From the Department of Public Health Sciences
Karolinska Institutet, Stockholm, Sweden

RISK FACTOR PATTERNS IN TYPE 2 DIABETES AND CARDIOVASCULAR DISEASE: EXPLORING METHODS FOR PRECISION MEDICINE IN PUBLIC HEALTH

Diego Yacamán Méndez

Karolinska Institutet

Stockholm 2023

All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2023

ISBN 978-91-8017-195-3
Cover illustration: AI generated through Magic Media.

Risk Factor Patterns in Type 2 Diabetes and Cardiovascular disease: Exploring Methods for Precision Medicine in Public Health

Thesis for Doctoral Degree (Ph.D.)

By

Diego Yacamán Méndez

The thesis will be defended in public at the lecture hall Ljung at the Center for epidemiology and community medicine, Solnavägen 1E, 4th floor, in Stockholm, Sweden. On Wednesday 6th, of December 2023 at 09:00 a.m.

Principal Supervisor:
Docent Anton Lager
Karolinska Institutet
Department of Global Public Health

Co-supervisor(s):
Docent Ylva Trolle Lagerros
Karolinska Institutet
Department of Medicine, Solna
Division of Clinical Epidemiology

Professor Antonio Ponce de Leon
Universidade do Estado do Rio de Janeiro
Department of Epidemiology

Professor Yvonne-Forsell
Karolinska Institutet
Department of Global Public Health

Opponent:
Docent Thomas Frisell
Karolinska Institutet
Department of Medicine, Solna
Division of Clinical Epidemiology

Examination Board:
Associate Professor Ingelise Andersen
University of Copenhagen
Department of Public Health
Division of Social Medicine

Docent Helena Nordenstedt
Karolinska Institutet
Department of Global Public Health

Adjunct professor Ulrik Sartipy
Karolinska Institutet
Department of Molecular Medicine and Surgery

Popular science summary of the thesis

Non-communicable diseases such as type 2 diabetes and heart disease are among the most important causes of death and disability around the world, and therefore a major public health challenge. These conditions are caused by a complicated combination of genetic and lifestyle factors and can develop slowly without symptoms, which makes their prevention, diagnosis and treatment challenging.

The road to disease is different for everyone, and individual characteristics play an important role in determining who becomes sick and when. At the same time, public health recommendations tend to be the same for everyone. Although these interventions have been very successful in reducing the impact of chronic diseases in the population, this progress seems to be slowing down.

A possible way forward is implementing precision medicine in public health. Precision medicine is a concept based on using individuals' information to subgroup a population based on their characteristics. The aim is to be able to offer the best therapy for these subgroups. In public health, it entails designing and implementing preventive strategies according to individuals' attributes. However, the use of precision medicine in public health has not yet been widely studied.

In this thesis, I explore the utility of methods that can be used to apply precision medicine to public health issues. These methods aim to group individuals based on how an exposure changes over time, how several factors combine at a single time point, or seek to estimate the importance of different causes of disease. The overall purpose is to provide examples and a discussion on how precision medicine can be useful to inform and improve public health practice.

Resumen

Las enfermedades crónicas son la principal causa de enfermedad a nivel global, y representan un importante reto para la salud pública. Las causas de estas enfermedades incluyen combinaciones complejas de factores genéticos, metabólicos y de estilo de vida, así como elementos económicos y sociales. A ello se suma la evolución progresiva y en muchos casos asintomática, que hacen de su prevención, diagnóstico oportuno y tratamiento un desafío considerable.

El desarrollo de las enfermedades crónicas varía de persona a persona, ya que las circunstancias individuales desempeñan un papel crucial en quién y cuándo manifiesta una enfermedad específica. Sin embargo, las estrategias de salud pública suelen basarse en recomendaciones generales, sin considerar esta variabilidad individual. A pesar del gran éxito logrado por estas iniciativas, especialmente en la reducción de enfermedades del corazón, existen indicios de que este progreso ha ido disminuyendo. Al mismo tiempo, el impacto de otras enfermedades como la diabetes tipo 2 ha aumentado constantemente.

Una herramienta que podría ser de gran utilidad para seguir avanzando hacia una mejor salud poblacional es utilizar el concepto de medicina de precisión en la práctica de la salud pública. La medicina precisión, o personalizada, es un concepto basado la identificación de subgrupos de individuos con características similares dentro de la población general. Con el propósito de ofrecer tratamientos óptimos, ajustados a dichas características. En el área de salud pública, este enfoque puede ser de gran utilidad para el diseño e implementación de intervenciones preventivas y políticas sanitarias más efectivas. Sin embargo, el uso de la medicina de precisión en el área de salud pública no ha sido ampliamente estudiado.

En esta tesis, exploro la utilidad de diferentes métodos que pueden ser utilizados para aplicar el concepto de medicina de precisión en salud pública. Algunos de estos métodos tienen como finalidad agrupar a los individuos de una población en subgrupos homogéneos basándose en factores de riesgo importantes. Otros, evalúan la importancia de los diferentes mecanismos de las enfermedades. El objetivo es proporcionar ejemplos y discutir la utilidad de estos métodos para mejorar la práctica de la salud pública.

Sammanfattning

Kroniska sjukdomar utgör idag en av de främsta orsakerna till den globala sjukdomsbördan och en betydande utmaning inom folkhälsan. De bakomliggande mekanismerna är komplexa och omfattar kombinationer av olika genetiska, metabola, livsstilsrelaterade samt ekonomiska och sociala faktorer. Dessutom sker sjukdomsutvecklingen gradvis och är ofta asymtomatisk, vilket gör förebyggande, tidig diagnos och behandling utmanande.

Utvecklingen av en kronisk sjukdom är olika för varje individ, och individuella faktorer spelar en betydande roll i att avgöra vilka personer som drabbas av specifika sjukdomar och när dessa inträffar. Trots detta bygger folkhälsostrategier ofta på generella rekommendationer och bortser från den individuella variationen. Även om betydande framsteg har uppnåtts, särskilt när det gäller att minska förekomsten av hjärt-kärlsjukdomar, finns det indikationer på att dessa framsteg har avtagit. Samtidigt blir andra sjukdomar, såsom typ 2-diabetes, vanligare.

En möjlig väg framåt är implementeringen av precisionsmedicin inom folkhälsan. Begreppet precisionmedicin bygger på principen att subgrupper av individer med liknande egenskaper kan identifieras i den allmänna befolkningen, med syfte att erbjuda optimala och anpassade behandlingar. Inom folkhälsoområdet kan detta vara användbart för mer effektiv utformning och genomförande av förebyggande åtgärder. Trots detta har användningen av precisionmedicin som ett verktyg inom folkhälsan inte studerats i särskilt hög omfattning.

I denna avhandling undersöker jag om olika metoder kan bara användbara för att tillämpa precisionmedicin inom folkhälsoområdet. Vissa av dessa metoder syftar till att gruppera individer i en befolkning i homogena delgrupper baserat på distributionen av betydande riskfaktorer. Andra metoder syftar till att utvärdera betydelsen av olika sjukdomsmekanismer. Målet är att ge exempel och diskutera för- och nackdelar med dessa metoder för att förbättra praktiken inom folkhälsoområdet.

Abstract

Non-communicable diseases, including type 2 diabetes and cardiovascular disease, are leading contributors to the global burden of disease and an important public health challenge. At an individual level, there is important variability in the risk of these conditions. However, public health interventions often adopt a generalized one-size-fits-all approach.

The overall aim of this thesis was to explore the utility of a precision medicine approach to public health and epidemiology, by applying different analytical methods to classify individuals into similar sub-populations based on their individual level characteristics.

In **study I**, I investigated the patterns of weight changes from childhood to early adulthood and how they relate to the occurrence of type 2 diabetes later in life. The results indicate that exposure to overweight/obesity during early adulthood explains a large proportion of the cases of type 2 diabetes, highlighting the importance of public health interventions during this period.

In **study II**, I used different methods for mediation analysis to study the importance of different mechanisms linking low socioeconomic status and type 2 diabetes. The findings show that around 50% of the association between socioeconomic status and type 2 diabetes could be reduced if unhealthy behaviors and metabolic exposures were removed. Interestingly, the results were similar across the different mediation methods.

Finally, in **studies III and IV**, I used data-driven methods to identify sub-groups of healthy adults based on simple clinical characteristics and laboratory values. The findings show that this method was equally effective, or even better, than those commonly used in clinical practice, and could improve the way we define who is at high risk of type 2 diabetes or cardiovascular disease.

In conclusion, these studies provide evidence that precision medicine can be a useful approach to guide development and implementation of public health interventions.

List of scientific papers

I. **Yacamán-Méndez D**, Trolle-Lagerros Y, Zhou M, Ponce de Leon A, Gudjonsdottir H, Tynelius P, Lager A. Life-course trajectories of weight and their impact on the incidence of type 2 diabetes. Sci Rep. 2021 Jun 14;11(1):12494. doi: 10.1038/s41598-021-91910-z.

II. **Yacamán Mendez D**, Trolle Lagerros Y, Ponce de Leon A, Tynelius P, Fors S, Lager A. Behavioral and metabolic mediators of socioeconomic inequalities in type 2 diabetes: comparing counterfactual and traditional mediation analysis. (submitted)

III. **Yacamán Méndez D**, Zhou M, Trolle Lagerros Y, Gómez Velasco DV, Tynelius P, Gudjonsdottir H, Ponce de Leon A, Eeg-Olofsson K, Östenson CG, Brynedal B, Aguilar Salinas CA, Ebbevi D, Lager A. Characterization of data-driven clusters in diabetes-free adults and their utility for risk stratification of type 2 diabetes. BMC Med. 2022 Oct 18;20(1):356. doi: 10.1186/s12916-022-02551-6.

IV. **Yacamán Méndez D**, Zhou M, Trolle Lagerros Y, Lager A. Cluster analysis for cardiovascular risk stratification. (submitted)

Scientific papers not included in the thesis

1. Berglind D, **Yacaman-Mendez D**, Lavebratt C, Forsell Y. The Effect of Smartphone Apps Versus Supervised Exercise on Physical Activity, Cardiorespiratory Fitness, and Body Composition Among Individuals with Mild-to-Moderate Mobility Disability: Randomized Controlled Trial. JMIR Mhealth Uhealth. 2020 Feb 4;8(2):e14615. doi: 10.2196/14615.

2. Manhica H, **Yacamán-Méndez D**, Sjöqvist H, Lundin A, Agardh E, Danielsson AK. Trajectories of NEET (Not in Education, Employment, and Training) in emerging adulthood, and later drug use disorder - a national cohort study. Drug Alcohol Depend. 2022 Apr 1;233:109350. doi: 10.1016/j.drugalcdep.2022.109350.

3. Gudjonsdottir H, Tynelius P, Fors S, **Yacamán Méndez D**, Gebreslassie M, Zhou M, Carlsson AC, Svefors P, Wändell P, Östenson CG, Brynedal B, Lager A. Cohort Profile: The Stockholm Diabetes Prevention Programme (SDPP). Int J Epidemiol. 2022 Dec 13;51(6):e401-e413. doi: 10.1093/ije/dyac147.

4. Manhica H, **Yacamán-Méndez D**, Sjöqvist H, Lundin A, Danielsson AK. Early substance use disorders and subsequent NEET-not in education, employment or training-a national cohort study. Eur J Public Health. 2023 Aug 1;33(4):633-639. doi: 10.1093/eurpub/ckad105.

Contents

List of abbreviations

ADA	American Diabetes Association
ACC	American College of Cardiology
AHA	American Heart Association
ARD	Absolute risk difference
ASCVD	Atherosclerotic cardiovascular disease
BMI	Body mass index
CVD	Cardiovascular disease
CDE	Controlled direct effect
DALY	Dissability adjusted life-year
DAG	Directed acyclic graph
EASD	European Association for the Study of Diabetes
eGFR	Estimated glumeral filtration rate
ESC	European Society of Cardiology
FMM	Finite mixture models
GBD	Global burden of disease
GBTM	Growth based trajectory model
GLP-1	Glucagon-like peptide-1
HbA1c	Glycosylated hemoglobin
HOMA	Homeostasis model assessment
HR	Hazard ratio
IDF	International Diabetes Federation
IFG	Impaired fasting glucose
IGT	Impaired glucose tolerance
IRR	Incidence rate ratio
LISA	Longitudinal integrated database for health insurance and labour market studies
MSC	Metabolic Syndrome Cohort
NCD	Non-communicable disease
NDE	Natural direct effect
NIE	Natural indirect effect
NPV	Negative predictive value
OGTT	Oral glucose tolerance test

PAF	Population attributable fraction
PCE	Pooled-Cohort Equations
PPV	Positive predictive value
RCT	Randomized controlled trial
RR	Risk ratio
SDG	Sustainable development goals
SCORE	Systematic Coronary Risk Evaluation
SDPP	Stockholm Diabetes Prevention Program
SES	Socioeconomic status
SUTVA	Stable unit treatment value assumption
TE	Total effect
UMAP	Uniform manifold approximation and projection
VAL	Regional Data Warehouse of Stockholm
WHO	World Health Organization

Introduction

Social and economic changes over the last two centuries have affected the structure and dynamics of society, and in turn the main causes of disease. Globally, health has improved and the life expectancy has increased from around 30 years in 1870 to the current 73 years.[1]

The increase in life expectancy, together with changes in environmental exposures, have resulted in an increased occurrence of non-communicable diseases (NCDs), which are the main drivers of disability and mortality, and a pressing public health challenge. These conditions, which constitute more than 60% of the global burden of disease,[2] usually present gradually and without clear symptoms, making them difficult to prevent. Furthermore, once diagnosed, they require long term treatment, involving substantial human and material resources, which is reflected in the increasing expenditure in health.[3]

The etiology of NCDs is complex, involving a combination of genetic susceptibility and exposure to environmental, behavioral, and metabolic risk factors.[4,5] Significant progress has been made in the field of NCD prevention through public health measures and policies targeting risk factors such as smoking and alcohol consumption.[6] However, efforts to reduce overweight and obesity, improve dietary habits or increase physical activity have been less successful.[7]

Conventionally, public health interventions and policies adopt a "one-size-fits-all" approach, assuming that exposure to a particular risk factor has the same effect across all individuals in the population. However, the reality is that the consequences of exposure to risk factors vary widely among individuals depending on several aspects, including the timing and duration of the exposure, the different underlying causal mechanisms, and the combination of different risk factors.[8]

A solution could be to implement precision medicine in public health, i.e., to deliver the right intervention, to the right individuals at the right time.[9] A key step towards this goal is to identify and describe specific subgroups of individuals from the general population who share important characteristics and could benefit from similar interventions.[10] This thesis and the accompanying articles provide evidence of how analytical methods can be used in epidemiology and public health towards this goal, using type 2 diabetes and cardiovascular disease (CVD) as examples.

As a background, I start by providing an overview of the epidemiology of type 2 diabetes, CVD, and their associated risk factors. This is followed by a summary of the current recommendations for their prevention and early detection. Next, I discuss the somewhat conflicting notion of precision medicine in public health. Finally, I summarize the methods used to identify sub-populations based on different aspects of the variability of risk factors.

1 Epidemiology of type 2 diabetes and cardiovascular disease

Both type 2 diabetes and CVD are among the most important causes of mortality, disability,[2] and health-care related costs at all income levels around the world.[3,11] Nevertheless, there is also an important contrast in the epidemiology of these conditions. The overall burden of type 2 diabetes has gradually increased in recent decades (see Figure 1), which is likely related to general increases in risk factors like overweight and obesity.[12] In contrast, the burden of CVD has largely decreased during the past decades (Figure 1), mirroring a reduction in important risk factors, such as smoking, and the implementation of effective treatments for high-cholesterol, hypertension, type 2 diabetes, and other comorbidities associated with greater CVD risk.[6]

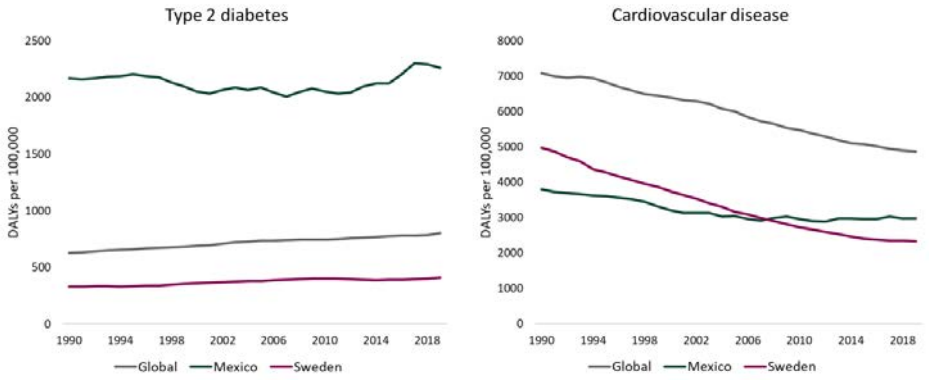

Figure 1. Burden of type 2 diabetes and cardiovascular disease in Sweden, Mexico, and the world from 1990 to 2019, measured as the age-standardized rate of disability adjusted life-years (DALYs).

1.1 Type 2 diabetes

Diabetes mellitus is the collective label given to a group of metabolic disorders characterized by insufficient insulin action resulting in persistent hyperglycemia and subsequent damage to blood vessels and nerves across the body.[13,14] Alterations at any point from the production and secretion of insulin to its uptake and action can lead to the clinical manifestations of diabetes.

Most cases (90% to 95%) are categorized as type 2 diabetes,[13] characterized by a combination of peripheral insulin resistance, and an inadequate compensatory insulin production or secretion from the pancreatic β- cells. According to the

World Health Organization (WHO) and the American diabetes Association (ADA), a diagnosis of type 2 diabetes is made according to the criteria summarized in Table 1.[13,15]

Table 1. Diagnostic criteria for Type 2 diabetes.

Fasting plasma glucose	≥7.0 mmol/L (≥126mg/dL)
HbA1c	≥48 mmol/L (≥6.5%)
OGTT	2-hr plasma glucose ≥11.1 mmol/L (≥200 mg/dL)
Random plasma glucose	≥11.1 mmol/L (≥200 mg/dL) and symptoms of hyperglycemia

Glycosylated hemoglobin (HbA1c), Oral glucose tolerance test (OGTT)

Estimates from the Global Burden of Disease study (GBD) suggest that almost half a billion people were living with type 2 diabetes in 2019, representing over 5% of the global population.[12] Furthermore, 20 million new cases and 1,5 million deaths due to type 2 diabetes were reported.[12]

During the last three decades, prevalence of type 2 diabetes has increased steadily, while the incidence seems to be stabilizing and even decreasing in large parts of the world. Indeed, the age-standardized prevalence increased from 3.8% in 1990 to around 5.5% in 2019. This increase might be expected because of important improvements in the treatment and survival of people living with diabetes.[16] Data about the incidence of type 2 diabetes are more scarce, but recent studies have reported an increasing trend up to around 2005, and since then a stable or declining incidence in many countries.[17,18] Unfortunately, data from most of the low and middle-income countries are not available. Overall, the global burden of type 2 diabetes continues to increase. In 1990, type 2 diabetes accounted for 628 disability adjusted life-years (DALYs) per 100,000 people, while in 2019 the burden increased to around 800 DALYs per 100,000 individuals globally.[12,16]

In addition, the economic costs related to diabetes pose an important burden to health care systems and individuals. Health expenditure due to type 2 diabetes accounts for over 10% of the total global spending. According to the International Diabetes Federation (IDF), 966 billion USD were spent on treating type 2 diabetes during 2021, mostly on managing long-term complications. Moreover, projections indicate that the cost will grow substantially in upcoming years.[11]

1.2 Cardiovascular disease

The term CVD is used to describe a group of disorders involving the heart or blood vessels including ischemic heart disease, stroke, hypertension, peripheral arterial diseases, and heart rhythm anomalies, among others.[6] Ischemic heart disease, stroke, and peripheral artery disease are caused by atherosclerosis, a chronic thickening and hardening of the blood vessels due to inflammation and accumulation of cholesterol, leading to their occlusion and are therefore collectively referred to as atherosclerotic cardiovascular disease (ASCVD).[19]

CVD remains the leading cause of death and overall burden of disease globally, accounting for 16% of the total burden of diseases.[2,20] The largest contributors to this are ischemic heart disease and stroke, which account for most of the burden of CVD (around 80%).[6] Yet, the age-standardized incidence rate of CVD has decreased by 14% since 1990 (from 790 cases per 100,000 in 1990 to around 680 cases per 100,000 population in 2019), and their overall burden has decreased by 31% (from 7,000 DALYs per 100,000 in 1990 to 4,800 per 100,000 in 2019).[2,6,20]

This substantial decline in the burden of CVD is an important public health achievement, which has mostly been attributed to the reduction of important risk factors such as smoking and to the increased availability and implementation of medical therapies.[21] However, recent estimates indicate the progress to reduce the public health impact of CVD has stabilized or even started to show signs of resurgence.[21,22]

The economic burden of CVD on health systems and individuals has increased consistently, mostly related to direct costs of care, but also due to productivity losses.[3] The treatment of CVDs and related risk factors has been estimated to account for around 10% of the total healthcare costs Europe[23] and around 5% in other middle and middle-high income countries.[24,25]

1.3 Risk factors

Type 2 diabetes and CVD share several well-established modifiable and non-modifiable risk factors.[26] Moreover, type 2 diabetes is in itself an important risk factor for CVD.[27] The risk of CVD is more than double among people with type 2 diabetes, and CVD accounts for over 50% of deaths in people with type 2 diabetes.[28,29]

Important non-modifiable risk factors include age, sex, ethnicity, and genetic susceptibility. Modifiable exposures include behavioral and metabolic factors

such as alcohol consumption, smoking, physical inactivity, unhealthy dietary habits, elevated blood glucose levels, overweight or obesity, high-blood pressure, and raised cholesterol levels (Figure 2).[30,31]

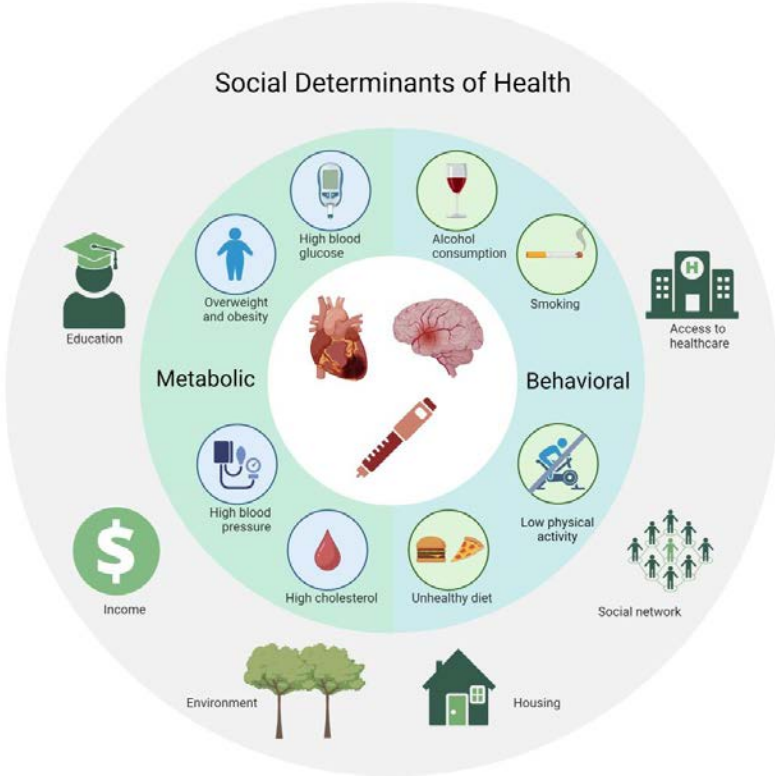

Figure 2. Summary of well-established modifiable risk factors for type 2 diabetes and cardiovascular disease. (Figure created with BioRender.com.)

