.M., Eliasson, B., Steering Committee of the Swedish National Diabetes, R. The national diabetes register in sweden: An implementation of the st. Vincent declaration for quality improvement in diabetes care. Diabetes Care. 2003;26(4):1270-6.
39. Sudlow, C., et al. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12(3):e1001779.

40. Sarker, I.H. Machine learning: Algorithms, real-world applications and research directions. SN Computer Science. 2021;2(3):160.
41. Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B., Heming, J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences. 2023;622:178-210.
42. Chen, T., Guestrin, C., editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.
43. Sundararajan, M., Najmi, A. The many shapley values for model explanation. In: Hal D, III, Aarti S, editors. Proceedings of the 37th International Conference on Machine Learning; Proceedings of Machine Learning Research: PMLR; 2020. p. 9269--78.
44. Sodium-glucose cotransporter protein-2 (sglt-2) inhibitors and glucagon-like peptide-1 (glp-1) receptor agonists for type 2 diabetes: Systematic review and network meta-analysis of randomised controlled trials. BMJ. 2022;376:o109.
45. Caruso, I., et al. Cardiovascular and renal effectiveness of glp-1 receptor agonists vs. Other glucose-lowering drugs in type 2 diabetes: A systematic review and meta-analysis of real-world studies. Metabolites. 2022;12(2).
46. DeRemer, C.E., et al. Comparing cardiovascular benefits between glp-1 receptor agonists and sglt2 inhibitors as an add-on to metformin among patients with type 2 diabetes: A retrospective cohort study. Journal of Diabetes and its Complications. 2021;35(9):107972.
47. Patorno, E., et al. Comparative effectiveness and safety of sodium-glucose cotransporter 2 inhibitors versus glucagon-like peptide 1 receptor agonists in older adults. Diabetes Care. 2021;44(3):826-35.
48. Nørgaard, C.H., et al. Cardiovascular outcomes with glp-1 receptor agonists vs. Sglt-2 inhibitors in patients with type 2 diabetes. European Heart Journal - Cardiovascular Pharmacotherapy. 2021;8(6):549-56.
49. Ahlqvist, E., et al. Novel subgroups of adult-onset diabetes and their association with outcomes: A data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol. 2018;6(5):361-9.
50. Udler, M.S., McCarthy, M.I., Florez, J.C., Mahajan, A. Genetic risk scores for diabetes diagnosis and precision medicine. Endocr Rev. 2019;40(6):1500-20.
51. Udler, M.S., et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. PLoS Med. 2018;15(9):e1002654.
52. Hodgson, S., et al. Population stratification in type 2 diabetes mellitus: A systematic review. Diabetic Medicine. 2022;39(1):e14688.

53. Lyssenko, V., Bianchi, C., Del Prato, S. Personalized therapy by phenotype and genotype. Diabetes Care. 2016;39 Suppl 2:S127-36.
54. Zghebi, S.S., et al. Using electronic health records to quantify and stratify the severity of type 2 diabetes in primary care in england: Rationale and cohort study design. BMJ Open. 2018;8(6):e020926.
55. Misra, S., et al. Precision subclassification of type 2 diabetes: A systematic review. Commun Med (Lond). 2023;3(1):138.
56. Herder, C., Roden, M. A novel diabetes typology: Towards precision diabetology from pathogenesis to treatment. Diabetologia. 2022;65(11):1770-81.
57. Kahkoska, A.R., et al. Validation of distinct type 2 diabetes clusters and their association with diabetes complications in the devote, leader and sustain-6 cardiovascular outcomes trials. Diabetes Obes Metab. 2020;22(9):1537-47.
58. Fedotkina, O., et al. Novel reclassification of adult diabetes is useful to distinguish stages of β-cell function linked to the risk of vascular complications: The dolce study from northern ukraine. Front Genet. 2021;12:637945.
59. Xing, L., et al. Clinical characteristics and risk of diabetic complications in data-driven clusters among type 2 diabetes. Front Endocrinol (Lausanne). 2021;12:617628.
60. Zaharia, O.P., et al. Risk of diabetes-associated diseases in subgroups of patients with recent-onset diabetes: A 5-year follow-up study. Lancet Diabetes Endocrinol. 2019;7(9):684-94.
61. Saito, K., et al. Usefulness of subclassification of adult diabetes mellitus among inpatients in japan. J Diabetes Investig. 2022;13(4):706-13.
62. Pigeyre, M., et al. Validation of the classification for type 2 diabetes into five subgroups: A report from the origin trial. Diabetologia. 2022;65(1):206-15.
63. Li, P.-F., Chen, W.-L. Are the different diabetes subgroups correlated with all-cause, cancer-related, and cardiovascular-related mortality? The Journal of Clinical Endocrinology & Metabolism. 2020;105(12):e4240-e51.
64. Fang, M., Wang, D., Echouffo-Tcheugui, J.B., Selvin, E. Prediabetes and diabetes screening eligibility and detection in us adults after changes to us preventive services task force and american diabetes association recommendations. JAMA. 2022;327(19):1924-5.
65. World Health Organization. Chronic Respiratory, D., Arthritis, T. Screening for type 2 diabetes : Report of a world health organization and international diabetes federation meeting. Geneva: World Health Organization; 2003.