Furthermore, the social determinants of health are important risk factors for both type 2 diabetes and CVD.[32–34] They are defined as "the condition in which people are born, grow, work, live and age, as well as the forces and systems shaping the conditions of daily life, which include economic policies and systems, social norms, social policies and political systems".[35] In general, behavioral and metabolic risk factors are affected by these broader determinants.[36]

Socioeconomic status (SES) is commonly used to assess the impact of social determinants of health.[35] However, SES is a complex and multidimensional construct that can be measured using different indicators, e.g. individual level factors like education, occupation, income, wealth, or combinations of these, and area level factors such as neighborhood of residence.[37] Although there is some clear overlap, different indicators measure different dimensions of SES. Education

is a measure of exposure during youth and captures knowledge-related assets of an individual, such as health literacy. Occupation, in turn, corresponds to exposure in adulthood, and is more directly associated with income along with physical and psychosocial working conditions.[37]

There is a clear social gradient in health, with individuals with lower SES being disproportionally affected. These health inequalities have been described using a variety of measures of SES such as income, occupation, educational attainment, race, and ethnicity.[32,38,39] The pattern holds between and within nations and through time[40,41], suggesting a causal effect of socioeconomic status on health. Yet, the mechanisms through which social determinants affect health remain elusive.

At an individual level, behavioral and metabolic risk factors combined are responsible for a large proportion of the cases and deaths due to NCDs.[7,42] According to the GBD study, modifiable risk factors accounted for roughly 50% of the global DALYs, and close to 70% of all deaths in 2019. Among the most important were high-blood pressure (19% of deaths and 9% of DALYs), tobacco use (15% of deaths and 8% of DALYs), dietary risks (13% of deaths and 7% of DALYs), high fasting plasma glucose (11% of deaths and 6% of DALYs), and high BMI (9% of deaths and 6% of DALYs). [7]

The trends since 1990 show mixed progress and a large proportion of these risk factors show no significant changes. A few have shown important reductions including tobacco use, while exposure to some important risk factors such as alcohol use, elevated blood glucose and high BMI has increased significantly. [7]

Reducing exposure to behavioral and metabolic risk factors is therefore an important public health goal. In a recent large observational study, exposure to high BMI, elevated blood pressure, raised cholesterol, smoking or to type 2 diabetes explained almost 60% of CVD cases[42]. In addition, studies have reported that reducing exposures to modifiable risk factors significantly lowers the risk of type 2 diabetes and of its complications.[43,44]

2 Prevention of type 2 diabetes and cardiovascular disease

There is a large potential to prevent type 2 diabetes and CVD through lifestyle changes or pharmacological interventions aimed at reducing the effect of behavioral and metabolic risk factors.[42,45,46] Public health recommendations for a healthy lifestyle target the general population, while other interventions are reserved for individuals at high-risk. However, identifying individuals at high risk remains an important challenge.[4,19,47] Furthermore, unlike for CVD, pharmacological interventions for prevention of type 2 diabetes remain controversial.[19,48]

2.1 Risk stratification and prediction models

Several risk prediction models for type 2 diabetes and CVD have been developed based on self-reported information and biomarkers.[49–51] However, risk prediction models have several limitations, mostly due to inaccurate performance when applied to different populations than the one used for their development.[52,53]

In the case of type 2 diabetes, the clinical utility of existing prediction models has been limited.[54] Prediabetes remains the most widely used method of risk stratification for type 2 diabetes. It is defined as an intermediate state in which blood glucose levels are elevated, but not enough to make a diagnosis of type 2 diabetes.[13] It is identified using the same laboratory tests as type 2 diabetes and it can be categorized according to the underlying glucose abnormalities as impaired fasting glucose (IFG), impaired glucose tolerance (IGT) or the coexistence of both. Although there is no consensus on the definition of prediabetes, the most widely used definitions include those from the ADA and the WHO,[55] shown in Table 2.

Table 2. Definition of prediabetes according to the American Diabetes Association (ADA) and the World Health Organization (WHO)

Prediabetes	Test	ADA	WHO
Impaired fasting glucose (IFG)	Fasting plasma glucose	5.6-6.9 mmol/l (100-125 mg/dl)	6.1-6.9 mmol/l (110-125 mg/dl)
	HbA1c	39-47 mmol/mol (5.7-6.4 %)	Not recommended
Impaired glucose tolerance (IGT)	OGTT	7.8-11.0 mmol/l (140-199 mg/dl)	7.8-11.0 mmol/l (140-199 mg/dl)

The utility of prediabetes to predict the risk of type 2 diabetes remains controversial. A recent meta-analysis reported that although individuals with prediabetes had a higher risk of type 2 diabetes, there were important differences depending on the definition used.[56] Furthermore, a systematic review described that a large proportion of individuals with prediabetes (around 40%) reverted to normal blood glucose levels without any specific treatment, even after 10 years.[57] In addition, among older adults, a recent study found that a diagnosis of prediabetes after the age of 60 did not have a clear clinical benefit.[58]

Risk stratification for the prevention of type 2 diabetes remains an important challenge for both public health and clinical practice. Recently, the ADA, the European Association for the Study of Diabetes (EASD) and the WHO have addressed the need for the development and standardization of risk scores for type 2 diabetes, and highlighted challenges regarding clinical variability in the risk and progression of type 2 diabetes.[4,14]

Prediction models for CVD, in contrast, play a central role in the prevention of CVD.[51] Several clinical guidelines have adopted the use of risk prediction models to guide interventions aimed to prevent CVD, such as defining targets for the management of high-blood pressure,[59] blood lipids[60,61] and glucose levels.[62] Two of the most widely used prediction models are the Systematic Coronary Risk Evaluation (SCORE) and the Pooled Cohort Equations (PCE).

The SCORE model was first introduced in 2003 by the European Society of Cardiology (ESC) and was initially developed to estimate the 10-year risk of CVD mortality for individuals aged 45 to 65 years.[63] It has since then been updated to include non-fatal CVD events and older individuals (up to 70 years) (SCORE2).[64] The SCORE2 model provides region specific predictions based on age, smoking status, systolic blood pressure, total- and HDL-cholesterol for individuals without previous type 2 diabetes or CVD. Furthermore, specific models have now been developed for individuals older than 70 years (SCORE2-OP),[65] or with type 2 diabetes (SCORE2-Diabetes).[66]

The PCE is a risk prediction tool first introduced in 2013 as part of the AHA/ACC guidelines for cardiovascular risk assessment.[67] It predicts the 10-year risk of CVD and provides sex and race specific estimates for individuals between 40 and 79 years of age. The PCE model is based on sex, age, systolic blood pressure, antihypertensive treatment, blood total cholesterol and high-density lipoprotein cholesterol concentrations, diabetes, and current smoking. The estimated risk in

both models can be used to stratify individuals into low-risk, moderate-risk, and high-risk groups, as shown in Table 3.

Table 3. Risk stratification based on the SCORE2, SCORE-OP and PCE prediction models.

	SCORE2 (<50 years)	SCORE2 (50-69 years)	SCORE2-OP	PCE
Low-risk	<2.5%	<5%	<7.5%	<7.5%
Moderate-risk	2.5 to <7.5%	5 to <10%	7.5 to <15%	7.5 to <20%
High-risk	≥7.5%	≥10%	≥15%	≥20%

Classification for SCORE2 is based on the moderate risk charts.

Since their publication, several studies have shown that CVD risk prediction models might lead to overestimation or underestimation of the risk among certain populations.[68–73] In a recent study comparing white and black individuals with similar risk-profiles, risk stratification using PCE led to different clinical decisions.[71] Similarly, a large study examining the use of SCORE2 across socioeconomic status and ethnicity, highlighted significant discrepancies.[72] Furthermore, previous studies have pointed out over- or underestimation of CVD risk among individuals with other chronic diseases such as obesity[69], diabetes[74], or HIV.[70]

This lack of generalizability has been a source of concern since inaccurate risk estimates might lead to overuse of medications or other interventions without a real clinical benefit, and thus risk harm. At the same time, inaccurate assessments may lead to missing beneficial interventions among other individuals.[75] This could in turn result in greater socioeconomic inequalities in the prevention and treatment of CVD.[76,77]

Therefore, despite recommendations in clinical guidelines, the clinical utility of risk prediction for CVD prevention remains uncertain.[78] Previous observational studies and RCTs have found that their use results in a moderate but significant reduction in the values of most risk factors, but with modest decreases in the rate of CVD cases and mortality due to CVD.[78,79] Their use also leads to a significant increase in the use of medications such as statins and antihypertensive medications.[78,79]

2.2 Screening

Screening strategies can be divided into opportunistic screening, which involves assessing individuals when they come in contact with healthcare for any reason,

or systematic screening, which is carried out as part of a structured program among the whole population or specific groups.[80-82]

The utility and cost effectiveness of population level screening for type 2 diabetes remains uncertain and studies have found contrasting results.[83] A cluster RCT in Europe found no benefit of population level screening using fasting plasma glucose over all-cause mortality, cardiovascular outcomes, or diabetes related outcomes.[84,85] But a more recent observational study reported a lower risk of CVD events among individuals who had been screened for type 2 diabetes in primary care.[86]

At the same time, epidemiological studies have estimated that around 50% of the cases of type 2 diabetes are undiagnosed.[87] This suggest a clear need for early detection and treatment of type 2 diabetes. The ADA and IDF recommend opportunistic screening which should be carried out in a non-invasive manner by informal assessment of risk factors or by using a validated risk calculator in all asymptomatic adults.[48,88] In addition, the ADA recommend that complementary diagnostic tests (fasting plasma glucose, HbA1c or OGTT) should be assessed at any age for adults with overweight or obesity and at least one additional risk factor, or for all adults from the age of 35, and repeated every 3-years.[48]

For CVD, studies have shown that both opportunistic and systematic screening programs are useful to identify and treat important risk factors, but their clinical utility to reduce the incidence of CVD remains uncertain.[82] A recent report from the WHO Regional Office for Europe concluded that screening for CVD in the general population did not seem to have an overall effect in reducing CVD cases or mortality.[80]

Currently, the ESC recommends systematic screening for individuals with important risk factors for CVD, such as family history of premature CVD (before age 60), smoking, hypertension, type 2 diabetes, hyperlipidemia, obesity, or other comorbidities increasing CVD risk. In addition, for the general population, the ESC recommends either opportunistic or systematic screening for men over 40 years and for women over 50 years.[47] The AHA/ACC, on the other hand, recommends opportunistic screening using the PCE to estimate the 10-year risk of ASCVD for all adults, every five years for people between 20 and 39 years of age and routinely for adults between 40 and 75 years.[19]

2.3 Interventions

Interventions to prevent type 2 diabetes and CVD can generally be divided into lifestyle modification and pharmacological interventions targeted at high-risk individuals, and public health interventions at a population level.

First and foremost, prevention is based on population level policies and changes in environment aiming to reduce exposure to risk factors among as many as possible.[89] Policies to improve physical activity, diet, smoking and to reduce the use of tobacco and air pollution play an important role.[90,91] Additionally, recommendations to maintain a healthy lifestyle should be made available at a population level.[92,93] Since the first reports of the effect of social determinants on health, public health policies have increasingly focused on reducing these inequalities. Unfortunately, gaps in income, and other social determinants are generally increasing around the world.[40,94]

In addition, at individual level, there is strong evidence indicating that a high proportion of cases of type 2 diabetes among people with impaired fasting glucose can be prevented or delayed with intensive lifestyle interventions or pharmacological therapy.[95] Subsequent studies have found that some positive effect of these interventions remain after 10 years,[95,96] and that such interventions seem to be cost-effective.[97,98] However, implementation in primary care settings has had more modest effects.[99,100]

Most lifestyle interventions aim to achieve a weight loss of at least 7% of the initial weight for participants with overweight or obesity, and weight maintenance and increased physical activity for all participants.[99] They are usually resource and time intensive and therefore difficult to generalize to the entire population. Simpler interventions or digital alternatives have shown some promise,[101] but are not yet widely available either.

The use of medications for the prevention of type 2 diabetes remains controversial, and not widely applied in clinical settings.[102] Metformin, a glucose lowering drug, is shown to be effective and is an option for individuals with prediabetes with a BMI ≥35 kg/m2, those <60 years of age, and women with a history of gestational diabetes mellitus, as well as to reduce CVD risk [48]. The recent development of effective medications to treat obesity such as GLP-1 inhibitors can also be effective to reduce the risk of type 2 diabetes and has recently been added to clinical guidelines for the prevention and treatment of type 2 diabetes.[103] However, GLP-1 analogs are not yet widely available for individuals without type 2

diabetes and their long-term effects and utility for preventing type 2 diabetes require further studies.

Bariatric surgery has also been shown to be effective in reducing type 2 diabetes. In a prospective study in Sweden, bariatric surgery reduced the risk of type 2 diabetes by almost 80%, compared to matched controls.[104] Nevertheless, it is a very invasive intervention and the proportion of people at risk for type 2 diabetes that are eligible for surgical therapy is limited.

Individual level interventions to prevent CVD include lifestyle interventions to implement and maintain healthy dietary habits, physical activity, weight loss and weight maintenance, and sleeping habits.[105,106] Measuring cardiovascular health instead of disease has been an area of interest to identify areas of opportunity and to improve public health. The AHA Life essential 8 provides a framework to measure cardiovascular health according to individuals' adherence to public health recommendations for physical activity, nicotine exposure, sleeping habits, BMI, blood lipids, and blood pressure. Studies have shown a low prevalence of optimal cardiovascular health (at least 5 metrics), indicating that there is room for improvement.[106]

Among individuals at high-risk, prevention is centered around pharmacological control of risk factors or comorbidities associated with CVD. This usually involves using antihypertensive medications, lipid-lowering medications such as statins and more recently anti-obesity medications, in combination with improving lifestyle with focus on physical activity.[93]

3 Precision Medicine and Public Health

Precision medicine aims to adapt interventions to individuals based on their unique characteristics[107], while public health and epidemiology focus on studying determinants of health, their distribution, and consequences at the population level.[108] Although these concepts might seem contrasting,[109] the use of precision medicine may have the potential to improve public health practice. To understand this, it is necessary to take a closer look at the evolution and current understanding of the concept of precision medicine.

In a sense, clinicians have always applied precision medicine, by recognizing the importance of adapting clinical recommendations to individuals, mostly based on clinical experience. A more systematic use of the concept of precision medicine began to develop in parallel to advancements in genetics, especially during the

early 2000's.[110] The human genome project generated an increased interest in identifying genetic markers that could be used as therapeutical targets.[110] Despite important achievements, especially for monogenic diseases, cancer and hematology, its utility for complex chronic diseases has been limited.[111] This is mostly due to the multiple pathophysiological mechanisms involved, including multiple genes, as well as external environmental factors and epigenetic mechanisms due to gene-environment interactions.[112]

Although there is not a universal definition of precision medicine, two widely used are those proposed by the National Research Council of the U.S in 2008,[107] and the one by the European Commission in 2015.[113] Both definitions highlight that the focus of precision medicine should be the ability to classify individuals into subgroups or phenotypes with differences in their susceptibility to a disease or response to a particular intervention. Furthermore, both organizations highlight prevention as an important application of precision medicine.[107,113]

Definitions specific to the fields of diabetes and CVD have also been proposed. The ADA and EASD define precision medicine in diabetes as a method to improve diagnosis, prediction, prevention, or treatment of diabetes with the use of multidimensional data of individual differences.[4] In cardiology, precision medicine has been defined as "an integrative approach to cardiovascular disease prevention and treatment that considers individuals' genetics, lifestyles and social and environmental exposures as determinants of cardiovascular risk and disease phenotypes".[5]

Despite this, applications of precision medicine in public health remain largely unexplored.[8] The increasing availability of data and the implementation of more advanced statistical methods for analysis represent an important area of research in precision medicine.[114] Prevision medicine can contribute to public health by identifying homogeneous subpopulations of individuals based on their unique patterns of exposure to relevant risk factors (Figure 3). This could be useful to guide the implementation of different interventions to those who will benefit the most.[9,111]

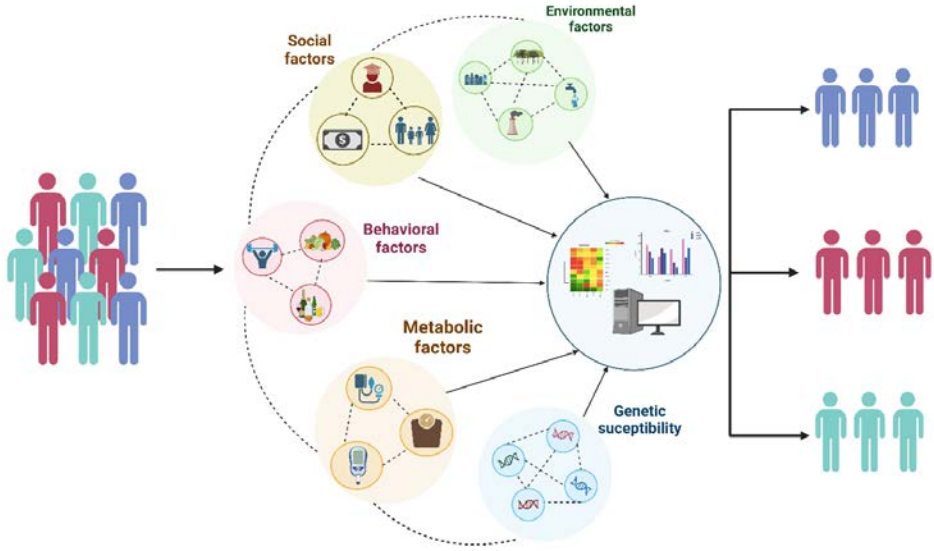

Figure 3. General framework for the use of precision medicine in public health. (Created with BioRender.com.)

3.1 Methods for precision medicine in public health

3.1.1 Life-course epidemiology: Group Based Trajectory Modelling

Exposure to some of the most important behavioral and metabolic risk factors changes with time, and their effect might be related to the accumulation or the timing of exposure.[115] Understanding the different patterns of exposure and their health consequences can have important public health implications, for example in optimizing the timing of different interventions.

Life-course epidemiology is the study of the health effects of exposure to risk factors through different periods of life, taking into consideration the biological, social, and environmental characteristics of different stages of development. A commonly used definition of life-course epidemiology is "the study of long-term effects on later health or disease risk of physical or social exposures during gestation, childhood, adolescence, young adulthood and later adult life".[116] Interest in the life-course approach to epidemiology is closely linked to the epidemiological transition. Namely, as the burden of NCDs increased, researchers began to be interested in weather early life factors could influence later health. An example is the observation that fetal exposures during pregnancy affected an individual's risk of diseases such as type 2 diabetes and CVD much later in life.[117,118]

The mechanisms through which exposure at different periods affects health outcomes are complex. Under the life-course framework, they are summarized as: 1) an accumulative effect model, in which the effect of an exposure increases proportional to the amount of time exposed; 2) a critical/sensitive period model, in which exposure during a certain window of time is associated with a more important effect; and 3) a chain of risk model, which points out that exposure to certain factors, for example low socioeconomic status, usually leads to exposure to other factors, such as smoking or unhealthy diet, increasing the risk of subsequent outcomes. In all these models, the effects of interaction or effect modification between exposure and time or between different exposures can be considered.[116,119]

However, for a given health outcome, several different patterns of exposure can result in an increased risk through different mechanisms[120]. For example, people exposed to overweight or obesity from childhood through to adulthood might develop cardiovascular disease due to the accumulative effect of BMI. Individuals exposed to overweight or obesity later in adulthood, in turn, might be at a sensitive period during which they are more vulnerable to the negative effects of exposure.

Methods such as Group-Based Trajectory Modeling (GBTM) can be used to describe the different patterns of development within a population and are useful tools in life-course epidemiology.[121–123] GBTM, together with latent class analysis, growth mixture models, and other methods, are categorized as finite mixture models (FMM). The general characteristic of FMM is the assumption that an observed distribution is formed by a mixture of unobserved, or latent, subgroups that can be described parametrically (Figure 4).

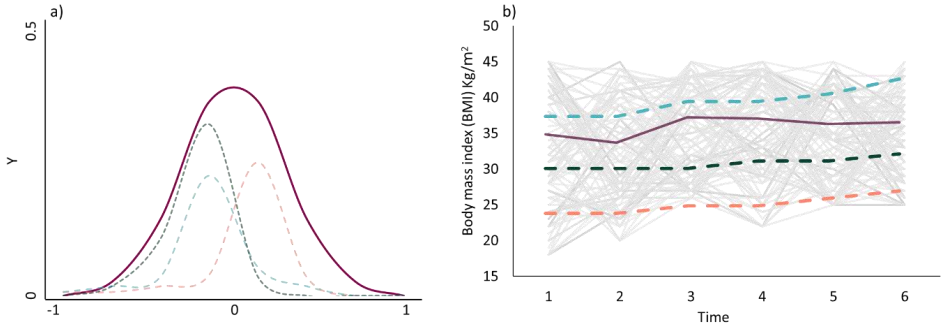

Figure 4. Finite mixture models. Panel a) shows the distribution of the observed population (solid plum line) and the unobserved underlying subpopulations (dashed lines). Panel b) shows the observed individual-level changes (in gray), the population mean of these changes (solid plum line), and the theoretical underlying trajectory groups (dashed colored lines).

GBTM can be used to model longitudinal data and assumes that there is no important variation between individuals in the same class. According to Nagin and collaborators,[122,123] the general formula to estimate GBTM is given by:

$$P(Y_i|\, t_i) = \sum_{j=1}^{J} \pi_j \times P\left(Y_i|t_i, j; \beta_j\right)$$

Y_i is the longitudinal vector of the measurements and t_i is the vector of time when the measurements were performed. π_j is the probability of membership to each of the (J) trajectory groups, and β_j are the parameters of the underlying distribution for this trajectory (for example mean and standard deviation for a normal distribution). To estimate the total probability of each longitudinal trajectory, we first multiply the product of probabilities at each time point, under the additional assumption that the values of Y at each time point are independent from previous observations, conditional on confounders.

$$P\left(Y_i|t_i, j; \beta_j\right) = \prod_{t=i}^{T} p(y_{it}|t_{it}, j; \beta_j)$$

These formulas include two unobserved components, the parameters of each underlying distribution (β_j), and the proportion of individuals in each trajectory group (π_j). The parameters at each time point can be estimated by fitting multiple regression models, where time can be modeled using different polynomial functions.[123] For example, if Y is a count variable, a Poisson regression model could be used.

$$\log(\lambda_{it}^{j}) = \beta_0 + \beta_1 t_{it} + \beta_2 t^2$$

The value of π_j is estimated from the results of a multinomial logistic regression with the different trajectories as outcome and the covariates of interest (x_i) as predictors.

$$\pi_j(x_i) = \frac{\exp(\gamma_j)}{\sum_j \exp(\gamma_j)}$$

Here, γ_j are the coefficients of the intercept from the multinomial logistic regression model. Since these are estimated after the model is fitted, they are called the posterior probabilities. Selecting the final model is typically done by repeating the process described above using different combinations of the number of underlying patterns and polynomial structures of time. The models are then compared based on their utility, interpretability and goodness of fit using criteria such as a minimum size of each trajectory or of the posterior probabilities, the Akaike information criteria (AIC), or the Bayes information criteria (BIC).[120]

Once individuals are assigned to one of the trajectory groups based on the values of the posterior probabilities, further analysis can be done to estimate the associations between the different trajectories and predictors or outcomes of interest.

3.1.2 Health inequalities: Mediation analysis

Understanding the magnitude of different mechanisms or pathways through which socioeconomic differences lead to type 2 diabetes and CVD is important for public health. A common analytical approach to study the contribution of different pathways is mediation analysis. The general aim is to divide the total effect of an exposure (X) on an outcome (Y) in its direct effect (X->Y) and its indirect effect though different mediators (X->M->Y).[124,125]

Using mediation analysis to study complex associations such as the effect of socioeconomic status is challenging. Limitations arise mainly due to the use of somewhat ambiguous exposures (i.e., socioeconomic status), the presence of interactions between the exposure and the mediator, and the importance of multiple mediators related to one another.[126] Below, I review the different methods, their limitations, and their use in social epidemiology.

3.1.2.1 *Conventional methods: The difference and the product of coefficients methods*

The difference method and the product of coefficients method, summarized in Figure 5, are two of the most common approaches to mediation analysis. Both are based in combinations of the coefficients from the following models.[108]

$$1)\ E[Y|x,c] = \alpha_0 + \alpha_1 x + \alpha_2 c$$

$$2)\ E[Y|x,m,c] = \beta_0 + \beta_1 x + \beta_2 m + \beta_3 c$$

$$3)\ E[M|x,c] = \gamma_0 + \gamma_1 x + \gamma_2 c$$

The difference method is a straightforward approach based on coefficients from regressions 1 and 2 above, i.e., a multiple regression for the outcome of interest including confounders as covariates, and a similar model adding the mediator to the covariates list.[108,124] The coefficient $\alpha_1 x$ is the total effect, the coefficient $\beta_1 x$ the direct effect, and the indirect effect is the difference between $\alpha_1 x - \beta_1 x$.

The product of coefficient method is based on combination of coefficients from equations 2 and 3, i.e., a full model for the outcome of interest including the exposure, mediator and confounders as covariates, and a model using the mediator as outcome, and adjusting for the exposure and confounders.[108,124] The direct effect is the coefficient $\beta_1 x$, and the indirect effect is the product of coefficients $\gamma_1 x$ and $\beta_2 m$.

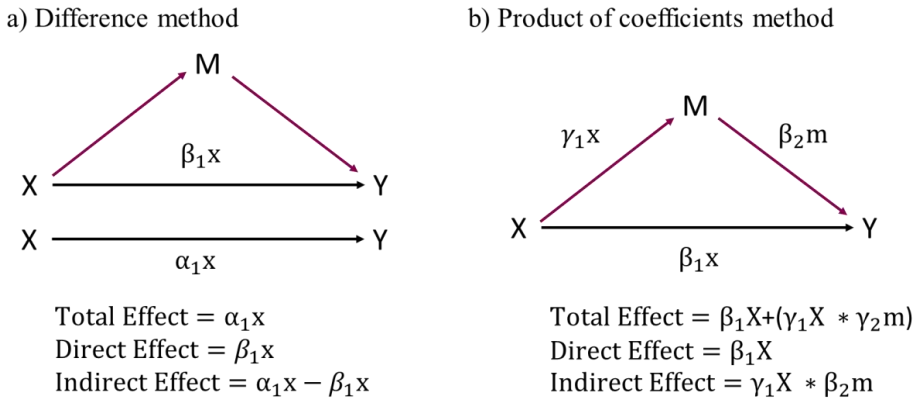

Figure 5. The difference and product of coefficient methods for mediation analysis.

3.1.2.2 *Counterfactual mediation analysis*

The counterfactual (or potential outcome) framework is a theory of causality based on hypothetical interventions that has gained popularity as an approach to

estimate causal effects from observational data. A causal effect is defined as the difference between the expected outcome had the individual been exposed and the expected outcome had the same individual not been not exposed.[127] However, only one of these scenarios is observed, and the opposite represents a hypothetical (counterfactual) state. Therefore this approach is based on the premise that, although the individual level counterfactuals are impossible to measure, it is possible to derive population level estimates under a specific set of assumptions, mentioned below.[128]

1) Consistency: It is assumed that there is only one well-defined version of the treatment or intervention, which always leads to the same potential outcome. The consistency assumption is sometimes included as part of the broader stable unit treatment values assumption (SUTVA), that further adds that there should be no interference between study units., i.e., an individual's counterfactual outcome is not affected by any other individual's exposure.

2) Conditional exchangeability: The probability of receiving the treatment of interest does not depend on any other confounders than the ones measured and included in the model.

3) Positivity: the probability of receiving every possible value of the treatment is larger than zero.

In mediation analysis, counterfactuals of the associations between exposure, mediator, and outcome can be used to describe different causal pathways. By defining hypothetical interventions for the exposure or mediator, we can estimate different measures of the indirect and direct effects. One approach is to decompose the total effect (TE) into the natural direct effect (NDE) and the natural indirect effect (NIE) by setting the mediator to its observed (natural) values among the exposed or unexposed. Using a simple example with a binary exposure (X), mediator (M), and outcome (Y), we can express this as follows[129]:

$$TE = NDE + NIE = E[Y_{x=1} - Y_{x=0}] = E\left[Y_{x=1,Mx=0} - Y_{x=0,Mx=0}\right] + E\left[Y_{x=1,Mx=1} - Y_{x=1,Mx=0}\right]$$

Where the NDE of X on Y is the difference between the estimated probability of the outcome if everyone was exposed and the value of the outcome given no exposure while holding the mediator to its observed value among unexposed $\left[Y_{x=1,Mx=0} - Y_{x=0,Mx=0}\right]$. The NIE, i.e., the effect of X on Y only though M, is the difference of the counterfactual outcomes if the mediator had been set to its

observed values among exposed or unexposed, and everyone had been exposed $\left[Y_{x=1,Mx=1} - Y_{x=1,Mx=0}\right]$.

Another commonly used definition is the controlled direct effect (CDE), which is defined as the difference between counterfactual outcomes had the exposure been present and not present, setting the value of the mediator constant at a preselected value $\left[Y_{x=1,M=m} - Y_{x=1,M=m}\right]$. Unlike the NDE, estimating the CDE requires two hypothetical interventions, one in the exposure and one in the mediator. A limitation of the CDE is that there is no analogous indirect effect, and the total effect cannot be easily decomposed. However, the CDE can be estimated with less assumptions than the NDE, and can provide useful information. What method to use depends largely on the research question.[125]

There has been a longstanding debate regarding the utility of counterfactual mediation analysis in social epidemiology.[130–132] The interpretation of causal effects from counterfactual models in social epidemiology remains a subject of controversy due to violations to the consistency assumption. These arise because common exposures used to measure SES, such as education, income, or occupation can be considered ambiguous or ill-defined interventions. That is, different versions of the exposure may exist (e.g., individuals could have a low income due to disability, disease, low education, unemployment, etc.).[133] However, there is a clear need to quantify health disparities and understand the causal pathways linking them to diverse health outcomes, for which the counterfactual framework could be useful.

Some authors argue that the violations to the theoretical assumptions are so severe that other methods should be sought, or that social epidemiologists should at least strive for more precise definitions of hypothetical interventions and prioritize the study of manipulable exposures.[130,132] Other authors, in contrast, argue that the consequences of violating the consistency assumption might be exaggerated, hindering progress in public health and health disparities research.[134,135] And that focusing on manipulable exposures would shift the focus towards determinants closer to the individual, and away from the broader social determinants of health.[134,136]

3.1.2.3 Comparing conventional and counterfactual mediation

In general, all methods for mediation analysis rely on the same basic assumptions and therefore share similar limitations.[108] However, counterfactual mediation analysis can be used to overcome some of these constrains.

The basic assumptions of all mediation analysis are i.) there are no unmeasured confounders in any of the associations (between exposure and outcome, exposure and mediator or mediator and outcome), ii.) there is no interaction between exposure and mediator, and iii.) none of the measured confounders of the mediator-outcome association are affected by the exposure (exposure induced confounders).[108]

The assumption of no unmeasured confounding is common for all observational methods. However, in mediation analysis, this assumption also includes confounders between mediator and outcome. As a result, mediation analysis tends to be more prone to this kind of bias. Consequently, mediation analysis often requires larger sample sizes to achieve an adequate power.[137] Methods for sensitivity analysis are available to estimate how strong an unobserved confounder must be to affect the findings of counterfactual mediation analysis.[138]

Interaction between the exposure and mediator is problematic because, if present, the total effect cannot longer be assumed to be equal to the addition of the direct and indirect effects. In fact, if the effect of the exposure on the outcome varies at different levels of the mediator, this later variable would technically be a moderator or effect modifier.[139] The counterfactual framework allows for more flexibility in modelling exposure-mediator interactions. For instance, in the three-way decomposition presented above, the effect of the exposure-mediator interaction is assigned to the NIE, but this could be easily adapted to instead be a part of the NDE.[129] A four-way decomposition has also been described, which divides the effect of the exposure-mediator interaction between the direct and indirect effect. Additionally, the CDE can be estimated at different values of the mediator which can be determined based on their utility to answer different research questions.[125]

Exposure induced confounders are a challenge in both conventional and counterfactual mediation, mostly related to the study of multiple mediators. The complexity arises because such variables would simultaneously be a second mediator, as well as a confounder of the mediator outcome association (Figure 6). Exposure induced confounders are frequent in the case of NCDs due to the presence of multiple related mediators and the long time between the occurrence of the exposure, mediator, and outcome.[125] A common approach to study multiple mediators using conventional methods is to add the estimates of the independent

effect of each mediator, ignoring the interactions between them. This approach can lead to biased estimates by duplicating the effect of common pathways.[125,140]

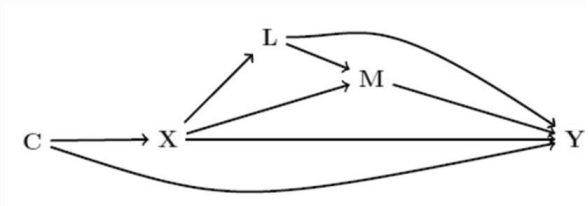

Figure 6. DAG of exposure induced confounders in mediation analysis. An exposure induced confounder (L) is a mediator between the exposure (X) and outcome (Y), and a confounder of the mediator-(M) outcome association. C represents confounders between X and Y.

Counterfactual mediation can be used to bypass this situation in different ways. First, if the main interest is the study of one mediator, the CDE can again be used to obtain valid estimates, even in the presence of exposure induced confounding.[129] However, if there is interest in studying several mediators, a set of mediators can be considered jointly to estimate an aggregated natural indirect effect.[141] Other approaches to study multiple mediators such as estimating pathway specific estimates or using interventional effects are also available.[125,141]

3.1.3 Risk profiles based on risk factors: Cluster analysis

Prediction models are usually developed by fitting multivariable regression models for the outcome of interest including a set of risk factors as predictors. From the modeled association, an equation can be derived and applied to external individuals.[76,142] The models therefore assume that the magnitude of the effect between the included variables and the outcome will remain constant across populations and levels of exposure to other variables. This is unlikely to be the case in the development of NCDs,[5,75,114] and might explain, to some extent, the observed lack of generalizability of prediction models.

To deal with these discrepancies, a commonly proposed solution has been to recalibrate the models to different populations,[142] which limits the direct implementation of the prediction models. More recently, there has been increased interest for the use of machine learning to improve risk prediction models.[143] However, the most common approach has been the use of supervised models. They do allow for more flexibility in the modeling of interactions between different risk factors but are also based on the association of risk factors with CVD to make

predictions, and therefore susceptible to the same limitation of assuming a constant effect of risk factors for all individuals.

Other methods, such as cluster analysis, have been suggested as an alternative [144]. The main advantage of cluster analysis is its ability to group individuals in homogeneous subgroups based on the patterns of risk factors present in the data, independently of the prevalence of these factors or their association with the outcome of interest. The assumption is, instead, that the risk factors included are important for the occurrence of the outcome, but the magnitude of their effect is allowed to vary between populations.[145,146] Therefore, the use of cluster analysis could help overcome limitations of current models.

Cluster analysis is a group of unsupervised machine learning methods used for grouping or segmenting data into subpopulations. In machine learning, methods can be divided into supervised or unsupervised learning based on the data input they require. Supervised machine learning needs to be trained to make predictions, while unsupervised methods aim to detect inherent patterns in the data without previous input or training.[145] The goals of cluster analysis include ascertaining whether the data consists of different sub-populations, and subsequently to group the objects based on a set of measurements for each object or of the relationship between objects.[147]

Methods for cluster analysis can be further divided into hierarchical and partitioning algorithms. Hierarchical methods use some measure of similarity between observations to group them together, while partitioning algorithms aim to find the most homogeneous way of dividing observations into a predefined number of subgroups (k).

Partitioning methods are based on a reiterative process following the general steps listed below:

1. A central point, or centroid, is assigned at random given the predefined number of clusters.
2. The mean distance (similarity) is calculated for each observation to each of the centroids.
3. Individual observations are grouped according to the minimum distance measurement.
4. These steps are repeated until assignment to the different clusters does not change any further.

A limitation of cluster analysis is that most methods allow for only continuous or categorical data.[145] K-prototype is a method that offers a solution by combining the dissimilarity measures used in k-means, a clustering algorithm useful for continuous data, and k-modes, used for categorical data.[148] See Figure 7 for an example.

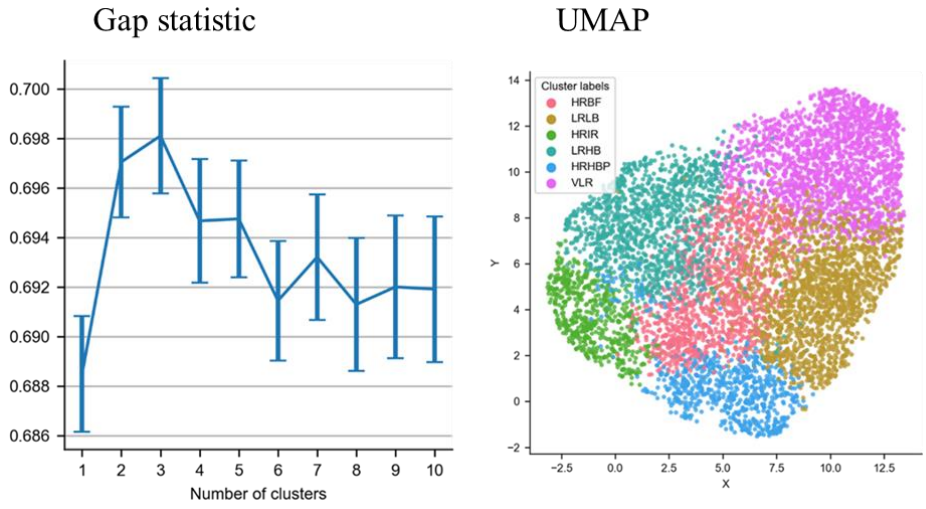

Figure 7. Example of the use of the Gap statistic to determine the number of clusters and of the results of cluster analysis using k-prototypes vizualized using UMAP for dimension reduction.

As mentioned previously, partitioning algorithms require the number of clusters (k) to be predefined, and it is therefore important how this selection is done. Although several methods exist to determine the optimal number of clusters, this decision is still somewhat subjective. The Gap statistic, used in studies III and IV, uses the within cluster variability compared to a reference distribution for which we know there are no clusters (Figure 7). The point at which the difference (or gap) between the observed and expected variability is maximized gives the best solution for the number of clusters in a particular data set.[147]

4 Research aims

The overall aim of this thesis was to provide a framework for the use of precision medicine in public health and epidemiology, and to explore the utility of different analytical methods to classify individuals into sub-populations that can be useful for risk-stratification, or to study the magnitude of the effects of different mechanisms susceptible to intervention.

Specific aims:

Study I: to characterize life-course trajectories of weight categories and estimate their impact on the incidence of type 2 diabetes.

Study II: to estimate the extent to which the association between low socioeconomic status and the incidence of type 2 diabetes is mediated through common metabolic and behavioral risk factors. Additionally, a secondary aim was to compare counterfactual and conventional methods for mediation analysis.

Study III: to determine whether cluster analysis could be used to identify homogeneous subgroups of diabetes-free adults based on the heterogeneity of known risk factors for type 2 diabetes, and asses their clinical utility to stratify the risk of type 2 diabetes, compared to that of prediabetes.

Study IV: to explore the use of risk stratification using cluster analysis for primary prevention of CVD and compare its performance with that of the SCORE2 and PCE models.

5 Materials and methods

The main data source for the four studies constituting this thesis work was the Stockholm Diabetes Prevention Program (SDPP), a longitudinal cohort study of Swedish individuals aiming to study the incidence of type 2 diabetes.[149] Study III additionally uses data from a similar cohort from Mexico, The Metabolic Syndrome Cohort,[150] for external validation.

5.1 The Stockholm Diabetes Prevention Program (SDPP)

The SDPP cohort was started in during the early 90's with the aim of studying risk factors for type 2 diabetes and to evaluate population-based interventions to prevent type 2 diabetes. It has since then been complemented with data linkage to regional and national healthcare and population registries, and has been divided into three sub-cohorts according to the available information: 1) a full cohort, comprising registry data of all individuals initially contacted, as well as their immediate family; 2) the survey responders, who returned the initial screening questionnaire; and 3) a clinical cohort, comprised of individuals who attended more in depth study visits. The studies in this thesis are based on the clinical cohort, described in more detail below.

Adults who were born in Sweden between 1938 and 1957 and were registered in one of the five participating municipalities of Stockholm County (Sigtuna, Tyresö, Upplands-Bro, Värmdö and Upplands Väsby) were identified using the total population registry of Sweden and invited to participate via post. In addition to providing informed consent, individuals were asked to fill and return an initial screening questionnaire. Next, to create the clinical cohort, individuals born in Sweden, without self-reported type 2 diabetes, with either strong or complete absence of family history of diabetes, and all women with history of gestational diabetes were invited to participate in the clinical examinations. A random sample of individuals was later excluded due to administrative reasons and one participant has asked to be removed from the study.

Of the 11,568 participants initially invited, data are available for 7,948 (69%). All participants who attended the baseline examination were invited to participate in a 10-year (n=5,734) and 20-year (n=3,627) follow-up. The study visits included extensive health questionnaires, anthropometric measurements, blood pressure measurement, fasting blood sample collection, and a standard oral glucose tolerance test (OGTTT) for individuals who did not report a new diagnosis of type

2 diabetes. All measurements and tests were carried out by trained personnel. Blood samples were taken during fasting and 2-hours after an oral challenge with 75 gr of a glucose solution. Glucose and insulin were measured immediately, and the blood samples were stored for future analysis. After the 20-year follow-up, blood samples of these participants were retrieved and extensive analysis including cholesterol, triglycerides, liver function tests, kidney function test and inflammatory markers were performed. In addition, data has been linked to regional and national registries using individuals' personal identity number.