66. Knowler, W.C., et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med. 2002;346(6):393-403.
67. Group, T.D.P.P.R. The diabetes prevention program (dpp): Description of lifestyle intervention. Diabetes Care. 2002;25(12):2165-71.
68. Collins, G.S., Mallett, S., Omar, O., Yu, L.-M. Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting. BMC Medicine. 2011;9(1):103.
69. Fregoso-Aparicio, L., Noguez, J., Montesinos, L., Garcia-Garcia, J.A. Machine learning and deep learning predictive models for type 2 diabetes: A systematic review. Diabetol Metab Syndr. 2021;13(1):148.
70. Asgari, S., Khalili, D., Hosseinpanah, F., Hadaegh, F. Prediction models for type 2 diabetes risk in the general population: A systematic review of observational studies. Int J Endocrinol Metab. 2021;19(3):e109206.
71. Lee, D.H., et al. Gamma-glutamyltransferase and diabetes--a 4 year follow-up study. Diabetologia. 2003;46(3):359-64.
72. Krishnan, E., Pandya, B.J., Chung, L., Hariri, A., Dabbous, O. Hyperuricemia in young adults and risk of insulin resistance, prediabetes, and diabetes: A 15-year follow-up study. Am J Epidemiol. 2012;176(2):108-16.
73. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2010;33 Suppl 1(Suppl 1):S62-9.
74. Krabbe, C.E.M., et al. Comparison of traditional diabetes risk scores and hba1c to predict type 2 diabetes mellitus in a population based cohort study. J Diabetes Complications. 2017;31(11):1602-7.
75. Dziopa, K., Asselbergs, F.W., Gratton, J., Chaturvedi, N., Schmidt, A.F. Cardiovascular risk prediction in type 2 diabetes: A comparison of 22 risk scores in primary care settings. Diabetologia. 2022;65(4):644-56.
76. Lassus, J., Harjola, V.P. Cystatin c: A step forward in assessing kidney function and cardiovascular risk. Heart Fail Rev. 2012;17(2):251-61.
77. Wang, Y., et al. Association between cystatin c and the risk of ischemic stroke: A systematic review and meta-analysis. Journal of Molecular Neuroscience. 2019;69(3):444-9.
78. Shlipak, M.G., et al. Cystatin c and the risk of death and cardiovascular events among elderly persons. N Engl J Med. 2005;352(20):2049-60.
79. Svensson-Färbom, P., et al. Cystatin c identifies cardiovascular risk better than creatinine-based estimates of glomerular filtration in middle-aged individuals without a history of cardiovascular disease. Journal of Internal Medicine. 2014;275(5):506-21.
80. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): Explanation and elaboration. Annals of Internal Medicine. 2015;162(1):W1-W73.

81. Collins, G.S., et al. Tripod+ai statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
82. Oikonomou, E.K., Khera, R. Machine learning in precision diabetes care and cardiovascular risk prediction. Cardiovasc Diabetol. 2023;22(1):259.
83. Austin, P.C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res. 2011;46(3):399-424.
84. Fry, A., et al. Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. American Journal of Epidemiology. 2017;186(9):1026-34.