5.2 The Swedish Regional and National Registries

Initially, data was available from the Regional Data Warehouse of Stockholm (VAL registry).[151] This is an administrative database including all inpatient and outpatient care, both private and public, financed by the Region of Stockholm. The VAL registry was used to gather data on date of diagnosis and comorbidities in all studies.

Later, data from National Registries was included from different sources. From Statistics Sweden,[152] the data used came from the Total Population Register, which contains data related to birth, date, citizenship, and migration both in Sweden and abroad; the Population and Housing Census, which were performed in Sweden every 5 years between 1960 and 1990 and has since been replace by electronic registries; and the Swedish Longitudinal Integrated Database for Health Insurance and Labor Market Studies (LISA), which compiles labor market data from several sources and includes information of individuals occupation, income, and level of education among other relevant socioeconomic variables, covering all individuals over 16 years from 1990.

From the National Board of Health and Welfare,[153] the data used included the Prescribed Drug Registry of Sweden, with information of all prescriptions written and dispensed in Swedish pharmacies from 2005; and the National Patient Registry, with information from all inpatients in public hospitals since 1987 and outpatient visits at both private and public care givers since 2001. Finally, data from the National Diabetes Registry,[154] managed by Region Västra Götaland, were used. It recompiles information from individuals with a diagnosis of type 2 diabetes since 1996.

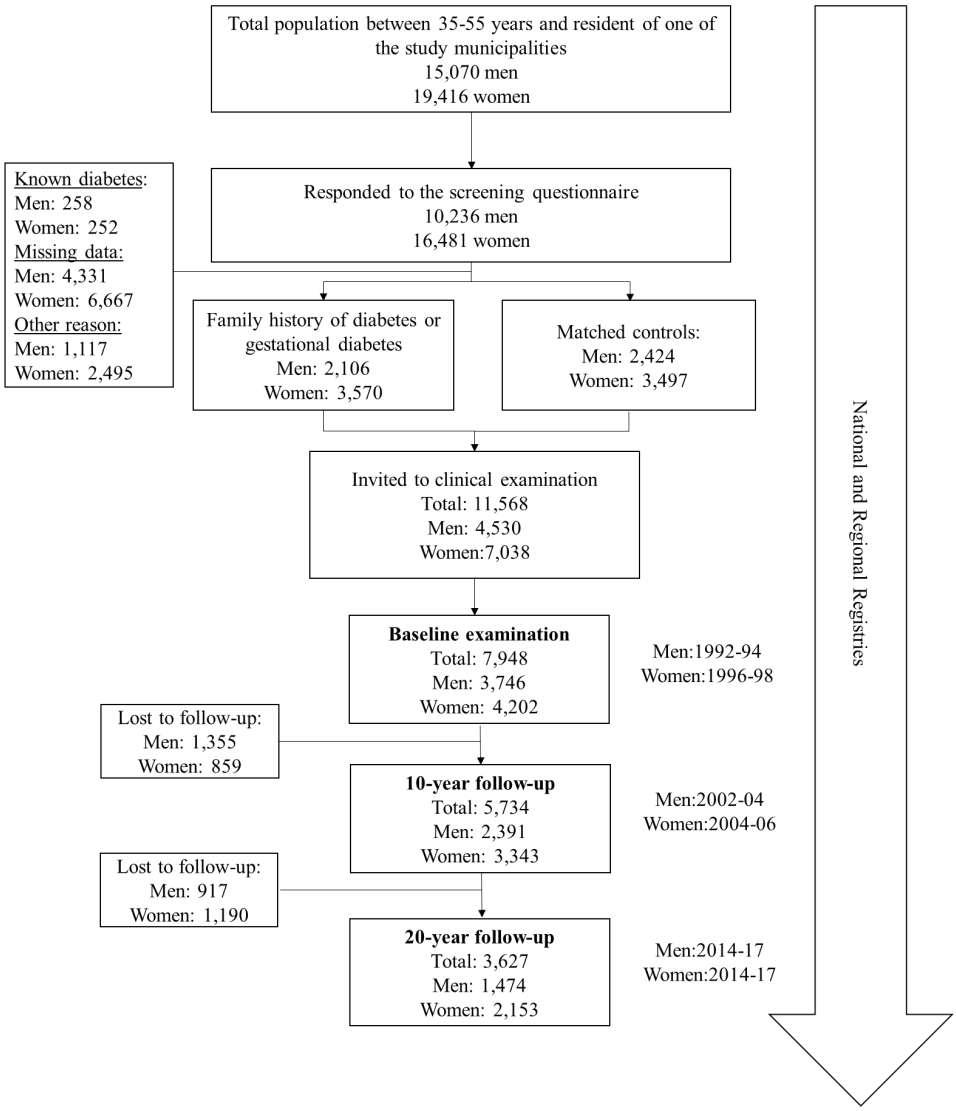

Figure 8. Flow chart of the Stockholm Diabetes Prevention Program study.

5.3 The Metabolic Syndrome Cohort

The MSC recruited healthy adults from 5 cities in central Mexico, with the aim to evaluate the incidence of type 2 diabetes, hypertension, and CVD, as well as to identify associated risk factors.[150] All participants were patients listed in one of the participating study centers. They were invited to participate if they were over 20 years of age, had a BMI ≥20 kg/m2, had resided for more than 6 months in the city where examination was to take place without intentions to move in the short term, were born in Mexico and had parents and grandparents that were also born in

Mexico. The exclusion criteria included previously diagnosed diabetes or cardiovascular disease, those unable to leave their home due to ill health or immobility, pregnancy, and self-report of high alcohol consumption.

An initial screening was conducted by study personal at the participants home, workplace or during their visit to a medical unit. After initial screening, 9,634 eligible participants accepted to participate and underwent a baseline examination including medical history, clinical examination, anthropometric measurements, standardized questionnaires, and fasting blood sample collection. After three years, participants were contacted via phone, e-mail, through friends or relatives, or via visits to their workplace and invited to undergo a follow-up evaluation. A total of 6,119 individuals attended the follow-up evaluation (participation rate 80.7%).

5.4 Study I: Life-course trajectories of weight and their impact on the incidence of type 2 diabetes

Data for this study came from the SDPP cohort and the VAL registry. The study sample included 7,203 individuals for whom weight or body size data at five time points were available.

The main exposure in this study were the life-course trajectories of body weight that were estimated from the five weight categories using Group Based Trajectory modeling, separately for women and men. Self-reported body size at age 7 and 18 were obtained from participants' answers to the questions "Compared to others of the same age, what was your weight status at age 7 and 18", which was available from the baseline questionnaire for women and from the 10-year follow-up questionnaire for men. The response options were: 1) very lean, 2) somewhat lean, 3) normal weight, 4) somewhat overweight, and 5) very overweight. We categorized answers 1 and 2 as lean, 3 as normal weight and 4 and 5 as overweight. In the baseline questionnaire, participants were asked to recall their weight 5 and 10 years before the study visit, when they were on average 42 and 37 years. The final time point was BMI measured during the baseline study visit when participants were 47 years on average.

The outcome was a new diagnosis of type 2 diabetes, which was recorded though OGTTs during any of the clinical examinations, self-reported in study questionnaires, or obtained from the VAL registry.

The baseline characteristics of the sample in each of the trajectory groups were described using mean and standard deviation for continuous values and proportions for categorical values. Modified Poisson regression was used to estimate the association between the different trajectories and incidence of type 2 diabetes. Then, population attributable fractions (PAFs) were calculated to assess the burden of type 2 diabetes attributable to the different trajectory groups. Covariates adjusted for included age at baseline, physical activity, family history of diabetes, comorbidities, general health, educational level, smoking and alcohol use, and history of gestational diabetes for women.

5.5 Study II: Behavioral and metabolic mediators of socioeconomic inequalities in type 2 diabetes

Data from SDPP was used in combination with data from LISA, the 1990 census, and the National Diabetes Registry. From the 7,948 total participants at baseline, we excluded 128 due to a diagnosis of type 2 diabetes during the baseline OGTT. A further 17 were excluded due to previous diagnosis of type 1 or LADA diabetes, 28 due to missing data on education of occupation, and 234 due to missing values in any of the covariates used in the analysis. The final sample for this study included data from 7,123 participants, of which 4,383 (62%) women and 2,740 (38%) men.

Educational attainment and occupational class were used as measures of socioeconomic status. Educational attainment was defined as the highest completed level of education: basic education, vocational education or upper secondary school, and university or higher. Low socioeconomic status was categorized as a binary variable comparing basic education to the rest. Occupational class was categorized according to the Swedish Socioeconomic Classification System into unskilled workers, semiskilled workers, skilled workers, non-manual employees, employed professionals, self-employed professionals, and higher civil servants.[155] Low occupational class was defined as a binary variable including unskilled and semiskilled manual workers.

The mediators of interest in this study were measured during the baseline examination and included well-established behavioral and metabolic risk factors. The behavioral risk factors were smoking, alcohol consumption, low physical activity, and diet low in vegetables. Metabolic risk factors included BMI, high fasting plasma glucose, and hypertension. Possible confounders included in the

analysis were age at baseline, sex, family history of type 2 diabetes, self-reported comorbidities, and self-reported health status.

Incident cases of type 2 diabetes were recorded from OGTT during any of the study visits, from the VAL registry, the National Diabetes Registry, the National Patient Registry, or based on self-report from the study questionnaires.

Baseline characteristics of the study sample were described according to individuals' socioeconomic status using mean and standard deviation for continuous values and proportion for categorical values. Poisson regression was used to estimate the association between measures of low SES and type 2 diabetes. For the main analysis, counterfactual mediation analysis, based on the DAG in Figure 9, was used to estimate the total effect through any of these pathways, the natural direct effect of low SES in the risk of type 2 diabetes, the natural indirect effect through all the mediators taken together, and the proportion mediated, defined as NIE/TE. The results were compared to those from mediation analysis using the difference and product of coefficient methods.

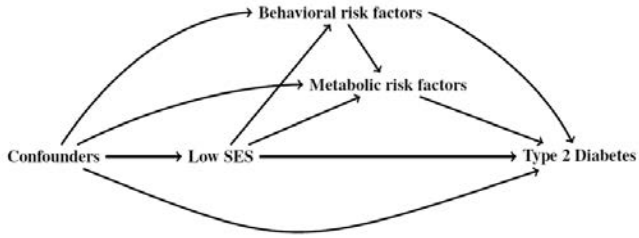

Figure 9. DAG of the association between low socioeconomic status (SES) and type 2 diabetes.

5.6 Study III: Characterization of data-driven clusters in diabetes-free adults and their utility for risk stratification of type 2 diabetes

The study sample included 7,317 individuals from SDPP, and 2,331 participants from the MSC study. In the SDPP dataset, from the initial 7,948 participants, 516 (6.5%) were removed due to missing information or extreme outliers in the variables used for cluster analysis and 115 (1.4%) were excluded due to a diagnosis of type 2 diabetes during the baseline examination, which resulted in a main analytical sample of 7,317 individuals. In a secondary analysis, data from 3,987 (50%) individuals who participated in all study follow-ups were used to test the stability of the clusters.

From the initial 9,637 participants in the MSC study, 1,839 (19.1%) younger than 30 years or older than 60 years were removed to ensure comparability with SDPP and to reduce the risk of type 1 and other forms of diabetes. Further 3,966 (40.5%) were removed due to missing information on a variable required for cluster analysis and 1,500 (15.6%) were lost to follow-up. The analytic sample for the MSC study thus included 2,332 (24%) participants.

The main exposures were the clusters of risk factors in each cohort. These were derived using k-prototype cluster analysis based on participants' fasting plasma glucose, insulin, estimations of insulin resistance and β-cell function based on the homeostatic model assessment (HOMA2-IR and HOMA2-B), BMI, systolic and diastolic blood pressure, sex, family history of type 2 diabetes, and educational attainment. Prior to analysis, continuous variables were standardized to a mean of zero and standard deviation of 1 to allow comparability. Cluster analysis was done independently in both cohorts and the gap statistic was used to determine the number of clusters in each dataset.

Incidence of type 2 diabetes was assessed in SDPP using the OGTT at each follow-up, retrieved from the VAL or the National Diabetes registries, or self-reported during any of the study visits. In the MSC, incidence of type 2 diabetes was determined from fasting plasma glucose during the follow-up examination, by self-report of a new diagnosis of type 2 diabetes, or self-report of starting a new treatment with glucose lowering drugs.

Baseline characteristics of the study sample among the derived clusters were explored using mean and standard deviation for continuous values and proportions for categorical values. Based on the crude incidence in each cluster, we further divided them into high-risk clusters and low-risk clusters.

The association between clusters and incidence of type 2 diabetes was assessed using Cox-proportional hazard models, using age as the underlying time variable. SDPP participants were followed-up from the date of baseline examination until the date of a new type 2 diabetes diagnosis, date of death or until March 31st, 2021. MSC participants were followed from the date of baseline examination until the self-reported date of type 2 diabetes diagnosis, start of glucose lowering therapy, or until February 28th, 2014. The survival models were adjusted for self-reported general health status, presence of chronic comorbidities, physical activity level, smoking status, and history of gestational diabetes among women.

The clinical utility of the clusters was assessed in the SDPP by studying the predictive accuracy and the long-term stability of the clusters. The results were compared to the ADA and WHO definitions of prediabetes. We estimated the sensitivity, specificity, and the concordance statistic as measures of predictive accuracy and the long-term stability using measures of intra-rater agreement.

5.7 Study IV: Cluster analysis for cardiovascular risk stratification

In this study we used data from the 20-year follow-up of SDPP due to the availability of laboratory data. Of the 3.627 participants who attended the 20-year study visit, 158 (4.4%) were removed due to a previous diagnosis of CVD, and 124 (3.4%) due to missing data or extreme outliers in the variables required for cluster analysis. The study population was formed by 3,345 participants with a mean age of 66 years and no previous history of CVD.

The exposures of interest were clusters of risk factors for CVD and risk categorization according to the SCORE and PCE models. Cluster analysis was estimated using K-prototypes using age, sex, smoking, educational attainment, use of antihypertensive medication, glucose lowering medication or statins, previous diagnosis of type 2 diabetes, fasting plasma glucose, systolic blood pressure, BMI, total cholesterol, HDL cholesterol, and estimated glomerular filtration rate (eGFR). Continuous variables were standardized to a mean of 0 and standard deviation of 1. The number of clusters was determined using the Gap statistic, and stability was assessed by estimating the Jaccard similarity index.

Risk categorization using the SCORE2, and SCORE2-OP models or the PCE model was done following the published protocols.[64,65,67] For the SCORE2 model, participants with prevalent type 2 diabetes were considered in the high-risk category.

The main outcome, incidence of CVD, was defined as a composite of fatal or non-fatal myocardial infarction or stroke and CVD-mortality, including cause specific mortality due to hypertensive disease, ischemic heart disease, arrhythmias, cerebrovascular disease, atherosclerosis, or sudden death. Diagnosis status, date of first recorded diagnosis, cause of death as well as date of death were ascertained from the VAL registry, the National Patient Registry, or the Total Population Registry.

The association between risk categories in each model and risk of CVD were studied using Poisson and Cox proportional hazards models to estimate relative

and absolute measures of association. Participants were followed from the date of their participation in the 20-year follow-up of SDPP until the date of first CVD event, date of death, or until administrative censoring on March 31st, 2021.

The performance of the risk stratification models was assessed estimating the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) comparing the high-risk categories of each model to the rest. The concordance statistic was used as a global measure of model discrimination. The Wald test was used to test the differences between the concordance statistic of the high-risk group in the clusters and the SCORE2 and PCE models.

5.8 Ethical considerations

All the studies in this thesis were conducted in accordance with the ethical principles for medical research involving humans stated in the declaration of Helsinki.[156] The SDPP was approved by the ethics review board of Region Stockholm, and the MSC was approved by the Intituto Nacional de Ciencias Medicas y Nutricion in Mexico City.

Overall, the benefits for society and knowledge generation overcome the risks to participants. As no intervention was administered in either study, the overall risk of participation was minimal. On the other hand, participants who received an early diagnosis of type 2 diabetes or CVD during their participation might have personally benefited. However, there are important ethical considerations regarding the use of personal and sensitive information and regarding the exclusion of individuals not born in the countries the studies were conducted.

All participants signed an informed consent when the data was collected, but their right to privacy might be threatened if the data are handled improperly. Furthermore, in the case of SDPP, data have been used after over 20 years of initial recruitment, and additional data has been gathered from linkage to regional and national registries. Extensions to SDPP have been submitted for additional review and approved by the ethics review board. All data used for research has been pseudonymised, and all measures were taken to ensure the security of personal and sensitive information according to local regulations.

Including only individuals born in Sweden (SDPP) or Mexico (MSC) was done for practical or financial issues when conducting the studies. However, this means that the results of these studies will not necessarily be applicable to the population excluded. Given that around 30% of the population of Stockholm and

around 10% of that of Mexico City were born outside the country, it is important to consider this when interpreting the findings of these studies or making public health decisions.

6 Results

6.1 Study I: Life-course trajectories of weight and their impact on the incidence of type 2 diabetes

The individual level life-course transitions of weight categories derived from the self-reported weight categories at ages 7 and 18, self-reported weight 10 and 5 years before the baseline examination, and measured weight during the study are summarized in Figure 10.

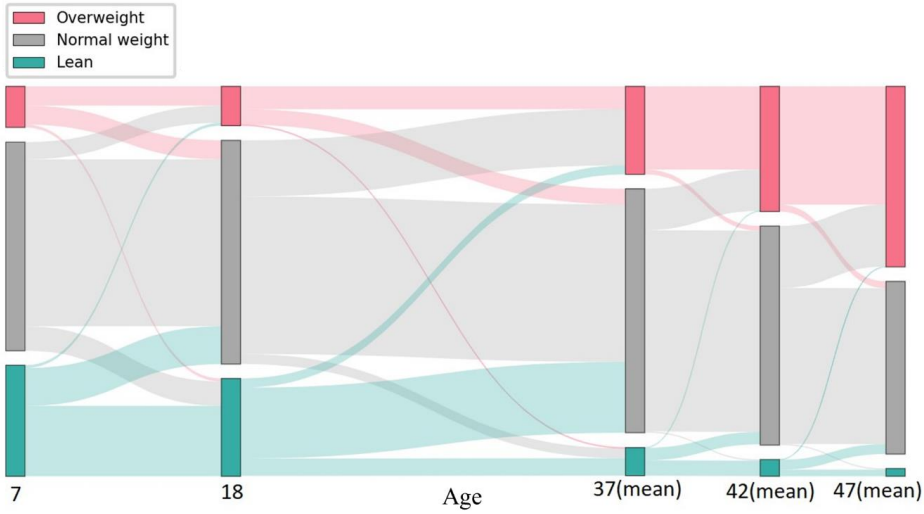

Figure 10. Individual level transitions of weight categories through life in the SDPP cohort.

The GBTM resulted in five distinct life-course trajectories of weight categories. They were descriptively named: 1) stable normal weight trajectory, 2) stable overweight trajectory, 3) lean increasing weight trajectory, 4) overweight from early adulthood trajectory and 5) overweight from late adulthood trajectory. The different trajectories are summarized in Figure 11.

A total of 981 new cases of type 2 diabetes were identified during the study period. The cumulative incidence was 13.74% in the total study population and was higher among men (18.7%) than women (11.3%).

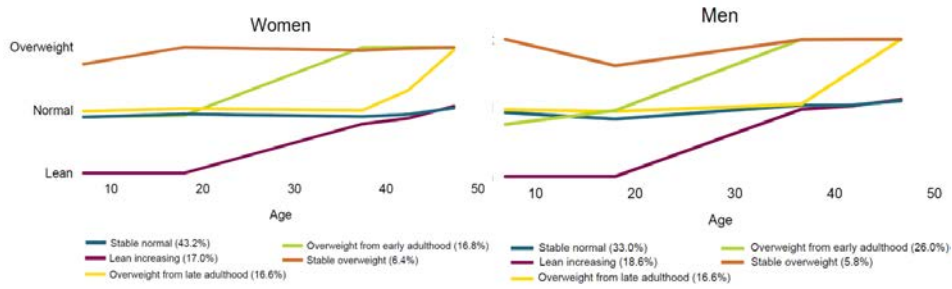

Figure 11. Results of the Group Based Trajectory Models by sex.

Results from the Poisson regressions indicated that the trajectory groups were associated with differences in the risk of type 2 diabetes in adulthood. Compared to the stable normal weight trajectory, both relative and absolute risk of type 2 diabetes was higher for all the other developmental trajectories, although not significantly so for the lean-increasing weight trajectory among men. The highest risk estimates were seen among the overweight from early adulthood and the stable overweight trajectories. That is, for weight gain during early adulthood and constant high weight since childhood (Table 4).

Table 4. Association between trajectory groups and type 2 diabetes in SDPP.

	Women		**Men**	
Risk Ratios (RR)	**Unadjusted RR (95%CI)**	**Adjusted RR (95%CI)**	**Unadjusted RR (95%CI)**	**Adjusted RR (95%CI)**
Stable normal weight	1 (ref)	1 (ref)	1 (ref)	1 (ref)
Stable overweight	3.69 (2.76-4.95)	2.77 (2.06-3.73)	2.94 (2.12-4.08)	2.68 (1.92-3.75)
Lean increasing weight	1.98 (1.50-2.60)	1.71 (1.31-2.24)	1.32 (0.96-1.81)	1.35 (0.98-1.85)
Overweight from early adulthood	4.07 (3.25-5.11)	3.43 (2.72-3.34)	3.03 (2.40-3.84)	2.77 (2.17-3.54)
Overweight from late adulthood	2.78 (2.13-3.63)	2.27 (1.75-2.95)	1.76 (1.29-2.40)	1.73 (1.26-2.37)
Absolute Risk Difference (ARD)	**Unadjusted ARD% (95%CI)**	**Adjusted ARD% (95%CI)**	**Unadjusted ARD% (95%CI)**	**Adjusted ARD% (95%CI)**
Stable normal weight	0 (ref)	0 (ref)	0 (ref)	0 (ref)
Stable overweight	10.90% (6.97%-14.83%)	7.94% (4.65%-11.24%)	15.05% (7.89%-22.21%)	13.11% (6.24%-19.99%)
Lean increasing weight	3.35% (1.34%-5.27%)	2.46% (0.57%-4.36%)	3.92% (0.05%-7.36%)	4.61% (0.95%-8.28%)
Overweight from early adulthood	9.03% (6.63%-11.40%)	7.76% (5.36%-10.17%)	15.23% (11.52%-18.95%)	14.05% (10.30%-17.79%)
Overweight from late adulthood	5.02% (2.66%-7.38%)	3.512% (1.36%-5.68%)	4.84% (0.85%-8.82%)	5.23% (1.03%-9.43%)

The population attributable fraction suggested that, among women, most of the cases of type 2 diabetes in the study population were attributable to exposure to the overweight from early adulthood trajectory group, followed by the overweight from late adulthood, the stable overweight, and the lean increasing trajectories. Among men, the higher PAF corresponded also to the overweight from early adulthood trajectory, followed by the stable overweight, the overweight from late adulthood, and the lean increasing trajectories (Figure 12).

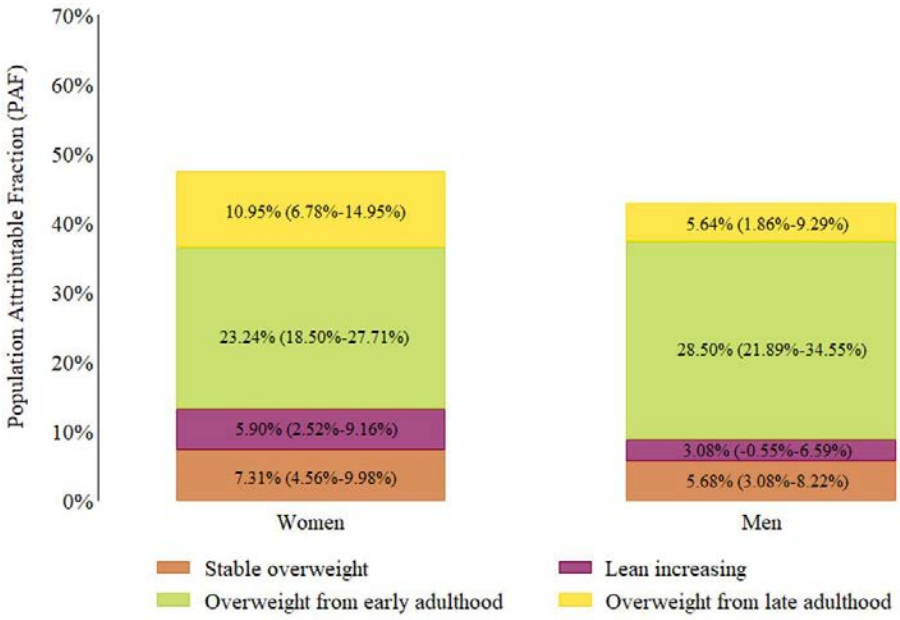

Figure 12. Population attributable fractions of the association between trajectory groups and incidence of type 2 diabetes.

6.2 Study II: Behavioral and metabolic mediators of socioeconomic inequalities in type 2 diabetes

Table 5 presents a summary of the baseline characteristics of the sample. In total there were 1,308 new cases of type 2 diabetes during the study. The cumulative incidence of type 2 diabetes was 18.4% and it was higher among men (24.5%) than women (14.5%).

Table 5. Baseline characteristics of the sample by socioeconomic status (SES).

	Education		Occupation	
	Middle/ high	Low	Middle/ high	Low
Age (SD)	46.77 (4.96)	47.83 (4.73)	47.16 (4.90)	46.49 (5.00)
Women n (%)	3,655 (62.9%)	728 (55.4%)	3,192 (63.4%)	1,191 (57.0%)
Men n (%)	2,153 (37.1%)	587 (44.6%)	1,842 (36.6%)	898 (43.0%)
Family history of type 2 diabetes n (%)	2,981 (51.3%)	760 (57.8%)	2,560 (50.9%)	1,181 (56.5%)
Comorbidities n (%)	1,567 (27.0%)	391 (29.7%)	1,295 (25.7%)	663 (31.7%)
Poor general health n (%)	104 (1.8%)	39 (3.0%)	75 (1.5%)	68 (3.3%)
Current smoking n (%)	1,310 (22.6%)	495 (37.6%)	1,112 (22.1%)	693 (33.2%)
High alcohol intake n (%)	1,099 (18.9%)	281 (21.4%)	992 (19.7%)	388 (18.6%)
Low physical activity n (%)	1,340 (23.1%)	382 (29.0%)	1,152 (22.9%)	570 (27.3%)
Diet low in fruits or vegetables n (%)	2,314 (39.8%)	606 (46.1%)	2,026 (40.2%)	894 (42.8%)
Body mass index (BMI) kg/m2 (SD)	25.41 (3.86)	26.33 (4.06)	25.34 (3.78)	26.15 (4.17)
BMI >30 kg/m^2 n (%)	626 (10.8%)	224 (17.0%)	525 (10.4%)	325 (15.6%)
Fasting plasma glucose mmol/L (SD)	4.68 (0.52)	4.81 (0.57)	4.69 (0.52)	4.75 (0.56)
Fasting plasma glucose >5.6 mmol/L n (%)	262 (4.5%)	98 (7.5%)	236 (4.7%)	124 (5.9%)
Systolic blood pressure (SBP) mmHg (SD)	121.84 (15.59)	124.88 (15.62)	122.04 (15.47)	123.29 (16.01)
Diastolic blood pressure (DBP) mmHg (SD)	76.52 (9.96)	78.10 (10.07)	76.62 (9.84)	77.28 (10.34)
Hypertension n (%)	722 (12.4%)	201 (15.3%)	621 (12.3%)	302 (14.5%)
Cases of type 2 diabetes n (%)	979 (16.9%)	239 (25.0%)	833 (16.5%)	475 (22.7%)

Data are presented as mean and standard deviation (SD) for continuous measures and number of observations and proportions for categorical variables.

Initial regression analyses showed that there was a significant unadjusted association between both measures of low SES and type 2 diabetes. The point estimates were reduced when adjustment for confounders was introduced but remained significant. Further adjustment for mediators resulted in important reduction of the estimates, which were no longer significant for the association between low educational attainment and type 2 diabetes (Table 6).

Table 6. Poisson regression for the association between low SES and type 2 diabetes.

	Model 1 IRR (95%CI)	Model 2	Model 3
Education	1.37 (1.23, 1.53)	1.29 (1.16, 1.43)	1.10 (0.99, 1.20)
Occupation	1.36 (1.23, 1.50)	1.27 (1.15, 1.41)	1.14 (1.04, 1.25)
Women			
Education	1.46 (1.24, 1.72)	1.36 (1.16, 1.60)	1.16 (0.98, 1.35)
Occupation	1.46 (1.26, 1.69)	1.33 (1.15, 1.54)	1.16 (1.01, 1.32)
Men			
Education	1.30 (1.13, 1.50)	1.23 (1.07, 1.42)	1.06 (0.92, 1.20)
Occupation	1.27 (1.11, 1.45)	1.22 (1.07, 1.39)	1.11 (0.98, 1.24)

Incidence risk ratio (IRR) and 95% confidence intervals (95%CI) are presented. Model 1: adjusted for age at baseline and sex, model 2: additionally adjusted for comorbidities, family history of type 2 diabetes and self-rated health, model 3: additionally adjusted for smoking, alcohol consumption, dietary factors, physical activity, high BMI, hyperglycemia, and hypertension.

The results of the mediation analysis using the three different methods are summarized in Figures 13 and 14. The main analysis, using counterfactual mediation, indicated that, compared to higher educational attainment, exposure to low educational attainment was associated with a risk ratio (RR) of 1.31 (95%CI 1.16, 1.45) higher risk of type 2 diabetes. The NDE was RR 1.12 (95%CI: 1.00, 1.24) and an the NIE mediated through behavioral or metabolic mediators was 1.17 (95%CI: 1.12, 1.22), representing 60% of the association. Similarly, exposure to low occupational class was associated with a TE of RR 1.30 (95%CI: 1.17, 1.44) higher risk of type 2 diabetes. The NDE was RR 1.16 (95%CI: 1.06, 1.28) and the NIE was RR 1.12 (95%CI: 1.08, 1.17), explaining 45% of the observed association. The findings of the sex stratified analyses were similar.

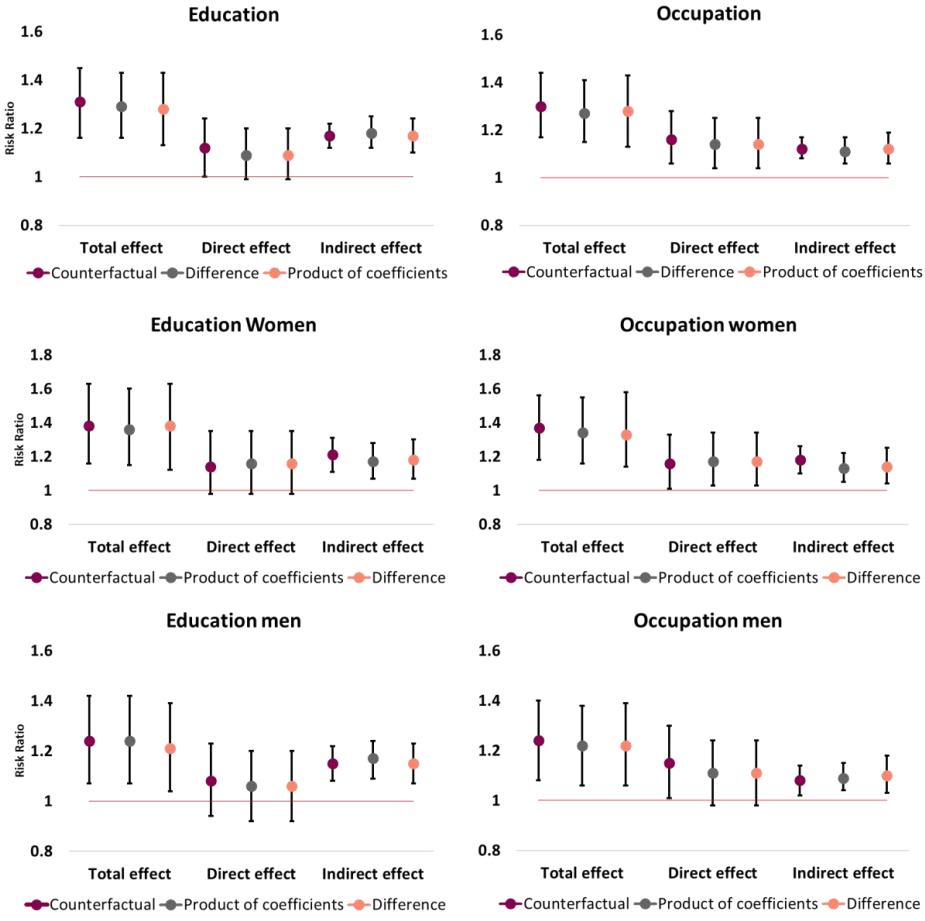

Figure 13. Comparison of the results of counterfactual mediation analysis with the difference method and the product of coefficient methods.

In comparison, using the difference method, the total effect of low educational attainment was 1.23 (95%CI: 1.07, 1.42), a direct effect of 1.06 (95%CI: 0.92, 1.20) and an indirect effect of 1.22 (95%CI: 1.13, 1.31), which accounted for 73% of the association. The total effect of low occupational class was 1.27 (95%CI: 1.15, 1.41), divided into a direct effect of 1.14 (95%CI: 1.04, 1.25) and an indirect effect of 1.11 (95%CI: 1.06, 1.17). The proportion mediated via the measured mediators was 44%.

The results according to the product of coefficients method resulted in a total effect for the association between low educational attainment and type 2 diabetes of 1.21 (95%CI: 1.04, 1.39), a direct effect equal to 1.06 (95%CI: 0.92, 1.20) and an indirect effect of 1.15 (95CI: 1.07, 1.23), which explained 72% of the association. Exposure to low occupational class was associated to a total effect

of 1.28 (95%CI: 1.13, 1.43) higher relative risk of type 2 diabetes, with a direct effect of 1-14 (95%CI: 1.04, 1.25), and an indirect effect of 1.12 (95%CI: 1.06, 1.19), which explained 46% of the observed association.

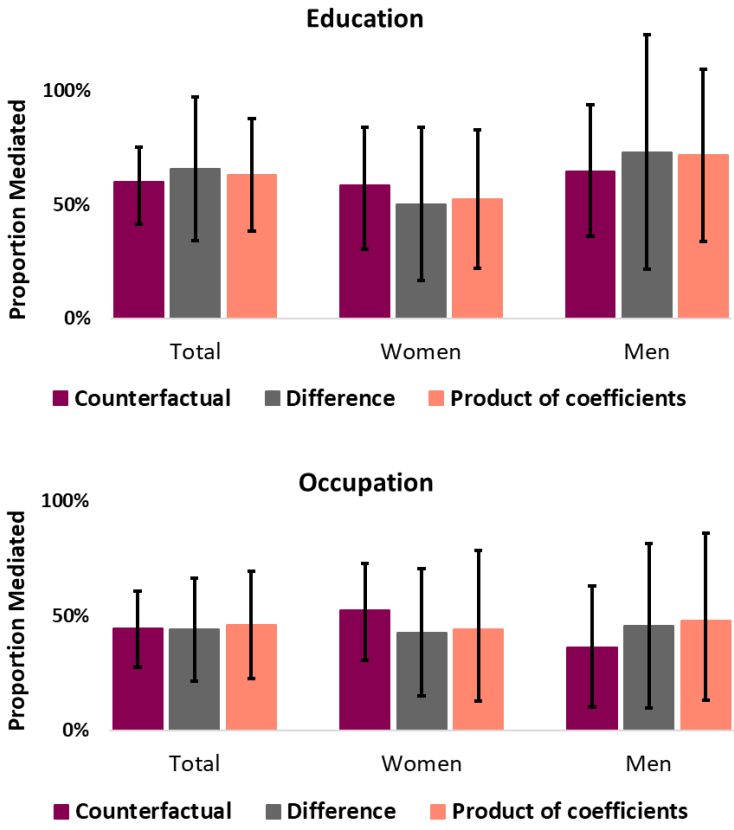

Figure 14. Comparison of the proportion mediated according to the counterfactual mediation analysis, the difference method, and the product of coefficients method.

6.3 Study III: Characterization of data-driven clusters in diabetes-free adults and their utility for risk stratification of type 2 diabetes

The baseline characteristics of the study samples are presented in Table 7. In SDPP, participants were followed for an average of 23 years during which 1,226 new cases of type 2 diabetes were documented. The incidence rate was 7.3 cases per 1,000 person-years of follow-up and was higher among men (9.0 cases per 1,000 person-years) than women (5.9 cases per 1,000 person-years). Participants of the MSC study were followed for a mean of 3.8 years and 131 new cases of type 2 diabetes were detected. The incidence rate among the overall MSC population was 23 cases per 1,000 person-years of follow-up, and slightly higher among men with 23.5 cases per 1,000 person years compared to women with 22.9 cases per 1,000 person-years.

Table 7. Baseline characteristics of the SDPP and MSC studies.

	SDPP (n=7,317)	MSC (n=2,332)
Mean age (SD)	47.10 (4.92)	42.60 (7.76)
Women	4,442 (60.7%)	1,663 (71.3%)
Mean BMI (kg/m²)	25.55 (3.83)	29.15 (4.59)
History of gestational diabetes (%)	169 (2.3%)	25 (1.1%)
Systolic blood pressure mmHg (SD)	122.51 (15.66)	114.69 (14.56)
Diastolic blood pressure mmHg (SD)	76.89 (9.99)	76.51 (10.22)
Fasting glucose mmol/L (SD)	4.71 (0.53)	4.91 (0.56)
Two-hour glucose mmol/L (SD)	4.74 (1.41)	NA
Fasting insulin µU/ml (SD)	14.33 (7.39)	12.15 (6.97)
Two-hour insulin µU/ml (SD)	46.68 (32.45)	NA
HOMA2-B	147.37 (56.39)	120.01 (45.97)
HOMA2-IR	1.56 (0.79)	1.34 (0.77)
Family history of type 2 diabetes (%)	4,278 (58.5%)	1,856 (79.6%)
Comorbidities (%)	1,846 (27.8%)	1,019 (43.7%)
Level of education (%)		
Primary education	2,249 (30.7%)	1,077 (46.2%)
Upper secondary level	2,920 (39.9%)	419 (18.0%)
University or higher	2,148 (29.4%)	836 (35.8%)
Low physical activity (%)	791 (10.8%)	1,285 (55.1%)
Current smoking (%)	1,925 (26.3%)	566 (29.9%)
Prediabetes (ADA/ WHO)	654 (8.9%) / 374(5.1%)	NA
IFG	374 (5.1%) / 94 (1.3%)	322 (13.8%) / 70 (3.0%)
IGT	191 (2.6%) / 248 (3.4%)	NA
IFG+IGT	89 (1.2%) / 32(0.4%)	NA

SDPP: Stockholm Diabetes Preventive Program, MSC: Metabolic Syndrome Cohort, SD: Standard deviation, ADA: American Diabetes Association, IFG: Impaired fasting glucose, IGT: Impaired glucose tolerance.

The results of the cluster analysis in both cohorts pointed to six distinctive clusters with comparable phenotypes (Figure 15). Compared to the average incidence of type 2 diabetes in each population, the clusters were further divided into low-risk clusters and high-risk clusters. Based on this and their characteristic features, the clusters were descriptively named: very low-risk (VLR), low-risk low beta cell function (LRLB), low-risk high beta cell function (LRHB), high-risk high blood pressure (HRHBP), high-risk beta cell failure (HRBF), and high-risk insulin resistance (HRIR).

Results of the Cox proportional hazards models were similar in both cohorts (Table 8 and Figure 16). In the SDPP sample, in comparison to the LRHB cluster, individuals in the VLR (HR: 0.38 95% CI: 0.28, 0.50) and the LRLB (HR: 0.71, 95% CI: 0.55, 0.90) clusters had a significantly lower risk of type 2 diabetes, and those in the HRHBP (HR: 2.34, 95 CI: 1.85, 2.96), the HRBF(HR: 3.22, 95% CI: 2.62, 3.96), or the HRIR (HR: 5.39, 95% CI: 4.30, 6.75) had a higher risk of incident type 2 diabetes. In the MSC study sample, in comparison to the LRHB cluster, the VLR (HR: 0.58, 95% CI: 0.51, 0.66) had a lower risk of type 2 diabetes, there was no significant difference with the LRLB cluster (HR: 1.24, 95% CI: 0.50, 3.12), and participants in the HRHBP (HR: 3.26, 95% CI: 1.49, 7.15), the HRBF (HR: 4.00, 95% CI: 2.05, 7.82) and the HRIR (HR: 4.52, 95% CI: 1.66, 12.32) clusters had a significantly higher risk of type 2 diabetes.

Table 8. Results of the Cox-proportional hazards models for the association between clusters and type 2 diabetes in the SDPP and MSC studies.

	SDPP			**MSC**		
	Cases	**Unadjusted HR (95% CI)**	**Adjusted HR (95% CI)**	**Cases**	**Unadjusted HR (95% CI)**	**Adjusted HR (95% CI)**
VLR	82	0.41 (0.32, 0.54)	0.38 (0.28, 0.50)	9	0.55 (0.51, 0.60)	0.58 (0.51, 0.66)
LRLB	141	0.75 (0.59, 0.95)	0.71 (0.55, 0.90)	23	1.23 (0.51, 2.97)	1.24 (0.50, 3.12)
LRHB	144	Ref (1.00)	Ref (1.00)	10	Ref (1.00)	Ref (1.00)
HRHBP	191	2.50 (2.01, 3.11)	2.34 (1.85, 2.96)	28	3.39 (1.55, 7.41)	3.26 (1.49, 7.15)
HRBF	418	3.58 (2.96, 4.33)	3.22 (2.62, 3.96)	42	3.96 (2.04, 7.68)	4.00 (2.05, 7.82)
HRIR	250	5.31 (4.32, 6.52)	5.39 (4.30, 6.75)	19	4.74 (1.70, 13.22)	4.52 (1.66, 12.32)
Total	**1226**			**131**		

Hazard Ratios (HR) and 95% confidence intervals (95%CI) presented. Very low-risk (VLR), low-risk low beta cell function (LRLB), low-risk high beta cell function (LRHB), high-risk high blood pressure (HRHBP), high-risk beta cell failure (HRBF), and high-risk insulin resistance (HRIR).

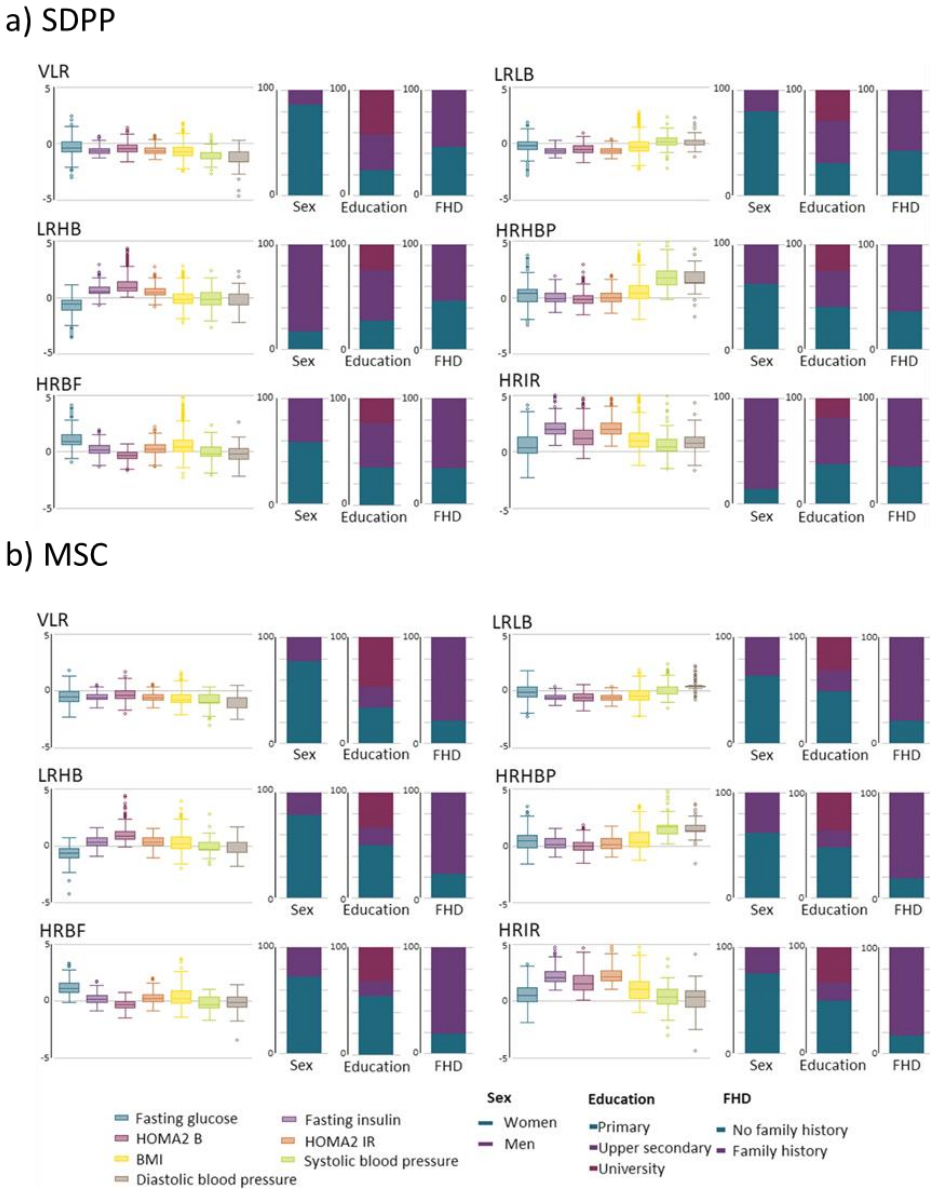

Figure 15. Box plot and bar charts of baseline characteristics among clusters. Panel "a)" shows the results of the Stockholm Diabetes Prevention Program (SDPP) and panel "b)" the results of the Metabolic Syndrome Cohort (MSC).

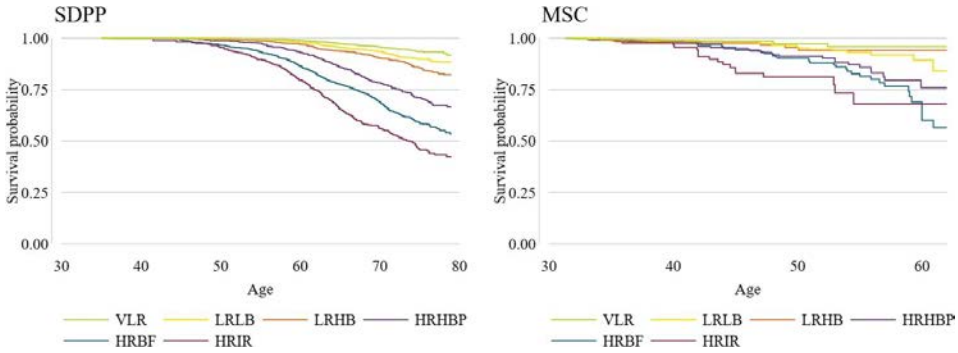

Figure 16. Kaplan–Meier estimates of the risk of type 2 diabetes by risk clusters in the Stockholm Diabetes Prevention Program (SDPP) and the Metabolic Syndrome Cohort (MSC).

Among SDPP participants with information for all study visits, the stability of the clusters ranged from fair to good, with around 73% of individuals in a high-risk cluster either remained stable or progressed to type 2 diabetes (Figure 17).

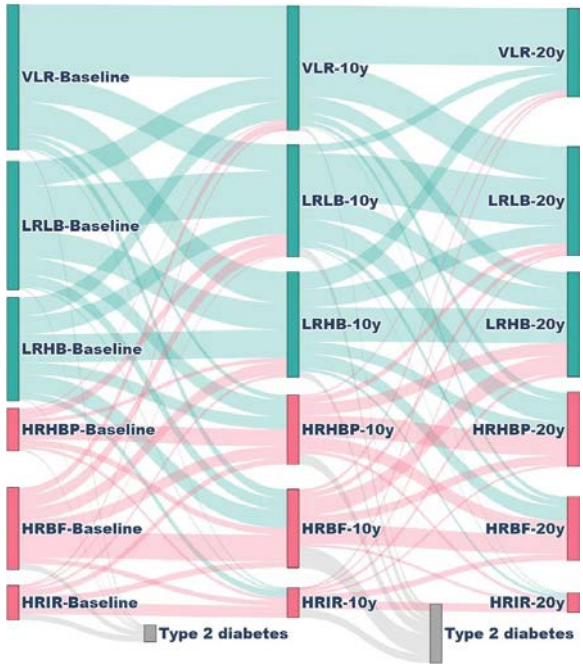

Figure 17. Transitions of the cluster labels at baseline, during the 10-year follow-up and the 20-year follow-up in the Stockholm Diabetes Prevention Program (SDPP).

The high-risk clusters had a high predictive accuracy to identify individuals at high-risk of type 2 diabetes. Compared to the ADA or WHO defined prediabetes, which included 8.9% and 5.1% of the SDPP baseline population, respectively, the high-risk clusters included close to one third of the population (34.2%), with a higher sensitivity, and area under the curve, maintaining a high specificity (Table 9). Furthermore, the high-risk clusters captured most cases of type 2 diabetes, in contrast to baseline prediabetes, which missed a large proportion of cases of type 2 diabetes (Figure 18).

Table 9. Sensitivity and specificity of the high-risk clusters and different definitions of prediabetes.

	High-risk clusters	Prediabetes ADA	Prediabetes WHO
Sensitivity %(95%CI)	70.1% (67.4%, 72.6%)	29.9% (27.4%, 32.6%)	19.5% (17.3%, 21.8%)
Specificity %(95%CI)	72.9%, (71.8%, 74.0%)	95.3% (94.8%, 95.9%)	97.8% (97.4%, 98.2%)
Area under the curve (95%CI)	0.71 (0.70, 0.73)	0.63 (0.61, 0.64)	0.59 (0.58, 0.60)
Concordance statistic (95%CI)	0.70 (0.68, 0.71)	0.63 (0.61, 0.64)	0.59 (0.58, 0.60)
Positive predictive value %(95%CI)	34.3% (32.5%, 36.2%)	56.5% (52.6%, 60.3%)	64.4% (59.3%, 69.3%)
Negative predictive value %(95%CI)	92.3% (91.6%, 93.1%)	87.1% (86.3%, 87.9%)	85.8% (84.9%, 86.6%)

American Diabetes Association (ADA), World Health Organization (WHO).

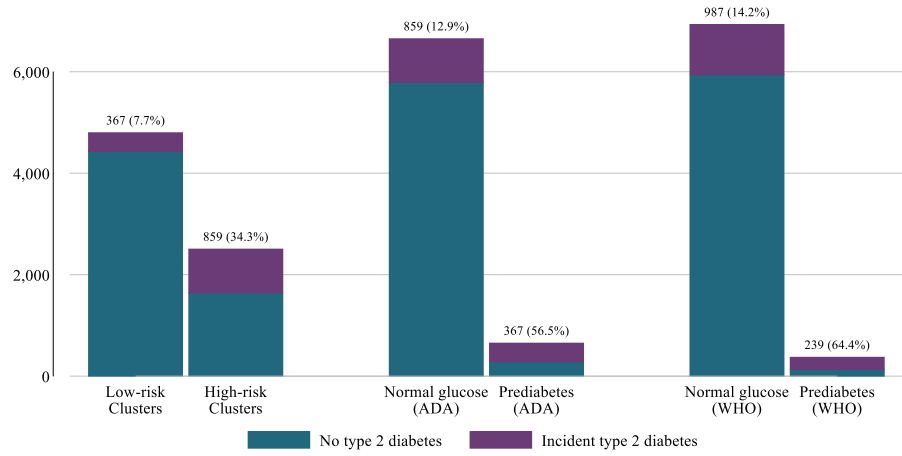

Figure 18. Cases of type 2 diabetes by risk cluster and prediabetes in the Stockholm Diabetes Prevention Program (SDPP).

6.4 Study IV: Cluster analysis for cardiovascular risk stratification

For this study, 3,345 SDPP participants with a mean age of 66 years were followed for an average of 5.2 years during which 155 new cases of CVD were documented. The incidence rate of CVD was 8.8 cases per 1,000 person-years of follow-up and was higher among men (13.9 cases per 1,000 person-years) than among women (5.1 cases per 1,000 person-years).

Cluster analysis resulted in three characteristic sub-populations based on the risk factors included in the analysis. These were named low-risk, moderate-risk, and high-risk clusters according to the predicted probabilities of incident CVD in each group (Figure 19). The characteristics of the clusters are shown in Figure 20.

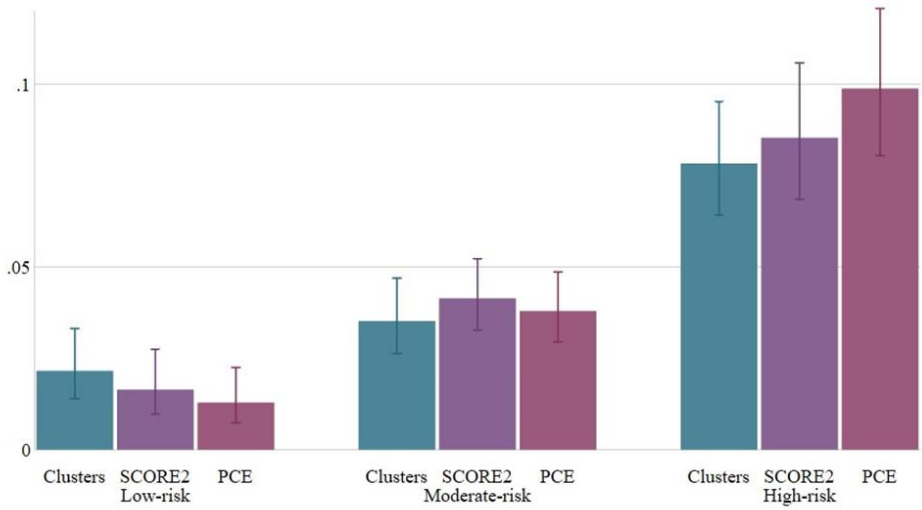

Figure 19. Probabilities of incident cardiovascular disease using the different risk stratification models in the SDPP study.

The main results are summarized in Table 10. Compared to the low-risk cluster, the absolute risk difference (ARD) of CVD was 1.37% (95%CI: 0.00%, 2.75%) higher in the moderate-risk cluster and 5.67% (95%CI: 3.87%, 7.48%) higher among participants in the high-risk cluster. On the relative scale, the moderate-risk category was associated with a HR of 1.69 (95%CI: 1.00, 2.87) and the high-risk cluster to a HR of 3.49 (95%CI: 2.15, 5.66).

In contrast, using the SCORE model, those in the moderate risk category had a 4.13% (95%C: 7.86%, 2.49%) higher absolute risk and a HR of 2.26 (95%CI: 1.27, 4.02). And for participants in the high-risk SCORE category, the ARD was 6.89% higher (96%CI: 4.84%, 8.93%) and the HR was 4.36 (95%CI: 2.45, 7.75), compared to the

low-risk category. Using the PCE model, the moderate-risk group was associated with an ARD of 2.50% (95%CI: 1.31%, 3.69%) and a HR of 2.69 (95%CI: 1.45, 5.01), while the high-risk category to an ARD of 8.59% (95%CI: 6.46%, 10.72%) and a HR of 6.55 (95%CI: 3.56, 12.04), compared to the low-risk group.

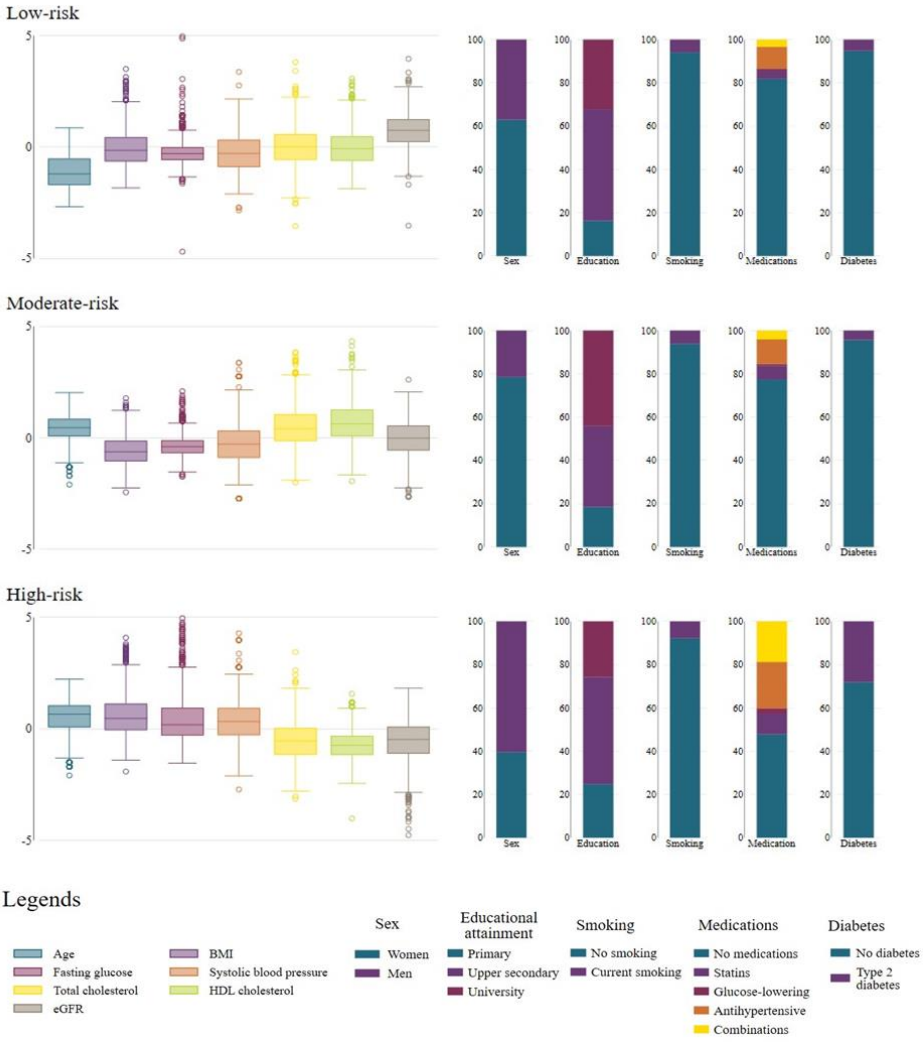

Figure 20. Box plot and bar charts of baseline characteristics by CVD risk clusters.

Table 10. Relative and absolute measures of association between the risk stratification models and CVD

	n	CVD events	predicted probabilities	Absolute risk difference (ARD)	Hazard ratio (HR)
Risk clusters					
Low-risk	930	20	2.15% (1.22%, 3.05%)	ref	ref
Moderate-risk	1,252	44	3.51% (2.49%, 4.53%)	1.36% (0.00%, 2.75%)	1.69 (1,00, 2.87)
High	1,163	91	7.82% (6.28%, 9.37%)	5.67% (3.87%, 7.48%)	3.49 (2.15, 5.66)
SCORE2					
Low-risk	856	14	1.64% (7.86%, 2.49%)	ref	ref
Moderate-risk	1,621	67	4.13% (3.16%, 5.10%)	2.50% (1.21%, 3.79%)	12.26 (1.27, 4.02)
High	868	74	8.53% (6.67%, 10.40%)	6.89% (4.85%, 8.93%)	4.36 (2.45, 7.75)
PCE					
Low-risk	924	12	1.28% (0.56%, 2.00%)	ref	ref
Moderate-risk	1,499	59	3.79% (2.84%, 4.73%)	2.50% (1.31%, 3.70%)	2.69 (1.45, 5.01)
High	767	84	9.87% (7.87%, 11.90%)	8.59% (6.46%, 10.70%)	6.55 (3.56, 12.08)

Point estimates and 95% confidence intervals in parentheses are provided. Systematic Coronary Risk Estimation (SCORE), Pooled Cohort Equations (PCE).

The risk clusters had a good predictive accuracy and discrimination, equivalent to those of the SCORE and PCE models (Figure 21). Out of the 155 total cases of CVD in the study population, the high-risk cluster correctly predicted 91 cases, with a resulting sensitivity of 58.71% (57.04%, 60.38%), a specificity of 66.39% (64.79%, 68.00%) and concordance statistic of 0.64 (0.60, 0,68). In comparison, the high-risk category derived from the SCORE model correctly identified 74 of the cases, with a sensitivity of 47.74% (46.05%, 49,43%), a specificity of 75.11% (73.64%, 76.57%) and concordance statistic of 0.63 (0.59, 0.67). In the high-risk group according to the PCE model, 84 cases of CVD were identified correctly with a sensitivity of 54.19% (52.51%, 55.88%), a specificity of 75.96% (74.51%, 77.40%) and a concordance statistic of 0.67 (0.63, 0.71). There were no significant differences in the concordance statistic between the high-risk cluster and the high-risk SCORE2 group (p-value=0.71) nor the high-risk PCE category (p-value=0.12).

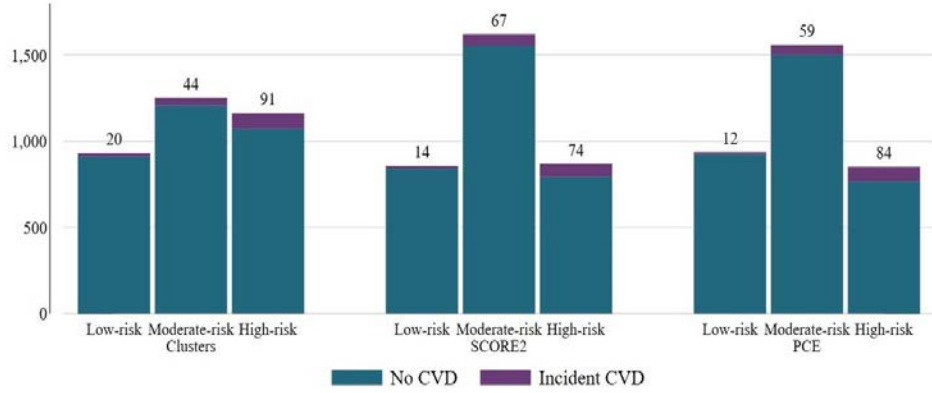

Figure 21. Cases of CVD by risk stratification models in the Stockholm Diabetes Prevention Program (SDPP).

7 Discussion

Overall, the work compiled in this thesis provides evidence for the utility of precision medicine in public health and epidemiology. Each of the component articles and manuscripts provides a specific example of how analytical methods can be used to inform decision making and implementation of preventive interventions.

In **study** I, results indicate that weight changes throughout the life-course could be generalized into five distinctive patterns, and that these patterns were associated with differences in the risk of developing type 2 diabetes in adulthood. At the population level, exposure during early adulthood explained a large proportion of the cases of type 2 diabetes, highlighting this as a period of life when public health interventions against overweight and obesity could be especially effective in preventing type 2 diabetes.

Existing evidence suggests that overweight or obesity tend to continue to adulthood, which in turn increases the risk of type 2 diabetes.[157-160] However, studies also indicate that children with high BMI who lose weight and maintain a normal weight during their teens and early adulthood have a similar risk of type 2 diabetes to people who were never exposed to overweight or obesity.[161] Furthermore, results from a recent meta-analysis suggest that although childhood obesity was associated with type 2 diabetes during adulthood, most of the cases of type 2 diabetes occurred among individuals who had a normal weight during childhood.[162] Therefore, interventions during adolescence and early adulthood might be more effective to reduce the prevalence of type 2 diabetes in the general population.

There were some important limitations in **study I**. First, retrospective assessment of the weight categories during childhood might have resulted in recall bias. Furthermore, the questions we used were an oversimplification of measured weight, which may have led to misclassification of the exposure. Additionally, individuals in the SDPP cohort were children between the 1940's and 1970's, when childhood overweight and obesity were less prevalent. In the SDPP cohort, the prevalence of childhood overweight and obesity was around 11%, whereas recent estimates indicate that this proportion has increased to around 30%.[163]

In **study II,** the results replicate the body of evidence pointing to an association between low SES and type 2 diabetes.[38] A much more challenging issue has been

to determine how much of this association can be attributed to different behavioral and metabolic risk factors, and studies have reported inconsistent findings.[164]

Furthermore, there has been increasing interest in social epidemiology regarding the use of emerging methods, such as counterfactual mediation analysis,[132] as there has been some concern that they could lead to different results than older methods. Interestingly our data yielded very similar results for the counterfactual mediation analysis and more conventional methods. However, important limitations of using the counterfactual framework in social epidemiology remain, and more research is needed in this area.

The findings of **study II** highlight the importance of addressing health inequalities in a more comprehensive way. Policies aimed at reducing inequalities could more directly address their impact in reducing exposure to risk factors.[165] Individual level interventions, in turn, should consider socioeconomic position, for example by adapting the intensity of interventions accordingly.[166]

Important limitations of this study concern the inherent complexity of assessing SES. However, education and occupation were used to attempt to cover different aspects and life-stages of SES. Measurements were dichotomized, which is likely an oversimplification and may have resulted in measurement bias.

Finding new ways to classify complex diseases such as type 2 diabetes and CVD has been a long-standing objective of precision medicine.[107] In **studies III and IV**, we applied cluster analysis to group individuals into more homogeneous sub-populations based on the patterns of relevant risk factors for type 2 diabetes (study III) and CVD (study IV). In both cases, we compared the risk stratification using these sub-populations to methods commonly used in clinical practice.

Cluster analysis in **study III**, resulted in six different subgroups with distinctive phenotypes, in two independent cohorts. According to their risk of type 2 diabetes, these subgroups were divided into three low-risk clusters and three high-risk clusters. In comparison to prediabetes, the high-risk clusters had better predictive accuracy. In **study IV**, the same methodology resulted in three characteristic sub-groups which were directly associated with low, middle, and high risk of CVD. Risk stratification based on these clusters had a similar predictive performance in comparison to models currently recommended for primary prevention.

Previous studies have used similar methods to examine the heterogeneity of type 2 diabetes and CVD. In a study from 2015, the authors use genetic data in combination with electronic medical records to describe three different sub-groups of individuals with type 2 diabetes.[167] More recently, a study using cluster analysis described five sub-classifications of type 2 diabetes.[168] This later approach has received considerable attention, and the findings have been replicated in several populations.[169,170] Research on clinical variability of CVD include a previous study that found four sub-groups of ischemic heart disease,[171] and four sub-groups after incidence of ischemic stroke[172,173] with differences in prognosis.

Although it can be hypothesized that the described phenotypes could be identified before the diagnosis of type 2 diabetes or CVD, only few studies have investigated the utility of approaches for risk stratification of type 2 diabetes or CVD.[174-176] A recent study applied cluster analysis in a selected sample of individuals at high-risk for type 2 diabetes and reported phenotypes with differences in risk of progression.[175] Studies have also used cluster analysis to identify individuals with type 2 diabetes or hypertension associated to higher CVD risk.[176,177] However these studies are based on individuals who were already at high-risk, and not the general population. Furthermore, most require data not usually available in primary care, limiting their applicability.

Besides, previous studies have not considered the role of SES in the described phenotypes, probably due to the difficulty of using categorical variables in combination with continuous data in cluster analysis. **Studies III** and **IV** contribute to the existing literature by including a measure of SES (educational attainment) as an important contributor to the variability in the risk of type 2 diabetes and CVD.

The results of these studies should be considered in the context of their limitations. An important limitation of both **study III and study IV** was the possibility of missing the occurrence of type 2 diabetes among participants who moved from Sweden, which might have led to biased estimates if the misclassification was non-differential. Additionally, in **study III** the smaller sample and shorter follow-up in the validation dataset could have affected the statistical power. However, the analysis in both samples generated similar findings, despite important differences in the incidence of type 2 diabetes between the two countries. The stability was assessed only among individuals who attended all

study visits in SDPP, which might lead to bias in these findings. Finally, an important limitation of **study IV** was the use of a more selected sample, including only individuals attending the 20-year follow-up of the SDPP study, which might limit the ability to detect existing associations due to lower power. This was done due to the availability of more extensive laboratory data in this follow-up.

8 Conclusions

In **study I,** the association of the patterns of weight changes through life with type 2 diabetes suggested that most cases were associated with exposure to overweight or obesity during early adulthood, implicating an important period for the implementation of targeted public health interventions. In **study II,** mediation analysis to study the magnitude of the different mechanisms linking low SES to incidence of type 2 diabetes, estimated that behavioral and metabolic risk factors, usual targets for public health interventions, explained roughly half of the observed social inequalities in the incidence of type 2 diabetes. Finally, in **studies III** and **IV,** cluster analysis was shown to be a useful alternative for the development of prediction models that outperformed or were at least as good as those currently used in clinical practice and could be used to guide the implementation of different preventive interventions.

In conclusion, these findings support the utility of a precision medicine approach to public health. The different methods were useful to identify and describe complex patterns of risk factors based on their changes through time, the importance of their underlying causal mechanism or the patterns of their combinations with other risk factors. The results highlight areas of opportunity to improve the development and implementation of public health interventions.

9 Points of perspective

The Sustainable Development Goals (SDGs) are a part of the United Nations 2030 Agenda for Sustainable Development and set the objectives for progress in 17 different interconnected areas. One of the main targets to achieve good health and well-being is to "reduce by one third premature mortality from non-communicable diseases".[178] To achieve this goal, it is important to improve public health policies, as well as early detection, and effective treatment of NCDs. The findings in this thesis can contribute to guide policy and further research in these areas.

The patterns and phenotypes described, as well as the insight into the importance of different pathways, may be useful to inform public health policies. For example, they could be taken into consideration during policy development to identify the target population and optimize the timing of different public health interventions.

From a more clinical perspective, data-driven risk stratification could lead to better-targeted interventions to prevent type 2 diabetes and CVD. Controversies such as preventive use of medications could be addressed by targeting those who might benefit more specifically. Lifestyle interventions could be adapted to improve their efficiency based on the characteristics of individuals.

The studies compiled in this thesis were mostly exploratory and open a variety of opportunities for further research. First, replication of these findings in different populations and different generations is important to assess their generalizability. Additionally, more studies are needed to explore the practical utility of these methods to improve preventive interventions. Furthermore, the subgrouping of individuals into more homogeneous clusters could provide interesting insights into the pathophysiology of type 2 diabetes and CVD. By using the trajectory groups or clusters as outcomes. Future studies could investigate their possible determinants, including genetic, biological, social, psychological, or environmental factors. Correspondingly, the described sub-groups could also be used as exposures in future research looking into their impact on other conditions or on the development of complications following a diagnosis of type 2 diabetes or CVD.

Other areas of future study include the inclusion of the social determinants of health in these types of analyses. These studies attempt to include SES directly into the detection of sub-groups. Future studies should strive to expand this by including other relevant factors such as environmental exposures.

10 Acknowledgements

I feel very lucky for the opportunity I've had to spend these years as a doctoral student at KI, especially because I've been surrounded by incredible people along the way. The work and ideas of many people are reflected in these studies, and I am thankful to all of you who've made this possible.

To begin with, I am extremely grateful to my main supervisor, **Anton Lager**. Thank you for placing your trust in me, and for always supporting me both in and outside of research, promoting my personal growth.

Thank you **Ylva Trolle Lageros,** my co-supervisor, for your guidance and support, and for all the doors you've opened for me. Thank you especially for the opportunity to work with patients again, and for your clinical mentorship.

I am also grateful to my co-supervisors **Yvonne Forssell**, for introducing me to research as a profession, and for helping me with the hard transition from master student to PhD student. Thank you, **Antonio Ponce de Leon,** for your help with the complex statistics and your helpful advice whenever I needed it. I would also like to thank my mentor, **Michael**, for your support and example, and for always taking the time for giving me advice when I found difficulties on the way.

I've also had the unique opportunity of working together with both public health professionals and clinicians. I would like to thank my epidemiologist and public health colleagues at CES, especially at the Unit of Health Care Analysis: **Anders, Andreas, Anton, David, Bo, Boel, Hrafnhildur, Marie, Minhao, Mihretab, Peeter, Per, Sara, Stefan,** and **Vicky**, as well as my clinical colleagues at the Center for Obesity**: Agneta, Anna, Anne, Birgit, Britt-Marie, Elizabeth, Gunilla, Jelena, Jessica, Josefin, Kajsa, Karolina, Maria, Martina, Sofia** and **Ylva**.

The studies wouldn't have been possible without everyone involved in the SDPP study. Thank you **Boel**, **Hrafnhildur** and **Per** for all your help with the studies and the work to improve the SDPP study. And special thanks to **Minhao** for your commitment and close collaboration with these projects.

I'm also thankful to all the members of the **PRIME Health** research group, especially to **Cecilia Magnusson,** for promoting an active research environment as group leader, and to my colleagues **Anton**, **Boel**, **Chu**, **Daniel**, **Elena**, **Lane**, **Lisa**, **Maki**, **Michael**, **Per** and **Viktor**.

There is something about being a PhD student that only other students can understand. So, I want to extend my gratitude to the current (**Anna**, **Chu**, **Fen**, **Filip**, **Hugo**, **Joseph**, **Robert**, **Veronica**, **Viktor**) and past (**Alicia**, **Dang**, **Dominika**, **Helio**, **Hua**, **Isidora**, **Jad**, **Maki**, **Martin**, **Megan**, **Melody**, **Nora**, **Pär, Rynaz**, **Yajun**, **Zakir**) PhD students or Postdocs I've met throughout the years at the Department of Global Public Health in Torsplan. It is hard to list everyone who left a mark on me during my time as a Ph.D. student, I am sorry if I forgot someone.

Many people have played an important role throughout the years, starting from the master's program in 2016-2018 and continuing until now. I'd like to thank **Jette** for her work as program director for the master's in public health, and to all the dedicated teachers that have been a part of my education here. **Nicola** for the chance to teach biostatistics. **Emilie** for her work as director of doctoral studies, as well as to the administrative team at KI for their support and assistance.

I am also very thankful for the support I've had from Mexico to pursue my Ph.D. studies. I am especially grateful to the National Council of Science of Technology (CONACyT), that supported me through a scholarship for my master and part of my doctoral education. And to my colleagues in Mexico for their interest and support with the collaboration, **Carlos Aguilar,** and everyone at his lab at the INCMNSZ, **Omar Bello** and his lab at UNAM and **Carolina Sepulveda**.

To all the friends I've made since I came to Sweden, including (but not limited to) **Alejandro, Anna, Camilla, Chris, Chu, Emma, Elena, Fabricio, Gustavo, Heloisa, Jonathan, Kerstin, Linn, Maarten, Megan, Minhao, Mirhetab, Patricio, Paul, Ruy, Sabsil, Salvador, Simon, Tali and Vicky.** For all the fun times and memories, and for helping me to maintain a better work-life balance. You have all helped me in one way or another, sometimes with support and words of encouragement, sometimes just having fun. Thank you all.

Above everything, I want to thank my family. My mom **Tere** and my dad **Salvador** for your overwhelming love, and for much more than words can express. I would not be standing here without your endless support and sacrifice. My grandfather **Jesus Kumate** for always being there for me during my upbringing and having set and incredible life example, instilling in me your love for medicine and public health, I miss you every day. My grandma **Bety** and **Tere** for your unconditional love and support through my life. My siblings **Gaby**, **Rodrigo**, **Mike** and **Tere,** as well as **Alejandro, Andy,** and **Paty,** for your love and support. And my nieces and nephews **Alexa, Emilia, Isabela**, **Mariano,** and **Mateo** for all the smiles.

My family in Iceland, **Aníta**, **Birkir**, **Breki**, **Daníel**, **Erna**, **Thora,** and **Sólveig,** thank you for making Iceland feel like home, for your encouragement, and for all the fun moments together.

And thank you **Hilda** for your love and for everything we share. Thank you for walking beside me. The journey would not have been as fun without you there to share the good times, and it would have been more challenging without you there to pick me up during the difficult moments. Out of everything that has happened over the past seven years, meeting you has been by far the best.

11 References

1. Bartelink V. Lager A. (editors). *Folkhälsorapport 2019*. Stockholm: Centrum för epidemiologi och samnhällsmedicin, Region Stockholm; 2019.
2. Vos T, Lim SS, Abbafati C, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*. 2020;396(10258):1204-1222.
3. Pogosova N. Costs associated with cardiovascular disease create a significant burden for society and they seem to be globally underestimated. *European Journal of Preventive Cardiology*. 2020;26(11):1147-1149.
4. Chung WK, Erion K, Florez JC, 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). *Diabetologia*. 2020;63(9):1671-1693.
5. Leopold JA, Loscalzo J. Emerging Role of Precision Medicine in Cardiovascular Disease. *Circ Res*. 2018;122(9):1302-1315.
6. Roth GA, Mensah GA, Johnson CO, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update From the GBD 2019 Study. *Journal of the American College of Cardiology*. 2020;76(25):2982-3021.
7. Murray CJL, Aravkin AY, Zheng P, et al. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*. 2020;396(10258):1223-1249.
8. Olstad DL, McIntyre L. Reconceptualising precision public health. *BMJ Open*. 2019;9(9):e030279.
9. Bilkey GA, Burns BL, Coles EP, Mahede T, Baynam G, Nowak KJ. Optimizing Precision Medicine for Public Health. *Front Public Health*. 2019;7:42.
10. Velmovitsky PE, Bevilacqua T, Alencar P, Cowan D, Morita PP. Convergence of Precision Medicine and Public Health Into Precision Public Health: Toward a Big Data Perspective. *Frontiers in Public Health*. 2021;9.
11. Williams R, Karuranga S, Malanda B, et al. Global and regional estimates and projections of diabetes-related health expenditure: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Res Clin Pract*. 2020;162:108072.
12. Ong KL, Stafford LK, McLaughlin SA, et al. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet*. 2023;402(10397):203-234.

13. ElSayed NA, Aleppo G, Aroda VR, et al. 2. Classification and Diagnosis of Diabetes: Standards of Care in Diabetes—2023. *Diabetes Care*. 2022;46(Supplement_1):S19-S40.

14. World Health Organization. Classification of diabetes mellitus. 2019.

15. World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation. 2006.

16. Safiri S, Karamzad N, Kaufman JS, et al. Prevalence, Deaths and Disability-Adjusted-Life-Years (DALYs) Due to Type 2 Diabetes and Its Attributable Risk Factors in 204 Countries and Territories, 1990-2019: Results From the Global Burden of Disease Study 2019. *Frontiers in Endocrinology*. 2022;13.

17. Magliano DJ, Chen L, Islam RM, et al. Trends in the incidence of diagnosed diabetes: a multicountry analysis of aggregate data from 22 million diagnoses in high-income and middle-income settings. *The Lancet Diabetes & Endocrinology*. 2021;9(4):203-211.

18. Magliano DJ, Islam RM, Barr ELM, et al. Trends in incidence of total or type 2 diabetes: systematic review. *Bmj*. 2019;366:l5003.

19. Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*. 2019;140(11):e596-e646.

20. Vaduganathan M, Mensah GA, Turco JV, Fuster V, Roth GA. The Global Burden of Cardiovascular Diseases and Risk. *Journal of the American College of Cardiology*. 2022;80(25):2361-2371.

21. Mensah GA, Wei GS, Sorlie PD, et al. Decline in Cardiovascular Mortality. *Circulation Research*. 2017;120(2):366-380.

22. Lopez AD, Adair T. Is the long-term decline in cardiovascular-disease mortality in high-income countries over? Evidence from national vital statistics. *International Journal of Epidemiology*. 2019;48(6):1815-1823.

23. Luengo-Fernandez R, Walli-Attaei M, Gray A, et al. Economic burden of cardiovascular diseases in the European Union: a population-based cost study. *European Heart Journal*. 2023.

24. Mendoza-Herrera K, Pedroza-Tobías A, Hernández-Alcaraz C, Ávila-Burgos L, Aguilar-Salinas CA, Barquera S. Attributable Burden and Expenditure of Cardiovascular Diseases and Associated Risk Factors in Mexico and other Selected Mega-Countries. *Int J Environ Res Public Health*. 2019;16(20).

25. Gheorghe A, Griffiths U, Murphy A, Legido-Quigley H, Lamptey P, Perel P. The economic burden of cardiovascular disease and hypertension in low- and middle-income countries: a systematic review. *BMC Public Health*. 2018;18(1):975.

26. Fernandes Silva L, Vangipurapu J, Laakso M. The "Common Soil Hypothesis" Revisited-Risk Factors for Type 2 Diabetes and Cardiovascular Disease. *Metabolites*. 2021;11(10).

27. Dal Canto E, Ceriello A, Rydén L, et al. Diabetes as a cardiovascular risk factor: An overview of global trends of macro and micro vascular complications. *European Journal of Preventive Cardiology*. 2020;26(2_suppl):25-32.

28. Yun JS, Ko SH. Current trends in epidemiology of cardiovascular disease and cardiovascular risk management in type 2 diabetes. *Metabolism*. 2021;123:154838.

29. Einarson TR, Acs A, Ludwig C, Panton UH. Prevalence of cardiovascular disease in type 2 diabetes: a systematic literature review of scientific evidence from across the world in 2007-2017. *Cardiovasc Diabetol*. 2018;17(1):83.

30. Bays HE, Taub PR, Epstein E, et al. Ten things to know about ten cardiovascular disease risk factors. *American Journal of Preventive Cardiology*. 2021;5:100149.

31. Weisman A, Fazli GS, Johns A, Booth GL. Evolving Trends in the Epidemiology, Risk Factors, and Prevention of Type 2 Diabetes: A Review. *Can J Cardiol*. 2018;34(5):552-564.

32. Mackenbach JP, Cavelaars AEJM, Kunst AE, Groenhof F. Socioeconomic inequalities in cardiovascular disease mortality. An international study. *European Heart Journal*. 2000;21(14):1141-1151.

33. Golden SH, Brown A, Cauley JA, et al. Health disparities in endocrine disorders: biological, clinical, and nonclinical factors--an Endocrine Society scientific statement. *J Clin Endocrinol Metab*. 2012;97(9):E1579-1639.

34. Hill-Briggs F, Adler NE, Berkowitz SA, et al. Social Determinants of Health and Diabetes: A Scientific Review. *Diabetes Care*. 2020;44(1):258-279.

35. WHO Commission on Social Determinants of Health. *Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health : Commission on Social Determinants of Health Final Report*. World Health Organization, Commission on Social Determinants of Health; 2008.

36. Teshale AB, Htun HL, Owen A, et al. The Role of Social Determinants of Health in Cardiovascular Diseases: An Umbrella Review. *Journal of the American Heart Association*. 2023;12(13):e029765.

37. Galobardes B, Lynch J, Smith GD. Measuring socioeconomic position in health research. *British Medical Bulletin*. 2007;81-82(1):21-37.

38. Agardh E, Allebeck P, Hallqvist J, Moradi T, Sidorchuk A. Type 2 diabetes incidence and socio-economic position: a systematic review and meta-analysis. *Int J Epidemiol*. 2011;40(3):804-818.

39. Lewer D, Jayatunga W, Aldridge RW, et al. Premature mortality attributable to socioeconomic inequality in England between 2003 and 2018: an observational study. *Lancet Public Health*. 2020;5(1):e33-e41.

40. Piketty T, Saez E. Inequality in the long run. *Science*. 2014;344(6186):838-843.

41. Rosengren A, Smyth A, Rangarajan S, et al. Socioeconomic status and risk of cardiovascular disease in 20 low-income, middle-income, and high-income countries: the Prospective Urban Rural Epidemiologic (PURE) study. *The Lancet Global Health*. 2019;7(6):e748-e760.

42. Magnussen C, Ojeda FM, Leong DP, et al. Global Effect of Modifiable Risk Factors on Cardiovascular Disease and Mortality. *N Engl J Med*. 2023;389(14):1273-1285.

43. Zhang Y, Pan XF, Chen J, et al. Combined lifestyle factors and risk of incident type 2 diabetes and prognosis among individuals with type 2 diabetes: a systematic review and meta-analysis of prospective cohort studies. *Diabetologia*. 2020;63(1):21-33.

44. Rawshani A, Rawshani A, Franzén S, et al. Risk Factors, Mortality, and Cardiovascular Outcomes in Patients with Type 2 Diabetes. *New England Journal of Medicine*. 2018;379(7):633-644.

45. Knowler WC, Barrett-Connor E, Fowler SE, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med*. 2002;346(6):393-403.

46. World Health Organisation. Global action plan for the prevention and control of NCDs 2013–2020. In: WHO Geneva; 2013.

47. Visseren FLJ, Mach F, Smulders YM, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J*. 2021;42(34):3227-3337.

48. ElSayed NA, Aleppo G, Aroda VR, et al. 3. Prevention or Delay of Diabetes and Associated Comorbidities: Standards of Care in Diabetes—2023. *Diabetes Care*. 2022;46(Supplement_1):S41-S48.

49. 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.

50. Abbasi A, Peelen LM, Corpeleijn E, et al. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. *BMJ : British Medical Journal*. 2012;345:e5900.

51. Rossello X, Dorresteijn JA, Janssen A, et al. Risk prediction tools in cardiovascular disease prevention: A report from the ESC Prevention of CVD Programme led by the European Association of Preventive Cardiology (EAPC) in collaboration with the Acute Cardiovascular Care Association (ACCA) and the Association of Cardiovascular Nursing and Allied Professions (ACNAP). *European Journal of Preventive Cardiology*. 2020;26(14):1534-1544.

52. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. 2016;353:i2416.

53. Lucaroni F, Modica DC, Macino M, et al. Can risk be predicted? An umbrella systematic review of current risk prediction models for cardiovascular diseases, diabetes and hypertension. *BMJ Open*. 2019;9(12):e030234.

54. Kengne AP, Beulens JWJ, Peelen LM, et al. Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models. *The Lancet Diabetes & Endocrinology*. 2014;2(1):19-29.

55. Echouffo-Tcheugui JB, Selvin E. Prediabetes and What It Means: The Epidemiological Evidence. *Annual Review of Public Health*. 2021;42(1):59-77.

56. Cai X, Zhang Y, Li M, et al. Association between prediabetes and risk of all cause mortality and cardiovascular disease: updated meta-analysis. *BMJ*. 2020;370:m2297.

57. Richter B, Hemmingsen B, Metzendorf MI, Takwoingi Y. Development of type 2 diabetes mellitus in people with intermediate hyperglycaemia. *Cochrane Database of Systematic Reviews*. 2018(10).

58. Shang Y, Marseglia A, Fratiglioni L, et al. Natural history of prediabetes in older adults from a population-based longitudinal study. *J Intern Med*. 2019;286(3):326-340.

59. Williams B, Mancia G, Spiering W, et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension: The Task Force for the management of arterial hypertension of the European Society of Cardiology (ESC) and the European Society of Hypertension (ESH). *European Heart Journal*. 2018;39(33):3021-3104.

60. Mach F, Baigent C, Catapano AL, et al. 2019 ESC/EAS Guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk: The Task Force for the management of dyslipidaemias of the European Society of Cardiology (ESC) and European Atherosclerosis Society (EAS). *European Heart Journal*. 2019;41(1):111-188.

61. Grundy SM, Stone NJ, Bailey AL, et al. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: A Report of the

American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*. 2019;139(25):e1082-e1143.

62. American Diabetes Association Professional Practice Committee. 10. Cardiovascular Disease and Risk Management: Standards of Medical Care in Diabetes—2022. *Diabetes Care*. 2021;45(Supplement_1):S144-S174.

63. Conroy RM, Pyörälä K, Fitzgerald AP, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *European Heart Journal*. 2003;24(11):987-1003.

64. SCORE2 working group, ESC Cardiovascular risk collaboration. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *European Heart Journal*. 2021;42(25):2439-2454.

65. SCORE2-OP working group, ESC Cardiovascular risk collaboration. SCORE2-OP risk prediction algorithms: estimating incident cardiovascular event risk in older persons in four geographical risk regions. *European Heart Journal*. 2021;42(25):2455-2467.

66. SCORE2-Diabetes working group, the ESC Cardiovascular Risk Collaboration. SCORE2-Diabetes: 10-year cardiovascular risk estimation in type 2 diabetes in Europe. *European Heart Journal*. 2023;44(28):2544-2556.

67. Goff DC, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. *Circulation*. 2014;129(25_suppl_2):S49-S73.

68. DeFilippis AP, Young R, McEvoy JW, et al. Risk score overestimation: the impact of individual cardiovascular risk factors and preventive therapies on the performance of the American Heart Association-American College of Cardiology-Atherosclerotic Cardiovascular Disease risk score in a modern multi-ethnic cohort. *Eur Heart J*. 2017;38(8):598-608.

69. Khera R, Pandey A, Ayers CR, et al. Performance of the Pooled Cohort Equations to Estimate Atherosclerotic Cardiovascular Disease Risk by Body Mass Index. *JAMA Netw Open*. 2020;3(10):e2023242.

70. Achhra AC, Lyass A, Borowsky L, et al. Assessing Cardiovascular Risk in People Living with HIV: Current Tools and Limitations. *Curr HIV/AIDS Rep*. 2021;18(4):271-279.

71. Vasan RS, van den Heuvel E. Differences in estimates for 10-year risk of cardiovascular disease in Black versus White individuals with identical risk factor profiles using pooled cohort equations: an in silico cohort study. *The Lancet Digital Health*. 2022;4(1):e55-e63.

72. Kist JM, Vos RC, Mairuhu ATA, 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.

73. van Trier TJ, Snaterse M, Boekholdt SM, et al. Validation of Systematic Coronary Risk Evaluation 2 (SCORE2) and SCORE2-Older Persons in the EPIC-Norfolk prospective population cohort. *European Journal of Preventive Cardiology*. 2023.

74. Dziopa K, Asselbergs FW, Gratton J, Chaturvedi N, Schmidt AF. Cardiovascular risk prediction in type 2 diabetes: a comparison of 22 risk scores in primary care settings. *Diabetologia*. 2022;65(4):644-656.

75. Lloyd-Jones DM, Braun LT, Ndumele CE, et al. Use of Risk Assessment Tools to Guide Decision-Making in the Primary Prevention of Atherosclerotic Cardiovascular Disease: A Special Report From the American Heart Association and American College of Cardiology. *J Am Coll Cardiol*. 2019;73(24):3153-3167.

76. Yadlowsky S, Hayward RA, Sussman JB, McClelland RL, Min YI, Basu S. Clinical Implications of Revised Pooled Cohort Equations for Estimating Atherosclerotic Cardiovascular Disease Risk. *Ann Intern Med*. 2018;169(1):20-29.

77. Farzadfar F. Cardiovascular disease risk prediction models: challenges and perspectives. *The Lancet Global Health*. 2019;7(10):e1288-e1289.

78. Karmali KN, Persell SD, Perel P, Lloyd-Jones DM, Berendsen MA, Huffman MD. Risk scoring for the primary prevention of cardiovascular disease. *Cochrane Database Syst Rev*. 2017;3(3):Cd006887.

79. Blue L, Kranker K, Markovitz AR, et al. Effects of the Million Hearts Model on Myocardial Infarctions, Strokes, and Medicare Spending: A Randomized Clinical Trial. *JAMA*. 2023;330(15):1437-1447.

80. Eriksen CU, Rotar O, Toft U, Jørgensen T, Network HE. *What is the Effectiveness of Systematic Population-level Screening Programmes for Reducing the Burden of Cardiovascular Diseases?* : WHO Regional Office for Europe, Health Evidence Network; 2021.

81. Lindholt JS, Søgaard R. Why and when to screen for cardiovascular disease in healthy individuals. *Heart*. 2021;107(12):1010-1017.

82. Dyakova M, Shantikumar S, Colquitt JL, et al. Systematic versus opportunistic risk assessment for the primary prevention of cardiovascular disease. *Cochrane Database Syst Rev*. 2016;2016(1):Cd010411.

83. Peer N, Balakrishna Y, Durao S. Screening for type 2 diabetes mellitus. *Cochrane Database Syst Rev*. 2020;5(5):Cd005266.

84. Simmons RK, Echouffo-Tcheugui JB, Sharp SJ, et al. Screening for type 2 diabetes and population mortality over 10 years (ADDITION-Cambridge): a cluster-randomised controlled trial. *Lancet*. 2012;380(9855):1741-1748.

85. van den Donk M, Sandbaek A, Borch-Johnsen K, et al. Screening for type 2 diabetes. Lessons from the ADDITION-Europe study. *Diabet Med*. 2011;28(11):1416-1424.

86. Ke C, Chu A, Shah BR, et al. Association of prior outpatient diabetes screening with cardiovascular events and mortality among people with incident diabetes: a population-based cohort study. *Cardiovasc Diabetol*. 2023;22(1):227.

87. Ogurtsova K, Guariguata L, Barengo NC, et al. IDF diabetes Atlas: Global estimates of undiagnosed diabetes in adults for 2021. *Diabetes Research and Clinical Practice*. 2022;183:109118.

88. Aschner P. New IDF clinical practice recommendations for managing type 2 diabetes in primary care. *Diabetes Res Clin Pract*. 2017;132:169-170.

89. Pearson TA. Public Policy Approaches to the Prevention of Heart Disease and Stroke. *Circulation*. 2011;124(23):2560-2571.

90. Joseph P, Leong D, McKee M, et al. Reducing the Global Burden of Cardiovascular Disease, Part 1. *Circulation Research*. 2017;121(6):677-694.

91. White M. Population Approaches to Prevention of Type 2 Diabetes. *PLoS Med*. 2016;13(7):e1002080.

92. ElSayed NA, Aleppo G, Aroda VR, et al. 5. Facilitating Positive Health Behaviors and Well-being to Improve Health Outcomes: Standards of Care in Diabetes—2023. *Diabetes Care*. 2022;46(Supplement_1):S68-S96.

93. Leong DP, Joseph PG, McKee M, et al. Reducing the Global Burden of Cardiovascular Disease, Part 2. *Circulation Research*. 2017;121(6):695-710.

94. United Nations. Department of Economic and Social Affairs. *World social report 2020: Inequality in a rapidly changing world*. . United Nations.; 2020.

95. Haw JS, Galaviz KI, Straus AN, et al. Long-term Sustainability of Diabetes Prevention Approaches: A Systematic Review and Meta-analysis of Randomized Clinical Trials. *JAMA Internal Medicine*. 2017;177(12):1808-1817.

96. Diabetes Prevention Program Research Group. Long-term Effects of Metformin on Diabetes Prevention: Identification of Subgroups That Benefited Most in the Diabetes Prevention Program and Diabetes Prevention Program Outcomes Study. *Diabetes Care*. 2019;42(4):601-608.

97. Herman WH. The cost-effectiveness of diabetes prevention: results from the Diabetes Prevention Program and the Diabetes Prevention Program Outcomes Study. *Clin Diabetes Endocrinol*. 2015;1:9.

98. Alouki K, Delisle H, Bermúdez-Tamayo C, Johri M. Lifestyle Interventions to Prevent Type 2 Diabetes: A Systematic Review of Economic Evaluation Studies. *J Diabetes Res*. 2016;2016:2159890.

99. Galaviz KI, Weber MB, Straus A, Haw JS, Narayan KMV, Ali MK. Global Diabetes Prevention Interventions: A Systematic Review and Network Meta-analysis of the Real-World Impact on Incidence, Weight, and Glucose. *Diabetes Care*. 2018;41(7):1526-1534.

100. Dunkley AJ, Bodicoat DH, Greaves CJ, et al. Diabetes prevention in the real world: effectiveness of pragmatic lifestyle interventions for the prevention of type 2 diabetes and of the impact of adherence to guideline recommendations: a systematic review and meta-analysis. *Diabetes Care*. 2014;37(4):922-933.

101. Joiner KL, Nam S, Whittemore R. Lifestyle interventions based on the diabetes prevention program delivered via eHealth: A systematic review and meta-analysis. *Prev Med*. 2017;100:194-207.

102. Moin T, Schmittdiel JA, Flory JH, et al. Review of Metformin Use for Type 2 Diabetes Prevention. *American Journal of Preventive Medicine*. 2018;55(4):565-574.

103. ElSayed NA, Aleppo G, Aroda VR, et al. 8. Obesity and Weight Management for the Prevention and Treatment of Type 2 Diabetes: Standards of Care in Diabetes—2023. *Diabetes Care*. 2022;46(Supplement_1):S128-S139.

104. Carlsson LMS, Peltonen M, Ahlin S, et al. Bariatric Surgery and Prevention of Type 2 Diabetes in Swedish Obese Subjects. *New England Journal of Medicine*. 2012;367(8):695-704.

105. Fegers-Wustrow I, Gianos E, Halle M, Yang E. Comparison of American and European Guidelines for Primary Prevention of Cardiovascular Disease: JACC Guideline Comparison. *Journal of the American College of Cardiology*. 2022;79(13):1304-1313.

106. Lloyd-Jones DM, Allen NB, Anderson CAM, et al. Life's Essential 8: Updating and Enhancing the American Heart Association's Construct of Cardiovascular Health: A Presidential Advisory From the American Heart Association. *Circulation*. 2022;146(5):e18-e43.

107. National Research Council. Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. 2011.

108. Lash TL. *Modern epidemiology*. 4th ed. Philadelphia, PA: Wolters Kluwer; 2021.

109. Bayer R, Galea S. Public Health in the Precision-Medicine Era. *N Engl J Med*. 2015;373(6):499-501.

110. Jørgensen JT. Twenty Years with Personalized Medicine: Past, Present, and Future of Individualized Pharmacotherapy. *Oncologist*. 2019;24(7):e432-e440.

111. Canfell OJ, Davidson K, Woods L, et al. Precision Public Health for Non-communicable Diseases: An Emerging Strategic Roadmap and Multinational Use Cases. *Frontiers in Public Health*. 2022;10.

112. McCarthy MI. Painting a new picture of personalised medicine for diabetes. *Diabetologia*. 2017;60(5):793-799.

113. Beccia F, Hoxhaj I, Castagna C, et al. An overview of Personalized Medicine landscape and policies in the European Union. *European Journal of Public Health*. 2022;32(6):844-851.

114. Kahkoska AR, Freeman NLB, Hassmiller Lich K. Systems-Aligned Precision Medicine—Building an Evidence Base for Individuals Within Complex Systems. *JAMA Health Forum*. 2022;3(7):e222334-e222334.

115. Nguena Nguefack HL, Pagé MG, Katz J, et al. Trajectory Modelling Techniques Useful to Epidemiological Research: A Comparative Narrative Review of Approaches. *Clin Epidemiol*. 2020;12:1205-1222.

116. Ben-Shlomo Y, Mishra G, Kuh D. Life Course Epidemiology. In: Ahrens W, Pigeot I, eds. *Handbook of Epidemiology*. New York, NY: Springer New York; 2014:1521-1549.

117. Eriksson JG, Osmond C, Kajantie E, Forsén TJ, Barker DJ. Patterns of growth among children who later develop type 2 diabetes or its risk factors. *Diabetologia*. 2006;49(12):2853-2858.

118. Robinson R. The fetal origins of adult disease. *BMJ*. 2001;322(7283):375-376.

119. Lynch J, Smith GD. A life course approach to chronic disease epidemiology. *Annual review of public health*. 2005;26(1):1-35.

120. Lennon H, Kelly S, Sperrin M, et al. Framework to construct and interpret latent class trajectory modelling. *BMJ Open*. 2018;8(7):e020683.

121. Nagin DS. Group-Based Trajectory Modeling: An Overview. *Annals of Nutrition and Metabolism*. 2014;65(2-3):205-210.

122. Nagin DS. Analyzing Developmental Trajectories: A Semiparametric, Group-Based Approach. *Psychological methods*. 1999;4(2):139-157.

123. Jones BL, Nagin DS. A Note on a Stata Plugin for Estimating Group-based Trajectory Models. *Sociological Methods & Research*. 2013;42(4):608-613.

124. MacKinnon DP, Fairchild AJ, Fritz MS. Mediation Analysis. *Annual Review of Psychology*. 2007;58(1):593-614.

125. VanderWeele T. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press; 2015.

126. Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation and bias. *International Journal of Epidemiology*. 2013;42(5):1511-1519.

127. Rubin DB. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*. 2005;100(469):322-331.

128. Hernan MA, Robins JM. *Causal Inference: What if*. Boca Raton: Chapman & Hall/CRC; 2020.

129. Pearl J. The Causal Mediation Formula—A Guide to the Assessment of Pathways and Mechanisms. *Prevention Science*. 2012;13(4):426-436.

130. Galea S, Hernán MA. Win-Win: Reconciling Social Epidemiology and Causal Inference. *American Journal of Epidemiology*. 2019;189(3):167-170.

131. VanderWeele TJ. Invited Commentary: Counterfactuals in Social Epidemiology-Thinking Outside of "the Box". *Am J Epidemiol*. 2020;189(3):175-178.

132. Kaufman JS. Commentary: Causal Inference for Social Exposures. *Annual Review of Public Health*. 2019;40(1):7-21.

133. VanderWeele TJ, Hernán MA. Causal Inference Under Multiple Versions of Treatment. *J Causal Inference*. 2013;1(1):1-20.

134. Greenland S. For and Against Methodologies: Some Perspectives on Recent Causal and Statistical Inference Debates. *European Journal of Epidemiology*. 2017;32(1):3-20.

135. Schwartz S, Prins SJ, Campbell UB, Gatto NM. Is the "well-defined intervention assumption" politically conservative? *Soc Sci Med*. 2016;166:254-257.

136. Glymour MM, Spiegelman D. Evaluating Public Health Interventions: 5. Causal Inference in Public Health Research-Do Sex, Race, and Biological Factors Cause Health Outcomes? *Am J Public Health*. 2017;107(1):81-85.

137. Fritz MS, Mackinnon DP. Required sample size to detect the mediated effect. *Psychol Sci*. 2007;18(3):233-239.

138. VanderWeele TJ, Chiba Y. Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. *Epidemiol Biostat Public Health*. 2014;11(2).

139. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*. 1986;51(6):1173-1182.

140. MacKinnon D. Contrasts in multiple mediator models. In: *Contrasts In Multiple Mediator Models.*2000:141-160.

141. VanderWeele TJ, Vansteelandt S. Mediation Analysis with Multiple Mediators. *Epidemiol Methods.* 2014;2(1):95-115.

142. Blaha MJ. The Critical Importance of Risk Score Calibration. *Journal of the American College of Cardiology.* 2016;67(18):2131-2134.

143. Krittanawong C, Virk HUH, Bangalore S, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific Reports.* 2020;10(1):16057.

144. Cooney MT, Dudina A, D'Agostino R, Graham IM. Cardiovascular Risk-Estimation Systems in Primary Prevention. *Circulation.* 2010;122(3):300-310.

145. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer New York; 2013.

146. Loftus TJ, Shickel B, Balch JA, et al. Phenotype clustering in health care: A narrative review for clinicians. *Frontiers in Artificial Intelligence.* 2022;5.

147. Tibshirani R, Walther G, Hastie T. Estimating the Number of Clusters in a Data Set Via the Gap Statistic. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* 2002;63(2):411-423.

148. Huang Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery.* 1998;2(3):283-304.

149. Gudjonsdottir H, Tynelius P, Fors S, et al. Cohort Profile: The Stockholm Diabetes Prevention Programme (SDPP). *International Journal of Epidemiology.* 2022;51(6):e401-e413.

150. Arellano-Campos O, Gómez-Velasco DV, Bello-Chavolla OY, et al. Development and validation of a predictive model for incident type 2 diabetes in middle-aged Mexican adults: the metabolic syndrome cohort. *BMC Endocr Disord.* 2019;19(1):41.

151. Region Stockholm. VAL-databaserna – Region Stockholm. https://www.folkhalsokollen.se/datakallor/val-databaserna/. Published 2023. Accessed 11-05, 2023.

152. Statistiska centralbyrån (SCB). Individregister och mikrodata. https://www.scb.se/vara-tjanster/bestall-data-och-statistik/bestalla-mikrodata/vilka-mikrodata-finns/individregister/. Published 2023. Accessed 11-05, 2023.

153. Socialstyrelsen. Register. https://www.socialstyrelsen.se/statistik-och-data/register/. Published 2023. Accessed 11-05, 2023.

154. Registercentrum Västra Götaland. Nationella Diabetesregistret. https://www.ndr.nu/#/english. Published 2023. Accessed 11-05, 2023.

155. Erikson R, Goldthorpe JH, Portocarero L. Intergenerational class mobility in three Western European societies: England, France and Sweden. *The British journal of sociology*. 1979;30(4):415-441.

156. World Medical Association. Declaration of Helsinki: ethical principles for medical research involving human subjects. *Jama*. 2013;310(20):2191-2194.

157. Zimmermann E, Bjerregaard LG, Gamborg M, Vaag AA, Sørensen TIA, Baker JL. Childhood body mass index and development of type 2 diabetes throughout adult life-A large-scale danish cohort study. *Obesity (Silver Spring)*. 2017;25(5):965-971.

158. Galuska DA, Gunn JP, O'Connor AE, Petersen R. Addressing Childhood Obesity for Type 2 Diabetes Prevention: Challenges and Opportunities. *Diabetes Spectrum*. 2018;31(4):330-335.

159. The GBD 2015 Obesity Collaborators. Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *New England Journal of Medicine*. 2017;377(1):13-27.

160. Olaiya MT, Knowler WC, Sinha M, et al. Weight tracking in childhood and adolescence and type 2 diabetes risk. *Diabetologia*. 2020;63(9):1753-1763.

161. Bjerregaard LG, Jensen BW, Ängquist L, Osler M, Sørensen TIA, Baker JL. Change in Overweight from Childhood to Early Adulthood and Risk of Type 2 Diabetes. *New England Journal of Medicine*. 2018;378(14):1302-1312.

162. Llewellyn A, Simmonds M, Owen CG, Woolacott N. Childhood obesity as a predictor of morbidity in adulthood: a systematic review and meta-analysis. *Obes Rev*. 2016;17(1):56-67.

163. World Health Organization. European Childhood Obesity Surveillance Initiative (COSI). In: *Report on the fifth round of data collection, 2018–2020*.2022.

164. Petrovic D, de Mestral C, Bochud M, et al. The contribution of health behaviors to socioeconomic inequalities in health: A systematic review. *Prev Med*. 2018;113:15-31.

165. Walker AF, Graham S, Maple-Brown L, et al. Interventions to address global inequity in diabetes: international progress. *The Lancet*. 2023;402(10397):250-264.

166. Chatzi G, Whittaker W, Chandola T, et al. Diabetes Prevention Programme and socioeconomic inequalities in Type 2 Diabetes in England. *Eur J Public Health*. 2022;32(Suppl 3).

167. Li L, Cheng W-Y, Glicksberg BS, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science Translational Medicine*. 2015;7(311):311ra174-311ra174.

168. Ahlqvist E, Storm P, Käräjämäki A, 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):361-369.

169. Deutsch AJ, Ahlqvist E, Udler MS. Phenotypic and genetic classification of diabetes. *Diabetologia*. 2022;65(11):1758-1769.

170. Sarría-Santamera A, Orazumbekova B, Maulenkul T, Gaipov A, Atageldiyeva K. The Identification of Diabetes Mellitus Subtypes Applying Cluster Analysis Techniques: A Systematic Review. *Int J Environ Res Public Health*. 2020;17(24).

171. Sharma A, Zheng Y, Ezekowitz JA, et al. Cluster Analysis of Cardiovascular Phenotypes in Patients With Type 2 Diabetes and Established Atherosclerotic Cardiovascular Disease: A Potential Approach to Precision Medicine. *Diabetes Care*. 2021;45(1):204-212.

172. Ding L, Mane R, Wu Z, et al. Data-driven clustering approach to identify novel phenotypes using multiple biomarkers in acute ischaemic stroke: A retrospective, multicentre cohort study. *eClinicalMedicine*. 2022;53.

173. Akyea RK, Ntaios G, Kontopantelis E, et al. A population-based study exploring phenotypic clusters and clinical outcomes in stroke using unsupervised machine learning approach. *PLOS Digital Health*. 2023;2(9):e0000334.

174. Cho SB, Kim SC, Chung MG. Identification of novel population clusters with different susceptibilities to type 2 diabetes and their impact on the prediction of diabetes. *Scientific Reports*. 2019;9(1):3329.

175. Wagner R, Heni M, Tabák AG, et al. Pathophysiology-based subphenotyping of individuals at elevated risk for type 2 diabetes. *Nat Med*. 2021;27(1):49-57.

176. Mao Z, Fukuma Y, Tsukada H, Wada S. Risk prediction of chronic diseases with a two-stage semi-supervised clustering method. *Preventive Medicine Reports*. 2023;32:102129.

177. Guo Q, Lu X, Gao Y, et al. Cluster analysis: a new approach for identification of underlying risk factors for coronary artery disease in essential hypertensive patients. *Scientific Reports*. 2017;7(1):43965.

178. Bennett JE, Kontis V, Mathers CD, et al. NCD Countdown 2030: pathways to achieving Sustainable Development Goal target 3.4. *The Lancet*. 2020;396(10255):918-934.